PARSEME – PARSing and Multiword Expressions within a European multilingual network

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, et al.

PARSEME – PARSing and Multiword Expressions within a European multilingual network

Agata Savary[*], Manfred Sailer[‡], Yannick Parmentier[°], Michael Rosner[⋆], Victoria Rosén[◇], Adam Przepiórkowski[†], Cvetana Krstev[♣], Veronika Vincze[♠], Beata Wójtowicz[†], Gyri Smørdal Losnegaard[◇], Carla Parra Escartín[▷], Jakub Waszczuk[*], Matthieu Constant[•], Petya Osenova[▽], Federico Sangati[⋈]

[*]Université François-Rabelais Tours, LI, France,
[‡]Goethe University Frankfurt, Institute for English and American Studies, Germany,
[°]Université d'Orléans, LIFO, France, [⋆]University of Malta, Department of Intelligent Computer Systems, Malta,
[◇]University of Bergen, Department of Linguistic, Literary and Aesthetic Studies, Norway,
[†]Polish Academy of Sciences, Institute of Computer Science, Poland,
[♣]University of Belgrade, Faculty of Philology, Serbia, [♠]University of Szeged, Department of Informatics, Hungary,
[▷]Hermes Traducciones, Spain, [•]Université Paris-Est, LIGM, CNRS, France,
[▽]Sofia University St. Kl. Ohridski and IICT-BAS, Bulgaria, [⋈]FBK, Digital Humanities, Italy

Abstract

The aim of this paper is to present PARSEME, a COST Action devoted to the issue of Multiword Expressions in parsing and in linguistic resources (corpora, lexicons). This is a "meta-paper" intended to be the main citation point for any future work referring to PARSEME: it does not describe in detail any single result of the Action, but rather summarises its multifarious activities and provides links to such results (both completed and in progress).

1. Background

Multiword expressions (MWEs) are linguistic objects containing two or more words and showing some degree of non-compositionality. For instance, the meaning of *to kick the bucket* (i.e. 'to die') cannot be predicted from the meaning of its components, while the (masculine) gender of *un peau rouge* ('redskin' in French) is not inherited from its nominal component (*peau* 'skin' is feminine). MWEs encompass versatile linguistic objects: compounds (*air brake*), complex terms (*random access memory*), multiword named entities (*European Bank for Reconstruction and Development*), light-verb constructions (*to take a nap*), phrasal verbs (*to make up for sth*), idioms (*to kick the bucket*), proverbs (*Fortune favours the bold.*), etc.

Basic facts about MWEs are that: (i) they are prevalent in natural language texts – up to 40% of text items belong to MWEs (Gross and Senellart 1998, Sag *et al.* 2002);(ii) they show unexpected behaviour at different levels: morphology (*attorney generals*, *attorneys general*), syntax (**the bucket was kicked* does not have an idiomatic meaning), semantics (*to spill the beans* = 'to reveal a secret'); (iii) most MWEs occur very rarely in corpora, so there is a problem with data sparseness (Baldwin and Villavicencio 2002); (iv) they are less ambiguous than simple words and can, therefore, be useful for information extraction, text classification, etc.; (v) they are underrepresented in language resources and tools.

While increasing attention is paid to MWEs in language technology, their integration into advanced levels of linguistic processing, notably deep parsing, is still largely unsatisfactory. To address these challenges, a European scientific network, PARSEME (PARSing and Multiword Expressions; `http://www.parseme.eu/`), was created in 2013, funded as the COST Action IC1207. Its objectives are threefold: (i) to focus on multilingualism in linguistic and technological studies, (ii) to establish a long-lasting cross-lingual, cross-theoretical and cross-methodological research network in NLP, and (iii) to bridge the gap between linguistic precision and computational efficiency in NLP applications. In September 2015, i.e. after two-thirds of its total duration (2013–2017), the Action has gathered a community of over 180 interdisciplinary members from 30 countries, representing 29 languages from 10 language families. It covers different parsing frameworks: Combinatory Categorial Grammar, Dependency Grammar, Transformational Grammar (TG), Head-driven Phrase Structure Grammar (HPSG), Lexical Functional Grammar (LFG), (Lexicalised) Tree Adjoining Grammar ((L)TAG), etc. It also addresses different methodologies (symbolic, probabilistic and hybrid parsing) and language technology applications (machine translation, information retrieval, etc.). This paper presents the scientific background, research programme, objectives, activities and results so far of the PARSEME Action.

2. Parsing and MWEs – state of the art

Multiword Expressions cross boundaries between traditional layers of linguistic processing, notably between lexicon, syntax and semantics. Even if some idiosyncrasies of MWEs call for description on the lexical level, other regular properties make them resemble well-formed syntactic structures. Therefore, one of the main challenges to be addressed is to find the most appropriate integration of MWE processing within parsing.

2.1. Symbolic MWE-aware parsing

Seminal proposals of representing MWEs within formal grammars have been put forward for major formalisms. Abeillé and Schabes (1989) show how French

MWEs are encoded as special kinds of elementary trees in LTAG. The idiomatic and the literal readings of a MWE obtain the same derived trees (showing syntactic structures of the sentence) but different derivation trees (showing which elementary trees have been combined and how). Within HPSG, Sag *et al.* (2002), Copestake *et al.* (2002), and Villavicencio *et al.* (2004) represent partly compositional English MWEs (e.g. *to spill the beans*) by paraphrasing and MWEs with opaque semantics (e.g. *to kick the bucket*) by separate semantic predicates. In LFG, Attia (2006) parses Arabic adjacent semi-fixed MWEs (e.g. *traffic light*) as single tokens, while syntactically compositional but semantically non-compositional MWEs (e.g. lit. *fiery bike* = 'motorbike') are handled by the grammar via lexical selection rules. In symbolic parsing based on TG, Wehrli *et al.* (2010) argue that the performance of both MWE identification and parsing, as well as of further parsing-based applications like machine translation, are systematically enhanced when both tasks are performed simultaneously. MWE lexical resources are used by their parser to promote MWE-oriented analyses in cases of ambiguity.

2.2. Statistical MWE-aware parsing

Statistical MWE-aware parsing has been considered either using a pipeline or a joint approach. The pipeline strategy consists in applying a standalone MWE recognizer before or after parsing. In most works implementing the pre-recognition strategy – e.g. (Cafferkey *et al.* 2007, Korkontzelos and Manandhar 2010, Constant *et al.* 2012, Kong *et al.* 2014) – the parser takes as input a sequence of lexical units that have been predicted at a preceding stage. For instance, each predicted MWE can be merged into a single token and treated as such in subsequent analysis steps. The disadvantage of methods using pre-recognition is that they are deterministic, so parsers cannot recover from their mistakes. A sentence like *He recognises her by the way she walks* cannot be analysed correctly if *by the way* is pre-analysed as a MWE adverb (by_the_way). In order to limit the drawback of a deterministic MWE analysis, a lattice of the most probable lexical segmentations might be provided as input to the parser, as in (Constant *et al.* 2013). The parser is then responsible for predicting the final lexical segmentation, as well as the syntactic structure. The MWE analysis stage can also take place after syntactic parsing, as MWE detection using syntactic relations leads to better results (Seretan 2011).

In order to cope with side effects caused by false MWE pre-recognition, various authors have employed a joint approach using different parsing frameworks. It consists in using a single statistical model for both MWE analysis and syntactic parsing. In the dependency parsing framework, specific edge labels are used for MWE components (Nivre and Nilsson 2004, Eryigit *et al.* 2011, Seddah *et al.* 2013, Vincze *et al.* 2013, Candito and Constant 2014, Nasr *et al.* 2015), which makes it possible to train MWE/syntactic parsers directly. In the constituent parsing framework, each MWE is annotated with a specific subtree (Arun and Keller 2005, Green *et al.* 2011, 2013).Such joint MWE-aware parsing results can be improved by using a discriminative, notably MWE-dedicated, reranker (Constant *et al.* 2012). Results are also improved by using a dual decomposition approach. For instance, Roux *et al.* (2014) combine several CRF-based sequential MWE labelers with several PCFG-LA MWE-aware joint parsers. All labelers and parsers are trained independently. The system iteratively penalises each parser and labeler until agreement on MWE segmentation is reached.

2.3. Lexical encoding and treebank annotation of MWEs

If deep linguistic processing is to be MWE-aware, language resources such as lexicons and treebanks containing fine-grained description of MWEs are necessary. While lexical approaches dedicated to a large variety of MWEs have a relatively long linguistic tradition, notably with Gross (1986) and Mel'čuk *et al.* (1988), NLP-applicable lexical encoding of MWEs has mainly concerned continuous MWEs – see (Savary 2008) for a survey on 11 formalisms applied to 7 languages. More recently, proposals have been put forward which also take verbal (largely non-continuous) MWEs into account. They may be divided roughly into: (i) morphosyntactic databases, e.g. (Grégoire 2010) for Dutch and (Al-Haj *et al.* 2014) for Hebrew, (ii) valence dictionaries such as (Hajič *et al.* 2003) for Czech and (Przepiórkowski *et al.* 2014) for Polish, (iii) ontological approaches with semantic calculus: (Marjorie McShane and Beale 2005) for English.

In treebanking, MWE annotation is increasingly taken into account but usually for a limited range of MWE classes, with notable efforts for including verbal MWE annotation in Czech (Bejček and Straňák 2010), Estonian (Kaalep and Muischnek 2008), and Hungarian (Vincze and Csirik 2010). See also §6.6. below.

3. Working Groups

Although considerable descriptive work has been done on MWEs, they are still not fully integrated into deep linguistic processing. Further steps into enhancing the state of the art are being taken by PARSEME via scientific activities organised within four working groups (WGs).

3.1. WG1: Lexicon/Grammar Interface

The activities of Working Group 1 aim at a better understanding of the linguistic properties of MWEs, in particular at the lexical and syntactic levels. There are two strongly interrelated subgroups within the working group. One focuses on the linguistic properties of MWEs and their possible classifications. A common point of departure and a major point of discussion within this subgroup are the classifications developed within formal and computational grammar, mainly based on observations about English. The second subgroup is primarily concerned with the computerised representation of MWEs. This representation, clearly, needs to be based on the linguistic properties as investigated in the other subgroup, but questions of encoding formalisms play an important role here, too.

Several initiatives of the working group have helped to promote a sensibility for language-specific differences

in the MWE inventory and for MWE-type-specific differences in the demands on lexical encoding. In the first two years the emphasis of WG1 was on surveys and overviews of existing research and resources to identify joint areas of interest within PARSEME (see §6.2.–§6.4.). This is now followed by an edited volume on the insights gained from a multi-lingual perspective on MWEs and by two workshops addressing questions of lexical representation (a hands-on workshop on lexical encoding and a joint workshop with experts in e-lexicography).

3.2. WG2: Parsing Techniques for MWEs

Working Group 2 focuses on the representation and parsing of MWEs. While statistical parsing of MWEs is a major focus of WG3 (see §3.3.), WG2 takes a closer look at deep parsing. In the traditional methodology this process is based on lexicons and grammars representing roughly properties and interactions of words in sentences, respectively. Several linguistically-motivated formal frameworks, such as HPSG, LFG, etc., have been proposed to encode these properties and interactions. They differ in terms of expressivity and complexity. Still, most of them already contain mechanisms for expressing properties of MWEs, which, however, need improvement in how they account for idiosyncrasies of MWEs on the one hand and their similarities to regular structures on the other. In this context WG2 studies how MWEs are represented and parsed in major grammar formalisms, allowing for better knowledge sharing. This study also considers MWEs from various language families. The objective is twofold: (i) designing best practices for MWE encoding within these grammar formalisms, and (ii) using the specific properties of MWEs to reduce parsing complexity.

During the first two years, WG2 activities were conducted mainly via (i) Short-Term Scientific Missions (STSMs), allowing researchers working on various languages and formalisms to share their expertise, and (ii) WG2 meetings, where more focused discussions on various aspects of MWE encoding and parsing, such as multilingualism, semantics and compositionality, took place. Hands-on sessions introducing tools for acquiring MWE lexicons and representing MWE grammatical properties also took place during these WG2 meetings. During the third and fourth year, WG2 activities will include the writing of a book summarizing research on MWE representation and parsing with formal grammars.

3.3. WG3: Statistical, Hybrid and Multilingual Processing of MWEs

It has become increasingly clear that no uniform approach will effectively handle the variety of different problems that arise with respect to the many different kinds of MWEs. WG3 was therefore conceived in recognition of the important role played by hybrid approaches which combine different methods to get the best results. Hybridity manifests itself in different ways. For instance it can involve different combinations of statistical and symbolic approaches: e.g. use of a grammatical formalism with an underlying statistical model for parsing or use of a lexicon to limit the search space of a statistically-based analyser. The objectives are to improve understanding of how hybrid methods may be applied to the processing of MWEs, and how they can be integrated in multilingual applications. WG3 also focuses on the central role of resources in the processing of MWEs (e.g. treebanks with MWE annotations for parsing, bilingual MWE lexicons for translation). Clearly, such resources are necessary in order to carry out MWE processing such as parsing. At the same time, the creation of such resources – particularly when complex grammatical relationships are involved – is supported by parsing. WG3 seeks to understand how to organise a research programme so that our currently imperfect and incomplete MWE resources and parsing/translation methods can be incrementally improved. It has also focused on different solutions to the problem of interleaving of parsing/translation with MWE recognition.

Faced with the wide variety of possible themes, WG3 decided to focus on the most prominent use cases dealing with MWE processing, namely discovery, machine translation and parsing, as well as their interactions. A practical consequence is the elaboration of a state-of-the-art survey with a specific focus on the interactions between the three themes and their common issues, in order to provide some recommendations for future research (see §6.5.).

3.4. WG4: Annotating MWEs in Treebanks

Working Group 4 studies the annotation of MWEs in treebanks, i.e. corpora annotated with syntactic and sometimes semantic information. Treebanks are crucial language resources which model linguistic phenomena on the basis of real-life and wide-coverage data. They are widely used in lexicography, language learning and linguistic research. They also constitute the core of rapidly progressing data-driven methods, including statistical parsing.

It is a weakness in many treebanks that some MWEs do not have any special annotation, or more specifically, the words that make up the MWE are annotated as if the constructions they are a part of can be analysed compositionally. In that case, MWEs are also difficult to search for and identify. The main objective of WG4 is to take a step towards enhanced MWE-aware methodologies of treebank construction, and their optimal usability in parsing. The expected outcomes are (1) annotation guidelines for representing MWEs in constituency and dependency treebanks, and (2) recommendations on how to use current and future treebanks to automatically extract lexicons and probability scores addressed in other WGs.

4. Instruments

COST instruments are meant for networking purposes. Thus, the major PARSEME events include:

(1) Bi-annual two-day **meetings**, organised so far in Brussels (Belgium), Warsaw (Poland), Athens (Greece), Frankfurt (Germany), and Valletta (Malta), and planned for Iași (Romania) and Struga (FYR Macedonia); they feature poster sessions, WG sessions and administrative meetings.

(2) **Workshops** usually co-organised with related communities and initiatives: (i) the Multi-Word Expressions Workshop at EACL 2014 in Gothenburg, Sweden, co-organised with the SIGLEX-MWE special interest group,

(ii) 2nd Workshop on Multi-word Units in Machine Translation and Translation Technology (MUMTTT 2015), co-organised with EUROPHRAS in Málaga, Spain, (iii) WG1 hands-on workshop on lexical encoding of MWEs in 2015, in Iași, Romania, (iv) Workshop on MWE e-lexicons, co-organised with the ENeL COST Action in 2016, in Skopje, FYR Macedonia.

(3) **Short-Term Scientific Missions** are a COST instrument aimed at fostering new collaboration and strengthening existing links. Within PARSEME, 19 missions have already taken place (far exceeding the usual COST Action expectations), with 4 more missions already approved. Most of the candidates that applied for STSMs were early stage researchers – 83%. Both researchers and hosts of STSMs come from variety of countries – 17 member countries and 2 international partner countries (see Fig. 1). Topics covered by STSMs include: extraction, description and parsing of MWEs using various formalisms, methods and tools, MWEs from the multilingual perspective, the role of MWEs in various NLP applications, etc. The reports of all completed STSMs are available at the PARSEME website.

(4) **Training schools** – see §6.7. below.

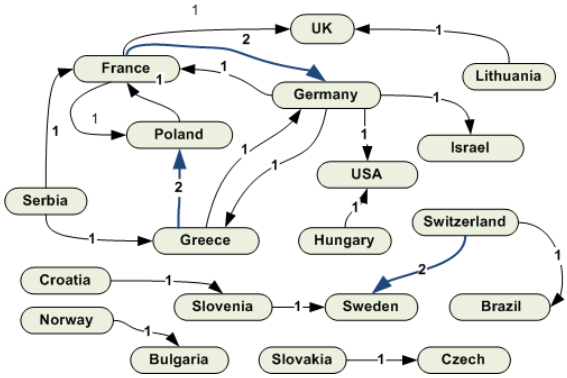

Figure 1: Source and destination countries of the past short-time scientific missions.

5. Policy

PARSEME is establishing a long-lasting collaboration platform of NLP experts, with a priority given to active Early-Stage Researchers (ESRs). It provides support for training and promoting ESRs through the organisation of training schools, workshops, STSMs, and WG meetings. ESRs make up 58% of the Action's WG members, and they also constitute 83% of the successful STSM candidates. In all Action activities the gender balance is also taken into account: 49% of the Action's WG members are women, and 50% of the STSM grants have been awarded to women. They also constitute 40% of the Action's management committee and 54% of the steering committee. To strengthen the COST Inclusiveness Policy,[1] priority has been given since 2014 to event locations in Inclusiveness Countries (the training school in the Czech Republic in 2015, general meetings in Malta and Romania in 2015, and in the FYR Macedonia in 2016). Of the 30 member countries in the PARSEME Action, 16 are on the Inclusiveness Target Countries list, and 43% (70 out of 164) of the WG members come from the Inclusiveness Countries. This balance is maintained while establishing the reimbursement lists (over 45% of the reimbursement funds are allocated to the participants from the Inclusiveness Countries).

[1] http://www.cost.eu/about_cost/strategy/excellence-inclusiveness

6. Results

6.1. Posters and joint papers

Ever since the second general meeting of PARSEME the poster sessions have been regarded as their central event. Researchers are invited to present their mature work (that may already have been presented elsewhere) on any topic relevant to the Action, particularly those that are relevant to current activities in its WGs. Abstracts, posters, and presentations of all accepted submissions are accessible from the PARSEME pages of respective general meetings. So far PARSEME poster sessions have attracted authors from as many as 26 (out of 30) member countries and from one International Partner Country (Brazil). A list of papers co-authored by PARSEME members and published in other venues is also accessible at the Action's website.

6.2. MWEs crosslinguistically

During the first year of the project a template was designed within WG1 that makes possible the documentation of MWEs in different languages along comparable dimensions of classification. These dimensions are: syntactic structure (i.e. VP-MWEs, NP-MWEs, etc.), syntactic flexibility (such as passivisation, modification), idiomaticity (lexical, syntactic, semantic, pragmatic, or statistic idiomaticity in the sense of Baldwin and Kim 2010). Other dimensions have been discussed but are not fully integrated yet. These are semantic relations among MWEs and rhetorical figures expressed in MWEs.

At present, there are richly elaborated templates for English, Greek, Macedonian, Norwegian, Polish, Serbian, Slovak, and Slovene. Each non-English example is transcribed, glossed, and translated. Comments on language-specific properties of syntactic categories or operations are added. The templates will be made public upon completion.

An important result of this work is that the strong correlation of semantic decomposability of a MWE and its syntactic flexibility (as emphasised for English in Nunberg *et al.* 1994) is not crosslinguistically valid. For this reason, classification according to semantic decomposability has been removed from the template.

This result has an immediate impact on the lexical representation and on MWE-sensitive parsing: Research based on (Nunberg *et al.* 1994) has tended to represent non-decomposable MWEs as phrasal units and decomposable MWEs as consisting of collocating words. This can now be replaced by a more uniform representation of MWEs (either as phrasal or as lexical, depending on the formalism), where the burden of restrictions on the syntactic flexibility is put on the language-specific properties of syntactic operations and on the semantics of the MWE.

6.3. Survey on MWE resources

In an effort towards consolidating past and ongoing research, PARSEME WG1 has conducted a survey of language resources (LRs) containing MWEs. Examples of such resources are monolingual and multilingual lists of MWEs, treebanks with MWE annotation, etc. (see Fig. 2). The survey collects information about language and linguality, LR size, linguistic features (are the MWEs continuous or discontinuous, are they represented as lemmas or also with inflected forms, etc.), lexical and grammatical frameworks used and various administration data.

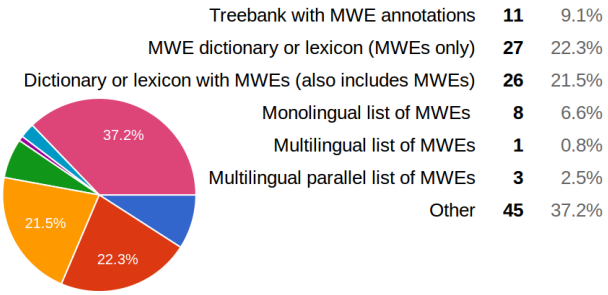

Figure 2: Types of MWE resources in the WG1 survey

The survey was launched in May 2014 and has since been advertised to relevant communities, including the Corpora and Linguist mailing lists. The impact is significant with around 100 unique responses. Almost half of the recorded LRs (45%) are freely available, while 46% are available under certain restrictions. Most resource owners (94%) have made, or are willing to make, their LRs available for use. As many as 28 languages are currently mentioned: Arabic, Bulgarian, Chinese, Czech, Danish, Dutch, Egyptian, English, Estonian, Finnish, French, German, Greek, Hebrew, Hungarian, Italian, Korean, Macedonian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Spanish, Swedish, and Turkish. The survey is still open (`http://goo.gl/iUTc06`) and the current anonymised results are publicly available (`http://goo.gl/WWZwzO`). When the survey is closed for further contributions, the results will be normalised and presented in a stable version on the PARSEME website.

6.4. Survey on lexical encoding of verbal MWEs

Addressing the WG1 objectives of (i) providing recommendations for best practices in lexical encoding and (ii) working towards the development of encoding standards, WG1 members have initiated a contrastive state-of-the-art survey on existing approaches to the lexical encoding of verbal MWEs. The survey describes a set of properties of verbal MWEs that are potentially problematic in lexical encoding, includes examples from several languages (currently French, German, Polish and Hebrew), and shows how different formalisms cover these challenges. This contrastive analysis might pave the way towards recommendations for an optimal lexical formalism.

6.5. Surveys on MWE discovery, translation and parsing

The WG3 survey, which is currently in preparation, will be grounded within a framework structured around the main use cases to which the notion of hybridity applies: parsing, machine translation and discovery, i.e. the automatic creation of resources that can subsequently be exploited for MWE processing. Certain combinations of the themes and associated symbiotic relationships will also be discussed. For example, the output of discovery, in the form of a MWE lexicon or annotated dataset, can clearly be used to support parsing. The survey should be completed by the end of 2015 and subsequently the work of the group will be devoted to realising the most feasible of its recommendations.

6.6. Survey on MWE annotation in treebanks

WG4 has conducted a survey to find out how MWEs are annotated in treebanks. The working group has established a wiki[2] that provides an overview of MWE annotations in 16 treebanks for 13 languages (Bulgarian, Czech, Dutch, English, Estonian, French, German, Hungarian, Latvian, Norwegian, Polish, Portuguese and Swedish). Several new treebanks are currently being added. The overview provides short descriptions of each treebank, including information such as name, author, formalism, license, links to documentation, history (how the treebank was constructed), whether it is static or dynamic, etc. The MWEs are classified as belonging to one of the following types: nominal MWEs (subtypes: named entities, noun–noun compounds and other nominal MWEs), verbal MWEs (subtypes: phrasal verbs, light verb constructions, VP idioms and other verbal MWEs), prepositional MWEs, adjectival MWEs, MWEs of other categories, and proverbs. For each MWE type that a treebank has a special annotation for, information is provided about what kind of analysis the treebank provides. This includes an example from the treebank (possibly simplified) together with a glossed version of the sentence and a prose description of the analysis. The examples and descriptions help to identify any cross-lingual inconsistencies and support a more language-aware typological comparison. The complete survey will be used as a basis for creating guidelines for best practice in treebank annotation of MWEs.

6.7. Resources from the Prague training school

The first training school organised by PARSEME took place in Prague in January 2015 (with another already planned for 2016). It consisted of four courses with the topics of treebanking and MWEs, MWEs in linguistic theories and their lexical encoding, MWEs in HPSG and MWEs in dependency parsing. Materials from the courses – i.e. slides and recordings from the lectures, as well as datasets and tools from the laboratory sessions – are publicly available at the website of the event.[3] The training school ended with a cross-module session, where participants discussed

[2] `http://clarino.uib.no/iness/page?page-id=MWEs_in_parseme`

[3] `https://ufal.mff.cuni.cz/events/parseme-1st-training-school`

issues concerning the multilingualism of MWEs. In addition, interesting and challenging examples of MWEs from several languages were analysed together by the audience. The lists of MWE examples and multilingual issues are also downloadable from the website, and can be used as a base for further research.

7. Conclusion

PARSEME, just like other COST Actions, is a scientific network rather than a full-fledged project; in particular, there are no funds for personal costs, i.e. for "real work". Despite that, the results of PARSEME are tangible: various surveys, poster presentations (including 2-page abstracts), and training school materials. PARSEME also fosters contacts between researchers interested in MWEs on the one hand and linguistic tools and resources on the other, not only via personal meetings, but also through joint work on surveys, discussions via dedicated mailing lists, common wiki spaces, etc. Apart from such intensive PARSEME-internal collaboration, links have been established with other COST Actions (IC1002 MUMIA, IS1006 SignGram, IC1302 KEYSTONE, IS1305 ENeL – with a joint workshop on MWE e-lexicons to be held in Skopje on 5–6 April 2016, IS1312 TextLink and IC1307 iV&L Net), with European projects (CLARIN, METANET, QTLeap, etc.), and with the ACL SIGLEX MWE section (which resulted in the joint organisation of The Tenth Workshop on MWEs in 2014). Most importantly, PARSEME has also already resulted in some spin-off and related national projects, including LD-PARSEME in the Czech Republic, VERBEL in Poland, JANES in Slovenia and PARSEME-FR in France. We hope that PARSEME will continue to catalyse work on linguistic and computational aspects of MWEs until the end of the project in 2017 and – hopefully – beyond.

References

Abeillé, A. and Schabes, Y. (1989). Parsing Idioms in Lexicalized TAGs. In H. L. Somers and M. M. Wood, eds., *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, pp. 1–9.

Al-Haj, H., Itai, A., and Wintner, S. (2014). Lexical Representation of Multiword Expressions in Morphologically-complex Languages. *International Journal of Lexicography*, **27**(2), 130–170.

Arun, A. and Keller, F. (2005). Lexicalization in Crosslinguistic Probabilistic Parsing: The Case of French. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 306–313, Stroudsburg, PA, USA.

Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, pp. 87–98, Berlin. Springer.

Baldwin, T. and Kim, S. N. (2010). Multiword Expressions. In N. Indurkhya and F. J. Damerau, eds., *Handbook of Natural Language Processing*, pp. 267–292. CRC Press, Boca Raton, 2 edition.

Baldwin, T. and Villavicencio, A. (2002). Extracting the Unextractable: A Case Study on Verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pp. 98–104.

Bejček, E. and Straňák, P. (2010). Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, **44**(1–2), 7–21.

Cafferkey, C., Hogan, D., and van Genabith, J. (2007). Multiword units in treebank-based probabilistic parsing and generation. In *Proceedings of the 10th International Conference on Recent Advances in Natural Language Processing (RANLP'07)*, Borovets, Bulgaria.

Candito, M. and Constant, M. (2014). Strategies for Contiguous Multiword Expression Analysis and Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 743–753.

Constant, M., Sigogne, A., and Watrin, P. (2012). Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, pp. 204–212, Stroudsburg, PA, USA.

Constant, M., Roux, J. L., and Sigogne, A. (2013). Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. *ACM Trans. Speech Lang. Process.*, **10**(3), 8:1–8:24.

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

Eryigit, G., Ilbay, T., and Can, O. A. (2011). Multiword Expressions in Statistical Dependency Parsing. In *Proceedings of the Second Workshop on Statistical Parsing of Morphologically Rich Languages (IWPT - 12th International Conference on Parsing Technologies)*, pp. 45–55, Dublin, Ireland.

Green, S., de Marneffe, M.-C., Bauer, J., and Manning, C. D. (2011). Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. In *EMNLP*, pp. 725–735.

Green, S., de Marneffe, M.-C., and Manning, C. D. (2013). Parsing Models for Identifying Multiword Expressions. *Computational Linguistics*, **39**(1), 195–227.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, **44**(1-2).

Gross, M. (1986). Lexicon-grammar: The Representation of Compound Words. In *Proceedings of the 11th Coference on Computational Linguistics*, pp. 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gross, M. and Senellart, J. (1998). Nouvelles bases statistiques pour les mots du français. In *Proceedings of JADT'98, Nice 1998*, pp. 335–349.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolářová,

V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In J. Nivre and E. Hinrichs, eds., *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Norway.

Kaalep, H.-J. and Muischnek, K. (2008). Multi-Word Verbs of Estonian: a Database and a Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pp. 48–51, Marrakech, Maroc.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A Dependency Parser for Tweets. In A. Moschitti, B. Pang, and W. Daelemans, eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1001–1012.

Korkontzelos, I. and Manandhar, S. (2010). Can Recognising Multiword Expressions Improve Shallow Parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 636–644, Stroudsburg, PA, USA.

Marjorie McShane, S. N. and Beale, S. (2005). The Description and Processing of Multiword Expressions in OntoSem. Working Paper 07-05, Institute for Language and Information Technologies University of Maryland Baltimore County.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N., and Mantha, S. (1988). *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques*. Presses de l'Univ. de Montréal.

Nasr, A., Ramisch, C., Deulofeu, J., and André, V. (2015). Joint Dependency Parsing and Multiword Expression Tokenisation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'15)*.

Nivre, J. and Nilsson, J. (2004). Multiword Units in Syntactic Parsing. In *Proceedings of MEMURA 2004 – Methodologies and Evaluation of Multiword Units in Real-World Applications, Workshop at LREC 2004, May 25, 2004, Lisbon, Portugal*, pp. 39–46, Lisbon, Portugal.

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

Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pp. 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Roux, J. L., Rozenknop, A., and Constant, M. (2014). Syntactic Parsing and Compound Recognition via Dual Decomposition: Application to French. In J. Hajic and J. Tsujii, eds., *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pp. 1875–1885.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expresions: A Pain in the Neck for NLP. In *Proceedings of CICLING'02*. Springer.

Savary, A. (2008). Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, **1**(2), 1–53.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., Foster, J., Goenaga, I., Gojenola Galletebeitia, K., Goldberg, Y., Green, S., Habash, N., Kuhlmann, M., Maier, W., Nivre, J., Przepiórkowski, A., Roth, R., Seeker, W., Versley, Y., Vincze, V., Woliński, M., Wróblewska, A., and Villemonte De La Clergerie, É. (2013). Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pp. 146–182, Seattle, Washington, United States.

Seretan, V. (2011). *Syntax-Based Collocation Extraction*. Springer, Dordrecht.

Villavicencio, A., Copestake, A., Waldron, B., and Lambeau, F. (2004). Lexical Encoding of MWEs. In *ACL Workshop on Multiword Expressions: Integrating Processing, July 2004*, pp. 80–87.

Vincze, V. and Csirik, J. (2010). Hungarian Corpus of Light Verb Constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1110–1118, Beijing, China. Coling 2010 Organizing Committee.

Vincze, V., Zsibrita, J., and T., I. N. (2013). Dependency Parsing for Identifying Hungarian Light Verb Constructions. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pp. 207–215.

Wehrli, E., Seretan, V., and Nerima, L. (2010). Sentence Analysis and Collocation Identification. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pp. 27–35, Beijing, China.