Multiword expressions at length and in depth

Extended papers from the MWE 2017 workshop

Edited by

Stella Markantonatou

Carlos Ramisch

Agata Savary

Veronika Vincze

Phraseology and Multiword Expressions 2

language
science
press

Phraseology and Multiword Expressions

Series editors

In this series:

1. Sailer, Manfred & Stella Markantonatou (eds.). Multiword expressions: Insights from a multilingual perspective.

2. Markantonatou, Stella, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.). Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop.

ISSN: 2625-3127

Multiword expressions at length and in depth

Extended papers from the MWE 2017 workshop

Edited by

Stella Markantonatou

Carlos Ramisch

Agata Savary

Veronika Vincze

language
science
press

Markantonatou, Stella, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.). 2018. *Multiword expressions at length and in depth*: *Extended papers from the MWE 2017 workshop* (Phraseology and Multiword Expressions 2). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/204

ISBN: 978-3-96110-123-8 (Digital)
978-3-96110-124-5 (Hardcover)

ISSN: 2625-3127
DOI:10.5281/zenodo.1469527
Source code available from www.github.com/langsci/204
Collaborative reading: paperhive.org/documents/remote?type=langsci&id=204

Cover and concept of design: Ulrike Harbort
Typesetting: Sebastian Nordhoff, Agata Savary
Proofreading: Amr Zawawy, Beatriz Sánchez-Cárdenas, Bojana Đorđević, Brett Reynolds, Conor Pyle, Eitan Grossman, Eleni Koutsomitopoulou, Felix Hoberg, Felix Kopecky, Glorianna Jagfeld, Harald Ulland, Ikmi Nur Oktavianti, Ivelina Stoyanova, Ivica Jeđud, Jeroen van de Weijer, Melanie Röthlisberger, Jean Nitzke, Steven Moran, Trinka D'Cunha, Vadim Kimmelman, Valeria Quochi
Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono
Typesetting software: XƎLaTeX

Language Science Press
Unter den Linden 6
10099 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Freie Universität Berlin

Contents

Preface

Stella Markantonatou

Institute for Language and Speech Processing, Athena RIC, Greece

Carlos Ramisch

Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Agata Savary

University of Tours, LIFAT, France

Veronika Vincze

University of Szeged, Hungary

In this introductory chapter we present the rationale for the volume at hand. We explain the origin and the selection process of the contributing chapters, and we sketch the contents and the organization of the volume. We also describe notational conventions put forward for citing and glossing multilingual examples of multiword expressions. We finally acknowledge the efforts which paved the way for setting up this book project, ensuring its quality and publication.

Multiword expressions (MWEs) belong to those language phenomena which pose the hardest challenges both in linguistic modelling and in automatic processing. This is due notably to their semantic non-compositionality, that is, the impossibility to predict their meaning from their syntactic structure and from the semantics of their component words in a way deemed regular for the given language. But MWEs also exhibit unpredictable behaviour on other levels of language modelling such as the lexicon, morphology and syntax, and call, therefore, for dedicated procedures in natural language processing (NLP) applications.

These challenges have been addressed by an ever-growing and increasingly multilingual community gathering at the Multiword Expressions Workshop, organized yearly since 2003, often jointly with major NLP conferences. The 13th

Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze. 2018. Preface. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, iii–xiii. Berlin: Language Science Press. DOI:10.5281/zenodo.1469549

edition of the Workshop, co-located with the EACL 2017 conference in Valencia, Spain, saw a major evolution of the topics and methods addressed by the community. This evolution resulted notably from the efforts coordinated by PARSEME, a European research network dedicated to parsing and MWEs, gathering, since 2013, researchers from 31 countries and working on as many languages.[1]

One of PARSEME's main outcomes was a corpus in 18 languages annotated for verbal MWEs (VMWEs), based on a unified methodology and terminology, and published under open licenses. This considerable collective and inclusive effort mobilized experts from different linguistic traditions, triggered cross-language and cross-domain discussions, and brought convergence to modelling and processing of MWE-related phenomena. The availability of this new open resource also made it possible to organize the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions, i.e. a competition of VMWE identification tools, whose culminating event was hosted by the MWE 2017 workshop in Valencia. The 7 participating systems covered jointly all 18 languages represented in the corpus. They also offered a large panorama of VMWE identification techniques. These assets, as well as some other contributions published in the main track of the MWE 2017 workshop, showed a growing awareness by the MWE community of specific challenges posed, in particular, by verbal MWEs, such as their discontinuity and high morpho-syntactic flexibility. The workshop programme addressed a large variety of MWE-dedicated tasks such as: lexical and grammatical encoding, annotation, tokenization, extraction, identification, classification, variation study, parsing, compositionality prediction, and translation. Finally, it testified that MWE research has reached a highly multilingual stage.

1 Organization and contents of the volume

This volume is a collection of selected extended papers from the MWE 2017 workshop in Valencia: 8 of them from the main track, and 5 from the shared task track. They address 19 languages from 9 language families, as shown in Figure 1. The chapter selection process was initiated by an open call, addressed to all co-authors of the workshop papers. The call included the requirement of extending the original contributions by at least 30% with unpublished content. An international programme committee reviewed 15 submissions and selected 14 of them for publication. One of the selected chapters was further withdrawn. As a result, the volume consists of 13 chapters covering a large variety of aspects related to MWE representation and processing, with a particular focus on verbal MWEs.

[1] http://www.parseme.eu

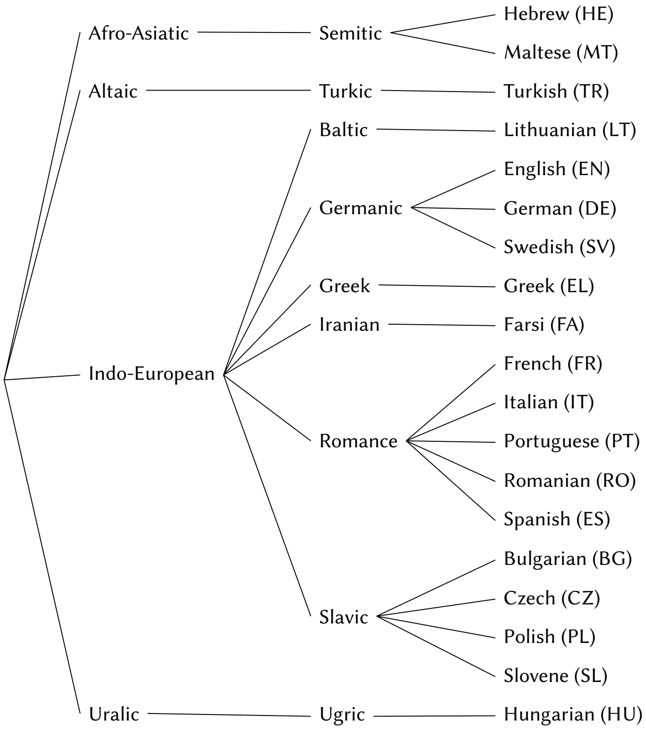


Figure 1: Languages addressed in the chapters of this volume, together with their two-letter language code, language families and genera (middle columns), according to WALS (World Atlas of Language Structures, Dryer & Haspelmath 2013).

Chapters 1 to 3 address outstanding linguistic properties of VMWEs and their automatic assessment. Geeraert et al. (2018 [this volume]) discuss idiomatic *variation* of several types of English verbal idioms. They draw on multimodal evidence, namely acceptability rating and eye-tracking measurements, to investigate comprehension mechanisms. Barančíková & Kettnerová (2018 [this volume]) deal with light-verb constructions and verbal idioms in Czech, and explore their paraphrasability by single verbs. They also propose a lexicographic scheme for VMWE paraphrase encoding, and show its usefulness in machine translation. Bhatia et al. (2018 [this volume]) focus on English verb-particle constructions, and estimate their compositionality degree, so as to further compute the semantics of sentences containing verb-particle constructions on the basis of lexical, grammatical and ontological resources.

Chapters 4 to 8 are dedicated to the PARSEME shared task on automatic identification of verbal MWEs. Savary et al. (2018 [this volume]) describe the PARSEME multilingual VMWE-annotated corpus underlying the shared task. In a first step, the annotation guidelines and methodology are presented, then the properties of the annotated corpus are analysed across the 18 participating languages. Maldonado & QasemiZadeh (2018 [this volume]) offer a critical analysis of the shared task organization and of its results across languages and participating systems. Chapters 6 to 8 are dedicated to three of the seven VMWE identification systems participating in the shared task. They show a representative panorama of recent techniques used to address the VMWE identification task. Moreau et al. (2018 [this volume]) model the task as sequence labelling with reranking. Al Saied et al. (2018 [this volume]) present a dedicated transition-based dependency parser, which jointly predicts a syntactic dependency tree and a forest of VMWEs. Finally, Simkó et al. (2018 [this volume]) rely on a generic dependency parser trained on a corpus with merged syntactic and VMWE labels.

Chapters 9 to 11 further discuss MWE identification issues in various settings and scopes. Brooke et al. (2018 [this volume]) show how comparing various annotations of the same MWE in an English corpus can help correct annotation errors, enhance the consistency of corpus annotation, and consequently increase the quality of downstream MWE identification systems. Scholivet et al. (2018 [this volume]) address identification of French continuous MWEs via sequence labelling, compare its results to more sophisticated parsing-based approaches, and show that feature engineering based on external lexical data (whether handcrafted or automatically extracted) systematically enhances the tagging performance. Taslimipoor et al. (2018 [this volume]), conversely, advocate modelling MWE identification as classification rather than tagging. They exploit word embeddings as classification features in Italian, Spanish and English, and put forward a MWE-specific methodology of train vs. test corpus split.

The last two chapters of the book, 12 and 13, are dedicated to multilingual MWE-oriented applications. Garcia (2018 [this volume]) describes automatic extraction of bilingual collocation equivalents in English, Spanish, and Portuguese, using syntactic dependencies, association measures and distributional models. Finally, Salehi et al. (2018 [this volume]) predict the compositionality of English and German MWEs on the basis of their translations extracted from highly multilingual lexical resources.

2 Conventions for citing and glossing multilingual MWE examples

As mentioned above, this volume addresses a large number of languages, particularly in the chapters related to the PARSEME corpus and shared task. Therefore, the editorial effort around this volume includes putting forward notational conventions which might become a standard for citing and glossing multilingual MWE examples. We illustrate the proposed conventions by the *numbered examples* (1) to (4). Each numbered example contains:

(i) a sample use of the VMWE, followed by the 2-letter ISO 639-1 language code (cf. Figure 1),

(ii) a transcription, if the language of the example is written with a script different from the one used for the main text,[2]

(iii) a gloss following the Leipzig Glossing Rules,[3]

(iv) a literal translation, followed by an idiomatic translation in single quotes.

For English examples, items (ii)–(iv) are irrelevant or optional but idiomatic translation might sometimes be useful to ease the comprehension by non-native readers. For right-to-left languages (e.g. Farsi or Hebrew), item (i) is spelled right-to-left, item (iv) left-to-right and items (ii)–(iii) left-to-right within components, and right-to-left from one component to another. Lexicalized components of the VMWE, i.e. those which are always realized by the same lexeme (cf. Savary et al. 2018 [this volume], §2, p. 92) are highlighted in bold face.

(1) She reluctantly **took on** this task. (EN)
‘She reluctantly agreed to be in charge of this task.’

(2) *Ida* ***skriva*** ***glavo*** ***v*** ***pesek***. (SL)
Ida hide.3.sg head in sand
Ida hides her head in the sand. ‘Ida pretends not to see a problem.’

[2] For instance, transcription is needed for Bulgarian, Greek, Farsi and Hebrew examples in this volume. Conversely, examples in English, or any other language using Latin script, would require transcriptions in texts written in Cyrillic, Greek, Arabic or Hebrew script.

[3] https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

(3) *Η Ζωή* ***παίρνει*** *μία* ***απόφαση***. (EL)
i Zoi perni mia apofasi
the.FEM.SG Zoe.FEM.SG take.3.SG a decision
Zoe takes a decision. ‘Zoe makes a decision.’

(4) به قدر کافی برای من خواب دیده است (FA)
ast **dide khab** man **baraye** kafi qadre be
is seen sleep me for enough quantity to
He had enough sleep for me. ‘He has many plans for me.’

In-line examples, used for brevity, are preceded by the 2-letter language code, contain items (i), (ii) if relevant, and (iv) only, and the idiomatic translation (if any) is introduced by a double arrow ‘⇒’. For instance, an in-line example corresponding to numbered example (2) would be the following: (SL) *Ida* ***skriva glavo v pesek*** ‘Ida hides her head in the sand’ ⇒ ‘Ida pretends not to see a problem’. If the language under study is written with a non-Latin alphabet, the inline example should not be in italics, and the transliteration should be included in parentheses, e.g. (EL) Η Ζωή **παίρνει** μία **απόφαση** (I Zoi perni mia apofasi) ‘The Zoe takes a decision’ ⇒ ‘Zoe makes a decision’. To keep such examples reasonably short, the first item can be omitted and only the transliterated example is kept: (EL) I Zoi perni mia apofasi ‘The Zoe takes a decision’ ⇒ ‘Zoe makes a decision’. The literal or the idiomatic translation are sometimes superfluous or too verbose, and can be skipped, as in: (EL) I Zoi perni mia apofasi ‘Zoe makes a decision’.

The typesetting commands for both numbered and in-line examples for LaTeX can be found in the GitHub repository containing the source codes of this volume, accessible from its webpage.[4]

3 Acknowledgements

Huge collective efforts paved the way towards the publication of this volume.

The PARSEME Shared Task on Automatic Identification of VMWEs was made possible by the European research network PARSEME[5], funded by the IC1207 COST[6] Action in 2013–2017. The PARSEME core group and eighteen language teams (cf. acknowledgements in Savary et al. 2018 [this volume]) prepared the annotation guidelines and tools, and created the multilingual corpus underlying

[4] http://langsci-press.org/catalog/book/204
[5] http://www.parseme.eu
[6] http://www.cost.eu/

the shared task. We are also grateful to the COST Officials, notably to Ralph Stübner, for their continuous support in the scientific and financial administration of the Action.

The 13th Workshop on Multiword Expressions (MWE 2017)[7] was organized and sponsored by PARSEME jointly with the Special Interest Group on the Lexicon (SIGLEX)[8] of the Association for Computational Linguistics. An international Programme Committee of over 80 researchers from 27 countries reviewed the workshop submissions and ensured a high-quality paper selection.

Our warm acknowledgements go also to the Editorial Staff of Language Science Press, and in particular to Sebastian Nordhoff, for his expert and friendly editorial assistance. We are also grateful to the editors of the *Phraseology and Multiword Expressions* book series for their support. In particular, Yannick Parmentier played the role of the editor-in-charge of the volume, and offered advice on technical editorial issues.

Finally, we thank the following reviewers, who provided insightful reviews to the chapters submitted to this volume:

- Dimitra Anastasiou (Luxembourg Institute of Science and Technology, Luxembourg)
- Doug Arnold (University of Essex, UK)
- Timothy Baldwin (University of Melbourne, Australia)
- Eduard Bejček (Charles University in Prague, Czech Republic)
- António Branco (University of Lisbon, Portugal)
- Marie Candito (Paris Diderot University, France)
- Fabienne Cap (Uppsala University, Sweden)
- Matthieu Constant (Université de Lorraine, France)
- Paul Cook (University of New Brunswick, Canada)
- Lucia Donatelli (Georgetown University, USA)
- Silvio Ricardo Cordeiro (Aix-Marseille University, France)

[7] http://multiword.sourceforge.net/mwe2017

[8] http://siglex.org/

- Béatrice Daille (University of Nantes, France)
- Gaël Dias (University of Caen Basse-Normandie, France)
- Voula Giouli (Institute for Language and Speech Processing/Athena RIC, Greece)
- Tracy Holloway King (eBay, USA)
- Philipp Koehn (Johns Hopkins University, USA)
- Dimitrios Kokkinakis (University of Gothenburg, Sweden)
- Yannis Korkontzelos (Edge Hill University, UK)
- Eric Laporte (Université Paris-Est Marne-la-Vallee, France)
- Timm Lichte (University of Düsseldorf, Germany)
- Gyri S. Losnegaard (University of Bergen, Norway)
- Héctor Martínez Alonso (Thomson Reuters Labs, Canada)
- Verginica Mititelu (Romanian Academy Research Institute for Artificial Intelligence, Romania)
- Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
- Joakim Nivre (Uppsala University, Sweden)
- Jan Odijk (University of Utrecht, Netherlands)
- Petya Osenova (Bulgarian Academy of Sciences, Bulgaria)
- Harris Papageorgiou (Institute for Language and Speech Processing/Athena RIC, Greece)
- Yannick Parmentier (Université de Lorraine, France)
- Carla Parra Escartín (Dublin City University, ADAPT Centre, Ireland)
- Agnieszka Patejuk (Institute of Computer Science, Polish Academy of Sciences, Poland)
- Pavel Pecina (Charles University in Prague, Czech Republic)

- Scott Piao (Lancaster University, UK)
- Martin Riedl (University of Stuttgart, Germany)
- Manfred Sailer (Goethe-Universität Frankfurt am Main, Germany)
- Nathan Schneider (Georgetown University, USA)
- Sabine Schulte Im Walde (University of Stuttgart, Germany)
- Ruben Urizar (University of the Basque Country, Spain)
- Aline Villavicencio (Federal University of Rio Grande do Sul, Brazil)
- Jakub Waszczuk (University of Tours, France)
- Shuly Wintner (University of Haifa, Israel)

We hope that the quality of this volume will be a valuable reward to all these contributors, and a source of information and inspiration for the international MWE community.

References

Al Saied, Hazem, Marie Candito & Matthieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 209–226. Berlin: Language Science Press. DOI:10.5281/zenodo.1469561

Barančíková, Petra & Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: The case of Czech light verbs and idioms. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 35–59. Berlin: Language Science Press. DOI:10.5281/zenodo.1469553

Bhatia, Archna, Choh Man Teng & James F. Allen. 2018. Identifying senses of particles in verb-particle constructions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 61–86. Berlin: Language Science Press. DOI:10.5281/zenodo.1469555

Brooke, Julian, King Chan & Timothy Baldwin. 2018. Semi-automated resolution of inconsistency for a harmonized multiword-expression and dependency-parse annotation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 245–262. Berlin: Language Science Press. DOI:10.5281/zenodo.1469565

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/, accessed 2018-7-25. Accessed on.

Garcia, Marcos. 2018. Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 319–342. Berlin: Language Science Press. DOI:10.5281/zenodo.1469571

Geeraert, Kristina, R. Harald Baayen & John Newman. 2018. "Spilling the bag" on idiomatic variation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 1–33. Berlin: Language Science Press. DOI:10.5281/zenodo.1469551

Maldonado, Alfredo & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–175. Berlin: Language Science Press. DOI:10.5281/zenodo.1469557

Moreau, Erwan, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel & Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 177–207. Berlin: Language Science Press. DOI:10.5281/zenodo.1469559

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2018. Exploiting multilingual lexical resources to predict MWE compositionality. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 343–373. Berlin: Language Science Press. DOI:10.5281/zenodo.1469573

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon

Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1471591

Scholivet, Manon, Carlos Ramisch & Silvio Cordeiro. 2018. Sequence models and lexical resources for MWE identification in French. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 263–297. Berlin: Language Science Press. DOI:10.5281/zenodo.1469567

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2018. Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 227–243. Berlin: Language Science Press. DOI:10.5281/zenodo.1469563

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 299–317. Berlin: Language Science Press. DOI:10.5281/zenodo.1469569

Chapter 1

"Spilling the bag" on idiomatic variation

Kristina Geeraert
KU Leuven

R. Harald Baayen
University of Tübingen & University of Alberta

John Newman
University of Alberta & Monash University

Recent corpus-based studies have shown that idioms can vary much more extensively than previously claimed (Moon 1998; Barlow 2000; Duffley 2013), but little research has been conducted on how we understand or regard these variants (Gibbs et al. 1989; McGlone et al. 1994). This study further examines idiomatic variation, specifically investigating the acceptability and processing of several types of variants through rating and eye-tracking experiments. Four types of idiom variants were included, in addition to the canonical form and a literal meaning of the expression (i.e. in a different context). The results show that altering an idiom does not render it completely unacceptable nor incomprehensible, but rather, certain variants are more preferred and easier to understand. These results show support for probabilistic accounts of language.

1 Introduction

Idioms have traditionally been regarded as multiword units whose meaning cannot be derived from the meaning of its parts (Bobrow & Bell 1973). The literal meaning of an idiom is the word-by-word compositional meaning of the words, whereas the idiomatic meaning is stored separately with the idiom, as if a large

Kristina Geeraert, R. Harald Baayen & John Newman. 2018. "Spilling the bag" on idiomatic variation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 1–33. Berlin: Language Science Press. DOI:10.5281/zenodo.1469551

word. Furthermore, if the idiom is stored whole, then idioms must also be structurally fixed. This rationale has led researchers to predominantly investigate idioms in their canonical form, and how they are understood in comparison to a literal paraphrase (Swinney & Cutler 1979; Gibbs 1980; Cacciari & Tabossi 1988; Titone & Connine 1999).

Recent corpus-based research however has shown that idioms can occur with a range of variation (Moon 1998; Barlow 2000; Langlotz 2006; Schröder 2013), such as syntactic variation (e.g. *her new-found reputation was a bubble that would burst* [from *burst one's bubble* 'shatter one's illusions about something'], *the question begged by* [*beg the question* 'invite an obvious question']), lexical variation (e.g., *set/start the ball rolling* 'set an activity in motion', *a skeleton in the closet/cupboard* 'a discreditable fact that is kept secret'), truncations (e.g., *make hay* [*while the sun shines*] 'take advantage of favourable conditions'), and even adverbial or adjectival modifications (e.g., *pulling political strings* [*pull strings* 'use one's power or influence to gain an advantage'], *rock the party boat* [*rock the boat* 'upset the status-quo']). This variation can even occur with nondecomposable idioms (Duffley 2013), or idioms thought to be semantically frozen and syntactically fixed, such as *kick the bucket* 'die' (e.g., *no buckets were kicked, reluctant to kick his brimming bucket of life*, and *my phone kicked the pail last week*). These studies demonstrate that idioms are not nearly as fixed as previously assumed. This variability in idioms, and MULTIWORD EXPRESSIONS (MWEs) more generally, is acknowledged as a key challenge in the automated identification of MWEs in corpora (cf. Savary et al. 2018 [this volume] and Scholivet et al. 2018 [this volume]).

Few studies have investigated idiomatic variation from an experimental perspective. Gibbs and colleagues (Gibbs et al. 1989; Gibbs & Nayak 1989) explored lexical and syntactic variation of decomposable and nondecomposable idioms using a semantic similarity rating task. They found that decomposable idioms (i.e. idioms whose constituents contribute to the meaning of the whole, as in *pop the question* 'propose marriage') were rated as more similar in meaning to a literal paraphrase than were nondecomposable idioms, or idioms whose constituents do not contribute meaning, as in *kick the bucket.* But as Duffley (2013) has shown, nondecomposable idioms can be modified in context and still retain their meaning. In addition, the semantic decomposability measure used by Gibbs and colleagues has not proven a reliable measure, with participants performing at chance in replication studies (Titone & Connine 1994b, Tabossi, Fanari & Wolf 2008). Replication studies have also shown inconsistent results – decomposable and nondecomposable idioms are not always found to be statistically different (Tabossi et al. 2008). Finally, a measure of semantic similarity between an idiom variant and a literal paraphrase may not be the best method for determining the comprehension of idiomatic variation. Semantic similarity has been shown to be predicted by the same local contexts as observed in corpora (Miller & Charles 1991), suggesting that this measure may simply be reflecting how interchangeable the variant is with its paraphrase, or how acceptable the variant may be at conveying the meaning in the paraphrase.

Meanwhile, McGlone et al. (1994) explored the semantic productivity of idiom variation. Variants in this study produced an enhanced idiomatic meaning based on the original (e.g. *shatter the ice*, from *break the ice*, meaning ‘to break an uncomfortable or stiff social situation in one fell swoop’). In a self-paced reading study, they measured the reaction time for participants to read the final sentence of a story, which contained idioms, variants, or literal paraphrases. They found that participants were significantly faster at reading the canonical form of the idiom, but that variants were read as fast as literal paraphrases. They suggest that canonical forms of idioms are accessed whole, while variants are processed like literal language and are therefore processed slower. However, they did not control for the type of variation used in their study. They included instances of lexical variation (e.g. *shatter the ice*), quantification (e.g. *not spill a single bean* [*spill the beans* ‘reveal secret information’]), and hyperboles (e.g. *it’s raining the whole kennel* [*rain cats and dogs* ‘rain very hard’]). It is unclear whether some types of variants are easier to comprehend than others.

The current study attempts to improve upon these previous studies. We explore the acceptability and processing of several types of idiom variants through a rating task and an eye-tracking experiment, respectively. While both of these experiments have been presented independently elsewhere – the eye-tracking study was presented in Geeraert, Baayen, et al. (2017) and part of the acceptability ratings study was presented in Geeraert, Newman, et al. (2017), but presented in full below – they have been brought together here in a multi-methodological study in order to tease apart and contrast speaker judgements from online processing. Previous research has sometimes conflated these two methods, making interpretation difficult (cf. Gibbs et al. 1989; Gibbs & Nayak 1989). But here we distinctly separate them, utilizing an online measure of processing in addition to a subjective evaluative measure, to determine any converging or diverging results between the two methods, which are important for understanding idioms and idiomatic variation. These two studies utilize the same data, yet are different perspectives. Thus, this chapter provides a complete reportage of the larger study, a discussion of the variables useful for predicting each modality (with suggestions as to why), as well as a unique perspective in the idiom literature.

Two main research questions are explored. First, how do variants compare with the canonical form? This question explores whether differences between the canonical form and the variants are still present when the type of variation is controlled. Second, how do variants compare with each other? This question explores whether any differences emerge between the variant types – are certain variant types more preferred or easier to process?

This study is largely exploratory, but we do have some predictions about the results. For example, formal idiom blends are often regarded in the literature as being errors or slips of the tongue (Fay 1982; Cutting & Bock 1997). We therefore hypothesized that blends would be least preferred and more difficult to process due to this perceived error-like nature. Meanwhile, some idioms have been shown to occur in "idiom sets" (Moon 1998), such as *shake/quake/quiver in one's boots* 'tremble with apprehension' or *down the drain/chute/tube/toilet* 'completely lost or wasted'. Given this, we predict that lexical variation will not be more difficult to understand than the canonical form, and may be considered an acceptable variant strategy. A modifier inserted into the idiom should take additional time to process due to the presence of the extra word, but given their relative frequency and overall productivity in corpora (Moon 1998; Schröder 2013), may be the most preferred variant. Lastly, a partial or truncated form will likely be faster to process, due to the omission of a word, but may not be widely accepted due to their limited occurrence in corpora (Moon 1998).

The remainder of the chapter proceeds as follows: We discuss each experiment in turn, beginning with the acceptability ratings, and then the eye-tracking experiment. We discuss the design of each experiment and the results obtained. We conclude with a discussion of our findings, how the results of the two experiments converge or diverge, as well as how the results fit into the larger discussion on idioms, and specifically within a probabilistic approach to language.

2 Acceptability rating experiment

2.1 Methodology

2.1.1 Materials

Sixty idioms were extracted from the Oxford Dictionary of English Idioms (Ayto 2009) and the Collins COBUILD Idioms Dictionary (Sinclair 2011), listed in Appendix A. These idioms varied in length and syntactic structure: 20 three-word idioms consisting of a verb and a noun phrase (V-NP, e.g. *rock the boat*); 20 four-word idioms consisting of a verb and a prepositional phrase (V-PP, e.g. *jump on* *the bandwagon* 'join others doing something fashionable'); and 20 five- or six-word idioms (10 each) consisting of a verb, noun phrase, and a prepositional phrase (V-NP-PP, e.g. *hear something through the grapevine* 'hear gossip'). Two contexts were created for each idiom: one literal and one figurative (e.g. *I used to pretend I could talk to plants, and I would hear things through the grapevine* = literal; and *I used to be a socialite, and I would hear things through the grapevine* = figurative). Both contexts had identical final clauses, with the idiom in sentence-final position. As syntactic variation is possible with idioms (Moon 1998; Schröder 2013), the contexts were not restricted to the present tense.

The form listed in the dictionary was regarded as the canonical form (for a different approach to establishing canonical forms of MWEs (see Savary et al. 2018 [this volume]). If more than one form was listed then the form most familiar to the first author was used, as she spoke the same variety as the participants in the study. In addition to the canonical form, these idioms were manipulated for four types of variation within the figurative context (i.e. the context was identical for all variants). First, lexical variation, where one of the lexical items within the expression was altered to a synonymous or near-synonymous word (e.g. *discover something through the grapevine*). Synonyms were selected based on their naturalness in the context to convey a similar meaning.[1] Second, partial form of the idiom, where only a portion of the idiom was presented, usually a key word or words (e.g. *use the grapevine*). In order for the sentence to still be grammatically correct, pronouns or lexically-vague words replaced the missing elements of the expression, such as *it, them, things* for nouns, or *have, be, do, use* for verbs. Third, integrated concept, where an additional concept was integrated into the idiom (e.g. *hear something through the judgemental grapevine*). These additional concepts expanded or emphasized the figurative contexts in which the idiom occurred. Finally, formal idiom blend, where two idioms were blended together (e.g. *get wind through the grapevine* – blending *hear something through the grapevine* with *get wind of something* 'hear a rumour'). Each "experimental" idiom (i.e. the 60 idioms selected) was paired with a non-experimental idiom for use in the idiom blend condition. These paired "blending" idioms were chosen for their intuitive plausibility, but controlled for their syntax and semantics (Cutting & Bock 1997). Four types of blends were created: same syntax, similar semantics (sSYN, sSEM); same syntax, different semantics (sSYN, dSEM); different syntax, similar semantics (dSYN, sSEM); and different syntax, different semantics (dSYN, dSEM), illustrated in Table 1. Five instances of each type of blend occurred with the three syntactic types (i.e. V-NP, V-PP, or V-NP-PP), totalling 15 of each blend

[1] An online thesaurus (http://www.thesaurus.com) was often utilized for synonymous words.

type. There is clearly a linguistic playfulness at work in the creation of the idiom blends in Table 1, just as there is in many of the non-canonical idiom forms found in naturally occurring usage (cf. Moon 1998; Duffley 2013). This playfulness, it should be noted, presents a special challenge to modelling of MWEs in the context of NLP work on multiword expressions or annotation of MWEs in corpora, as noted in Savary et al. (2018 [this volume]). Indeed, Savary et al. (2018 [this volume]) consider "wordplay proneness", as they call it, a challenge that "goes far beyond the state of the art in semantic modelling and processing of VMWEs [verbal MWEs]".

Table 1: Four types of blends used in the idiom blend condition.

Type of blend	Example	Source idioms	Total
sSYN, sSEM	*rock the applecart*	*rock the boat* *upset the applecart*	15
sSYN, dSEM	*shoot your tongue*	*shoot the breeze* *hold your tongue*	15
dSYN, sSEM	*pass the mustard*	*cut the mustard* *pass muster*	15
dSYN, dSEM	*face the wringer*	*face the music* *put through the wringer*	15

Half of the idioms had the beginning portion of the expression altered (verb), while the other half had alternations made to the final portion of the expression (noun). In total, there are six conditions: one in a literal context and five in a figurative context (i.e. one canonical form and four variants). The experiment utilized a Latin-square design, where every participant saw each idiom once in one of the six conditions. Six versions of the experiment were created, each one containing 10 idioms in each of the six conditions.

Conditions:

1. **Literal meaning** of the idiom in its canonical form
(*While the guys were reshingling, they suddenly* went through the roof.)
2. **Canonical form** of the idiom in a figurative context
(*Although these were new stocks, they suddenly* went through the roof.)
3. **Lexical variation** of the idiom in a figurative context
(*Although these were new stocks, they suddenly* went through the ceiling.)

4. **Partial form** of the idiom in a figurative context
(*Although these were new stocks, they suddenly* went through it.)
5. **Integrated concept** within the idiom in a figurative context
(*Although these were new stocks, they suddenly* went through the investment roof.)
6. **Idiom blend** of two idioms in a figurative context
(*Although these were new stocks, they suddenly* went through the charts.)

Since the blending idioms only occurred in one condition (i.e. idiom blend), they were used as fillers in their canonical form in the other five versions of the experiment, occurring in either a figurative or literal context. Each blending idiom was excluded as a control in the version of the experiment where it occurred in the idiom blend condition in order to avoid a bias in the materials. Therefore, in each version of the experiment, 10 of the blending idioms occurred in the idiom blend condition, while the remaining 50 appeared as fillers. Of these fillers, 20 occurred in a figurative context and 30 occurred in a literal context. This was done to increase the number of literal contexts in the experiment so that they were not so underrepresented. In sum, each participant saw 110 items: 60 experimental idioms (10 in each condition) and 50 blending idioms as fillers.

Finally, six practice sentences were created using a different six idioms. These idioms all occurred in their canonical form. Three were in a figurative context and three in a literal context. These were the same for all participants.

2.1.2 Procedure

Using E-prime 2.0 standard edition software, each sentence was presented in random order at the top of the computer screen. The text was presented in a black, bold, 24-point Courier New font, centered on a white background. Below each sentence was a VISUAL ANALOGUE SCALE (VAS), which is a continuous graphical rating scale that allows fine gradations to be measured (Funke & Reips 2012).

Participants were explicitly told that they would be reading sentences containing English expressions, but that some of the expressions had been modified in various ways. They were asked to rate the acceptability of the expression, as it occurred in the sentence, by clicking the mouse anywhere on the provided scale, which was labelled with "acceptable" on the right and "unacceptable" on the left. The mouse was repositioned to the centre of the scale on each trial. Participants were encouraged to use the whole scale before the experiment began, and were given an opportunity to take a short break halfway through the experiment.

After the participants had completed the rating task, they were asked whether they knew each idiom. As different speakers are familiar with different subsets of idioms, this information allowed us to control, at the level of the individual, whether they knew the idiom (Cacciari et al. 2005), while maximizing the number of idioms in the study. Each idiom appeared, in its canonical form, in a black, bold, 22-point Courier New font, centered on a white background. Above the idiom was the question "Do you know this expression?" and below were two boxes, one labelled "yes" and the other labelled "no". Using the mouse, participants clicked on the appropriate box to respond. The mouse repositioned itself to the center of the screen on each trial.

At the end of this second task, participants were presented with a few additional questions pertaining to their idiom usage (e.g. How often do you use these expressions?, Do you like using these expressions?). Participants responded to these questions using the same VAS scale as the rating task, this time labelled with a thumbs-up image on the right for a positive response and a thumbs-down image on the left for a negative one. Lastly, participants were asked to rate the acceptability of seven prescriptively "incorrect" sentences (LQs), shown below, also using this VAS scale. These sentences attempted to elicit a measure of the participant's flexibility with language and non-standard usage.

Language questions (LQs):

1. The only option the school board has is to lay off a large *amount* of people.
2. Slot machines are thought to be more *addicting* than table games.
3. The document had to be signed by both Susan and *I*.
4. While cleaning the kitchen, Sally looked up and saw a spider on the *roof*.
5. I thought it could*'ve went* either way.
6. She *could care* less what he had to say about it.
7. You have to balance your life, *irregardless* of what anybody thinks.

2.1.3 Participants

Forty-eight undergraduate linguistics students from the University of Alberta participated in this experiment. All participants were native speakers of English. There were 37 female and 11 male participants, ranging from 17–43 years of age. All participants were reimbursed for their time with course credit.

2.2 Results

The results were analyzed with mixed-effects linear regression using the `lme4` package (Bates et al. 2015) in `R` (R Core Team 2012). We focus on two analyses:

the rating responses and the rating reaction times. Only the 60 experimental idioms were included in these analyses (i.e. the fillers were not included outside of the idiom blend condition).

Six predictor variables are discussed below. `Condition` is a factor indicating the condition in which the idiom occurred (e.g. canonical form, lexical variation, idiom blend). `Length` specifies the number of words within the idiom’s canonical form. `PairedSemantics` is a factor specifying whether the two idioms used in the formal idiom blend have similar or different semantics (e.g. *spill the beans & let the cat out of the bag* ‘reveal a secret’ = similar; *shoot the breeze* ‘have a casual conversation’ & *hold your tongue* ‘remain silent’ = different). Meanwhile, `KnowIdiom` is a factor indicating the participant’s knowledge of the idiom (i.e. yes or no). And `Trial` is the scaled (i.e. standardized) order of presentation of the stimuli in the experiment. Since the stimuli was presented randomly, this order will be different for each participant.

`meanTransparencyRating` is the scaled average rating for the transparency (or clarity) of the idiom’s meaning as a whole. Since speakers differ in how they interpret the decomposability (i.e. compositionality) of idioms, as evidenced by the low reliability of the decomposability classification task (Titone & Connine 1994b, Tabossi, Fanari & Wolf 2008), we were interested in a measure for how clear or obvious people find the meaning of the idiom as a whole. This measure then, may provide some indication of how literal or figurative people consider an idiom. These ratings were collected in a separate experiment, specifically designed to elicit ratings of transparency. Those participants saw each idiom, along with a definition and an example sentence, and were asked to rate the transparency of the idiom (for further details, see Geeraert 2016). The average rating for each idiom was included as a separate predictor to determine whether transparency influences people’s preferences of variation.

2.2.1 Rating responses

The first model examines the rating responses. The fixed effects of this model are shown in Table 2. This model has three significant interactions with `Condition`. The first, between `Condition` and `KnowIdiom`, is shown in the top-left panel of Figure 1. As expected, participants are not sensitive to variation when an idiom is unfamiliar. But when the idiom is known, there is a clear preference for the canonical form. Two variants types, integrated concepts and lexical variation, are rated as more acceptable than the others, with a slight preference for variants which have an additional concept inserted into the idiom. The remaining variants: idiom blends, partial forms, and a literal reading of the idiom, are all rated as the least preferred variants.

Table 2: Fixed effects for the acceptability rating responses.

	Estimate	Std. Error	t-value	ΔAIC
Intercept	88.81	6.54	13.59	
Condition=Concept	-19.68	7.34	-2.68*	187.74
Condition=Blend	-37.56	7.36	-5.10*	
Condition=Lexical	-22.69	7.36	-3.08*	
Condition=Literal	-46.46	7.36	-6.31*	
Condition=Partial	-45.25	7.37	-6.14*	
KnowIdiom=No	-30.17	3.58	-8.43*	30.24
Length	-2.02	1.49	-1.35	1.98
meanTransparencyRating	3.82	1.91	2.00	18.30
Trial	1.77	0.69	2.58	4.26
I(KnowIdiom=No|Condition=Concept)	13.66	4.92	2.78*	52.10
I(KnowIdiom=No|Condition=Blend)	31.76	4.74	6.71*	
I(KnowIdiom=No|Condition=Lexical)	23.31	4.94	4.71*	
I(KnowIdiom=No|Condition=Literal)	31.26	4.74	6.59*	
I(KnowIdiom=No|Condition=Partial)	22.63	4.85	4.66*	
I(Length|Condition=Concept)	1.09	1.72	0.64	2.63
I(Length|Condition=Blend)	2.52	1.71	1.48	
I(Length|Condition=Lexical)	0.52	1.71	0.31	
I(Length|Condition=Literal)	4.63	1.71	2.71*	
I(Length|Condition=Partial)	4.11	1.71	2.40	
I(meanTransparencyRating|Condition=Concept)	0.72	2.25	0.32	1.32[a]
I(meanTransparencyRating|Condition=Blend)	2.01	2.22	0.90	
I(meanTransparencyRating|Condition=Lexical)	3.59	2.31	1.56	
I(meanTransparencyRating|Condition=Literal)	6.35	2.27	2.80*	
I(meanTransparencyRating|Condition=Partial)	0.37	2.31	0.16	

* = Factors that remain significant after a Bonferroni correction

[a] An ANOVA test run during model comparison indicates that the inclusion of this interaction is significant (p = 0.045).

`Length` also occurs in a significant interaction with `Condition`, shown in the top-centre panel of Figure 1. Participants tend to rate idioms as less acceptable in their canonical form if they are longer. This pattern holds for most variants as well: integrated concepts, lexical variation, and formal idiom blends have slopes which are not significantly different from the canonical form and are therefore depicted in grey. Literal meanings and partial forms however are rated as more acceptable if the idiom is longer. Apparently, literal interpretations (which likely may also characterize partial forms) benefit from the presence of many words.

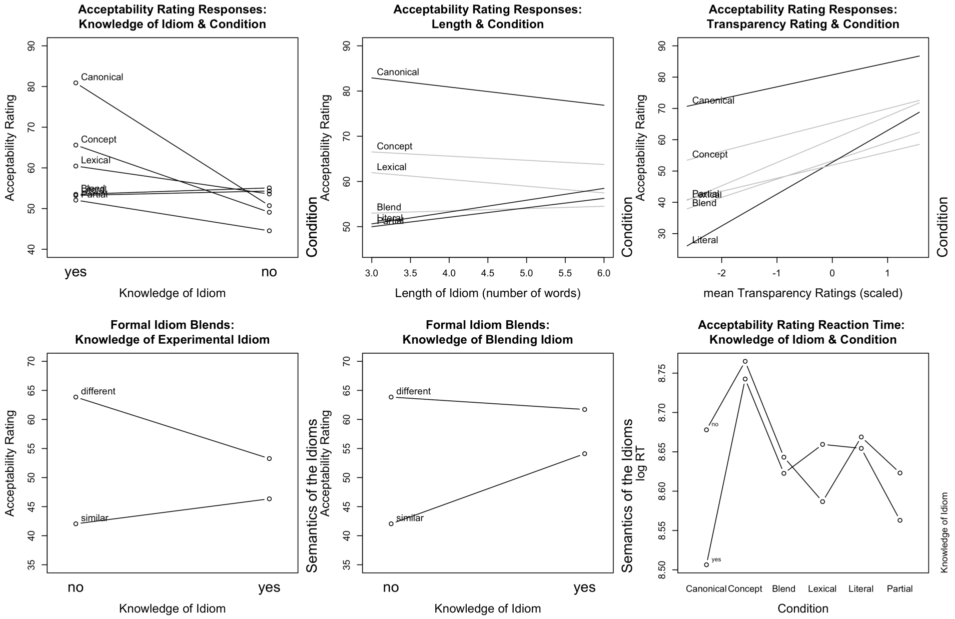


Figure 1: Interactions in the mixed-effects linear regression models for the acceptability rating responses and reaction times. Lines in grey represent factors levels which are not significantly different.

This might suggest that as expressions become longer, the non-idiomatic reading becomes stronger and begins to interfere with the idiomatic reading.

The last interaction, between `meanTransparencyRating` and `Condition`, is illustrated in the top-right panel of Figure 1. Higher ratings of acceptability are given to idioms judged to be more transparent. For the condition in which the context enforced a literal reading, the effect of transparency was stronger than for any of the other idiom variants. This result is not unexpected, given that not all idioms have a plausible literal meaning (Titone & Connine 1994b).

`Trial` was significant as a main effect; participants became more accepting of the stimuli (both variants and the canonical form) the further they advanced through the experiment. But participants differed in exactly how much more accepting they became, as evidenced by the by-Subject random slopes for Trial. These slopes in the random effects structure are in addition to `Subject` and `Idiom` included as random effects.[2] Finally, it is interesting to note that the frequency or syntax of the idiom, as well as whether modifications were made to the verb or the noun, did not affect the acceptability of idioms or variants.

[2] The rating response model and the RT model show the same random effects structure.

We also looked specifically at formal idiom blends, given their error-like status in the literature (Fay 1982; Cutting & Bock 1997), in order to explore whether any factors influence their acceptability. Two interactions appear significant, shown in Table 3: both between the participant's knowledge of an idiom and the paired semantics of the two idioms involved in the blend. The bottom-left panel in Figure 1 shows the interaction with knowledge of the experimental idiom, while the bottom-centre panel shows the knowledge of the blending idiom. Both interactions indicate that blends are rated as more acceptable when the meanings of the two idioms differ, and less acceptable when they are similar. Participants significantly rate blends with similar semantics with a lower acceptability if one of the idioms is unknown. A three-way interaction between these variables (i.e. knowledge of both idioms and the semantic similarity of the idioms) is not significant, suggesting that speakers only need to be unfamiliar with one of the idioms to regard semantically similar idiom blends as less acceptable. The noticeability of the unknown idiom likely causes this increase in unacceptability, which is perhaps not as noticeable for those who are familiar with both blended idioms – presumably, they are able to access the meaning of the blend easier, as they are familiar with both idioms from which the parts belong, and therefore are not as surprised or unimpressed by the blend. Finally, `meanTransparencyRating` is also significant in this model – speakers prefer idiom blends that are more transparent and clearer in meaning.

Table 3: Fixed effects for the acceptability ratings of idiom blends.

	Estimate	Std. Error	t-value	ΔAIC
Intercept	63.62	6.06	10.50	
meanTransparencyRating	5.21	2.30	2.26	3.00
KnowExpIdiom=Yes	-10.58	4.64	-2.28*	1.02
KnowBlendingIdiom=Yes	-2.13	4.77	-0.45	0.59
Semantics=Similar	-21.80	7.43	-2.93*	2.00
I(Semantics=Similar|KnowExpIdiom=Yes)	14.87	6.47	2.30*	3.23
I(Semantics=Similar|KnowBlendingIdiom=Yes)	14.19	6.50	2.18*	2.74

* = Factors that remain significant after a Bonferroni correction

2.2.2 Rating reaction times

We also analyzed the REACTION TIMES (RTs) for how quickly the participants responded to the acceptability rating task. Faster reaction times are associated with easier judgements of acceptability. The fixed effects for this model are shown in Table 4. Only one interaction, between KnowIdiom and Condition, is significant in this model, illustrated in the bottom-right panel in Figure 1. The RTs associated with each condition are similar for both those who know the idiom and those who do not. Significantly longer RTs are observed with integrated concepts, while significantly shorter RTs are observed with partial forms. These results may simply reflect the fact that the integrated concept condition has an additional word inserted into the idiom, whereas the partial form condition has a word omitted from the expression. This RT difference likely corresponds to length of the expression. The most notable observation is that participants are significantly faster rating the canonical form of the expression if the idiom is known. If the idiom is unknown, the RT to rate the canonical form does not differ significantly from the other variants. These results illustrate that the canonical form has an advantage if it is familiar, but that variants of the same length as the canonical are rated as quickly as if one does not know the expression.

Table 4: Fixed effects for the acceptability rating reaction times.

	Estimate	Std. Error	t-value	ΔAIC
Intercept	8.51	0.04	226.18	
Condition=Concept	0.24	0.02	9.65*	93.85
Condition=Blend	0.12	0.02	4.70*	
Condition=Lexical	0.15	0.02	6.22*	
Condition=Literal	0.15	0.02	5.92*	
Condition=Partial	0.06	0.02	2.27	
KnowIdiom=No	0.17	0.04	4.40*	1.69
Trial	-0.08	0.01	-8.16	39.96
I(KnowIdiom=No|Condition=Concept)	-0.15	0.05	-2.82*	13.10
I(KnowIdiom=No|Condition=Blend)	-0.15	0.05	-2.91*	
I(KnowIdiom=No|Condition=Lexical)	-0.24	0.05	-4.67*	
I(KnowIdiom=No|Condition=Literal)	-0.16	0.05	-3.08*	
I(KnowIdiom=No|Condition=Partial)	-0.11	0.05	-2.18	

* = Factors that remain significant after a Bonferroni correction

In sum, this experiment explored the acceptability of idiomatic variation, using several types of variants. The canonical form is the most preferred and participants are quicker at rating this form, but only when the expression is known. Modifying an idiom makes it less acceptable, but the decrease in acceptability varies according to the type of alternation – integrating an additional element (*go through the investment roof*) or replacing a word with a near-synonym (*go through the ceiling*) were considered more acceptable variants. We now turn our attention to how these variants are understood.

3 Eye-tracking experiment

3.1 Methodology

3.1.1 Materials

This experiment utilized the same materials as the previous experiment.

3.1.2 Procedure

This experiment used the Eye-Link 1000, desk-top mounted video-based eye-tracking device, manufactured by SR Research. The eye-tracker sampled the pupil location and size at a rate of 1000Hz, and was calibrated using a 9-point calibration grid. Calibration occurred at the beginning of the experiment, after the practice, and again after every 22 sentences, for a total of five blocks. The computer screen resolution was set to 1920 x 1080 pixels.

The stimuli were presented in two parts. Participants first saw the "context clause" (e.g., *Although these were new stocks,*), followed by the "idiom clause" (e.g. *they suddenly went through the roof.*) on a separate screen. Each trial began with a fixation cross presented for 1,000 msec on the left side of a light-grey screen. Next, they saw the context clause, also on a light-grey background, in a bold, black, Courier New 30-point font. Every clause was displayed in full and fit on one line. To exit this screen, participants had to trigger an invisible boundary in the bottom right corner. A blank, light-grey screen was presented for 1,000 msec before the fixation cross preceding the idiom clause appeared. The sequence of screens for the idiom clause was identical to the context clause.

Ten percent of the stimuli were followed by a true/false comprehension question, which pertained to the immediately preceding sentence, and were presented randomly throughout the experiment. Participants pushed one of two buttons on a game controller to answer these questions, which were clearly labelled on the question screen. The experiment began with a practice session, which consisted of six practice sentences and three questions. These were the same for all participants, although their order varied.

All participants had normal or corrected-to-normal vision. The right eye of each participant was tracked. Participants sat approximately 85cm from the computer screen, with the camera placed on the desk about 35cm in front of the computer screen. The participants sat in a sound-proof booth, while the experimenter sat outside the booth, running the experiment. The lights were kept on. The experiment was self-paced and took about 45 minutes to complete. Each participant was given an opportunity for a short break half-way through the experiment.

After the participants had completed the eye-tracking portion, they were asked to complete three additional tasks: (1) to indicate their knowledge of each expression; (2) to answer questions pertaining to their idiom usage; and (3) to rate the acceptability of the seven prescriptively "incorrect" sentences (LQs). These tasks were identical to the ones in the acceptability rating experiment.

3.1.3 Participants

Sixty linguistics undergraduate students from the University of Alberta participated in this experiment. All were native speakers of English, and all were different participants than those who participated in the previous study. There were 43 female and 17 male participants, ranging from 17–29 years of age. All participants were reimbursed for their time with course credit.

3.2 Results

The results were analyzed using mixed-effects linear regression. We focus on the total fixation duration (i.e. the total amount of time spent fixating on the AREA OF INTEREST, or AOI) within two AOIs: the idiom as a whole (i.e. the summed fixations on all words within the idiom) and the altered word within the idiom (i.e. the synonymous word in lexical variation, the additional word in the integrated concept, the semantically vague "replacement" word in partial forms, and the word from another idiom in the idiom blend). As above, the analyses only include the 60 experimental idioms.

Ten predictor variables appeared significant in the models. `Condition`, `Know Idiom`, `Length`, `meanTransparencyRating`, and `Trial` are the same variables used in the previous experiment. `Gender` is a factor specifying whether the participant is male or female. `PortionAltered` is a factor specifying which part of the idiom (i.e. beginning/verb or ending/noun) was manipulated in the variant. And meanVariationRating is a scaled mean measure of acceptability for a particular idiom with a each type of variation – these averaged ratings were extracted from the previous experiment and included here to determine if participants' preferences influence their ease of comprehension.

Two measures reflecting the semantic contribution of the constituents were utilized in analyzing these results. meanTransparencyRating (described above) and LSA.Score.Paraphrase, which is a measure of similarity using LATENT SEMANTIC ANALYSIS (LSA), between the words in the idiom and its paraphrase (e.g., *spill the beans* 'reveal a secret'). This score was obtained from a pairwise comparison of two texts (i.e. an idiom and its paraphrase), which compares the local contexts in order to obtain a value of similarity (Landauer et al. 1998).[3] This measure allows us to control for the idiom's compositionality. If the exact words in the idiom have little to do with the expression's meaning, then the LSA score will be small (e.g., *cut the mustard* 'be acceptable' = 0.07). But if the words used share meaning or contribute to the idiom's meaning, then the LSA score will be larger (e.g., *stop something in its tracks* 'stop something' = 0.87).

As idioms are MWEs, multiple frequency measures were obtained: the frequency of the idiom, frequencies of the individual words, and all possible combinations of adjacent words (e.g. word1 and word2; word2 and word3; word1 and word2 and word3). To avoid collinearity, a PRINCIPAL COMPONENTS ANALYSIS (PCA) was conducted on these frequency measures. Only the first principal component (henceforth PC1.logFrequency) is significant. Finally, a second PCA was conducted on the rating responses for the seven LQs above. Only PC2 (henceforth PC2.LQ) was significant. This latent variable may reflect the participant's flexibility with language usage.

3.2.1 Idiom as AOI

The first model examines the summed fixation durations on the idiom as a whole. The fixed effects for this model are shown in Table 5. The first interaction, between Condition and KnowIdiom, is shown in the left panel of Figure 2. The canonical form, and the majority of variants, show the same general pattern: shorter fixation durations on known idioms. These variants (except integrated concepts) are therefore shown in grey, as they do not significantly differ from the canonical form. Partial forms however show a different pattern. Fixation durations are relatively similar regardless of whether the participant is familiar with the ex-

[3] The LSA scores were obtained from the English Lexicon Project (Balota et al. 2007), available at http://lsa.colorado.edu.

pression or not; thus a facilitation effect for knowing the idiom is not observed as it is with the other variants. This particular variant is fixated upon less than the canonical form, likely due to it being shorter in length (i.e. fewer number of words). This is in line with longer fixations observed on integrated concepts – an additional word is integrated into the idiom, making it longer in length and requiring additional fixations.

Table 5: Fixed effects for the idiom as AOI.

	Estimate	Std. Error	t-value	ΔAIC
Intercept	6.71	0.09	75.97	
Condition=Concept	0.49	0.10	5.04*	130.12
Condition=Blend	0.08	0.10	0.75	
Condition=Lexical	0.01	0.10	0.05	
Condition=Literal	-0.19	0.10	-1.94	
Condition=Partial	-0.75	0.16	-4.80*	
KnowIdiom=Yes	-0.18	0.04	-4.32*	34.84
Length	0.11	0.02	6.76	40.19
PortionIdiomAltered=Ending	-0.06	0.02	-2.52*	3.50
PC2.LQ	-0.07	0.03	-2.42	3.60
LSA.Score.Paraphrase	0.24	0.07	3.49	8.21
meanVariationRating	-0.06	0.01	-7.23	43.80
Gender=Male	-0.17	0.08	-2.17	2.53
TrialScaled	-0.04	0.01	-3.78	10.80
I(KnowIdiom=Yes|Condition=Concept)	0.06	0.05	1.16	1.26
I(KnowIdiom=Yes|Condition=Blend)	0.08	0.06	1.42	
I(KnowIdiom=Yes|Condition=Lexical)	0.08	0.06	1.52	
I(KnowIdiom=Yes|Condition=Literal)	0.03	0.06	0.55	
I(KnowIdiom=Yes|Condition=Partial)	0.17	0.06	2.75*	
I(Length|Condition=Concept)	-0.05	0.02	-2.62	14.11
I(Length|Condition=Blend)	-0.01	0.02	-0.36	
I(Length|Condition=Lexical)	0.00	0.02	0.20	
I(Length|Condition=Literal)	0.02	0.02	1.04	
I(Length|Condition=Partial)	0.08	0.03	2.48	

* = Factors that remain significant after a Bonferroni correction

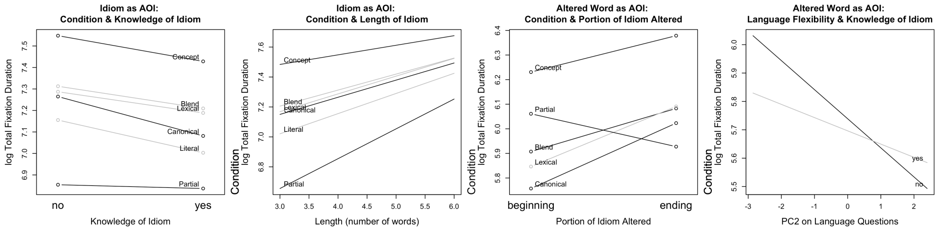


Figure 2: Interactions in the mixed-effects linear regression models for the summed total fixation duration on the whole idiom and the altered word as an AOI. Lines in grey represent factor levels which are not significantly different or slopes which are not significant.

The second interaction, shown in the second panel of Figure 2, is between Condition and Length. Longer idioms show longer summed fixation durations, as expected, due to the increased number of words in the idiom. Lexical variation, formal idiom blends, and the literal meaning of the idiom are not significantly different from the canonical form (shown in grey). The other two variants show a pattern that is significantly different from the canonical form. Integrated concepts show a slight inhibitory effect of length, where an additional concept is more difficult to integrate into shorter idioms (i.e. extra time is required). Whereas partial forms of shorter idioms have even fewer words to fixate upon and therefore show considerably shorter fixation durations. Thus, durations on integrated concepts and partial forms are more comparable to the canonical form when the idiom is longer.[4]

Interestingly, the literal meaning of the idiom shows shorter fixation durations than the canonical form, albeit not quite significantly shorter (t = -1.94). The literality of the expression (Titone & Connine 1994a) may be contributing to this result. Nevertheless, a general pattern is evident based on these two above interactions with Condition: variants of the same length as the canonical form are not processed significantly different from this canonical form.

Six main effects are observed in this model. Longer fixation durations are observed on the whole idiom if the beginning (the verb) was altered (i.e. Portion Altered). This is not dependent on the type of variation; all variants are easier to process if the change comes later in the idiom. This is a different result than

[4] PC1.logFrequency was also significant in the idiom as AOI model. However, this variable is strongly correlated with Length (r = -0.9). This correlation is unsurprising given that PC1.logFrequency was created using adjacent co-occurrence frequencies. Model comparison shows that Length is the more significant predictor in this model, producing a considerably lower AIC value, and therefore was retained at the expense of PC1.logFrequency.

that of Gibbs and colleagues (Gibbs et al. 1989; Gibbs & Nayak 1989) who found no difference with similarity ratings in whether the noun or verb was altered.

MeanVariationRating is also significant. Variants which received higher acceptability ratings are fixated on less long, suggesting preferred variants are easier to understand and interpret (or perhaps variants easier to interpret are preferred). Longer fixation durations appear on idioms which have higher LSA scores for the idiom's paraphrase (i.e. LSA.Score.Paraphrase). This finding seems initially surprising, as previous analyses on the comprehension of idioms suggest that idioms are easier to understand when the individual components contribute meaning to the whole (Gibbs et al. 1989). However, the LSA scores indicate how similar the local contexts are between the idiom and its paraphrase (i.e. how interchangable is the idiom and its paraphrase). When the LSA score is high (i.e. the paraphrase is easily interchangable), looking time increases as the contexts are not distinctive for the idiom. But if the LSA score is low, then the idiom and its paraphrase are less interchangable, making the context more distinctive and the idiom more predictable. Interestingly, meanTransparencyRating is not significant. The degree to which the idiom is considered obvious in meaning does not seem to influence the comprehension of idioms or variants.

A main effect is also observed for PC2.LQ, a latent variable representing the participants' "flexibility" with language (i.e. the more they consider nonstandard or erroneous forms acceptable). Shorter fixations are observed on idioms, both the canonical form and variants, if speakers are more flexible with language. It is interesting to note that this finding is not restricted to only the variants. Gender also shows a significant main effect – males tend to fixate less long on the idiom than females, although we are not quite sure why. Finally, a main effect of Trial is also significant; participants fixate less long on the idiom the further into the experiment they get. But the degree to which each participant is affected by the order of presentation varies, as evidenced by significant by-Subject random slopes for Trial.[5] By-Item random slopes for Condition with correlation parameters are also significant in this model. These slopes indicate that participants' fixation durations vary depending on which idiom occurred in which condition – participants found certain idioms easier or more difficult to understand depending on the condition in which they occurred.

3.2.2 Altered word as AOI

We next investigate the fixation duration on the altered word (i.e. the word in the idiom that was manipulated). The fixed effects for this model are shown in Table 6.

[5] Both idiom as AOI and altered word as AOI models have the same random effects structure.

Table 6: Fixed effects for the altered word as AOI.

	Estimate	Std. Error	t-value	ΔAIC
Intercept	5.70	0.06	98.48	
Condition=Concept	0.47	0.06	8.28*	58.40
Condition=Blend	0.15	0.06	2.67*	
Condition=Lexical	0.09	0.06	1.54	
Condition=Partial	0.30	0.07	4.61*	
PortionIdiomAltered=Ending	0.27	0.06	4.49*	17.88
KnowIdiom=Yes	-0.04	0.03	-1.29	0.21
PC2.LQ	-0.10	0.03	-3.12	2.59
PC1.logFrequency	0.03	0.01	4.70	15.43
meanVariationRating	-0.07	0.02	-4.27	14.09
TrialScaled	-0.04	0.01	-2.79	5.28
I(PortionIdiomAltered=Ending|Condition=Concept)	-0.12	0.08	-1.46	9.81
I(PortionIdiomAltered=Ending|Condition=Blend)	-0.09	0.08	-1.17	
I(PortionIdiomAltered=Ending|Condition=Lexical)	-0.02	0.08	-0.26	
I(PortionIdiomAltered=Ending|Condition=Partial)	-0.40	0.09	-4.42*	
I(PC2.LQ|KnowIdiom=Yes)	0.06	0.02	2.27*	3.09

* = Factors that remain significant after a Bonferroni correction

As there is no altered word in the literal condition, this section focuses on the four idiom variants: lexical variation, partial forms, idiom blends, and integrated concepts, and how they compare to the canonical form.

The interaction between `Condition` and `PortionAltered` is seen in the third panel of Figure 2. The overall pattern is that longer fixation durations occur at the end of the idiom, which is also true for the canonical form. Since the idiom occurs at the end of a sentence, these longer fixations may reflect a sentence wrap-up effect (Rayner et al. 2000; Hirotani et al. 2006). Nevertheless, the altered word for most variants shows significantly longer fixations than the canonical form. This is not true of lexical variation, which is the only variant that does not have significantly longer fixations than the canonical form (t = 1.54). Thus, a lexically altered variant is just as easy to process as the canonical form. Partial forms however, appear considerably different from the canonical form. Longer fixations are observed on the altered word when the beginning has been altered, as in *use the grapevine*. But when the ending is altered (e.g., *spilled it*), fixations on the altered word are not significantly different from the canonical form (t = -1.44). Altering the verb does not always result in significantly longer fixations (cf. the non-significantly different lexical variant when the beginning is altered), however altering the verb to a semantically vague verb (i.e. *be*, *do*, *used* – in order to make the sentence grammatical) does significantly inhibit processing.

The second interaction, shown in the last panel of Figure 2, is between knowledge of the idiom (i.e. `KnowIdiom`) and the participant's flexibility with language (i.e. `PC2.LQ`). Flexibility with language only appears facilitative for those who do not know the idiom, illustrated by the non-significant slope (in grey) for those who know the expression. Other strategies are apparently relied upon to interpret the idiom when knowledge of it is not available.

Additional main effects are also observed. Fixation durations are longer on the altered word when the co-occurrence frequencies of the idiom are higher. Thus, altering part of a more frequent sequence causes greater processing costs. In addition, participants have shorter fixation durations when the variant is rated as more acceptable (i.e. `meanVariationRating`). The more the variation strategy is preferred with a particular idiom, the easier it is to interpret. Finally, the further the participants get into the experiment (i.e. `Trial`), the shorter their fixation durations on the altered word.

We also specifically looked at idioms blends, to determine whether the syntax or the semantics of the two merged idioms affects the processing of this variant. Interestingly, neither of these variables were predictive of fixation duration – we can understand idiom blends regardless of the syntax or the semantics of the two idioms used in the blend.

Some of these alternations may have been surprising to the participants, resulting in effects that continued beyond the altered word. We therefore ran a model to explore any spillover effects from the altered word, shown in Table 7. As the idiom occurred in sentence-final position, spillover effects from an altered noun (i.e. the end of the idiom) are not able to be determined; thus, this model only focuses on spillover effects from an altered verb. We examined the fixation duration on the first content word after the verb when the verb was manipulated (i.e. the alternation occurred at the beginning of the idiom).

Spillover effects are observed for all variant types (i.e. `Condition`), but the longest durations are for integrated concepts and partial forms. Incorporating an additional word into an idiom results in a processing cost likely due to the surprisal of this extra word. Integrating this additional information into the idiom and context requires extra time. The largest spillover effect is with partial forms. It appears that the semantically vague words used in these sentences (to make them grammatical) make these partial forms more difficult to comprehend and cause considerable spillover effects. It remains to be determined whether partial forms from more naturalistic language produce this same effect.

Table 7: Fixed effects for the first content word after the verb.

	Estimate	Std. Error	t-value	ΔAIC
Intercept	5.95	0.08	73.41	
Condition=Concept	0.27	0.07	3.76*	11.6
Condition=Blend	0.17	0.06	2.75*	
Condition=Lexical	0.14	0.05	2.92*	
Condition=Partial	0.30	0.06	4.62*	
PC1.logFrequency	0.04	0.01	3.54	6.38
KnowIdiom=Yes	-0.11	0.05	-2.32*	3.20

* = Factors that remain significant after a Bonferroni correction

The last two effects are `PC1.Frequency` and `KnowIdiom`. The higher the co-occurrence frequencies of the idiom, the longer the fixation duration on the first content word after the alternation. Modifying a frequent multiword sequence inhibits processing. However, these spillover effects are reduced if the idiom is familiar (i.e. `KnowIdiom`).

4 Discussion

This study employed a multi-methodological approach to investigate the acceptability and processing of idiomatic variation. One advantage of using multiple methods is that they can reveal greater insights, by contrasting converging and diverging results between the different methods. Converging results can provide greater confidence that a particular result or predictor variable is robust; whereas diverging results can uncover differences due to a specific modality or shed light on findings concerning the larger picture that would otherwise be overlooked or thought contradictory (Arppe & Järvikivi 2007). The findings between the acceptability rating and eye-tracking experiments presented together in this chapter do in fact show converging and diverging results worthy of discussion.

Interestingly, the findings between the experiments primarily show diverging results with regard to our two research questions. For example, our first research question asks how the variants compare with the canonical form, and we see from the acceptability ratings that the canonical form is rated as more acceptable than variants or a literal reading, with speakers clearly preferring this form. However, the processing differences are not nearly as straightforward. Some variants are processed differently than the canonical form. The variant showing the greatest difference from the canonical form is the partial form of the idiom (e.g., *use the grapevine*). This idiom variant is fixated on less than the canonical form, as expected, largely due to the omission of a word (or words) from the expression. Yet despite this shorter fixation on the whole idiom, participants fixated significantly longer on the “replacement” verbs (i.e. the semantically vague verbs used to connect the idiom to the sentence) and significant spillover effects were observed on the first content word after these verbs. A similar inhibitory effect was not observed if the ending was modified (e.g., *spilled it*). These results are likely due to the design of the experiment. The tightly controlled stimuli used in this study made these partial forms unnatural and difficult to interpret. A study investigating partial forms in naturally occurring language may shed more light on the degree of difficulty for processing this variant.

Idioms with additional concepts integrated into the expression are also processed differently from the canonical form. These variants require additional processing time, as anticipated, but this longer reading time is largely attributable to the extra word in the expression. The longer duration on the whole idiom is very similar to the altered word AOI, suggesting that this variant experiences very little processing costs over and above having to read an extra word.

However, modification of an idiom’s form does not always result in a processing disadvantage. Some variants – lexical variation, formal idiom blends, and a literal reading of the idiom – are not processed significantly slower than the canonical. Differences between these variants and the canonical form are observed, such as longer fixations on the altered word (at least for idiom blends) or some spillover effects if the verb was altered, but these differences do not result in longer reading times for the idiom as a whole. These findings are partly in line with our predictions. Only idiom blends were predicted to be processed slower than the canonical form, due to the potential surprisal at or unrecognizability with this so-called error. But as observed, they do not present difficulties in comprehension. Thus, intentional or not, altering a word within an idiom to a synonymous or non-synonymous word does not result in a processing cost.

Our second research question asks how variants compare with each other. Once again, diverging results between the two methods are evident. Lexical variation and idiom blends are processed quite similarly, showing comparable fixation durations, to each other and to the canonical form. The length of the original idiom is maintained in these variants, possibly explaining these comparable durations. However, they do not share similar acceptabilities. Lexically modified variants are considered much more acceptable than idiom blends. In fact, idiom blends are even less preferred when the two idioms used to make the blend share similar semantics, possibly explaining why blends are often viewed as errors (Fay 1982; Cutting & Bock 1997). Meanwhile, integrated concepts, which add extra information into the idiom, show longer reading times than the other variants, yet are the most preferred. This higher acceptability was expected, given their relatively frequent occurrence in corpora (Moon 1998; Schröder 2013), and leaves us wondering whether semantically productive lexical variants (cf. McGlone et al. 1994) would show higher levels of acceptability (on par with integrated concepts) compared with the synonymous lexical variants utilized here (following Gibbs et al. 1989). Finally, partial forms and a literal reading of the idiom are not acceptable variation strategies, even though they have comparable (or shorter) reading times to the canonical form.

The findings from these two research questions present two main observations. First, variants which add an extra element or are truncated in some way show longer or shorter reading times, respectively, while modifications that maintain the same length as the canonical form show comparable reading times to the canonical form. Second, variants that preserve more of the canonical form (e.g. integrated concepts, lexical variation) are considered more acceptable, although preference remains with the canonical form (which likely facilitates the learning of idioms and leads to faster recognition).

One cautionary note must be made. These aggregated results show patterns and preferences, but they do not imply that all idioms can be altered using all variation strategies. Much variability, particularly when it comes to comprehension, is also observed. Including the mean acceptability for each idiom in each condition as a control variable in the comprehension models resulted in preferred variants showing shorter reading times. In other words, the way in which an idiom is modified can affect how easy it is to understand. Variability is also observed in the random effects structure of the comprehension models, which have by-Item random slopes with correlation parameters for `Condition`, indicating that specific idioms can be easier or more difficult to process depending with which condition they occurred. Thus, while variation is possible and general patterns can be observed, there are also idiom-specific preferences that factor into how an idiom is altered, understood, and appreciated.

Converging and diverging results are also observed with the predictor variables in the analyses. Two variables converge between the two methods. Length is shown to be an important predictor and yet is rarely included in the idiom literature (cf. Fanari et al. 2010). Longer idioms require additional processing time, as expected, and there can be some facilitation or inhibitory processing effects depending on the type of variation encountered. A literal reading gains additional approval when the idiom is longer (and perhaps also more transparent), while shorter idioms are even more preferred with the idiomatic reading. Perhaps the extra words in longer idioms clearly identify the metaphorical links associated with the idiom, making a literal reading also more interpretable.

The participant's knowledge of the expression is another important predictor that converges between the two studies. Participants fixated less on the idiom (i.e. shorter reading times) and were faster to rate the expression when they knew the idiom. They also considered the idiom and its variants as more acceptable when it was familiar. Yet surprisingly, research on idioms tends to include an average measure of familiarity (cf. Titone & Connine 1994b), as a control for frequency or as a measure of subjective familiarity. This study demonstrates that a speaker-specific measure of familiarity is important for idiom research, as it incorporates speaker-specific experiences into the model.

Not all participant-related variables show converging results. The language questions (LQs) that were collected to provide a latent measure of the participant's flexibility with language only appeared significant in predicting the comprehension of idioms, and not their acceptability. Participants who are more flexible (i.e. more accepting of non-standard or erroneous forms) have an easier time processing idioms and variants. This of course makes sense; these speakers are not distracted by the specific form used, but focus solely on the message being conveyed.

Frequency was also not predictive of acceptability. Even highly frequent idioms can be regarded as acceptable when altered. But frequency is predictive of comprehension. This variable only appeared in the altered word model (due to the high correlation with length in the idiom model), and revealed that alternations made to frequent idioms result in a processing cost. When a sequence of words that typically occur together has been modified, additional time is required to interpret the new sequence, as the advantage it once received due to its predictability is no longer available. The opposite pattern is observed for the semantics of formal idiom blends – a significant predictor of acceptability, but not comprehension. Speakers find blends unacceptable when the merged idioms share similar semantics, but appear to have no difficulty interpreting them.

A divergence is also seen with the variable `PortionAltered` (i.e. where in the idiom the alternation occurred: beginning/verb or ending/noun). This variable is not predictive of acceptability – participants' judgements were not affected by where in the idiom the alternation occurred (i.e. modifications to nouns and verbs are equally acceptable). However, this variable is predictive of comprehension –

alternations made earlier in the idiom (the verb) result in greater processing costs. Gibbs et al. (1989) found no difference between modifications made to nouns or verbs in their similarity rating task, providing further confirmation that a subjective rating of similarity is not measuring comprehension. In addition, these results may also provide support for a time-dependent nature of idiom processing (Titone & Libben 2014). As one advances through the idiom, the predictability of the idiom becomes greater and the idiomatic meaning accumulates resulting in greater priming effects for later words. It seems reasonable then that changes made later in the expression will be less costly – the meaning is more predictable even if changes have been made.

Finally, a divergence is also evident between which compositionality measure was determined to be predictive for each modality. A measure of transparency is predictive of the acceptability rating responses; speakers prefer idioms that are transparent and clear in meaning. But an objective measure of contextual similarity is predictive of comprehension. Idioms are faster to process in unique or distinctive contexts (i.e. lower LSA scores), because they are more predictable. Thus, evaluative judgements are influenced by the clarity of the expression, whereas comprehension is affected by the local contexts in which the idiom occurs.

These (largely diverging) results, in regards to the predictor variables, nicely capture patterns between the two methods. Clarity of the expression and motivation for the alternation are important for the acceptability of idioms and variants, whereas the placement of the alternation and the local context (i.e. distinctiveness, as well as disruptions to this context) are important for comprehension.

This study has shown that not all variants are processed significantly different than the canonical form and that the predictability of idioms is important, especially during processing. Yet these findings conflict with traditional views of idioms, which claim that idioms cannot be modified without losing their idiomatic meaning, or that idioms are stored and accessed whole along with their idiomatic meaning, since they do not equal the sum of their parts. These traditional approaches proposed a dual-route model to account for the processing of idioms – literal language would be understood incrementally through ordinary linguistic processing and idioms would be accessed directly along with their meaning (cf. Swinney & Cutler 1979; Cacciari & Tabossi 1988). For instance, they could be activated by accessing the "idiom key" (Cacciari & Tabossi 1988), which is the idiomatic configuration indicating that sufficient input has been received. But how does one receive sufficient input if the form has been altered? One proposal is to store each variant, but this inefficient method of handling variation would result in a large burden being placed on the mental lexicon (Baayen et al.

2013). McGlone et al. (1994) proposed that idioms are accessed whole for faster processing, but that they could be understood through ordinary linguistic processing, while variants are understood like literal language (which is why they are processed slower), using various strategies in order to understand them. But this study showed that not all variants are processed slower – variants of the same length as the canonical are processed just as quickly.

This study also shows the importance of predictability in understanding idioms – when the local context is distinctive, idioms are faster to process; when alternations are made later in the expression, variants are easier to process; when frequent sequences are altered, variants are slower to process; and even the more flexible a speaker is with language, the easier idioms are to process. These results are in line with other aspects of predictability or probability seen elsewhere in language. Idioms that have a higher cloze probability have an idiomatic meaning that is available earlier (Cacciari & Tabossi 1988; Titone & Connine 1994a). The combination of words can lead to certain predictions or expectations (Elman 2011): subject-verb combinations lead to predictions about the upcoming object (e.g. *lumberjack cuts* primes *wood*, whereas *surgeon cuts* primes *bone*), and the type of semantic theme can be predicted based on voice (e.g. *she arrested* primes *crook*, but *she was arrested* primes *cop*). Speakers have even been shown to make accurate probabilistic predictions about the type of syntactic choice made by others; for example, in the dative alternation: *because he brought the pony to my children* vs. *because he brought my children the pony* (Bresnan 2007).

These predictions would not be possible if language was understood in a truly compositional way. Some scholars are therefore challenging the traditional view of the mental lexicon, as a list of dictionary entries, and instead are proposing probabilistic approaches to language, where words do not possess meaning but are simply cues to meaning, modulated by context and experience (Elman 2004; 2011; Ramscar & Baayen 2013). These approaches highlight the vast amount of information speakers have available to them, besides simply the meaning of the word – speakers are able to draw upon past experience, cultural norms, event and world knowledge and even the feelings of the speaker to interpret the meaning being communicated.

In one such framework, Implicit Grammar (Baayen & Ramscar 2015), learning a language is about learning which cues (i.e. sounds, morphemes, words, contexts) are informative, or discriminative, for a particular outcome (i.e. meaning). Thus, learning occurs when cues successfully predict outcomes, but also when predictions fail to result in those outcomes. Under this view, idioms and their variants would be processed similarly to literal language: being a sequence of words which are cues to the intended meaning (cf. Geeraert, Newman, et al. 2017). This is likely why speakers prefer the canonical form – using perfectly good cues for accessing the intended idiomatic meaning. But altering these cues is still possible. Integrated concepts still use the canonical form, but with an additional word inserted into the expression. This extra information takes additional time to integrate, but does not alter the already established cues, making this the most preferred variant. Whereas lexical variation and idiom blends alter one of the cues in the idiom, causing relearning in order to discriminate the new cue with the intended meaning, and making these variants less appreciated. This approach may also explain why idioms can become shorter or truncated over time: certain cues are better at discriminating the intended meaning, while others become irrelevant and are eventually dropped. But before this natural development happens, omitting (potentially useful) cues is considerably less appreciated.

Acknowledgements

This study was generously supported by an Izaak Walton Killam memorial scholarship, awarded to the first author. In addition, we would like to thank Lauren Rudat for her suggestions on improving the stimuli, and to the editors and anonymous reviewers for their suggestions on improving the text.

Abbreviations

AOI	area of interest
LQ	language question
LSA	latent semantic analysis
MWE	multiword expression
PCA	principal components analysis
RT	reaction time
VAS	visual analogue scale

Appendix A: Canonical form for the 60 idioms used in the two studies

beat around the bush
bend the rules
burn a hole in your pocket
burn your bridges
bury the hatchet
chomp at the bit
cut the mustard
dragged through the mud
drink someone under the table
drive someone up the wall
drown your sorrows
face the music
fall by the wayside
flip your lid
fly off the handle
foot the bill
get the show on the road
get under someone's skin
give up the ghost
give up the ship
go against the grain
go behind someone's back
go through the roof
go with the flow
grind to a halt
hang up your boots
have a card up your sleeve
have many irons in the fire
have someone over a barrel
have your finger on the pulse
hear something through the grapevine
jump on the bandwagon
keep someone on their toes
keep your eye on the ball
keep your nose to the grindstone
lie through your teeth
line your pockets
lose your marbles
nip something in the bud
paint yourself into a corner
pick your brain
pull someone's leg
pull the strings
pull up your socks
put your foot in your mouth
rock the boat
run the gauntlet
shake in your boots
shoot the breeze
shoot the messenger
skate on thin ice
spill the beans
spin your wheels
stick to your guns
swallow your pride
sweep something under the rug
tear a strip off someone
wash your hands of something
wear your heart on your sleeve
wrap someone around your finger

References

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts – combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3. 131–160.

Ayto, John (ed.). 2009. *From the horse's mouth: Oxford dictionary of English idioms*. Oxford: Oxford University Press.

Baayen, R. Harald, Peter Hendrix & Michael Ramscar. 2013. Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech* 56. 329–347.

Baayen, R. Harald & Michael Ramscar. 2015. Abstraction, storage and naive discriminative learning. In Ewa Dabrowska & Dagmar Divjak (eds.), *Handbook of cognitive linguistics*, 100–120. Berlin: Mouton de Gruyter.

Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson & Rebecca Treiman. 2007. English lexicon project. *Behavior Research Methods* 39(3). 445–459.

Barlow, Michael. 2000. Usage, blends and grammar. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 315–345. Stanford, CA: CSLI Publications.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. DOI:10.18637/jss.v067.i01

Bobrow, Samuel A. & Susan M. Bell. 1973. On catching on to idiomatic expressions. *Memory & Cognition* 1. 343–346.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 75–96. Berlin/New York: Mouton de Gruyter.

Cacciari, Cristina, Paola Corradini & Roberto Padovani. 2005. Speed of processing effects on spoken idioms comprehension. In Bruno G. Bara, Lawrence Barsalou & Monica Bucciarelli (eds.), *Proceedings of the XXVII annual meeting of the cognitive science society*, 21–23. New Jersey: Lawrence Erlbaum.

Cacciari, Cristina & Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language* 27. 668–683.

Cutting, J. Cooper & Kathryn Bock. 1997. That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory & Cognition* 25(1). 57–71.

Duffley, Patrick J. 2013. How creativity strains conventionality in the use of idiomatic expressions. In Mike Borkent, Barbara Dancygier & Jennifer Hinnell (eds.), *Language and the creative mind*, 49–61. Stanford, CA: CSLI Publications.

Elman, Jeffrey L. 2004. An alternative view of the mental lexicon. *Trends in Cognitive Sciences* 8(7). 301–306.

Elman, Jeffrey L. 2011. Lexical knowledge without a lexicon. *The Mental Lexicon* 6. 1–33.

Fanari, Rachele, Cristina Cacciari & Patrizia Tabossi. 2010. The role of idiom length and context in spoken idiom comprehension. *European Journal of Cognitive Psychology* 22(3). 321–334.

Fay, David. 1982. Substitutions and splices: A study of sentence blends. In Anne Cutler (ed.), *Slips of the tongue and language production*, 163–195. Amsterdam: Mouton de Gruyter.

Funke, Frederik & Ulf-Dietrich Reips. 2012. Why semantic differentials in web-based research should Be made from visual analogue scales and not from 5-point scales. *Field Methods* 24(3). 310–327.

Geeraert, Kristina. 2016. *Climbing on the bandwagon of idiomatic variation: A multi-methodological approach*. University of Alberta Doctoral Dissertation.

Geeraert, Kristina, R. Harald Baayen & John Newman. 2017. Understanding idiomatic variation. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE2017)*, 80–90. Association for Computational Linguistics. April 4, 2017.

Geeraert, Kristina, John Newman & R. Harald Baayen. 2017. Idiom variation: Experimental data and a blueprint of a computational model. *Topics in Cognitive Science* 9(3). 1–17. DOI:10.1111/tops.12263

Gibbs, Raymond W. 1980. Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition* 8(2). 149–156.

Gibbs, Raymond W. & Nandini P. Nayak. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology* 21. 100–138.

Gibbs, Raymond W., Nandini P. Nayak, John L. Bolton & Melissa E. Keppel. 1989. Speakers' assumptions about the lexical flexibility of idioms. *Memory & Cognition* 17(1). 58–68.

Hirotani, Masako, Lyn Frazier & Keith Rayner. 2006. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language* 54. 425–443.

Landauer, Thomas K., Peter W. Foltz & Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes* 25. 259–184.

Langlotz, Andreas. 2006. *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English.* Amsterdam: John Benjamins.

McGlone, Matthew S., Sam Glucksberg & Cristina Cacciari. 1994. Semantic productivity and idiom comprehension. *Discourse Processes* 17. 167–190.

Miller, George A. & Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1). 1–28.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English.* Oxford: Oxford University Press.

R Core Team. 2012. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/. ISBN 3-900051-07-0.

Ramscar, Michael & R. Harald Baayen. 2013. Production, comprehension, and synthesis: A communicative perspective on language. *Frontiers in Psychology* 4. 233. DOI:10.3389/fpsyg.2013.00233

Rayner, Keith, Gretchen Kambe & Susan A. Duffy. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology* 53A(4). 1061–1080.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1471591

Scholivet, Manon, Carlos Ramisch & Silvio Cordeiro. 2018. Sequence models and lexical resources for MWE identification in French. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 263–297. Berlin: Language Science Press. DOI:10.5281/zenodo.1469567

Schröder, Daniela. 2013. *The syntactic flexibility of idioms: A corpus-based approach.* Munich: AVM.

Sinclair, John (ed.). 2011. *Collins COBUILD idioms dictionary.* Harper Collins.

Swinney, David A. & Anne Cutler. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behaviour* 18. 523–534.

Tabossi, Patrizia, Rachele Fanari & Kinou Wolf. 2008. Processing idiomatic expressions: Effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(2). 313–327.

Titone, Debra A. & Cynthia M. Connine. 1994a. Comprehension of idiomatic expressions: Effects of predictability and literality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20(5). 1126–1138.

Titone, Debra A. & Cynthia M. Connine. 1994b. Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbolic Activity* 9(4). 247–270.

Titone, Debra A. & Cynthia M. Connine. 1999. On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics* 31. 1655–1674.

Titone, Debra A. & Maya R. Libben. 2014. Time-dependent effects of decomposability, familiarity and literal plausibility on idiom meaning activation: A crossmodal priming investigation. *The Mental Lexicon* 9. 473–496.

Chapter 2

Paraphrases of verbal multiword expressions: the case of Czech light verbs and idioms

Petra Barančíková
Charles University

Václava Kettnerová
Charles University

In this chapter, we deal with two types of Czech verbal MWEs: light verb constructions and verbal idiomatic constructions. Many verbal MWEs are characterized by the possibility of being paraphrased by single words. We explore paraphrasability of Czech verbal MWEs by single verbs in a semiautomatic experiment using word embeddings. Further, we propose a lexicographic representation of the obtained paraphrases enriched with morphological, syntactic and semantic information. We demonstrate one of its practical application in a machine translation experiment.

1 Introduction

Multiword expressions (MWEs) are widely acknowledged as a serious challenge for both foreign speakers and many NLP tasks (Sag et al. 2002). Out of various MWEs, those that involve verbs are of great significance as verbs represent the syntactic center of a sentence. Baldwin & Kim (2010) distinguish the following four types of verbal MWEs:

- verb-particle constructions (also referred to as particle verbs, or phrasal verbs), e.g., *catch up*, *put on*, *swallow down*;

Petra Barančíková & Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: the case of Czech light verbs and idioms. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 35–59. Berlin: Language Science Press. DOI:10.5281/zenodo.1469553

- prepositional verbs, e.g., *come across*, *refer to*;
- light-verb constructions (also referred to as verb-complement pairs, or support verb constructions), e.g., *do a report*, *give a kiss*, *make an attempt*;
- verb-noun idiomatic constructions (also referred to as VP idioms), e.g., *spill the beans*, *pull strings*, *shoot the breeze*.

In this chapter, we focus on two particular types of Czech verbal MWEs: light-verb constructions (LVCs) and idiomatic verbal constructions (IVCs) as they also represent MWEs in Czech in contrast to the first two types that are primarily expressed as single prefixed verbs.

We explore the possibility of expressing these two types of MWEs by single synonymous verbs, which is considered to be one of their prototypical features, see e.g. Chafe (1968) and Fillmore et al. (1988). The motivation for this work lies in the fact that paraphrases greatly assist in a wide range of NLP applications such as information retrieval (Wallis 1993), machine translation (Madnani & Dorr 2013; Callison-Burch et al. 2006; Marton et al. 2009) or machine translation evaluation (Kauchak & Barzilay 2006; Zhou et al. 2006; Barančíková et al. 2014).

The content of this chapter is an extended version of Barančíková & Kettnerová (2017). In addition, it is further explored with IVCs and linguistic properties of LVCs and IVCs relevant to the paraphrasing task are discussed in detail. The new version of the dictionary of paraphrases is larger and it provides a more elaborated set of morphological, syntactic and semantic features, including information on aspects and aspectual counterparts of verbs.

This chapter is structured as follows. First, linguistic properties of LVCs and IVCs are discussed (§2) and related work on their paraphrases is introduced. Second, a paraphrasing model is proposed, namely the selection of LVCs and IVCs, an automatic extraction of candidates for their paraphrases and their manual evaluation are described in detail (§3). Third, the resulting data and their representation in a dictionary of paraphrases are introduced (§4). Finally, in order to present one of the many practical applications of this dictionary, a random sample of paraphrases of LVCs is used in a machine translation experiment (§5).

2 Linguistic properties of LVCs and IVCs

Both LVCs and IVCs represent verbal multiword units: they are composed of separate words that, however, refer to an extralinguistic reality as a whole. Their linguistic properties relevant for their paraphrasability by single verbs are introduced below.

2.1 Light-verb constructions

The theoretical research on light-verb constructions is characterized by an enormous diversity in terms and analyses used, see esp. Amberber et al. (2010) and Alsina et al. (1997). Here, we use the term LVC for a multiword unit within which the verb – not retaining its full semantic content – provides grammatical functions and to which the main predicative content is contributed by a noun; as a result, such a multiword unit serves as a single predicative unit, see e.g. Algeo (1995), Alsina et al. (1997) and Butt (2010).[1] In contrast to IVCs, predicative nouns in LVCs have the same meanings as in nominal structures, meanings of light verbs are rather impoverished when compared with their full verb counterparts, see §2.2.

In the Czech language, the central type of LVCs are represented by LVCs in which predicative nouns are expressed as a direct or indirect object of a light verb (e.g., *dostat strach* 'to get fear' ⇒ 'to become afraid', *vzdát úctu* 'to pay tribute', and *vyvolat pobouření* 'to provoke indignation' ⇒ 'to cause uproar'). The LVCs in which a predicative noun occupies an adverbial of the light verb, (e.g., *dát do pořádku* 'to put in order', *mít pod kontrolou* 'to have under control', *mít na starosti* 'to have on care' ⇒ 'to be responsible') are more syntactically and morphologically fixed than the central type of LVCs (Radimský 2010).

As single predicative units, most LVCs have their single predicative counterparts by which they can be paraphrased. A single verb paraphrase can be either morphologically related, or non-related with the predicative noun representing the nominal component of the paraphrased LVC. For example, the LVCs *dát polibek* and *dát pusu* 'give a kiss' can be both paraphrased by the verb *políbit* 'to kiss', which is morphologically related only with the nominal component of the first LVC. There is no synonymous verb morphologically related to the nominal component of the second LVC.

In contrast to their single predicative paraphrases, LVCs manifest greater flexibility in their modification, compare e.g. adjectival modifiers of the LVC *dát polibek* 'give a kiss' and the corresponding adverbial modifiers of its single verb paraphrase *políbit* 'to kiss': *dát vášnivý/něžný/letmý/manželský/májový/smrtící polibek* 'give a passionate/tender/fleeting/marriage/May/fatal kiss' vs. *vášnivě/*

[1] Besides predicative nouns, adjectives, adverbs and verbs can also serve as predicative elements. These cases are left aside here.

*něžně/letmo/*manželsky/*májově/*smrtelně políbit* 'to kiss passionately/tenderly/fleetingly/*marriagely/*Mayly/?fatally'. Easier modification of LVCs is often considered a motivation for their use (Brinton & Akimoto 1999).

Another motivation lies in the possibility to structure the expressed event in a more subtle way than what single verbs allow. For example, in Czech various combinations of the grammatical aspect of light verbs and the number of predicative nouns allow for the expression of several meanings that cannot be expressed with single verbs; these cases require lexical modification, see Table 1.

Finally, in many cases, the selection of different light verbs allows for per-

Table 1: Possible combinations of the grammatical aspect of the light verbs *dát*^{pf}, *dávat*^{pf} 'to give' and the number of the noun *polibek* 'kiss' and their paraphrasability by the perfective and imperfective single verbs *políbit*^{pf} and *líbat*^{impf} 'to kiss', respectively.

LVC	Single verb paraphrase	Lexical modification	Example[a]
sg & pf	pf	no	*Petr dal Janě polibek.* 'Peter gave a kiss to Jane.' ∼ *Petr Janu políbil.* 'Peter kissed Jane.'
pl & impf	impf	no	*Petr dával Janě polibky.* 'Peter gave kisses to Jane.' ∼ *Petr Janu líbal.* 'Peter was kissing Jane.'
pl & pf	pf	yes	*Petr dal Janě polibky.* 'Peter gave several kisses to Jane.' ∼ *Petr Janu několikrát políbil.* 'Peter kissed Jane several times.'
sg & impf	impf	yes	*Petr Janě dával polibek.* 'Peter was giving a kiss to Jane.' ∼ *Petr Janu právě líbal.* 'Peter was just kissing Jane.'

[a] Let us emphasize that the single verb paraphrases of the last two combinations require to be lexically modified – by the words *několikrát* 'several times' and *právě* 'just', respectively.

spectivization of the expressed event from the point of view of its different participants, see esp. Kettnerová & Lopatková (2015). For example, besides the light verb *dát* 'to give', the noun *polibek* 'kiss' can select the light verb *dostat* 'to get' as well. The LVC *dát polibek* 'to give a kiss' promotes a kisser in the subject position while the LVC *dostat polibek* 'to get a kiss' puts a kissee into this position. Both these LVCs are paraphrasable by a single verb *políbit* 'to kiss', however, with different values of the grammatical voice: the LVC *dát polibek* 'to give a kiss' can be paraphrased by the verb *políbit* 'to kiss' in the active voice (e.g., *Petr dal Janě polibek.* 'Peter gave a kiss to Jane.' ∼ *Petr Janu políbil.* 'Peter kissed Jane.') while the LVC *dostat polibek* 'to get a kiss' requires the passive voice of the verb *políbit* 'to kiss' (e.g., *Jana dostala od Petra polibek.* 'Jane got a kiss from Peter.' ∼ *Jana byla políbena od Petra.* 'Jane was kissed by Peter'.)

LVCs in NLP. One of the trending topics concerning LVCs in the NLP community is their automatic identification. In this task, various statistical measures often combined with information on syntactic and/or semantic properties of LVCs are employed, see e.g. Bannard (2007) and Fazly et al. (2005). The automatic detection benefits especially from parallel corpora representing valuable sources of data in which LVCs can be automatically recognized via word alignment, see e.g. Chen et al. (2015), de Medeiros Caseli et al. (2010), Sinha (2009), Zarrieß & Kuhn (2009). However, work on paraphrasing LVCs is still not extensive. For example, a paraphrasing model has been proposed within the Meaning↔Text Theory (Žolkovskij & Mel'čuk 1965); its representation of LVCs by means of lexical functions and rules applied in the paraphrasing model are thoroughly described in Alonso-Ramos (2007). Further, Fujita et al. (2004) presents a paraphrasing model which takes advantage of semantic representation of LVCs by lexical conceptual structures. As with our method proposed in §3, their model also takes into account several morphological and syntactic features of LVCs, which have turned out to be highly relevant for the paraphrasing task.

2.2 Idiomatic Verbal Constructions

Despite their low frequency, IVCs form a substantial part of a lexis, see e.g. Baldwin & Kim (2010), Sag et al. (2002) and Cowie (2001). Similarly to LVCs, definitions of idioms vary depending on diverse purposes of their description, see e.g. Healy (1968), Fraser (1970), van der Linden (1992) and Nunberg et al. (1994).

Here, we define an IVC as a verbal multiword unit that exhibits strong lexical co-occurrence restrictions so that at least one of its parts cannot be used with the same meaning outside the given multiword unit. The idiomatic meaning of individual components of IVCs is reflected in the fact that they are only rarely interchangeable with words of similar meanings. IVCs thus represent highly conventionalized multiword units, see e.g. Everaert et al. (2014), Granger & Meunier (2008) and Cowie (2001). IVCs can exhibit the following specific properties, see e.g. Burger et al. (2007), Čermák (2001) and Everaert et al. (2014):

- markedness at the syntactic and/or morphological level: e.g., *vzít za své* 'take as one's own' ⇒ 'to be no more' (syntactically marked as the reflexive adjective *své* does not modify any noun), and *nalít někomu čistého vína* 'to pour someone pure wine' ⇒ 'to tell someone the honest truth' (morphologically marked due to the partitive genitive of the noun *víno* 'wine', which is highly restricted in contemporary Czech);
- figuration: e.g., *vstát z mrtvých*, 'raise the dead' (as it involves a metaphor), *pověsit se někomu na krk* 'to hang around someone's neck' (as it involves a metonymy);
- fixedness at syntactic and/or morphological level: e.g., *postavit někoho na nohy* 'to put someone back on his feet' (syntactically fixed as it cannot be transformed into the passive structure), and *přijít na jiné myšlenky* 'to come to different ideas' ⇒ 'to find something else to think about' (morphologically fixed as the noun *myšlenka* 'idea' can have only the plural form);
- proverbiality: IVCs are typically used for recurrent socially significant situations, implying often their subjective evaluation (e.g., *vidět někomu do duše* 'to see right through someone');
- informality: IVCs are typically of informal register (e.g., *strčit si něco za klobouk* 'to put something behind a hat' ⇒ 'to stick it up one's jumper').

Some IVCs can be paraphrased by a single word verb, see e.g. the IVC *podat někomu pomocnou ruku* 'to give someone helping hand' and its single verb paraphrase *pomoci* 'to help'. However, many IVCs are paraphrasable rather by a whole syntactic structure, see e.g. the IVC *mít slovo* 'to have a word' ⇒ 'to be someone's turn to speak'.

IVCs in NLP. There is considerable work focused on automatic identification of idioms in the text and their extraction (Cook et al. 2007; Li & Sporleder 2009;

Muzny & Zettlemoyer 2013; Peng et al. 2015; Katz 2006). However, little attention has been paid to paraphrases of idioms. Let us introduce two works focused on paraphrases of idioms. First, Pershina et al. (2015) identifies synonymous idioms based on their dictionary definitions and their occurrences in tweets. Similarly, Liu & Hwa (2016) generate paraphrases of idioms using dictionary entries. However, there are no lexical resources available for NLP applications providing information on idioms in Czech.

3 Paraphrase model

In this section, the process of extracting paraphrases is described in detail. First, we present the selection of LVCs and IVCs (§3.1). For their paraphrasing, we had initially intended to use some of the existing resources, however, they turned out to be completely unsatisfactory for our task.

First, we used the *ParaPhrase DataBase* (PPDB) (Ganitkevitch & Callison-Burch 2014), the largest paraphrase database available for the Czech language. PPDB was created automatically from large parallel data. Unfortunately, there were only 54 candidates for single verb paraphrases of LVCs present. A manual analysis of these candidates showed that only 2 of them were detected correctly, the rest was noise in PPDB. Similarly for idioms, PPDB contained a correct single verb paraphrase for only 6 IVCs from our data (i.e. about 1%). As this number is clearly insufficient, we chose not to use parallel data for paraphrasing.

Therefore, we adopted another approach to the paraphrasing task applying *word2vec* (Mikolov et al. 2013), a neural network model. *Word2vec* is a group of shallow neural networks generating word embeddings, i.e. representations of words in a continuous vector space depending on the contexts in which they appear. In line with the distributional hypothesis (Harris 1954), semantically similar words are mapped close to each other (measured by the cosine similarity) so we can expect LVCs and IVCs to have similar vector space distribution to their single verb paraphrases.

Word2vec computes vectors for single tokens. As both LVCs and IVCs represent multiword units, their preprocessing was thus necessary: each LVC and IVC had to be first identified and connected into a single token (§3.2). Particular settings of our model for an automatic extraction of candidates for single verb paraphrases are described in §3.3.

The advantage of this approach is that only monolingual data – generally easily obtainable in a large amount – is necessary for word embeddings training. The disadvantage is that not only paraphrases can have similar word embeddings.

Antonyms and words with more specific or even different meaning can appear in similar contexts as well. Therefore, a manual evaluation of the extracted candidates is necessary (§3.4).

3.1 Data selection

3.1.1 LVCs selection

Three different datasets of LVCs – containing together 2,389 unique LVCs[2] – were used in our experiment. As all the datasets were manually created, they allow us to achieve the desired quality of the resulting data.

The first dataset resulted from the experiment examining the native speakers' agreement on the interpretation of light verbs (Kettnerová et al. 2013). This dataset consists of both LVCs in which predicative nouns are expressed as a direct or indirect object by a prepositionless case (e.g. *položit otázku* 'put a question') and LVCs in which predicative nouns are expressed as an adverbial by a simple prepositional case (e.g., *dát do pořádku* 'put in order') or by a complex prepositional group (e.g., the verb *přejít* 'go' plus the complex prepositional group *ze smíchu do pláče* 'from laughing to crying').

The second dataset resulted from a project aiming to enhance the high coverage valency lexicon of Czech verbs VALLEX[3] with the information on LVCs (Kettnerová et al. 2016). In this case, only the predicative nouns expressed as the direct object by the prepositionless accusative were selected. For identification of LVCs, the modified test of coreference was applied (Kettnerová & Bejček 2016). As the frequency and saliency have been taken as the main criteria for their selection, the resulting set represents a valuable source of LVCs for Czech.

The third small dataset is represented by LVCs in which the predicative noun is expressed as an adverbial. These LVCs were obtained from the VALLEX lexicon as a result of manual analysis of verbal multiword units marked as idioms. As these multiword units were treated inconsistently in the annotation, including not only IVCs but sometimes also LVCs with predicative nouns in adverbial positions, the obtained dataset had to be manually selected.

As in the VALLEX lexicon, information on aspectual counterparts of the given verbs is available, we have used it to expand these datasets by adding missing aspectual counterparts. The overall number of LVCs in the datasets is presented below in Table 2. The union of LVCs from these datasets has been used in the paraphrase candidates extraction task.

[2] When counting aspectual counterparts separately, the number increases to 3,509 unique LVCs

[3] http://ufal.mff.cuni.cz/vallex/3.0/

3.1.2 IVCs selection

The dataset of IVCs was extracted from the VALLEX lexicon after the manual filtering of LVCs with predicative nouns in adverbial positions, see the third dataset in §3.1.1. From the obtained IVCs, those IVCs that include the highly polysemous pronoun *to* 'it' were removed as their automatic identification could be unreliable. The final set consists of 595 IVCs (counting aspectual counterparts separately 621 IVCs), see the statistics provided in Table 2.

Table 2: The number of LVCs and IVCs, verbs and nominal components in the three datasets described in §3.1.1, before (first number) and after (second number) the aspectual counterparts expansion.

Dataset	LVCs	IVCs	Verbs	Nominal components
First	726/1,167	0/0	49/84	612
Second	1,640/2,366	0/0	126/131	699
Third	104/106	595/621	310/324	324
Union[a]	2,389/3,509	595/621	417/446	1444

[a] The numbers do not add up due to a small overlap among the datasets.

3.2 Data preprocessing

We used four large lemmatized and POS-tagged corpora of Czech texts: SYN2000 (Čermák et al. 2000), SYN2005 (Čermák et al. 2005), SYN2010 (Křen et al. 2010) and CzEng 1.0 (Bojar et al. 2011). These corpora were further extended with the data from the Czech Press – a large collection of contemporary news texts containing more than 2,000 million lemmatized and POS-tagged tokens. The overall statistics on all datasets is presented in Table 3.

To generate LVCs and IVCs paraphrases, all the selected LVCs and IVCs (§3.1) had to be automatically identified in the given corpora. For their identification, we started with verbs. First, all verbs in the corpora were detected. From these verbs, only those verbs that represent parts of the selected LVCs and IVCs were further processed. For each selected verb, each noun phrase in the context ± 4 words from the given verb was identified based on POS tags and extracted in case the verb and the given noun phrase can combine in some of the selected LVCs or IVCs.

Further, as word embeddings are generated for single words, each detected noun phrase was connected with its respective verb into a single word unit. In

Table 3: Basic statistics of datasets (numbers in millions of units).

Corpus	Sentences	Tokens
CNK2000	2.78	121.81
CNK2005	7.95	122.99
CNK2010	8.18	122.48
Czeng 1.0	14.83	206.05
Czech Press	57.03	2447.68
Total	90.77	3021.01

cases where some verb could combine with more than one noun phrase into LVCs or IVCs, or in cases where a particular noun phrase could be connected with more than one verb, we followed the principle that every verb should be connected to at least one noun phrase in order to maximize the number of identified LVCs and IVCs. For example, if there were two verbs v_1 and v_2 in a sentence and v_1 had a candidate noun phrase c_1, while v_2 had two candidate noun phrases c_1 and c_2, v_1 was connected with c_1 and v_2 with c_2. In case this principle was not sufficient, a verb was assigned the closest noun phrase on the basis of word order. When each noun phrase was connected maximally with one verb and each verb was connected maximally with one noun phrase, we have joined the noun phrases to their respective verbs into single word units with the underscore character and deleted the noun phrases from their original positions in sentences.

Further, to compensate sparsity of LVCs and IVCs in the data, after identifying a verb from the selected LVCs and IVCs in the data, its aspectual counterpart – if relevant – has been automatically added. For example, after detecting the imperfective verb *vcházet*$^{\mathrm{impf}}$ 'enter' in the data and the prepositional noun phrase *do dějin* 'to history' in its context, not only the given imperfective verb, but also its perfective counterpart *vejít*$^{\mathrm{pf}}$ have been connected with the given noun phrase into the resulting unit *vcházet_vejít_do_dějin*. We refer to such an artificially constructed unit as an *abstract unit* from now on. The abstract unit *vcházet_vejít_-do_dějin* then replaced the verb *vcházet* in the sentence, while the noun phrase *do dějin* was deleted from the sentence. Each LVC and IVC identified in the data is thus represented by a single abstract unit representing also its relevant aspectual counterparts.

On this basis, almost 7 million instances of LVC and IVC abstract units were generated in the corpora, see Table 4. The rank and frequency of the most and the least common ones are presented in Table 5.

Table 4: The number of LVCs and IVCs detected in the data. The first row shows the total number of LVC and IVC abstract units identified in the data. The second row represents the number of their unique instances. The third row provides the number of those unique units with higher frequency than 100 occurrences. The last row shows the number of unique LVCs and IVCs without aspectual counterparts expansion, i.e. after splitting the generated abstract units back to a single verb–a single noun phrase pairs.

	LVCs	IVCs
abstract units	6,541,394	374,493
unique abstract units	1,776	211
unique abstract units > 100	1,361	153
unique MWEs	2,954	353

Table 5: The ranking of LVC and IVC abstract units identified in the data, based on their frequency.

rank	type	abstract unit	frequency
1.	LVC	*mít_problém* ‘have a problem’	211,296
2.	LVC	*mít_možnost* ‘have a possibility’	207,330
⋮	⋮	⋮	⋮
29.	IVC	*mít_na_mysli* ‘have in mind’	43,521
⋮	⋮	⋮	⋮
1986.	IVC	*chytnout_chytat_chytit_za_špatný_konec* ‘get hold of the wrong end of the stick’	1
1987.	LVC	*přechodit_přecházet_přejít_ze_smíchu_do_pláče* ‘go from laughing to crying’	1

3.3 Word2vec model

To the resulting data, we applied *gensim*, a freely available *word2vec* implementation (Řehůřek & Sojka 2010). In particular, we used a model of vector size 500 with continuous bag of word (CBOW) training algorithm and negative sampling.

As it is impossible for the model to learn anything about a rarely seen word, we set a minimum number of word occurrences to 100 in order to limit the size of the vocabulary to reasonable words. Even though we increased frequencies of LVCs and IVCs by the unified representation for their aspectual counterparts, this limit still filtered more than 300 rarely used LVC and 50 IVC abstract units; the resulting number is provided in the third row of Table 4.

After training the model, for each of 1,361 LVC and 153 IVC abstract units with more than 100 occurrences we extracted 30 words with the most similar vectors. From these 30 words, we selected up to 15 single verbs closest to a given LVC or IVC abstract unit. These verbs were taken as candidates for single verb paraphrases of LVCs or IVCs in that abstract unit. On average, there were 7 candidates for each LVC abstract unit and 10 candidates for each IVC abstract unit.

Before the manual evaluation of the candidates, the abstract units were divided back to individual IVCs or LVCs and their paraphrase candidates were again enriched with their aspectual counterparts from the VALLEX lexicon. This way, annotators could select a paraphrase with a proper aspect for each verbal MWE.

3.4 Annotation process

In this section, the annotation process of the candidates for single verb paraphrases of LVCs and IVCs is thoroughly described. Let us repeat that *word2vec* generates semantically similar words depending on the context in which they appear. However, not only words having the same meaning can have similar space representations, but words with an opposite meaning, more specific meaning or even different meaning can be extracted as they can appear in similar contexts as well. Manual processing of the extracted single verbs was thus necessary for evaluating the results of the adopted method.

In the manual evaluation, two annotators were asked to indicate for each instance of the unique paraphrase candidates of an LVC or IVC whether it represents a single verb paraphrase of the given LVC or IVC, or not. For example, the single word verbs *upřednostňovat* and *preferovat* 'to prefer' were indicated as paraphrases of the LVC *dávat přednost* 'to give a preference'. Similarly, for the IVC *prásknout do bot* 'to bang to the shoes' ⟹ 'to take to one's heels', the single verbs *utéci* 'to run away' and *zdrhnout* 'to make off' among others were chosen as paraphrases.

Moreover, single verbs antonymous to LVCs or IVCs were marked as well since they can also function as paraphrases in a modified context. For example, for the LVC *vypovídat pravdu* 'to tell the truth' the antonymous verb *lhát* 'to lie' was selected as well, as the sentence *Nevypovídá pravdu.* 'He is not telling the truth.' can be paraphrased as *Lže.* 'He is lying.'.

Further, when the annotators determined a certain candidate as a single verb paraphrase of an LVC or IVC, they took into account the following four morphological, syntactic and semantic aspects.

First, they had to pay special attention to the morphosyntactic expression of arguments. As Czech encodes syntactic relations via morphological forms, changes in the morphological expression of arguments reflect different perspectives from which the event denoted by an LVC or IVC on the one hand and its single verb paraphrase on the other hand is viewed. For example, the single verb *potrestat* 'to punish' paraphrases the LVC *dostat trest* 'to get a punishment', however, the morphological forms of the punisher and the punishee, two semantic roles evoked by the given LVC and the single verb, differ. In the LVC *dostat trest* 'to get punishment', the punishee (*Petr* 'Peter') is expressed by the nominative and the punisher (*otec* 'father') has the form of the prepositional group *od*+genitive (e.g., $Petr_{nom}$ *dostal od* $otce_{od+gen}$ *trest.* 'Peter got punishment from his father.'), while with its single verb paraphrase *potrestat* 'to punish' the nominative encodes the punisher and the accusative expresses the punishee (e.g., $Otec_{nom}$ $Petra_{acc}$ *potrestal.* 'Father punished Peter.').

Second, the annotators had to take into account differences between the syntactic structure of a sentence created by an LVC or IVC and by its respective paraphrase. Particularly, the difference between sentences with a subject and subjectless sentences had to be indicated. For example, the LVC *dojít k oddělení* 'to happen to the separation' ⇒ 'the separation happens' is paraphrasable by the single verb *oddělit se* 'to separate', although the LVC forms a subjectless structure, the syntactic structure of its single verb paraphrase needs a subject.

Third, in some cases the reflexive morpheme *se/si*, marking usually intransitive verbs, has to be added to a single verb paraphrase so that its meaning corresponds to a meaning of its respective multiword counterpart. For example, the IVC *vejít do dějin* 'to come into history' ⇒ 'to go down in history' can be paraphrased by the verb *proslavit* only on the condition that the reflexive morpheme *se* is attached to the verb lemma *proslavit se* 'to achieve fame'.

Lastly, some verbs function as paraphrases of particular LVCs or IVCs only if nouns in these LVCs or IVCs have certain adjectival modifications. These paraphrases were paired with appropriate adjectives during the annotation. For example, if the LVC *provozovat praxi* ‘to run a practice’ is to be paraphrased by the single verb *ordinovat* ‘to see patients’, the adjective *lékařský* ‘medical’ has to modify the noun *praxe* ‘practice’.

The above given four features are not mutually exclusive – they can combine. For example, the verb *zaměstnat* ‘to hire’ is a paraphrase of the LVC *nalézt uplatnění* ‘to find an use’ but both the reflexive morpheme *se* and the adjectival modification *pracovní* ‘working’ are required.

To summarize, for each identified single verb paraphrase *v* of an LVC or IVC *l*, the annotators have chosen from the following options:

- *v* is a paraphrase of *l*
  e.g., *mít zájem* ‘to be interested’ and *chtít* ‘to want’;

- *v* is an antonym of *l* (the modification of the context is necessary)
  e.g., *zaznamenat propad* ‘to experience a drop’ and *stoupnout* ‘to rise’;

- *v* is a paraphrase of *l* but changes in the morphosyntactic expression of arguments are necessary
  e.g., *dostat nabídku* ‘to get an offer’ and *nabídnout* ‘to offer’;

- *v* is a paraphrase of *l* but the change in a sentence structure is required
  e.g., *dojít k poruše* ‘to happen to the failure’ ⇒ ‘the failure happens’ and *porouchat se* ‘to breakdown’;

- *v* is a paraphrase of *l* but the modification of the verb lemma by the reflexive morpheme *se*/*si* is necessary
  e.g., *nést název* ‘bear a name’ and *nazývat se* ‘to be called’;

- *v* is a paraphrase of *l* only if a noun component of *l* is modified by a particular adjectival modification
  e.g., *podat oznámení* ‘to make an announcement’ can be paraphrased as *žalovat* ‘to sue’ only if the noun *oznámení* is modified with the adjective *trestní* ‘criminal’;

- *v* is a not a paraphrase of *l*.

As a result of the annotation, for 1,421 of 2,954 LVCs identified in the data (48,1%) and for 200 of 353 IVCs (56,6%) at least one single verb paraphrase was found. The highest number of single verb paraphrases indicated for one multiword unit was nine and that was the LVC *provést řez* ‘to make an incision’ and the LVC *dát do pořádku* ‘to put in order’. The total number of the indicated single verb paraphrases of LVCs and IVCs was 2,912 and 498, respectively, see Table 6 providing results of the annotation including the frequency of the additional morphological, syntactic and semantic features used in the annotation.

Table 6: The basic statistics on the annotation.

	LVC	IVC
no constraints	2063	336
+ antonymous	115	47
+ reflexive morpheme	473	85
+ morphosyntactic change	270	38
+ syntactic change	43	0
+ an adjective	30	1
total[4]	2912	498

4 Dictionary of paraphrases

3,410 single verbs indicated by the annotators as paraphrases or antonyms of 1,421 LVCs and 200 IVCs (§3.4) form the lexical stock of *ParaDi* 2.0, a dictionary of single verb paraphrases of Czech multiword units of the selected types.[5]

The format of *ParaDi* 2.0 has been designed with respect to both human and machine readability. The dictionary is thus represented as a plain table in the TSV format, as it is a flexible and language-independent data format.

Each lexical entry in the dictionary describes an individual LVC or IVC, providing the following information:

(i) *type* – the type of the given verbal multiword expression with the following three possible values: LVC (indicating an LVC with the predicative noun in the direct or indirect object position), ILVC (representing an LVC with the predicative noun in the adverbial position), or IVC;

(ii) *verb* – a lemma of the verbal component of the given multiword unit;

(iii) *reflexive* – the reflexive morpheme of the lemma, if relevant;

[4] The columns do not add up as the features are not mutually exclusive as mentioned earlier.

[5] *ParaDi* 2.0 is freely available at the following URL: http://hdl.handle.net/11234/1-2377.

(iv) *aspect* – a value of the grammatical aspect of the verb;

(v) *aspectual counterpart* – the aspectual counterpart of the verb, if relevant;

(vi) *noun phrase* – the nominal component of the given multiword unit;

(vii) *morphology* – the morphemic form of the given noun phrase;

(viii) *lemmatized noun phrase* – a lemma representing the noun phrase;

(ix) *synonyms* – a list of synonymous single verb paraphrases;

(x) *antonyms* – a list of antonymous single verbs;

(xi) *adj-modification* – a list of single verb paraphrases and adjectival modifications of the nominal component of the LVC or IVC;

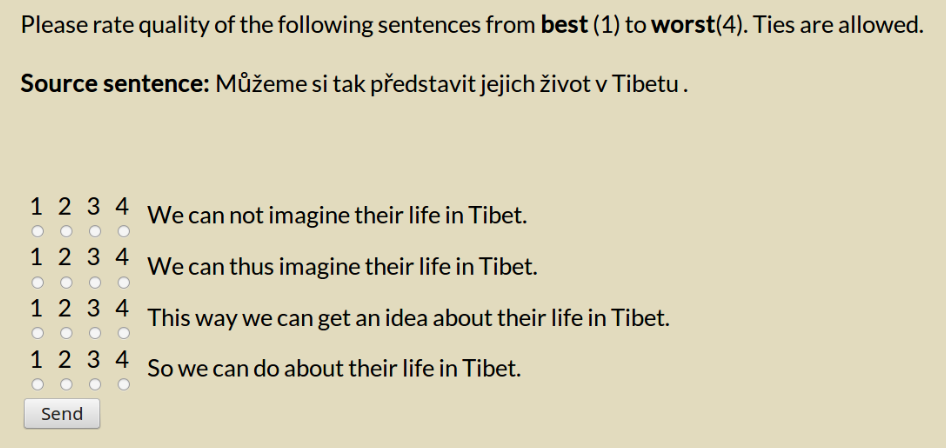
(xii) *structural_change* – a list of single verb paraphrases requiring a change in their sentence structure;

(xiii) *voice_change* – a list of single verb paraphrases requiring changes in the morphosyntactic expression of arguments.

While the information provided in the columns (i)-(viii) concerns multiword units, the information given in (ix)-(xiii) is relevant for their single verb paraphrases. A single verb paraphrase can appear in several columns if it is relevant. For example, the verb paraphrase *zalíbit se* 'to find appealing' of the LVC *nalézt zalíbení* 'to find a delight' ⇒ 'to find appealing' is present in both columns *reflexive* and *voice_change* as it represents the verb paraphrase, which requires both adding the reflexive morpheme *se* to the verb lemma and changes in the morphosyntactic expression of its arguments.

5 Machine translation experiment

In this section, we show how the dictionary providing high quality data can be integrated into an experiment with improving statistical machine translation quality. If translated separately, multiword expressions often cause errors in machine translation. For example, IVCs have been reported to negatively affect statistical machine translation systems which might achieve only half of the BLEU score (Papineni et al. 2002) on the sentences containing IVCs compared to those that do not (Salton et al. 2014).



Please rate quality of the following sentences from **best** (1) to **worst**(4). Ties are allowed.

Source sentence: Můžeme si tak představit jejich život v Tibetu .

1 2 3 4
○ ○ ○ ○ We can not imagine their life in Tibet.

1 2 3 4
○ ○ ○ ○ We can thus imagine their life in Tibet.

1 2 3 4
○ ○ ○ ○ This way we can get an idea about their life in Tibet.

1 2 3 4
○ ○ ○ ○ So we can do about their life in Tibet.

Figure 1: Example of the annotation interface for the MT experiment.

We took advantage of the *ParaDi* dictionary in a machine translation experiment in order to verify its benefit for one of the key NLP tasks. We experimented only with LVCs as we expected quality of LVC translations higher than those of IVCs due to their weaker lexical markedness and their more common use as their higher frequencies in the data suggested (see Table 4).

We selected 50 random LVCs from the dictionary. For each of them, we randomly extracted one sentence from our data containing the given LVC. This set of sentences is referred to as BEFORE. By substituting the LVC for its first paraphrase, i.e. the closest paraphrase in the vector space, we have created a new dataset, referred to as AFTER. We have translated both these datasets – BEFORE and AFTER – to English using two freely available MT systems – *Google Translate*[6] (GT) and *Moses*.[7]

We used crowdsourcing for evaluation of the resulting translations. Six annotators were presented randomly a Czech source sentence either from the dataset BEFORE or from AFTER and their English translations in a randomized order. The annotation interface is displayed in Figure 1. For each translated sentence, the annotators had to indicate its quality, allowing for the same ranking of more than one translated sentences.

We collected almost 300 comparisons. The inter-annotator agreement measured by Krippendorff's alpha (Krippendorff 2007), a reliability coefficient developed to measure the agreement between judges, has achieved 0.58, i.e. a moder-

[6] http://translate.google.com

[7] http://quest.ms.mff.cuni.cz/moses/demo.php

ate agreement. The results of replacing the selected verbal MWEs by their single verb paraphrases in machine translation are very promising: annotators clearly preferred translations of AFTER (i.e. the translations with single verbs) to BEFORE (i.e. with LVCs), in 45% of cases for Moses and in 44% of cases for Google Translate. The results are consistent for both translation systems, see Table 7.

Table 7: Results of the manual evaluation of the MT experiment. The first column shows the source of the better ranked sentence in the pairwise comparison within one translation model or whether they tied.

Source	Moses	GT
BEFORE	30%	33%
AFTER	45%	44%
TIE	25%	23%

However, the example in Table 8 illustrates that even minimal change in a source sentence can substantially change its translations as both the translation models are phrase-based.[8] Based on this fact, we can expect that the evaluation of the translations was not affected only by differences between translations of LVCs and their respective single verb paraphrases but by overall low quality of the translations, which is inevitably reflected in the lower inter-annotator agreement, typical of machine translation evaluation (Bojar et al. 2013). The judges unanimously agreed that the translations of the AFTER source sentence are better than the translations of the BEFORE source sentence. Both systems exhibited a tendency to translate the LVC *dát branku* literally word by word, resulting in incorrect translations of the BEFORE source sentence.

6 Conclusion

We have explored the paraphrasability of Czech light-verb constructions and idiomatic verbal constructions. We have shown that their single verb paraphrases are automatically obtainable from large monolingual data with a manual verification in a significantly larger scale than from paraphrase tables generated from parallel data. Our semiautomatic experiment further revealed that although these verbal multiword units exhibit different linguistic properties, the possibility to

[8] The translations were performed on 9th July 2016, i.e. before a massive expansion of neural translation systems.

Table 8: An example of translated sentences.

Source	BEFORE	*Fotbalisté Budějovic opět* ***nedali*** ***branku*** Footballers Budějovice again did.not.give gate ‘Footballers of Budějovice didn’t make a goal again.’
	AFTER	*Fotbalisté Budějovic opět* ***neskórovali*** Footballers Budějovice again did.not.score ‘Footballers of Budějovice didn’t score again.’
GT	BEFORE	Footballers Budejovice again not given goal
	AFTER	Footballers did not score again Budejovice
Moses	BEFORE	Footballers Budějovice again gave the gate
	AFTER	Footballers Budějovice score again

paraphrase them is very similar; for about one half of the selected light-verb constructions and idiomatic verbal constructions single verb paraphrases have been detected.

The results of our experiment form the lexical stock of a new version of the freely available *ParaDi* dictionary. We have demonstrated one of its possible applications, namely an experiment with improving machine translation quality. However, the dictionary can be used in many other NLP tasks (text simplification, information retrieval, etc.). We have used largely language independent methods, a similar dictionary can be thus created for other languages as well.

Acknowledgments

The research reported in this chapter was supported by the Czech Science Foundation GA ČR, grant No. GA15-09979S. This work used language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education, Youth and Sports of the Czech Republic, project No. LM2015071.

Abbreviations

GT	Google Translate	LVCS	light-verb constructions
IVCS	idiomatic verbal constructions	MT	Machine Translation

References

Algeo, John. 1995. Having a look at the expanded predicate. In Bas Aarts & Charles F. Meyer (eds.), *The verb in contemporary English: Theory and description*, 203–217. Cambridge: Cambridge University Press.

Alonso-Ramos, Margarita. 2007. Towards the synthesis of support verb constructions: Distribution of syntactic actants between the verb and the noun. In Leo Wanner (ed.), *Selected lexical and grammatical issues in the Meaning-Text Theory*, 97–137. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Alsina, Alex, Joan Bresnan & Peter Sells (eds.). 1997. *Complex predicates.* Stanford: CSLI Publications.

Amberber, Mengistu, Brett Baker & Mark Harvey (eds.). 2010. *Complex predicates in cross-linguistic perspective.* Cambridge: Cambridge University Press.

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

Bannard, Colin James. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions* (MWE '07), 1–8. Association for Computational Linguistics.

Barančíková, Petra & Václava Kettnerová. 2017. ParaDi: Dictionary of paraphrases of Czech complex predicates with light verbs. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 1–10. Association for Computational Linguistics. April 4, 2017.

Barančíková, Petra, Rudolf Rosa & Aleš Tamchyna. 2014. Improving evaluation of English-Czech MT through paraphrasing. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard & Joseph Mariani (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), 596–601. European Language Resources Association (ELRA).

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the 8th Workshop on Statistical Machine Translation*, 1–44. Association for Computational Linguistics.

Bojar, Ondřej, Zdeněk Žabokrtský, Ondřej Dušek, Petra GalušČáková, Martin Majliš, David MareČek, Jiří Maršík, Michal Novák, Martin Popel & Aleš Tamchyna. 2011. *Czech-English parallel corpus 1.0 (CzEng 1.0).* LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Brinton, Laurel & Minoji Akimoto (eds.). 1999. *Collocational and idiomatic aspects of composite predicates in the history of English*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Burger, Harald, Dmitrij O. Dobrovol'skij, Peter Kühn & Neal R. Norrick (eds.). 2007. *Phraseologie / phraseology: Ein internationales Handbuch zeitgenössischer Forschung / an international handbook of contemporary research: Volume 1*. Berlin/ New York: Walter de Gruyter.

Butt, Miriam. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker & Mark Harvey (eds.), *Complex predicates in cross-linguistic perspective*, 48–78. Cambridge: Cambridge University Press.

Callison-Burch, Chris, Philipp Koehn & Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (HLT-NAACL '06), 17–24. Association for Computational Linguistics.

Čermák, František. 2001. Substance of idioms: Perennial problems, lack of data or theory? *International Journal of Lexicography* 14(1). 1–20.

Čermák, František, Renata Blatná, Jaroslava HlaváČová, Jan Kocek, Marie Kopřivová, Michal Křen, Vladimír PetkeviČ, Věra Schmiedtová & Michal šulc. 2000. *SYN2000: Balanced corpus of written Czech*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Čermák, František, Jaroslava HlaváČová, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Marie Kopřivová, Michal Křen, Renata Novotná, Vladimír PetkeviČ, Věra Schmiedtová, Hana Skoumalová, Johanka Spoustová, Michal šulc & Zdeněk Velíšek. 2005. *SYN2005: Balanced corpus of written Czech*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Chafe, Wallace L. 1968. Idiomaticity as an anomaly in the chomskyan paradigm. *Foundations of Language* 4(2). 109–127.

Chen, Wei-Te, Claire Bonial & Martha Palmer. 2015. English light verb construction identification using lexical knowledge. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence* (AAAI'15), 2375–2381. AAAI Press.

Cook, Paul, Afsaneh Fazly & Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on a Broader Perspective on*

Multiword Expressions (MWE '07), 41–48. Association for Computational Linguistics.

Cowie, Anthony (ed.). 2001. *Phraseology: Theory, analysis, and applications.* Oxford, UK: Oxford University Press.

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

Everaert, Martin, Erik-Jan van der Linden, André Schenk & Rob Schreuder. 2014. *Idioms: Structural and psychological perspectives.* New York, USA & East Sussex, UK: Psychology Press.

Fazly, Afsaneh, Ryan North & Suzanne Stevenson. 2005. Automatically distinguishing literal and figurative usages of highly polysemous verbs. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition* (DeepLA '05), 38–47. Association for Computational Linguistics.

Fillmore, Charles J., Paul Kay & Mary Catherine O'Connore. 1988. Regularity and idiomaticity in grammatical constructions: The case of *Let Alone*. *Language* 64(3). 501–538.

Fraser, Bruce. 1970. Idioms within a transformational grammar. *Foundations of Language* 6(1). 22–42. http://www.jstor.org/stable/25000426.

Fujita, Atsushi, Kentaro Furihata, Kentaro Inui, Yuji Matsumoto & Koichi Takeuchi. 2004. Paraphrasing of Japanese light-verb constructions based on lexical conceptual structure. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE 2004)* (MWE '04), 9–16. Association for Computational Linguistics.

Ganitkevitch, Juri & Chris Callison-Burch. 2014. The multilingual paraphrase database. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Joseph Mariani Bente Maegaard, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014). European Language Resources Association (ELRA).

Granger, Sylviane & Fanny Meunier (eds.). 2008. *Phraseology: An interdisciplinary perspective.* Amsterdam / Philadelphia: John Benjamins.

Harris, Zellig. 1954. Distributional structure. *Word* 10(23). 146–162.

Healy, Adam. 1968. English idioms. *Kivung (Journal of the Linguistic Society of the University of Papua New Guinea* 1(2). 71–108.

Katz, Graham. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-*

06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (MWE '06), 12–19. Association for Computational Linguistics.

Kauchak, David & Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (HLT-NAACL '06), 455–462. Association for Computational Linguistics.

Kettnerová, Václava, Petra Barančíková & Markéta Lopatková. 2016. Lexicographic description of Czech complex predicates: Between lexicon and grammar. In George Meladze Tinatin Margalitadze (ed.), *Proceedings of the 17th EURALEX international congress.* Tbilisi, Georgia: Ivane Javakhishvili Tbilisi University Press. September 6-10, 2016.

Kettnerová, Václava & Eduard Bejček. 2016. Distribution of valency complements in Czech complex predicates: Between verb and noun. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 515–521. Paris, France: European Language Resources Association.

Kettnerová, Václava & Markéta Lopatková. 2015. At the lexicon-grammar interface: The case of complex predicates in the functional generative description. In Eva Hajičová & Joakim Nivre (eds.), *Proceedings of depling 2015*, 191–200. Uppsala, Sweden: Uppsala University.

Kettnerová, Václava, Markéta Lopatková, Eduard Bejček, Anna Vernerová & Marie Podobová. 2013. Corpus based identification of Czech light verbs. In Katarína Gajdošová & Adriána Žáková (eds.), *Proceedings of the 7th International Conference Slovko 2013; Natural Language Processing, Corpus Linguistics, E-learning*, 118–128. Lüdenscheid, Germany: RAM-Verlag.

Křen, Michal, Tomáš Bartoň, Václav CvrČek, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Renata Novotná, Vladimír PetkeviČ, Pavel Procházka, Věra Schmiedtová & Hana Skoumalová. 2010. *SYN2010: Balanced corpus of written Czech.* LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Krippendorff, Klaus. 2007. *Computing krippendorff's alpha reliability.* Tech. rep. University of Pennsylvania, Annenberg School for Communication. http://repository.upenn.edu/asc_papers/43.

Li, Linlin & Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP '09), 315–323. Singapore: Association for Computational Linguistics.

Liu, Changsheng & Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *HLT-NAACL*, 363–373.

Madnani, Nitin & Bonnie J. Dorr. 2013. Generating targeted paraphrases for improved translation. *ACM Transactions on Intelligent Systems and Technology* 4(3). 40:1–40:25.

Marton, Yuval, Chris Callison-Burch & Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1* (EMNLP '09), 381–390. Association for Computational Linguistics.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*. arXiv preprint arXiv:1301.3781.

Muzny, Grace & Luke S. Zettlemoyer. 2013. Automatic idiom identification in Wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2013), 1417–1421. Association for Computational Linguistics.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (ACL '02), 311–318. Stroudsburg, PA, USA: Association for Computational Linguistics. DOI:10.3115/1073083.1073135

Peng, Jing, Anna Feldman & Hamza Jazmati. 2015. Classifying idiomatic and literal expressions using vector space representations. In *Proceedings of recent advances in natural language processing*, 507–511. Association for Computational Linguistics.

Pershina, Maria, Yifan He & Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Workshop on linking models of lexical, sentential and discourse-level semantics* (LSDSem), 76–82. Association for Computational Linguistics.

Radimský, Jan. 2010. *Verbonominální predikáty s kategoriálním slovesem*. České Budějovice: Editio Universitatis Bohemiae Meridionalis.

Řehůřek, Radim & Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for*

NLP Frameworks, 45–50. European Language Resources Association (ELRA). May 22, 2010.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Salton, Giancarlo, Robert J. Ross & John D. Kelleher. 2014. An Empirical Study of the Impact of Idioms on Phrase Based Statistical Machine Translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation, HyTra@EACL 2014, April 27, 2014, Gothenburg, Sweden*, 36–41. http://aclweb.org/anthology/W/W14/W14-1007.pdf.

Sinha, R. Mahesh K. 2009. Mining complex predicates in Hindi using a parallel Hindi-English corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications* (MWE '09), 40–46. Association for Computational Linguistics.

van der Linden, Erik-Jan. 1992. Incremental processing and the hierarchical lexicon. *Computational Linguistics* 18(2). 219–238.

Wallis, Peter. 1993. Information retrieval based on paraphrase. In *PACLING '93, 1st Pacific Association for Computational Linguistics Conference](formerly JA-JSNLP, the Japan-Australia Joint Symposia on Natural Language Processing).*

Zarrieß, Sina & Jonas Kuhn. 2009. Exploiting translational correspondences for pattern-independent MWE identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications* (MWE '09). Association for Computational Linguistics.

Zhou, Liang, Chin-Yew Lin & Eduard Hovy. 2006. Re-evaluating Machine Translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (EMNLP '06), 77–84. Association for Computational Linguistics.

Žolkovskij, Alexander K. & Igor A. Mel'čuk. 1965. O vozmožnom metode i instrumentax semantičeskogo sinteza. *Naučno-texničeskaja informacija* 6. 23–28.

Chapter 3

Identifying senses of particles in verb-particle constructions

Archna Bhatia

Florida Institute for Human and Machine Cognition, Ocala, FL, USA

Choh Man Teng

Florida Institute for Human and Machine Cognition, Pensacola, FL, USA

James F. Allen

University of Rochester, Rochester, NY, USA
Florida Institute for Human and Machine Cognition, Pensacola, FL, USA

In order to attain broad coverage understanding, a system need not only identify multiword expressions such as verb-particle constructions (VPCs), but must compute their meaning. It is not plausible to hand enumerate all possible combinations, although WordNet is an admirable start. This chapter focuses on the identification of senses of particles in VPCs in order to compute the meanings of VPCs – using information obtained from existing lexical resources such as WordNet, and augmenting it with additional knowledge based on linguistic investigation of VPCs identified in terms of generalizations encoded in the TRIPS ontology. The approach consists of first determining compositionality of a VPC based on the information present in WordNet, and then assigning a relevant sense to the particle in a compositional VPC based on the sense classes we have identified and encoded in the TRIPS' computational lexicon. Contributions of the described work are twofold: (1) A discussion of senses of particles in VPCs and corresponding generalizations makes a linguistic contribution. (2) We show how linguistic knowledge can be used to automatically parse sentences containing VPCs and obtain a semantic representation of them. An advantage of the described approach is that VPCs not explicitly found in lexica can be identified and semantically interpreted.

Archna Bhatia, Choh Man Teng & James F. Allen. 2018. Identifying senses of particles in verb-particle constructions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 61–86. Berlin: Language Science Press. DOI:10.5281/zenodo.1469555

1 Introduction

To compute deep semantic representations of sentences, we need to pay attention to the richness of lexical meaning. Multiword expressions (MWEs) constitute a significant proportion of the lexicon in any natural language (Moreno-Ortiz et al. 2013). In fact, Jackendoff (1997) estimated the number of MWEs in a speaker's lexicon to be of the same order of magnitude as the number of single words. Thus, it is important to get a good interpretation of MWEs.

This chapter builds on and extends the work reported in Bhatia et al. (2017) with focus on a specific type of MWEs, namely verb-particle constructions (VPCs). VPCs consist of a verb and an adverbial or prepositional particle, e.g., *eat up*, *fade out*, *go on*, *show off*, and *walk down*.[1] Adding every single occurrence of such verb-particle combinations to a lexicon is not efficient nor ideal since knowledge about individual parts (verb and particle) can be leveraged for many of these VPCs as they are interpretable compositionally, e.g., *fly up*.

Other VPCs that indeed are noncompositional require special interpretation, and hence need to be added into the lexicon, e.g., *bring off* 'achieve a goal' and *egg on* 'urge someone for an action that might not be a good idea'. Our work on compositionality of VPCs, described in Bhatia et al. (2017) and developed further here, helps identify VPCs of each type for a proper treatment.

For an inventory of senses for verbs, many lexical resources, such as WordNet (Miller 1995; Fellbaum 1998) and the TRIPS lexicon (Allen & Teng 2017), are available that can be leveraged for interpreting compositional VPC types. In contrast, there is not much for particles except for a few attempts at the semantics of a few particles, such as *up* (Cook & Stevenson 2006) and *out* (Tyler & Evans 2003). However, particles seem to add their own semantics in compositional VPCs and are found to be regular when occurring with verbs in specific verb classes. For example, the particle *up* has a DIRECTION sense when it appears in resultative VPCs with verbs of motion, such as *wander/stroll/go/run up* (Villavicencio 2006). In this chapter, we provide a refined set of senses for particles in VPCs originally presented in Bhatia et al. (2017). We discuss and demonstrate how these senses are identified in compositional VPCs in order to compute meanings of sentences containing VPCs.

For computation of meaning, we use a broad coverage deep semantic parser, TRIPS (Allen et al. 2007), which combines semantic, ontological, and grammati-

[1] Note that we focus on the particle usage here, not on the prepositional usage, i.e., a verb followed by a particle not a prepositional phrase. However, there may be an overlap in lexical semantic content (i.e., senses) of the homophonous particles and prepositions, see §4.1.

cal information to produce semantic representations.[2] We encode the semantics of particles mentioned above in the TRIPS ontology.[3] The ontology encodes semantic types, the set of word senses and semantic relations that can be used in logical form (LF) graphs. Word senses are defined based on subcategorization patterns and selectional restrictions driven by linguistic considerations. The semantic types in the ontology are, to a large extent, compatible with FrameNet (Johnson & Fillmore 2000). The ontology uses a rich set of semantic features. Unlike WordNet, our ontology does not attempt to capture all possible word senses, but rather focuses on the level of abstraction that affects linguistic processing.

This chapter builds on the work described in Bhatia et al. (2017) in the following ways: Improvements have been made in the classification of compositionality and sense types as well as in the heuristics to automatically identify the types. The evaluation is made more robust with a larger test data set with full coverage of the heuristics. A more thorough analysis has led to the identification of generalizations regarding sense types.

The chapter is organized as follows: Previous work on VPCs is discussed in §2. A classification of VPCs based on their compositionality is discussed in §3 with its feasibility using inter-annotator agreement in §3.1. A set of heuristics to identify different classes of VPCs are presented in §3.2 and their evaluation in §3.3. In §4, we discuss the semantics of particles in VPCs. An inventory of general sense classes for particles used in VPCs is provided in §4.1, which is followed by a brief discussion of manual sense annotations as well as the use of compositionality heuristics for sense identification. In §5, we present various generalizations corresponding to the identified sense classes for the particles, and briefly discuss how a computational lexicon (including a lexicon for particles) is built for the computation of meaning for VPCs. Through examples, we demonstrate the general procedure to compute the meaning of sentences involving compositional VPCs and that the linguistic generalizations are reasonably helpful in the accurate identification of particle senses in VPCs. §6 concludes the chapter.

[2] For a more detailed overview of the TRIPS system, refer to Allen & Teng (2017) and Allen et al. (2008).
The TRIPS parser can be accessed at: http://trips.ihmc.us/parser/cgi/parse

[3] The TRIPS ontology can be accessed at: http://www.cs.rochester.edu/research/cisd/projects/trips/lexicon/browse-ont-lex-ajax.html

2 Related work

A lot of computational literature on VPCs focuses on the identification or extraction of VPCs, or on the compositionality of VPCs, as discussed below. There are a few articles dealing with different senses of particles, but they usually focus on only one or two specific particles rather than on a broader coverage of particles. For example, Niinuma (2014) discusses grammaticalization of the particle *away* in English, specifying directional, completive, and continuing or iterative usages of the particle. Ishizaki (2010; 2012) also studies grammaticalization of the particle *away* together with the particle *out*, presenting a classification of VPCs into three categories of fully compositional, partially idiomatic, and fully (or highly) idiomatic.

Vincze et al. (2011) presents the Wiki50 corpus that has 446 VPCs (342 unique types) annotated. Bannard (2002) makes an attempt to identify different types of VPCs in terms of compositionality and builds a (decision tree) classifier to identify the four types. Bannard et al. (2003) also adopt a similar approach for compositionality. As an annotation experiment, they investigate various VPCs to see whether the sense is contributed by the verb and/or the particle. They build four classifiers for automatic semantic analysis of VPCs. Patrick & Fletcher (2004) also have a similar approach, but they focus on automatic classification of different types of compositionality. Unlike our work, in all these works, the focus is on compositionality only, not on the identification of actual senses of the particles.

Cook & Stevenson (2006) discuss various senses for the particle *up* in a cognitive grammar framework, annotate a dataset and perform some classification experiments to identify the senses of *up* in unseen data. As a linguistic study, Jackendoff (2002) provides a very nice discussion of various types of VPCs involving particles such as directional particles, aspectual particles, time-AWAY constructions, and some idiomatic constructions. Our work differs from theirs in having a broader coverage of particles and/or strong emphasis on ontology with respect to the sense classes of the particles and how different particle sense classes relate to verbal ontological classes.

Fraser (1976) mentions semantic properties of verbs affecting patterns of verb-particle combinations, for instance semantically similar verbs *bolt/cement/clam/glue/paste/nail* all can combine with the particle *down* and specify the objects that can be used to join material. Our approach is based on the similar assumption that there are generalizations, such as particles with certain sense classes combine with specific verb classes or ontological classes. Villavicencio (2003) also adopts the same approach where she tries to encode the information in terms of lexical rules and restrictions, etc. However, her focus is on obtaining productive patterns in VPCs rather than on their interpretation.

Our work also differs from the previous work mentioned above in the following respect: We emphasize the building of complete semantic representations of the sentences, not just the particles' semantics or just the classification of VPCs. Similar to our criteria for compositionality, McCarthy et al. (2003), Baldwin et al. (2003), and Bannard et al. (2003) have looked at distributional similarity as a measure of compositionality of VPCs. In contrast to the approaches focusing on statistical classification based on word/syntax features, our approach (both symbolic and statistical) uses information obtained from existing lexical resources, such as WordNet, for the classification of VPCs. We augment it with additional knowledge based on the linguistic investigation of VPCs identified in terms of generalizations, which we encode into an ontology, in order to compute the semantics of the compositional classes.

3 Classification of VPCs

VPCs have often been classified in terms of their compositionality (i.e., whether all constituents of a VPC, the verb and the particle, contribute their simplex meanings to the overall semantic content of the VPC). The classes fall somewhere between fully compositional VPCs, e.g., *fly up*, and fully idiomatic VPCs, e.g., *egg on*. For example, see Fraser (1976), Chen (1986), O'Dowd (1998), Dehé (2002), and Jackendoff (2002).

We also classify VPCs into two types, compositional VPCs and noncompositional VPCs. The difference between the two classes is that the meaning of a compositional VPC is the sum of the meanings of its parts (the verb and the particle) whereas a noncompositional VPC has semantic content which is not contributed by the individual constituents (i.e., the verb and the particle).

The compositional VPCs can be further classified into three subtypes: symmetrically compositional VPCs, light particle compositional VPCs (LP-compositional VPCs), and light verb compositional VPCs (LV-compositional VPCs), based on the type of semantic content contributed by each of the constituents. Symmetrically compositional VPCs refer to VPCs in which both constituents, the verb and the particle, contribute their simplex meanings (their lexical-semantic content). For example, in *Debris **flew up** and hit the window in the furthest unit*, the senses for the verb *fly* (e.g., in WordNet, sense fly%2:38:00) as well as the particle *up* (e.g., in WordNet, sense up%4:02:00) combine together to provide the meaning of the VPC *fly up*.[4] We distinguish the other two compositional VPC types from the symmetrically compositional VPCs only in the aspect that in the other two types, the particle or the verb have a relatively lighter contribution than the other constituent which adds its regular lexical-semantic content.[5]

LP-compositional VPCs involve particles which, instead of contributing a preposition like lexical-semantic content, contribute aspectual information to the VPCs. Verbs contribute most of the lexical-semantic content in such VPCs. For example, in *Susan **finished up** her paper* (Bannard & Baldwin 2003), the verb *finish* contributes its regular lexical content (e.g., in WordNet, sense finish%2:30:02). However, the particle *up*, instead of contributing its regular lexical-semantic content, adds aspectual information that the action was completed (i.e., the COMPLETELY sense in our sense inventory). See §4.1 for the specific senses of particles.

LV-compositional VPCs involve light verbs which carry bleached meaning compared to regular verbs, e.g., CAUSE and BECOME. The particles contribute their regular lexical-semantic content to the VPCs's semantics. For example, in *The thief **made away** with the cash*, the particle *away* contributes its regular meaning (e.g., WordNet sense away%4:02:00), but the verb *make*, instead of contributing its regular meaning (e.g., WordNet sense make%2:36:01), adds a bleached meaning (e.g., cause to be). For details on the procedure to compute meanings of sentences with compositional VPCs, see §5.

In a noncompositional VPC, the sum of meanings of individual parts (the verb and the particle) may not completely account for the meaning of the VPC or may not account for it at all. Let's consider a few examples. In *They **turned away** hundreds of fans*, the VPC ***turned away*** is noncompositional despite the fact that the individual constituents' semantic content is reflected in the semantics of the VPC, since the VPC has additional content ('refuse entrance or membership') besides those contributed by the constituents. This additional content is not inferable from the individual parts and needs to be included in the lexical entry for the VPC. Another example of a noncompositional VPC is idiomatic usages. For example, the VPC **egged on** in *John wouldn't have done the dangerous experiment if his brother hadn't **egged** him **on*** involves idiosyncratic content that is not inferable from the parts and hence needs to be encoded in the lexicon. Some idiomatic usages may involve certain generalizations which may aid in interpre-

[4] For this study, we have used WordNet version 3.0. The numbers appearing after the symbol % in the WordNet senses represent the sense keys in WordNet.

[5] The term "light particle" is used in analogy with the term "light verb" which is commonly used in the literature for verbs with bleached content.

tation of the VPC. For example, two generalizations regarding the interpretation of the noncompositional VPC ***take up*** are (i) it takes the sense 'starting to do an activity' when it appears with activities as direct objects, e.g., *She* ***took up*** *photography/swimming* [activities] and (ii) it takes the sense 'assume a responsibility' when it appears with positions/responsibilities as direct objects, e.g., *She* ***took up*** *her position* [responsibility/position].

For identification of the compositionality type for a given VPC, one may adopt tests as in Figure 1.

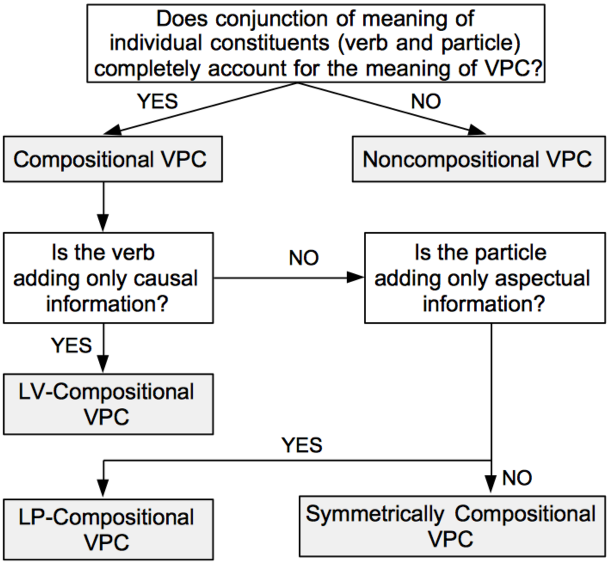

Does conjunction of meaning of individual constituents (verb and particle) completely account for the meaning of VPC?
YES
NO
Compositional VPC
Noncompositional VPC
Is the verb adding only causal information?
NO
Is the particle adding only aspectual information?
YES
LV-Compositional VPC
YES
NO
LP-Compositional VPC
Symmetrically Compositional VPC

Figure 1: Tests to identify compositionality type

3.1 Human agreement on coarse-grained classification of VPCs

From among all the VPCs for which WordNet has an entry, we automatically extracted 50 random VPCs such that four particles, namely *up*, *down*, *out*, and *away*, were represented in the extracted VPCs. Since a VPC may have both compositional and noncompositional usages in different contexts (represented by different word senses or synsets in WordNet), we restricted the assignment of annotation label for a specific VPC to only one label by considering a single synset from WordNet for annotations. Some of the WordNet synsets do not have an example with the exact VPCs.[6] We restricted the automatic extraction of VPCs to only those VPCs for which WordNet had a synset which included the VPC in its example. This synset was presented to the annotators together with the VPC to be annotated and the example usage.

These test VPCs were manually annotated by three annotators for compositionality labels. Fleiss' kappa score (Fleiss 1971) was used to test inter-annotator reliability. The three sets of annotations achieved a score of 0.651, an intermediate-good score.[7] We also created the gold annotations from the three sets of manual annotations. In cases of disagreement, the three annotators discussed reasons for their decisions and arrived at a consensus to create the gold annotations for the VPCs. The distribution of compositional vs. noncompositional VPCs in the automatically extracted 50 test VPCs was 60% vs. 40%. In terms of the specific compositionality types, the distribution was 30% for symmetrically compositional cases, 30% for LP-compositional, and 40% for noncompositional cases. Note that the 50 randomly extracted test VPCs did not have an instance of an LV-compositional VPC.

In the manual annotations, there were 37 VPCs out of 50 for which there was full agreement among all three annotators on the coarse-grained labels for compositionality type. Since there were only two coarse-grained labels (compositional and noncompositional), for the rest of the 13 VPCs, there was agreement between two out of three annotators for the compositionality label. The annotators disagreed more on the compositional cases (76.92% of the disagreements were on the compositional VPCs) than on the noncompositional ones (23.08%).

In terms of VPCs with certain particles, VPCs with the particle *down* were found to be the most challenging for the annotators. For 42.86% of the usages of VPCs with *down* (3 out of 7 usages) among the 50 test cases, the annotators did not agree on the annotation label. On the other hand, VPCs with the particle *up* were found to be the least challenging. For only 14.29% of the cases (3 out of 21 usages), the annotators disagreed on the compositionality type. This may be due to the fact that VPCs with the particle *up* have a higher frequency than VPCs with other particles (Villavicencio 2006), such as *down*, and hence users have relatively better intuitions about VPCs with the particle *up* compared to VPCs with the particle *down*.

[6] It may have examples which use related verbs or related VPCs instead, or it may not have an example at all. For example, the WordNet entry *turn_away%2:38:02* for the VPC ***turn away*** does not have any example with the VPC itself.

[7] NLTK's agreement package was used to calculate the Fleiss' kappa score which showed intermediate-good agreement. Cohen's kappa score (Cohen 1960) was also calculated using the same package showing substantial agreement, also with a score of 0.651.

3.2 Heuristics for compositionality of VPCs

As a first step toward an interpretation of VPCs, we need to determine whether a given VPC is compositional or not. To perform this task automatically, we employ a number of heuristics that make use of the rich inventory of hierarchically organized word senses (i.e., synsets) in WordNet which contains over 100,000 words including 64,188 multiwords. Heuristics 1–6 below are used to identify compositional VPCs, whereas heuristic 7 indicates non-compositionality.

1. If the verb is among the list of light verbs, and WordNet does not have an entry for the VPC, it is LV-compositional. For example, the VPC *make away* uses the light verb *make* and the VPC does not have an entry in WordNet.
2. Given a VPC, if heuristic 1 does not apply and WordNet has an entry for the verb as well as for the particle, but no entry for the VPC, VPC is compositional (LP-compositional or symmetrically compositional). For example, *fly* with the sense key fly%2:38:01 as well as *up* with the sense key up%4:02:00 appears in WordNet, but *fly up* does not appear in any synset in WordNet.
3. If WordNet has the VPC as well as the verb in the same synset, VPC is LP-compositional. For example, the VPC *sort out* (sort_out%2:31:00) and the verb *sort* (sort%2:31:00) both appear in the same synset in WordNet.
4. If WordNet has the verb in the VPC as a hypernym for the VPC, VPC is either symmetrically compositional or LP-compositional. For example, compositional VPC *push up* (push_up%2:38:00) has the verb *push* (push%2:38:00) as its direct hypernym.
5. If WordNet has the verb in the definition (in its base or inflected form) of the synset where the VPC appears, the VPC is either symmetrically compositional or LP-compositional. For example, the compositional VPC *move up* (move_up%2:38:00) has the verb *move* in its definition *move upwards.*
6. If WordNet has the relevant VPC as well as another VPC with the verb replaced with another verb in the same synset, the VPC is compositional (either symmetrically compositional or LP-compositional or LV-compositional). For example, the compositional VPCs *pull out* (pull_out%2:35:00) and *rip out* (rip_out%2:35:00) appear in the same WordNet synset.

7. If none of the above heuristics apply, the VPC is noncompositional. For example, none of the above heuristics apply to the idiomatic VPC *catch up* (catch_up%2:38:0).

3.3 An evaluation of the heuristics for compositionality of VPCs

For an evaluation of the heuristics, we used two test sets: (i) Test Set 1: the test set consisting of the same 50 randomly extracted VPCs that were used to calculate inter-annotator agreement scores mentioned in §3.1, and (ii) Test Set 2: a test set consisting of 653 VPCs created using the VPCs on the first page of each of the four English wiktionary entries with the title "Category: English phrasal verbs with particle" for particles *up*, *out*, *down*, and *away* respectively.

A Python implementation of the heuristics was applied to the test VPCs in both test sets to assign them a compositionality label. Heuristics 1–6 identify compositional VPCs, whereas heuristic 7 identifies noncompositional VPCs. Note that together these heuristics have full coverage, i.e., a prediction is made for each of the VPCs in the test sets. Heuristics 1 and 2 apply to VPCs for which WordNet does not have an entry. Heuristics 3-6 apply to VPCs for which WordNet has an entry. Heuristic 7 applies when none of the heuristics 1-6 apply.

Since Test Set 1 consisted of VPCs that were randomly extracted from WordNet, only heuristics 3–7 could be evaluated on Test Set 1. Heuristics 1 and 2 could not be evaluated using this test set since these two heuristics apply to VPCs that are not included in WordNet. In order to test heuristics 1 and 2, we used Test Set 2 which has VPCs that may or may not be included in WordNet.

Test Set 1 has manually created gold annotations. Hence, the labels assigned by the heuristics were tested against them for the VPCs in Test Set 1. For Test Set 2, on the other hand, we examined the heuristics-assigned labels corresponding to heuristics 1 and 2 manually.

An evaluation of the heuristics on Test Set 1 is presented in Table 1. Heuristics 3, 4, and 5 (used to identify compositional VPCs) performed perfectly in identifying compositionality of VPCs whereas heuristic 6 (also intended to identify compositional VPCs) did not have as high precision. Two out of the three cases where heuristic 6 made an incorrect prediction, however, involved noncompositional VPCs for which the individual constituents' meaning was also reflected in the semantics of the VPCs even though they also had additional content that was not inferable based on the constituents alone. The fact that the constituents' semantics was reflected in the VPCs' semantics was what heuristic 6 had captured. Heuristic 7 (the only heuristic used to identify noncompositional VPCs) also had a lower but better than chance performance in comparison to the other heuristics.

Table 1: Evaluation of heuristics 3–7 using Test Set 1 (50 test cases)[a]

Heuristic #	Precision for the heuristic (%)	Coverage of the heuristic (%)
3	100	16
4	100	6
5	100	12
6	76.92	26
7	58.62	58
Overall (heuristics 3–7)	70	100

[a] We define precision as Cn/Tn and coverage as Tn/N, where N is the corpus size, Tn is the sample size that heuristic n is applicable to, and Cn is the number of correct assignments it makes.

Overall, heuristics 3–7 had a precision of 70% in assigning a label (compositional or noncompositional) to a VPC. Whenever multiple heuristics applied to a VPC (14% of the test cases), they were always correct in their prediction. This suggests that presence of multiple characteristics of compositionality (represented by the heuristics) may be a reliable indicator of compositionality. In 86% of the cases, a single heuristic identified the VPC as representative of its own category (compositional or noncompositional). Regarding the relatively lower performance of heuristic 7, since noncompositonal VPCs carry additional information not contributed by individual constituents (common sense knowledge or idiosyncratic information), it may not be observable and hence is not captured using a heuristic as easily or directly.

More investigation is required into heuristics 6 and 7 to see how they can be modified to make them better usable for identification of the compositionality type of a VPC. For example, a future step may be to examine if certain generalizations exist in terms of verb types and particle types that can help us further determine when the cases identified by these heuristics are compositional or noncompositional.

As a result of the examination of the heuristics' output, we noticed a few more indicators which could also be incorporated into the heuristics to improve their performance in the future. For example, if the word *completely* or *thoroughly* appears in the definition of a synset of a VPC, the particle in the VPC may carry the aspectual sense COMPLETELY. Hence, it could be labelled as a compositional VPC (specifically, as an LP-compositional VPC).

As mentioned above, for an evaluation of the heuristics 1 and 2, we used Test Set 2 that consisted of 653 VPCs, a subset of the VPCs mentioned in the English wiktionary. The Python implementation of the heuristics mentioned above was used to assign compositionality labels to the VPCs in Test Set 2. Out of the 653 VPCs, heuristic 1 applied to only 2 VPCs (0.3% of the test items) and heuristic 2 applied to 280 VPCs (42.88% of the test items). Heuristic 7 applied to 9 other VPCs that did not have a WordNet entry. Overall, 44.56% of the test cases did not have an entry in WordNet. Heuristic 2 covered 96.22% of these cases with reasonable precision (80%, evaluated on 15 randomly selected test VPCs to which heuristic 2 applied).

Next, we move on to the semantics of particles in VPCs. We will focus mostly on the compositional VPCs for the rest of this chapter.

4 Semantics of particles in VPCs

As mentioned in §3, particles contribute to the overall semantics of compositional VPCs. In order to study the contribution of particles in VPCs, in our prior work (Bhatia et al. 2017), we conducted an investigation of VPCs consisting of verbs in the ontology class ONT::EVENT-OF-CAUSATION in the TRIPS ontology. This class consisted of 1383 words with verb senses (and a total of 1784 verb senses of those words). Our investigation consisted of combinations of these verbs with the following particles (wherever the combinations appeared as VPCs): *across*, *away*, *by*, *down*, *in*, *into*, *off*, *on*, *out*, *over*, *through*, and *up*.[8] We searched for examples for each of the combinations using Google and manually went through each of the examples to check for a number of properties. For example, we checked if any of the verb or the particle contributed to the overall meaning of the VPC. We identified the senses particles had in the VPCs if any. We checked: (i) if the particle could be taken out without a major change in the meaning, (ii) if the particle expressed RESULT or could be replaced with a RESULT-Prepositional Phrase,[9] (iii) if a corresponding VPC consisting of the particle with the opposite

[8] While we do find prepositional phrasal verb constructions with *into*, we did not find any VPCs involving intransitive usages of *into*.

[9] RESULT is one of the argument roles identified in the TRIPS ontology. The argument roles signal different argument positions for predicates as well as have their own inferential import, some other examples are AGENT, AFFECTED, MANNER, LOCATION, and FIGURE.

polarity was also possible (e.g., *take in* vs. *take out*), (iv) if specific argument types (e.g., MANNER, RESULT, LOCATION, AFFECTED) were instantiated in the sentence, etc. In the rest of this section, we present the sense classes particles in compositional VPCs tend to fall into.

4.1 Sense classes for particles in VPCs

While particles may encode subtle nuances of meanings in each of their occurrences in (compositional) VPCs, they may also display some general senses across many VPCs. WordNet attempts to capture the nuances by storing each of the VPCs as a separate lexical item. However, this approach results in having as many sense categories as there are VPCs and we lose information about the common contributions made by the particles in VPC semantics which can be useful while producing semantic representation of sentences with new VPCs not stored in WordNet or another lexical resource. Hence, we focus on the general senses particles display across compositional VPCs.

We identified two general sense classes for the particles in compositional VPCs, namely DIRECTION and ASPECTUAL. The DIRECTION sense class has a number of subclasses, each instantiated by a specific directional particle, such as *away*, *down*, *in*, *off*, *on*, *out*, and *up* denoting a specific direction sense. For example, the directional particle *away* instantiates the subclass DIRECTION-AWAY in *I took one last look at the house and* ***walked away***. Similarly, the particle *out* in *My mom never* ***threw*** *it* ***out*** and the particle *up* in *The magic ketchup should sink when you squeeze the bottle and* ***float up*** *when you release it* instantiate subclasses DIRECTION-OUT and DIRECTION-UP respectively. The DIRECTION sense class particles assume the RESULT role in relation to the verb. For example, in the DIRECTION-AWAY example, the particle *away* denotes the RESULT of a *walking* event in terms of the direction the AFFECTED entity is walking to.

The ASPECTUAL sense class has two subclasses, namely COMPLETELY and CONTINUING, where the particle modifies the verb by providing aspectual information. In terms of the TRIPS semantic roles, the aspectual sense particles assume the MANNER role in relation to the verb. The COMPLETELY sense is used to express that the activity denoted by the verb is performed to the full extent or with thoroughness. Particles with COMPLETELY sense may also emphasize the telicity of an action. For example, the particle *out* in *He* ***sorted out*** *every scrap of manuscript, every map, and the native letters* emphasizes that each of the items mentioned were thoroughly and completely sorted. Particles *up*, *out*, and *down* are used more often than other particles to convey this aspectual sense.

The CONTINUING sense is used to emphasize the durative nature of the event denoted by the verb. For example, in *Day after day she* ***worked away*** *…* the particle *away* conveys that the activity continued for a duration. The particles in this class may also convey an iterative sense when they are used with semelfactive verbs, e.g., note the use of the particle *away* in *Start your explorations here,* ***click away*** *all you want.* Usually particles *away* and *on* are used to convey this sense in VPCs.

These sense classes correspond to the VPC classes based on their compositionality types mentioned in §3. For example, the DIRECTION sense class is generally instantiated by the symmetrically compositional or LV-compositional VPCs. The ASPECTUAL sense class, on the other hand, is instantiated by the LP-compositional VPCs. Besides these two sense classes, there are a few other senses that particles may express in VPCs corresponding to the symmetrically compositional and LV-compositional VPCs. For example, the senses IN-WORKING-ORDER-VAL and NOT-IN-WORKING-ORDER-VAL mentioned in Table 2 and Table 4 are expressed by particles in LV-compositional VPCs. These senses are usually conveyed by the particles *up*, *down*, and *out*. Similarly, the sense DISAPPEARANCE, conveyed by the particle *away*, may be taken to correspond to the particle in symmetrically compositional VPCs. Table 2 contains information about each of these senses and corresponding sense classes for the particles *up*, *down*, *away*, and *out*. Table 3 and Table 4 present generalizations corresponding to these senses and sense classes.

We did not get very high agreement on human annotations for the sense labels for the particles in VPCs from Test Set 1. The three annotators had full agreement in 20 cases (40% of the cases). 75% of these 20 cases had DIRECTION sense, 20% had COMPLETELY, and 5% IDIOMATIC. The disagreement cases involved labels such as DIRECTION vs. COMPLETELY, DIRECTION vs. IDIOMATIC, or COMPLETELY vs. IDIOMATIC. For a more reliable identification of senses for the particles in VPCs, one may adopt tests as in Figure 2.

We also explored the use of compositionality heuristics to determine the senses of particles in VPCs, the general sense labels being (DIRECTION, ASPECTUAL, and IDIOMATIC). For example, if heuristic 3 applied to a VPC, the particle was assumed to have an ASPECTUAL sense. If heuristic 7 applied, the VPC was taken to be noncompositional and hence with an IDIOMATIC sense. In 30% of the cases, a correct sense label was assigned unambiguously. In another 32% of the cases, multiple sense labels were identified and one of them was correct. In the rest of the 19 cases where the sense label based on the application of specific heuristics was incorrect, the DIRECTION sense was the most misclassified sense (57.89%), followed by the COMPLETELY sense (31.58%), followed by the IDIOMATIC sense (10.53%).

Table 2: Sense classes for particles *up*, *down*, *away*, and *out* in compositional VPCs.

Sense class		
Sense type	Particle	Example
DIRECTION:		
DIRECTION-UP	up	He took the carpet up.
DIRECTION-DOWN	down	He skied down.
DIRECTION-AWAY	away	He ran away.
DIRECTION-OUT	out	He cried out.
ASPECTUAL:		
COMPLETELY	up	Clean up the room!
	down	London nightclub closed down over fights with knives and bottles.
	out	He sorted out every scrap of manuscript, every map, and the native letters.
ASPECTUAL:		
CONTINUING	away	Night and day they hammered away, coming on like great waves.
		He scrubbed away at the floor.
IN-WORKING-ORDER-VAL	up	Bring the browser up!
NOT-IN-WORKING-ORDER-VAL	down	The computer went down again.
	out	The national electric grid went out.
DISAPPEARANCE	away	The echo died away.
		The music faded away.

Table 3: Generalizations about sense classes DIRECTION and ASPECTUAL for particles in compositional VPCs.

Sense class: Sense type	Generalizations
DIRECTION: DIRECTION-UP DIRECTION-DOWN DIRECTION-AWAY DIRECTION-OUT	verb trajectory = + Particle relative to some scale/domain PP alternative is possible for the particles *up*, *down*, and *out*. The particle *away* can take a PP-location as its GROUND argument. The senses tend to appear with motion and movement-related verbs in the TRIPS ontology. Cases pass two tests: (1) Is there a change in physical location/on a scale? (2) Does the entity exist anywhere? DIRECTION-UP is also possible with verbs in the TRIPS ontology classes apply-force and acquire. DIRECTION-DOWN is also possible with verbs in the TRIPS ontology classes hitting. DIRECTION-AWAY is also possible with event-of-change verbs where grammatical resultative construction is used. DIRECTION-OUT is also possible with verbs of vocalization (and the verbs in the TRIPS ontology classes locution and manner-say) and some perception verbs.
ASPECTUAL: COMPLETELY	verb trajectory = - PP alternative not possible Default sense with the event-of-change verbs in the TRIPS ontology COMPLETELY for the particle *up* is also possible with verbs in the TRIPS ontology classes protecting, joining, acquire-by-action, herd, and arrange-text. COMPLETELY for the particle *down* is also possible with verbs in the TRIPS ontology classes change, consume, protecting, put, and pursue. COMPLETELY for the particle *out* is also possible with verbs in the TRIPS ontology classes evoke-tiredness, arranging, and put.
ASPECTUAL: CONTINUING	verb trajectory = - Usually with atelic verbs that have extended duration Iterative usage with semelfactive verbs PP alternative is not possible, i.e., particle cannot take NP object, but can take PP object in conative constructions. CONTINUING is also possible with verbs in the TRIPS ontology class event-of-action.

Table 4: Generalizations about senses IN-WORKING-ORDER-VAL, NOT-IN-WORKING-ORDER-VAL, and DISAPPEARANCE for particles in compositional VPCs. Note that NP* refers to the object argument of a transitive verb, e.g., *the carpet* in *he took the carpet up*, or the subject of an intransitive verb, e.g., *he* in *he skied down*.

Sense class: Sense type	Generalizations
IN-WORKING-ORDER-VAL	NP* is phys-obj object-function = provides-service-up-down i.e., the entity is a device with some functionality. PP alternative not possible The verbs tend to be light/causal verbs.
NOT-IN-WORKING-ORDER-VAL	NP* is phys-obj object-function = provides-service-up-down i.e., the entity is a device with some functionality. PP alternative not possible NOT-IN-WORKING-ORDER-VAL is also possible for the particle *down* with verbs in the TRIPS ontology class event-of-undergoing-action. The verbs tend to be light/causal verbs.
DISAPPEARANCE	verb trajectory = - Usually with verbs of disappearance PP alternative not possible Cases fail two tests: (1) Is there a change in physical location/on a scale? (2) Does the entity exist anywhere? DISAPPEARANCE is also possible with verbs in the TRIPS ontology classes event-of-undergoing-action and change.

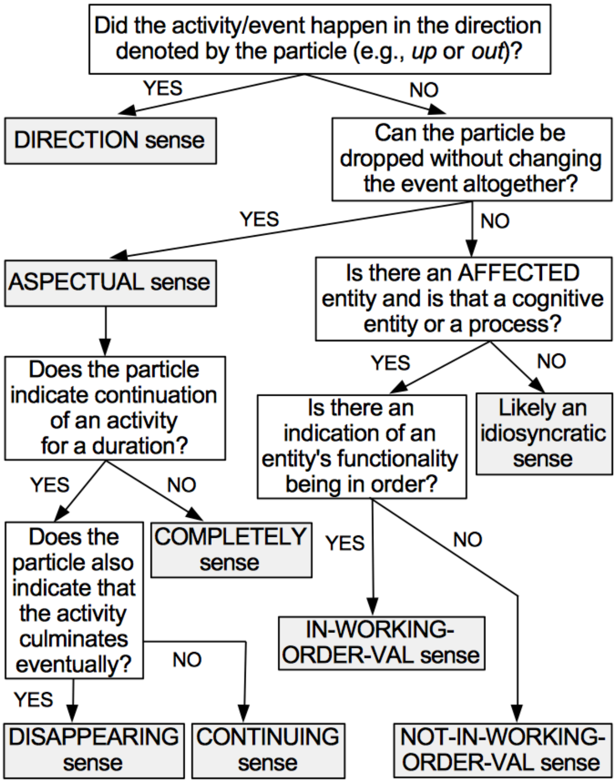

Did the activity/event happen in the direction denoted by the particle (e.g., *up* or *out*)?
YES
NO
DIRECTION sense
Can the particle be dropped without changing the event altogether?
YES
NO
ASPECTUAL sense
Is there an AFFECTED entity and is that a cognitive entity or a process?
Does the particle indicate continuation of an activity for a duration?
YES
NO
Is there an indication of an entity's functionality being in order?
Likely an idiosyncratic sense
YES
NO
Does the particle also indicate that the activity culminates eventually?
COMPLETELY sense
YES
NO
IN-WORKING-ORDER-VAL sense
NO
YES
DISAPPEARING sense
CONTINUING sense
NOT-IN-WORKING-ORDER-VAL sense

Figure 2: Tests to identify particle senses

5 Computing semantics of sentences with VPCs

For the task of interpreting sentences with VPCs, we first need to determine if the VPC is compositional or not. We use the heuristics mentioned in §3.2 to determine the compositionality of a VPC. For the compositional cases, we get the senses for the verb and the particle from the TRIPS ontology and/or WordNet. The senses for the particles as well as relevant linguistic generalizations to identify these senses are encoded in the TRIPS lexicon and ontology. In this section, we briefly discuss some of the generalizations encoded in the ontology and demonstrate the process of computing the semantics of sentences containing compositional VPCs using three example sentences involving VPCs with different senses for the particle *up*: *She cleaned up her room*, *She pushed the ball up*, and *The network came up*. The logical forms (LFs) produced for these sentences using the TRIPS parser, a broad coverage deep semantic parser driven by the TRIPS ontology, are presented in Figure 3 to Figure 5.

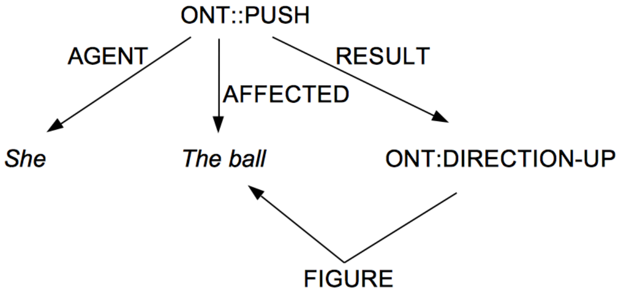


Figure 3: LF for the sentence: *She pushed the ball up.*

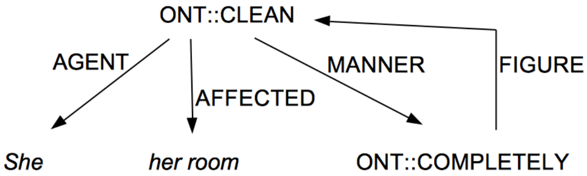


Figure 4: LF for the sentence: *She cleaned up her room.*

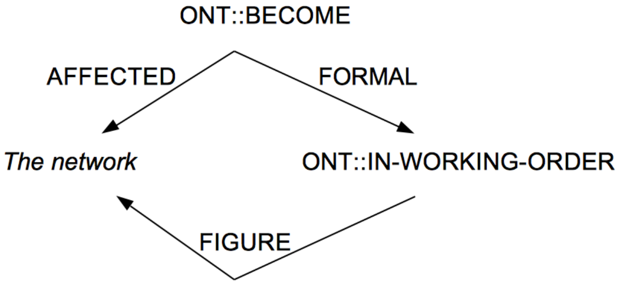


Figure 5: LF for the sentence: *The network came up.*

Particles in compositional VPCs can express the senses mentioned in Tables 2–4, and in Figure 2. This information is encoded in the TRIPS ontology by adding the ontology types corresponding to these senses in the particle's lexicon. For example, the lexical entry for the particle *up* lists sense ontology types ONT::DIRECTION-UP, ONT::COMPLETELY, and ONT::IN-WORKING-ORDER-VAL, used in Figure 3 to Figure 5 respectively, among other possible senses.[10]

[10] For a better idea of what information the lexical entries and semantic/ontology classes carry in the TRIPS lexicon and ontology, see http://www.cs.rochester.edu/research/cisd/projects/trips/lexicon/browse-ont-lex-ajax.html.

WordNet sense keys corresponding to the particle may be added in the ontology entries for these sense ontology types. For example, WordNet sense keys up%4:02:00 and up%4:02:05 are added in the entry for the sense ontology type ONT::DIRECTION-UP.

The senses that particles express in a VPC may depend on the verb type they combine with in the VPC. That is, a particle may convey the same sense when it appears with any of the verbs in a specific verb ontology class.[11] For example, particles tend to get DIRECTION sense with verbs of motion and CONTINUING sense with atelic verbs that have extended durations (e.g., activity type verbs). With verb ontology types corresponding to continuous-change verbs (which appear under change verbs in the TRIPS ontology), particles tend to get the COMPLETELY sense. Specific verb ontology classes are also identified for specific particles, e.g., particle *down* exhibits COMPLETELY sense with the verbs in the TRIPS ontology class ONT::PURSUE, as can be seen in *The internet* ***tracked down*** *this guy's stolen car (...)* and *A motorist* ***chased down****, slapped, and threatened a boy (...)*.

In addition, we observed an interesting fact that the particles *up* and *out* seem to be in complementary distribution with respect to various verb ontology classes for the COMPLETELY sense. That is, for the COMPLETELY sense, either *up* or *out* is used, but not both with verbs from a specific verb ontology class.[12] For example, with the verb ontology class ONT::ACQUIRE, *up* is used with COMPLETELY sense, *out* cannot be used with the verbs in this ontology class with the same sense. Note the COMPLETELY sense in the VPC *acquire up* in *Techstars has* ***acquired up*** *Global*, but we do not observe a VPC *acquire out* with the same sense. Similarly, with the verb ontology class ONT::EVOKE-TIREDNESS, *out* is used with the COMPLETELY sense, but *up* cannot be used. Note that we can say, *Someone's a bit* ***tuckered out***, but not ***tuckered up***.

Similarly there are generalizations observed for specific senses of particles corresponding to the semantic relation labels. For example, the verb takes a particle with an ASPECTUAL sense as its MANNER argument and the ASPECTUAL sense particle takes the verb as a FIGURE, as is observed in the LF in Figure 4. For the DIRECTION sense class particles, the verb assigns a RESULT argument

[11] We find that different verb ontology types that were distinguished for other reasons in the TRIPS ontology (Allen et al. 2007) also line up with the particle distinctions.

[12] This observation about the complementary distribution of usage between *up* and *out* may not be accidental. The Law of Differentiation (Paul 1890; Bréal 1900), and the Avoid Synonymy principle (Kiparsky 1983; Clark 1987) have been proposed in the lexico-semantic sphere which suggest that languages prefer not to have a given semantic slot be filled by two distinct lexical items.

role instead of a MANNER role to the particle, as illustrated in Figure 3. The particle, on the other hand, assigns the FIGURE role to the AFFECTED entity as in Figure 3. For the IN-WORKING-ORDER-VAL sense, the particle takes the AFFECTED or NEUTRAL entity as its FIGURE argument as is illustrated in Figure 5. Thus, we see that different semantic relations may be involved in sentences with different senses for the particles.

In order to get the correct semantic relations in different cases, we encode this information in the ontology. The sense ontology type ONT::COMPLETELY, used in Figure 4, for example, specifies for its FIGURE argument all the verb ontology types with which a particle gets this sense. This information is presented in the Generalizations column in Table 3. Note that under ONT::COMPLETELY, we list all the verb ontology types with which we get the COMPLETELY sense irrespective of the specific particles used.[13] Hence, ONT::COMPLETELY would specify for its FIGURE argument the ontology types ONT::PURSUE, ONT::ACQUIRE as well as ONT::EVOKE-TIREDNESS, for example.

Each of the verb ontology types mentioned above specifies which particles can take a specific semantic relation label. For example, the verb ontology type ONT::PURSUE would specify for its MANNER argument particle *down*, verb ontology type ONT::ACQUIRE would specify particle *up*, and verb ontology type ONT:EVOKE-TIREDNESS would specify particle *out*.

For the DIRECTION senses, the particle could be replaced with a RESULT-Prepositional Phrase (RESULT-PP) in the sentence. For example, the particle *off* in *The officer got **off** when he spotted an illegally parked car* can be replaced with a RESULT-PP *off his motorcycle* as in *The officer got **off his motorcycle** when he spotted an illegally parked car.*

Depending on the compliance or violations of constraints such as the ones described above, the parser assigns scores for various parse options involving various senses of the particle in a VPC. The parse with the highest score is selected as the semantic representation of the sentence involving the VPC.

In *She cleaned up her room* (Figure 4), the verb *clean* (ONT::CLEAN which appears under ONT::CHANGE-STATE in the ontology) is not among the list of relevant verb ontology types with which a DIRECTION-UP sense is licensed for the particle *up*. Additionally, a restriction on the verb argument for the DIRECTION-UP sense is that the argument have a semantic feature [+moveable] which is also violated in the given sentence, since *the room* is generally not a moveable entity. Hence, the parser assigns a low score to the parse which involves the

[13] Note that since the ontology is hierarchical, there is no need to list all the children ontology types as well if the parent ontology types are included.

DIRECTION-UP sense for the particle *up* in this sentence. The verb *push*, however, in the sentence in Figure 3, is a movement-related verb and the affected entity *ball* is a moveable object. Hence, the parser assigns a higher score to the parse that involves the DIRECTION-UP sense for the particle *up* in this sentence.

The IN-WORKING-ORDER-VAL sense requires restrictions on the verbs that they take cognitive entities, devices or processes as their AFFECTED arguments which provide service, i.e., they have some functionality associated with them. The AFFECTED argument for the verb *clean*, namely *the room*, in Figure 4 does not satisfy this restriction. Similarly, The AFFECTED argument for the verb *push*, namely *the ball*, in Figure 3 does not satisfy this restriction. Hence, the parser assigns a low score for the parses involving an IN-WORKING-ORDER-VAL sense for the particle *up* in the two sentences. On the other hand, the entity *The network* in the sentence in Figure 5 satisfies the restriction and hence the sentence in Figure 5 gets a higher score for the parse that contains IN-WORKING-ORDER-VAL sense for the particle *up*.

The constraints for the COMPLETELY sense of the particle *up* are satisfied for the sentence in Figure 4. The verb *clean* is among the set of verbs in the ontology type (ONT::CHANGE-STATE) with which the relevant particle has been identified in the ontology to get this sense. Hence, the parser assigns a higher score to the parse for the sentence with the COMPLETELY sense for the particle *up*. The other two sentences get a lower score for a parse with the COMPLETELY sense for the particle *up*.

Thus, as mentioned above, the parse with the highest score is selected as a semantic representation for each of the sentences involving the VPC. Hence, the sentence in Figure 3 has the DIRECTION-UP sense for the particle *up*, the sentence in Figure 4 has the COMPLETELY sense, and the sentence in Figure 5 has the IN-WORKING-ORDER-VAL sense for the particle *up*.

6 Conclusion

Identification of MWEs, such as verb-particle constructions, as well as the computation of their meaning is necessary for a broad coverage understanding. It is not plausible to hand enumerate all the possible combinations, although WordNet is an admirable start. We have described an approach where the meaning of a wide range of VPCs is computed compositionally, with the advantage that VPCs not explicitly found in the lexicon can be both identified and semantically interpreted. To accomplish this, we identified the core senses of particles that have broad application across verb classes. This information is used while building computational lexica. We encoded the generalizations corresponding to various senses of particles in the TRIPS ontology and used them to identify these senses. We found that the generalizations corresponding to grammatical/semantic/ontological information help us identify appropriate senses of the particle. The procedure adopted enables compositional parsing by helping differentiate between particle senses, which is then used to obtain full semantic representations of sentences.

While we have outlined our approach to compute the semantics of sentences with VPCs in English using resources such as the TRIPS lexicon and ontology and WordNet, a similar approach can be adopted to incorporate more lexical or other linguistic resources to further improve semantic parsing of sentences involving VPCs in English. Also, such an approach can be adopted for semantic parsing of sentences involving other MWEs, as well as for developing semantic parsing capabilities for other languages using the resources available for those languages.

Abbreviations

LF	logical form
LP-compositional	light particle compositional
LV-compositional	light verb compositional
MWE	multiword expression
RESULT-PP	RESULT-prepositional phrase
VPC	verb-particle construction

References

Allen, James F., Myroslava Dzikovska, Mehdi Manshadi & Mary Swift. 2007. Deep linguistic processing for spoken dialogue systems. In *Proceedings of the ACL 2007 Workshop on Deep Linguistic Processing*, 49–56. Association for Computational Linguistics. June 28, 2007.

Allen, James F., Mary Swift & Will de Beaumont. 2008. Deep semantic analysis of text. In *Symposium on semantics in systems for text processing (STEP)*, 343–354.

Allen, James F. & Choh Man Teng. 2017. Broad coverage, domain-generic deep semantic parsing. In *Proceedings of the AAAI Spring Symposium, Computational Construction Grammar and Natural Language Understanding 2017*, 108–115.

Baldwin, Timothy, Colin James Bannard, Takaaki Tanaka & Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18* (MWE '03), 89–96. Stroudsburg, PA, USA: Association for Computational Linguistics. DOI:10.3115/1119282.1119294

Bannard, Colin James. 2002. *Statistical techniques for automatically inferring the semantics of verb-particle constructions.* Tech. rep. University of Liverpool.

Bannard, Colin James & Timothy Baldwin. 2003. Distributional models of preposition semantics. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, 169–180. Association for Computational Linguistics. http://lingo.stanford.edu/pubs/tbaldwin/sigsemprep2003-prepsem.pdf.

Bannard, Colin James, Timothy Baldwin & Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18* (MWE '03), 65–72. Stroudsburg, PA, USA: Association for Computational Linguistics. DOI:10.3115/1119282.1119291

Bhatia, Archna, Choh Man Teng & James F. Allen. 2017. Compositionality in verb-particle constructions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 139–148. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1719.

Bréal, Michel. 1900. *Semantics.* Trans. by Mrs. H. Cust. New York: Henry Holt.

Chen, Ping. 1986. Discourse and particle movement in English. *Studies in Language* 10. 79–95.

Clark, Eve V. 1987. The principle of contrast. In Brian MacWhinney (ed.), *Mechanisms of language acquisition*, 1–33. New York: Academic Press.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20. 37–46.

Cook, Paul & Suzanne Stevenson. 2006. Classifying particle semantics in English verb-particle constructions. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (MWE '06), 45–53. Association for Computational Linguistics. July 23, 2006.

Dehé, Nicole. 2002. *Particle verbs in English: Syntax, information structure and intonation.* Amsterdam: John Benjamins.

Fellbaum, Christiane. 1998. *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Fleiss, Jacob L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5). 378–382.

Fraser, Bruce. 1976. *The verb-particle combination in English*. New York: Academic Press.

Ishizaki, Yasuaki. 2010. Some notes on the grammaticalization of *away*. In Hirozo Nakano, Masayuki Ohkado, Tomoyuki Tanaka, Tomohiro Yanagi & Azusa Yokogishi (eds.), *Synchronic and diachronic approaches to the study of language: A collection of papers dedicated to the memory of Professor Masachiyo Amano*, 71–83. Tokyo: Eichosha.

Ishizaki, Yasuaki. 2012. A usage-based analysis of phrasal verbs in early and late modern English. *English Language and Linguistics* 16(2). 241–260. DOI:10.1017/S1360674312000020

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Jackendoff, Ray. 2002. English particle constructions, the lexicon, and the autonomy of syntax. In Nicole Dehé, Ray Jackendoff, Andrew McIntyre & Silke Urban (eds.), *Verb-Particle Explorations* (Interface Explorations [IE]), 67–94. New York: Mouton de Gruyter.

Johnson, Christopher & Charles J. Fillmore. 2000. The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. In *Proceedings of the First Conference on North American Chapter of the Association for Computational Linguistics*, 56–62. Morgan Kaufmann Publishers Inc.

Kiparsky, Paul. 1983. Word-formation and the lexicon. In Frances Ingemman (ed.), *Proceedings of the 1982 Mid-America Linguistics Conference*, 47–78. University of Kansas, Dept. of Linguistics.

McCarthy, Diana, Bill Keller & John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18* (MWE '03), 73–80. Stroudsburg, PA, USA: Association for Computational Linguistics. DOI:10.3115/1119282.1119292

Miller, George A. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38(11). 39–41.

Moreno-Ortiz, Antonio, Chantal Pérez-Hernández & M. ángeles Del-Olmo. 2013. Managing multiword expressions in a lexicon-based sentiment analysis system for Spanish. In *Proceedings of the 9th Workshop on Multiword Expressions* (MWE '13), 1–10. Association for Computational Linguistics.

Niinuma, Fumikazu. 2014. Grammaticalization of the particle *away* in English: A cartographic approach. *Interdisciplinary Information Sciences* 20(2). 163–188.

O'Dowd, Elizabeth M. 1998. *Prepositions and particles in English: A discourse-functional account*. Oxford: Oxford University Press.

Patrick, Jon & Jeremy Fletcher. 2004. Differentiating types of verb particle constructions. In *Proceedings of Australasian Language Technology Workshop 2004* (ALTW 2004). http://aclweb.org/anthology-new/U/U04/U04-1022.pdf. December 8, 2004.

Paul, Hermann. 1890. *Principles of the history of language*. Trans. by H. A. Strong from the 2nd German Edition. Reprinted College Park, MD. 1970: McGrat.

Tyler, Andrea & Vyvyan Evans. 2003. *The semantics of English prepositions: Spatial scenes, embodied meaning, and cognition*. New York: Cambridge University Press.

Villavicencio, Aline. 2003. Verb-particle constructions and lexical resources. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE 2003)- *Volume 18*, 57–64. Association for Computational Linguistics. DOI:10.3115/1119282.1119290

Villavicencio, Aline. 2006. Verb-particle constructions in the World Wide Web. In Patrick Saint-Dizier (ed.), *Syntax and semantics of prepositions*, 115–130. The Netherlands: Springer.

Vincze, Veronika, István Nagy T. & Gábor Berend. 2011. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 289–295. RANLP 2011 Organising Committee. http://aclweb.org/anthology/R11-1040.

Chapter 4

PARSEME multilingual corpus of verbal multiword expressions

Agata Savary[1], Marie Candito[2], Verginica Barbu Mititelu[3], Eduard Bejček[4], Fabienne Cap[5], Slavomír Čéplö[6], Silvio Ricardo Cordeiro[7], Gülşen Eryiğit[8], Voula Giouli[9], Maarten van Gompel[10], Yaakov HaCohen-Kerner[11], Jolanta Kovalevskaitė[12], Simon Krek[13], Chaya Liebeskind[11], Johanna Monti[14], Carla Parra Escartín[15], Lonneke van der Plas[6], Behrang QasemiZadeh[16], Carlos Ramisch[7], Federico Sangati[17], Ivelina Stoyanova[18] & Veronika Vincze[19]

[1]Université de Tours (France), [2]Université Paris Diderot (France), [3]Romanian Academy Research Institute for Artificial Intelligence (Romania), [4]Charles University (Czech Republic), [5]Uppsala University (Sweden), [6]University of Malta (Malta), [7]Aix Marseille University (France), [8]Istanbul Technical University (Turkey), [9]Athena Research Center in Athens (Greece), [10]Radboud University in Nijmegen (Netherlands), [11]Jerusalem College of Technology (Israel), [12]Vytautas Magnus University in Kaunas (Lithuania), [13]Jožef Stefan Institute in Ljubljana (Slovenia), [14]"L'Orientale" University of Naples (Italy), [15]ADAPT Centre, Dublin City University (Ireland), [16]University of Düsseldorf (Germany), [17]independent researcher (Italy), [18]Bulgarian Academy of Sciences in Sofia (Bulgaria), [19]University of Szeged (Hungary)

Multiword expressions (MWEs) are known as a "pain in the neck" due to their idiosyncratic behaviour. While some categories of MWEs have been largely studied, verbal MWEs (VMWEs) such as *to take a walk*, *to break one's heart* or *to turn off* have been relatively rarely modelled. We describe an initiative meant to bring about substantial progress in understanding, modelling and processing VMWEs. In this joint effort carried out within a European research network we elaborated

Agata Savary et al. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop, 87-147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471591

a universal terminology and annotation methodology for VMWEs. Its main outcomes, available under open licenses, are unified annotation guidelines, and a corpus of over 5.4 million words and 62 thousand annotated VMWEs in 18 languages.

1 Introduction

One of the basic ideas underlying linguistic modelling is compositionality (Baggio et al. 2012), seen as a property of language items (Janssen 2001; Partee et al. 1990) or of linguistic analyses (Kracht 2007). Counterexamples which challenge the compositionality principles (Pagin & Westerståhl 2001) include multiword expressions (MWEs) (Sag et al. 2002; Kim 2008), and notably verbal MWEs (VMWEs), such as (1–4).[1]

(1) *Ida* ***skriva glavo v pesek****.* (SL)
Ida hide.3.SG head in sand
Ida hides her head in the sand. 'Ida pretends not to see a problem.'

(2) *Er* ***legt*** *die Prüfung* ***ab****.* (DE)
he lay.3.SG the exam PART
He lays the exam PART. 'He takes the exam.'

(3) *Η Ζωή* ***παίρνει*** *μία* ***απόφαση****.* (EL)
i zoi perni mia apofasi
the Zoe take.3.SG a decision
Zoe takes a decision. 'Zoe makes a decision.'

(4) *Alina* ***se face*** *doctor.* (RO)
Alina REFL.3.SG make.3.SG doctor
Alina REFL makes doctor. 'Alina becomes a doctor.'

VMWEs pose special challenges in natural language processing (NLP):

1. Semantic non-compositionality: The meaning of many VMWEs cannot be deduced in a way deemed grammatically regular on the basis of their syntactic structure and of the meanings of their components. For instance, the meaning of sentence (1) cannot be retrieved from the meanings of its component words (SL) *glava* 'head' and *pesek* 'sand', except when very specific interpretations of these words and of their combination are admitted.

[1] See the preface for the description of the conventions used to present multilingual examples.

2. Lexical and grammatical inflexibility: VMWEs are frequently subject to unpredictable lexical or syntactic constraints. For instance, when the individual lexemes in (EN) *to **throw** somebody **to the lions*** are replaced by their synonyms or the noun is modified by an adjective, the expression loses its idiomatic meaning:[2] (EN) *#to fling sb to the lions*, *#to throw sb to the hungry lions*. Similarly, the predicative noun in the light-verb construction (EN) *she **took** a **glance** at the headline* cannot take a modifier denoting an agent, especially if different from the verb's subject (**she **took** Paul's **glance** at the headline*).

3. Regular variability: Despite this inflexibility the VMWEs can still exhibit some regular variability, e.g.: (i) inflection or passivisation, as in (EN) *he was **thrown to the lions***, (ii) a restricted lexical replacement and an adjectival modification of the predicative noun, as in (EN) *he **took/had** a quick **glance** at the headline*, (iii) omission of components without change in meaning, as in (EL) ***meno me ti glika (sto stoma)*** 'I stayed with the sweetness (in.the mouth)' ⇒ 'I was very close to enjoy something desired but I failed to'.

4. Discontinuity: The components of a VMWE may not be adjacent, e.g. (EN) *a **mistake** was frequently **made***, *never **turn** it **off***.

5. Categorical ambiguity: VMWEs of different categories may share the same syntactic structure and lexical choices. For instance, (EN) *to **make** a **mistake*** and (EN) *to **make a meal** of something* 'to treat something as more serious than it really is' are combinations of the same verb with a direct object but the former is a light-verb construction (since the verb is semantically void and the noun keeps its original predicative meaning), while the latter is an idiom (since the noun loses its original sense).

6. Syntactic ambiguity: Occurrences of VMWEs in text may be syntactically ambiguous, e.g. (EN) *on* is a particle in *to **take on** the task* 'to agree to be in charge of the task', while it is a preposition in (EN) *to **sit on the fence*** 'not to take sides in a dispute'.

7. Literal-idiomatic ambiguity: A VMWE may have both an idiomatic and a literal reading. For instance the VMWE (EN) *to **take the cake*** 'to be the

[2] Henceforth, an asterisk (*) preceding a sentence will mean that the sentence is ungrammatical, while a dash (#) will signal a substantial change in meaning with respect to the original expression.

most remarkable of its kind' is understood literally in (EN) *to take the cake out of the fridge.*

8. Non-literal translatability: Word-for-word translation of VMWEs is usually incorrect, e.g. (EN) *to* ***take the cake*** 'to be the most remarkable of its kind' does not translate to (FR) *prendre le gâteau* 'to take the cake'.

9. Cross-language divergence: VMWEs behave differently in different languages and are modelled according to different linguistic traditions. For instance, functional tokens, such as (EN) *off*, have a status of stand-alone words and can form verb-particle constructions in Germanic languages, e.g. (EN) *to* ***turn off***. In Slavic languages, conversely, they function as prefixes, as in (PL) *wyłączyć* 'part.connect' ⇒ 'turn off', and are seen as inherent parts of verbal lexemes. Therefore, they cannot trigger MWE-related considerations (cf. §8). Also, the scope of light (or support) verb constructions may greatly vary from one linguistic tradition to another, e.g. depending on whether the copula *to be* is considered a light verb or not (cf. §9.1).

10. Wordplay proneness: In particular contexts, VMWEs can be a subject of ad hoc creativity or a playful usage, as in (EN) *they want us to put the cat back inside the bag* 'they want us to pretend that the revealed secret remains unrevealed'.

Due to these unpredictable properties, the description, identification, analysis and translation of VMWEs require dedicated procedures. For example, due to 2 and 3, the description of VMWEs can be constrained neither to the level of the lexicon nor to the one of the syntax only. Challenge 4 hinders VMWE identification with traditional sequence labelling approaches and calls for syntactic analysis. Challenges 5, 6 and 7, however, mean that their identification and categorisation cannot be based on solely syntactic patterns. Challenges 1, 2, 7 and 8 constitute central issues in machine translation. Challenge 9 affects cross-lingual VMWE modelling. Finally, challenge 10 goes far beyond the state of the art in semantic modelling and processing of VMWEs.

A consistent linguistic and NLP terminology is required in order to better understand the nature of VMWEs, compare their properties across languages, hypothesise linguistic generalisations, model VMWEs according to common principles, develop cross-language VMWE identifiers and compare results obtained by different authors on different datasets. Such a consistency is, however, largely missing: different authors assign different names to the same phenomena or call different phenomena by the same name, be it from a linguistic or an NLP point of view. This situation is similar to other areas of linguistic modelling, where universalism-driven efforts have been undertaken – such as the Universal Dependencies (UD) project dedicated to standardising morphological and syntactic annotations for dozens of languages (Nivre et al. 2016), or the normalisation of uncertainty cue annotation across languages, genres and domains (Szarvas et al. 2012).

This chapter describes an initiative taken by the European PARSEME network,[3] towards bringing about substantial progress in modelling and processing MWEs. Its main outcomes include unified definitions and annotation guidelines for several types of VMWEs, as well as a large multilingual openly available VMWE-annotated corpus. Eighteen languages are addressed (note that the last 4 are non-Indo-European):

- *Balto-Slavic*: Bulgarian (BG), Czech (CS), Lithuanian (LT), Polish (PL) and Slovene (SL);
- *Germanic*: German (DE) and Swedish (SV);
- *Romance*: French (FR), Italian (IT), Romanian (RO), Spanish (ES) and Portuguese (PT);[4]
- *Others*: Farsi (FA), Greek (EL), Hebrew (HE), Hungarian (HU), Maltese (MT) and Turkish (TR).

The corpus gave rise to the PARSEME shared task on automatic identification of VMWEs, whose organisation and results are described by Savary et al. (2017). See also Taslimipoor et al. (2018 [this volume]) and Maldonado & QasemiZadeh (2018 [this volume]) who address the use of the PARSEME corpus in VMWE identification and its evaluation, as well as Moreau et al. (2018 [this volume]), Al Saied et al. (2018 [this volume]) and Simkó et al. (2018 [this volume]) who describe 3 of the 7 systems participating in the shared task.

This chapter builds upon those sections of the PARSEME shared task description paper (Savary et al. 2017), presented in the MWE 2017 workshop, which describe the corpus construction. Each of these sections has been substantially extended, except the descriptions of the corpus format and inter-annotator agreement, which required few additions and updates. Many new analyses and examples have been added, conclusions drawn from the PARSEME annotation campaign have been addressed and the state of the art has been thoroughly revised. As a result, the chapter is organised as follows. We give the definitions underly-

[3] http://www.parseme.eu

[4] In this chapter we address the Brazilian dialect of Portuguese. All examples cited here are taken from this dialect.

ing the scope of our work (§2), and the VMWE typology (§3). We describe the annotation principles, including the VMWE identification and categorisation tests, and the deviations from the unified guidelines applied in some languages (§4). We discuss the annotation methodology and tools (§5). We present the resulting corpus and a cross-language quantitative analysis of some phenomena relevant to challenges 1–10 (§6). We describe some language-specific studies based on the corpus (§7) and discuss interesting problems which occurred during the project (§8). We analyse the state of the art in MWE modelling and annotation, and compare it to our approach (§9). We finally conclude and discuss future work (§10).

2 Definitions and scope

While the definition of a MWE inherently relies on the notion of a WORD (i.e. a linguistically motivated unit), identification of VMWEs is performed on pragmatically defined TOKENS. The relation between tokens and words can be threefold:

(1) A token coincides with a word, e.g. (MT) *ferħ* 'happiness', (SV) *förvåning* 'surprise'.

(2) Several tokens build up one MULTITOKEN WORD (MTW), if punctuation marks are considered token boundaries, as in (EN) *Pandora's*, (PL) *SMS-ować* 'to write an SMS'. Note that the latter example is not a VMWE as it contains only one word.

(3) One MULTIWORD TOKEN (MWT) contains several words, as in contractions, e.g. (IT) *della* 'of.the', or detachable pre-verbal particles, e.g. (DE) ***ausmachen*** 'PART.make' ⇒ 'to turn off'. Note that the latter example is a (one-token) VMWE. A MWT is not always a simple concatenation of words, e.g. (IT) *della* is a contraction of *di* 'of' and *la* 'the.FEM'.

In this work, MULTIWORD EXPRESSIONS (MWEs) are understood as (continuous or discontinuous) sequences of words which:

- contain at least two component words which are **lexicalised**, i.e. always realised by the same lexemes (see below for a more precise definition), including a head word and at least one other syntactically related word,
- display some degree of lexical, morphological, syntactic and/or semantic idiosyncrasy, formalised by the annotation procedures in §4.1–§4.2.

This definition relatively closely follows the one by Baldwin & Kim (2010). Two notable exceptions are that we impose syntactic constraints on the lexicalised components (one of them must be the head word), and that Baldwin & Kim (2010) include pragmatic and statistical idiosyncrasy in the set of the MWE definition criteria. For us, conversely, COLLOCATIONS, i.e. word co-occurrences whose idiosyncrasy is of pragmatic or statistical nature only (e.g. *all aboard*, *the graphic shows*, *drastically drop*) are disregarded.

Note that there is no agreement on the understanding of the border between the scopes of MWEs and collocations. For Sag et al. (2002), collocations are any statistically significant word co-occurrences, i.e. they include all forms of MWEs. For Baldwin & Kim (2010), collocations form a proper subset of MWEs. According to Mel'čuk (2010), collocations are binary, semantically compositional combinations of words subject to lexical selection constraints, i.e. they intersect with what is here understood as MWEs. This chapter puts forward yet another point of view: MWEs and collocations are seen as disjoint sets of linguistic objects.

Our definition of a MWE is also relatively close to the notion of non-compositional semantic phrasemes in Mel'čuk (2010), but we include light-verb constructions in our scope. It is compatible as well with the one by Sag et al. (2002), where a MWE is seen as an "idiomatic interpretation that crosses word boundaries". The major differences between our approach and these seminal works are its multilingual context and the fact that, within the restricted scope of verbal MWEs (see below), we delimit the MWE phenomenon by a relatively precise and complete MWE identification and categorisation procedure, given in the form of decision trees built upon linguistic tests (§4). Note that this approach does not focus on another salient property of MWEs which is their variable degree of idiosyncrasy (Gross 1988), that is, the fact that various MWEs exhibit more or less unexpected lexical, syntactic and semantic properties. A scale-wise modelling of MWEs is hard to implement in the task of MWE annotation, which is our major operational objective. Instead, we assume that decisions on MWE-hood are binary, and the decision trees are designed so as to make them reproducible.

VERBAL MWEs (VMWEs) are multiword expressions whose canonical form (see below) is such that: (i) its syntactic head is a verb V, (ii) its other lexicalised components form phrases directly dependent on V. Boundary cases for condition (i) include at least two types of VMWEs. Firstly, those with irregular syntactic structures may hinder the identification of the headword as in (EN) ***short-circuited***, where the verb is atypically prefixed by an adjective. Secondly, for those with two coordinated lexicalised verbs there is no consensus as to which component – the conjunction or the first verb – should be considered the head, as in (5). Condition (ii) requires that the lexicalised components of a VMWE form a connected dependency graph. For instance, in (EN) *to **take on** the task* 'to agree to be in charge of the task' the particle *on* directly depends on the verb, thus ***take***

on fulfils the syntactic requirements to be a VMWE. Conversely, if the lexicalist hypothesis in syntax is followed (de Marneffe et al. 2014),[5] the preposition *on* in (EN) *to rely on someone* does not directly depend on the verb, thus, *rely on* cannot be considered a VMWE.

(5)	*wo*	*man*	***lebt***	*und*	***leben***	***lässt***	(DE)
	where	one	lives	and	live	lets	

where one lives and lets live 'where one is tolerant'

Just like a regular verb, the head verb of a VMWE may have a varying number of arguments. For instance, the direct object and the prepositional complement are compulsory in (EN) *to **take** someone **by surprise***. Some components of such compulsory arguments may be lexicalised, that is, always realized by the same lexemes. Here, *by surprise* is lexicalised while *someone* is not.

Note that lexicalisation is traditionally defined as a diachronic process by which a word or a phrase acquires the status of an autonomous lexical unit, that is, "a form which it could not have if it had arisen by the application of productive rules" (Bauer 1983 apud Lipka et al. 2004). In this sense all expressions considered VMWEs in this work are lexicalized. Our notion of lexicalisation extends this standard terminology, as it applies not only to VMWEs but to their components as well. The reason is that, in the context of the annotation task, we are in need of specifying the precise span of a VMWE, i.e. pointing at those words which are considered its inherent, lexically fixed components. Precisely these components are referred to as lexicalized within the given VMWE. Throughout this chapter, the lexicalised components of VMWEs are highlighted in bold.

A prominent feature of VMWEs is their rich morpho-syntactic variability. For instance, the VMWE (EN) *to **take** someone **by surprise*** can be inflected (*they **took** him **by surprise***), negated (*they did not **take** him **by surprise***), passivised (*he will be **taken by surprise***), subject to extraction (*the **surprise** by which I was **taken***), etc. Neutralizing this variation is needed when applying the linguistic tests defined in the annotation guidelines (§4), which are driven by the syntactic structure of the VMWE candidates. We define a prototypical verbal phrase as a minimal sentence in which the head verb *V* occurs in a finite non-negated form and all its arguments are in singular and realized with no extraction. For instance, (EN) *Paul made/makes a pie* is a prototypical verbal phrase while *Paul did not make a pie*, *the pie which Paul made* and *the pie was made by Paul* are

[5] The lexicalist hypothesis strongly inspired the PARSEME annotation guidelines, and is expected to be even more thoroughly followed in the future versions of the corpus.

not. If a VMWE can occur as a prototypical verbal phrase while keeping its idiomatic meaning, then such a phrase is its CANONICAL FORM. Otherwise, its least marked variation is considered canonical (a non-negated form is less marked than a negated one, active voice is less marked then passive, and a form with an extraction is more marked than one without it). For instance, a canonical form of (EN) *a bunch of **decisions** which were **made** by him* is (EN) *he **made** a **decision***. But since (6) and (7) lose their idiomatic readings in active voice – (PL) *#wszyscy rzucili kości* 'everyone threw dies' – and with no negation – (BG) *#tya **iska i da chue*** 'she wants to also hear' – their canonical forms are passive and negated, respectively. Whenever a VMWE candidate is identified in a sentence, the linguistic tests are to be applied to one of its canonical forms (whether it is a prototypical verbal phrase or not).

(6) ***Kości zostały rzucone.*** (PL)
dies were cast
The dies were cast. 'The point of no-return has been passed.'

(7) *Тя **не иска и да чуе**.* (BG)
Tya **ne iska i da chue**
she not want and to hear
She does not even want to hear. 'She opposes strongly.'

(8) ***Пиле не може да прехвръкне.*** (BG)
pile ne mozhe da prehvrakne
Bird not can to PART.fly
A bird cannot fly across something. 'Something is very strictly guarded.'

Throughout this chapter examples of VMWEs will always be given in their canonical forms, possibly accompanied by adjuncts, if the subject is lexicalised as in (8). Otherwise, their canonical forms may alternate – for brevity – with infinitive forms, or – rarely – with other variants when particular phenomena are to be illustrated.

MWEs containing verbs but not functioning as verbal phrases or sentences are excluded from the scope of annotation, e.g. (FR) *peut-être* 'may-be' ⟹ 'maybe', *porte-feuille* 'carry-sheet' ⟹ 'wallet'.

Let us finally comment on the notion of universalism. Formally, this term should only be used when a property or a phenomenon has been proven relevant to all languages, which is practically out of range of any endeavour, however multilingual and inclusive. Therefore, in this chapter we use the adjective 'universal' in the sense of a scientific hypothesis rather than of a proven fact. When we speak about a universal category or property, it is to be understood that we deem them universal, based on the evidence from the languages currently in our scope. Since our framework is meant to continually evolve by including new languages and MWE types, we hope our definitions and findings to approximate the truly universal properties increasingly well.

3 VMWE typology

The typology of VMWEs, as well as linguistic tests enabling their classification, were designed so as to represent properties deemed universal in a homogeneous way, while rendering language-specific categories and features at the same time. The 3-level typology consists of:

1. *Universal* categories, valid for all languages participating in the task:
    a) light-verb constructions (LVCs), as in (9):

    (9) *Eles* ***deram*** *uma* ***caminhada****.* (PT)
    they gave a walk
    They gave a walk. ‘They took a walk.’

    b) idioms (ID), as in (10):

    (10) به قدر کافی برای من خواب دیده است. (FA)
    ast **dide khab** man **baraye** kafi qadre be
    is seen sleep me for enough quantity to
    He had enough sleep for me. ‘He has many plans for me.’

2. *Quasi-universal* categories, valid for some language groups or languages, but not all:
    a) inherently reflexive verbs (IReflVs), as in (11):

    (11) *Ils ne* ***s’apercevront*** *de rien.* (FR)
    they not REFL.3.PL’perceive.3.PL.FUT of nothing
    They will REFL-perceive nothing. ‘They will not realise anything.’

    b) verb-particle constructions (VPCs), as in (12):

    (12) *Sie* ***macht*** *die Tür* ***auf****.* (DE)
    she makes the door PART
    She makes PART the door. ‘She opens the door.’

3. *Other* verbal MWEs (OTH), not belonging to any of the categories above (due to not having a unique verbal head) e.g. (EN) *he never* ***drinks and drives****, she* ***voice acted****, the radio* ***short-circuited****.*

Table 1: Examples of various categories of VMWEs in four non-Indo-European languages.

Lang.	ID	LVC	Quasi-universal / OTH
HE	**אבד עליו כלח** ‘Kelax is lost on him.’ ‘He is outdated.’	**הגיע למסקנה** ‘to come to a conclusion’ ‘to conclude’	**לא הבישן למד** ‘the bashful does not learn’ ‘one should dare ask questions’
HU	**kinyír** ‘to out-cut’ ‘to kill’	**szabályozást ad** ‘to give control’ ‘to regulate’	**feltüntet** (VPC) ‘to PART-strike’ ‘to mark’
MT	**Għasfur żgħir qalli.** ‘A small bird told me.’ ‘I learned it informally.’	**ħa deċizjoni** ‘to take a decision’ ‘to make a decision’	**iqum u joqgħod** (OTH) ‘to jump and stay’ ‘to fidget’
TR	**yüzüstü bırakmak** ‘to leave (sb) face down’ ‘to forsake’	**engel olmak** ‘to become obstacle’ ‘to prevent’	**karar vermek** (OTH) ‘to give a decision’ ‘to make a decision’

While we allowed for language-specific categories, none emerged so far. Table 1 and Table 2 show examples of VMWEs of different categories in the 18 languages in our scope (4 non-Indo-European and 14 Indo-European). None of those languages seems to possess VMWEs of all 5 terminal categories (LVC, ID, IReflV, VPC and OTH).

We thoroughly considered introducing another universal category of inherently prepositional verbs (IPrepVs), such as (EN) *to rely on*, *to refer to*, or *to come across*. However, the IPrepV-related linguistic tests used in the pilot annotation proved not sufficiently reliable to distinguish such expressions from compositional verb-preposition combinations, such as (EN) *to give something to someone*. Therefore, we abandoned this category, considering that prepositions belong to the area of verb valency and should be handled by a regular grammar (combined with a valency lexicon). Reconsidering this category experimentally belongs to future work (§10).

Table 2: Examples of various categories of VMWEs in 14 Indo-European languages.

Lang.	ID	LVC	Quasi-universal / OTH
BG	**бълвам змии и гущери** ‘to spew snakes and lizards’ ‘to shower abuse’	**държа под контрол** ‘to keep under control’ ‘to keep under control’	**усмихвам се** (IReflV) ‘to smile REFL’ ‘to smile’
CS	**házet klacky pod nohy** ‘to throw sticks under feet’ ‘to put obstacles in one’s way’	**vyslovovat nesouhlas** ‘to voice disagreement’ ‘to disagree’	**chovat se** (IReflV) ‘to keep REFL’ ‘to behave’
DE	**schwarz fahren** ‘to drive black’ ‘to take a ride without a ticket’	eine **Rede halten** ‘a hold a speech’ ‘to give a speech’	**sich enthalten** (IReflV) ‘to contain REFL’ ‘to abstain’
EL	**χάνω τα αυγά και τα καλάθια** ‘to lose the eggs and the baskets’ ‘to be at a complete and utter loss’	**κάνω** μία **πρόταση** ‘to make a proposal’ ‘to propose’	**μπαίνω μέσα** (VPC) ‘to get PART’ ‘to go bankrupt’
ES	**hacer de tripas corazón** ‘to make heart of intestines’ ‘to pluck up the courage’	**hacer** una **foto** ‘to make a picture’ ‘to take a picture’	**coser y cantar** (OTH) ‘to sew and to sing’ ‘as easy as pie’
FA	دست گل به آب دادن ‘to give a flower bouquet to water’ ‘to mess up, to do sth. wrong’	امتحان کردن ‘to do an exam’ ‘to test’	به خود آمدن ‘to come to REFL’ ‘to gain focus’
FR	**voir le jour** ‘to see the daylight’ ‘to be born’	**avoir** du **courage** ‘to have courage’ ‘to have courage’	**se suicider** (IReflV) ‘to suicide REFL’ ‘to commit suicide’
IT	**entrare in vigore** ‘to enter into force’ ‘to come into effect’	**fare** un **discorso** ‘to make a speech’ ‘to give a speech’	**buttare giù** (VPC) ‘to throw PART’ ‘to swallow’
LT	**pramušti dugną** ‘to break the bottom’ ‘to collapse’	**priimti sprendimą** ‘to take on a decision’ ‘to make a decision’	
PL	**rzucać grochem o ścianę** ‘to throw peas against a wall’ ‘to try to convince somebody in vain’	**odnieść sukces** ‘to carry-away a success’ ‘to be successful’	**bać się** (IReflV) ‘to fear REFL’ ‘to be afraid’
PT	**fazer das tripas coração** ‘make the tripes into heart’ ‘to try everything possible’	**fazer** uma **promessa** ‘to make a promise’ ‘to make a promise’	**se queixar** (IReflV) ‘to complain REFL’ ‘to complain’
RO	a **trage pe sfoară** ‘to pull on rope’ ‘to fool’	a **face** o **vizită** ‘to make a visit’ ‘to pay a visit’	a **se gândi** (IReflV) ‘to think REFL’ ‘to think’
SL	**spati kot ubit** ‘to sleep like killed’ ‘to sleep soundly’	**postaviti vprašanje** ‘to put a question’ ‘to ask a question’	**bati se** (IReflV) ‘to fear REFL’ ‘to be afraid’
SV	att **plocka russinen ur kakan** ‘to pick raisins out of the cake’ ‘to choose only the best things’	**ta** ett **beslut** ‘to take a decision’ ‘to make a decision’	**det knallar och går** (OTH) ‘it trots and walks’ ‘it is OK/as usual’

4 Annotation guidelines

Given the definitions in §2 and a text to annotate, each iteration of the annotation process starts with: (i) selecting a candidate sequence, i.e. a combination of a verb with at least one other word which could form a VMWE, (ii) establishing the precise list of its lexicalised components and its canonical forms. These steps are largely based on the annotator's linguistic knowledge and intuition.

Once a candidate sequence has been selected, its status as a VMWE is tested in two steps: identification and categorisation. Each step is based on linguistic tests and examples in many languages, organised into decision trees, so as to maximise the determinism in decision making.

4.1 Identification tests

Five generic non-compositionality tests were defined in order to identify a VMWE (of any category):

> **Test 1** [CRAN]: Presence of a cranberry word, e.g. (EN) *it* ***goes astray***;
>
> **Test 2** [LEX]: Lexical inflexibility, e.g. (EN) *they #allowed the feline out of the container* (*they* ***let the cat out of the bag***); *_to give a stare_ (to **give** a **look**);
>
> **Test 3** [MORPH]: Morphological inflexibility, e.g. (EN) *to #take a turn* (*to* ***take turns***);
>
> **Test 4** [MORPHOSYNT]: Morpho-syntactic inflexibility, e.g. (EN) *#I give you his word for that* (*I* ***give*** *you my* ***word*** *for that*);
>
> **Test 5** [SYNT]: Syntactic inflexibility, e.g. (EN) *#Bananas are gone* (*he* ***went bananas***).

If none of these tests apply, an additional hypothesis covers the LVC candidates, which usually fail Tests 1 and 3–5 and for which Test 2 is hard to apply due to their relatively high, although restricted, productivity.

> [LVC hypothesis]: In a verb+(prep)+noun candidate the verb is a pure syntactic operator and the noun expresses an activity or a state, e.g. (EN) ***makes*** *a* ***speech***.

Passing any of Tests 1–5 is sufficient for a candidate sequence to be identified as a VMWE, while the LVC hypothesis has to be confirmed by the LVC-specific tests.[6]

4.2 Decision tree for categorisation

Once a VMWE has been identified or hypothesised following the tests in the preceding section, its categorisation follows the decision tree shown in Figure 1. Tests 6–8 are structural, the others are category-specific.

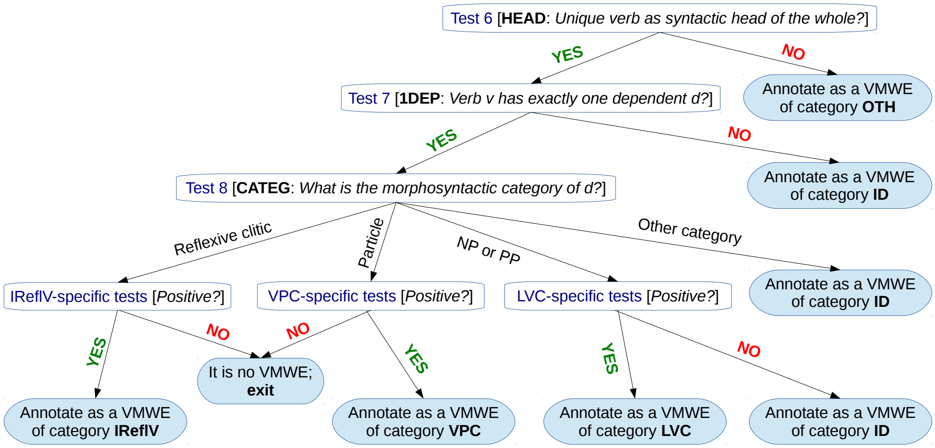

Test 6 [HEAD: Unique verb as syntactic head of the whole?]
YES
NO
Annotate as a VMWE of category OTH
Test 7 [1DEP: Verb v has exactly one dependent d?]
YES
NO
Annotate as a VMWE of category ID
Test 8 [CATEG: What is the morphosyntactic category of d?]
Reflexive clitic
Particle
NP or PP
Other category
IReflV-specific tests [Positive?]
VPC-specific tests [Positive?]
LVC-specific tests [Positive?]
Annotate as a VMWE of category ID
YES
NO
NO
YES
YES
NO
It is no VMWE; exit
Annotate as a VMWE of category IReflV
Annotate as a VMWE of category VPC
Annotate as a VMWE of category LVC
Annotate as a VMWE of category ID

Figure 1: Decision tree for VMWE categorisation.

4.2.1 Structural tests

Categorisation of a VMWE depends on the syntactic structure of its canonical form determined by the following three tests:

> **Test 6** [HEAD]: Presence of a unique verb functioning as the syntactic head of the whole expression, like in (13) and unlike in (14).

[6] As explained in §10, feedback from the large-scale annotation of version 1.0 of the corpus led us to questioning the correctness of the two-stage VMWE annotation. In edition 1.1 we transformed the identification tests into ID-specific tests and performed VMWE identification simultaneously to their categorisation.

(13) *Je **laisse tomber**.* (FR)
I let fall
I let fall. ‘I let go, I abandon.’

(14) *wo man **lebt** und **leben lässt** (DE)*
where one lives and live lets
where one lives and lets live ‘where one is tolerant’

Test 7 [1DEP]: Among the phrases dependent on the head verb exactly one contains lexicalised components, as in (EN) ***made it up***, and unlike in (EN) ***made up** her **mind**.*

Test 8 [CATEG]: Morphosyntactic category of the verb’s dependent. Contrary to most other tests, the result of this test is not binary but taken from a closed list of values: (i) reflexive clitic (REFL), as in (15), (ii) particle (PART), as in (16); (iii) nominal or prepositional phrase, as in (17); (iv) other (including a verb, an adverb, a non-reflexive pronoun, etc.), as in (18).

(15) *Той **се** **страхува**.* (BG)
toy **se** **strahuva**
he REFL fears
He fears REFL. ‘He is afraid.’

(16) *Der Film **fängt** **an**.* (DE)
the film catches PART
The film catches PART. ‘The film begins.’

(17) *Mój bratanek **buja** **w** **obłokach**.* (PL)
my nephew swings in clouds
My nephew swings in the clouds. ‘My nephew fantasizes.’

(18) *Uma ajudinha **cai** muito **bem**.* (PT)
a help.DIM falls very well
A little help falls very well. ‘A little help comes at the right moment.’

When a VMWE fails Test 6 or 7, it is automatically classified as OTH and ID, respectively. This means that we do not allow cumulative categories. For instance, in (20) the reflexive clitic considerably changes the meaning of the base VPC from (19), which might qualify the whole as an IReflV. However, due to the presence of two lexicalised syntactic arguments of the verb, such cases are necessarily classified as IDs (here: with a nested VPC).

(19) *Er **stellte** mir seine Freundin **vor**.* (DE)
he put me his friend PART
He put his friend PART to me. 'He presented his friend to me.'

(20) *Er **stellte sich** die Reise **vor**.* (DE)
he put REFL.3.SG the travel PART
He put the travel PART to REFL. 'He imagined the travel.'

Test 8, with return values (i)-(iii), triggers the category-specific tests for IReflVs, VPCs and LVCs, respectively. For other categories the candidate automatically qualifies as an ID.

4.2.2 Light-verb constructions

Light-verb constructions (LVCs) gave rise to a vast literature since first introduced by Jespersen (1965), possibly because there is no consensus on their exact definition and scope. We consider a candidate sequence an LVC if it consists of a verb *V* and a nominal complement *N*, possibly introduced by a preposition, provided that it passes all of the following tests:

Test 9 [N-EVENT]: *N* denotes an event or a state, as in (21);

(21) *Οι συσκευές **έχουν** τη **δυνατότητα** σύνδεσης.* (EL)
I siskieves eχun ti δinatotita sinδesis
the devices have the ability connection.SG.GE.
The devices have the ability to connect. 'The devices can connect.'

Test 10 [N-SEM]: *N* has one of its original senses, as in (22) and unlike in (23);

(22) *Steffi **rend visite** à Monica.* (FR)
Steffi returns visit to Monica
Steffi returns a visit to Monica. 'Steffi pays a visit to Monica.'

(23) *Je* ***jette*** *l'****éponge****.* (FR)
I throw the'sponge
I throw the sponge. 'I give up.'

Test 11 [V-LIGHT]: *V* only contributes morphological features (tense, mood, person, number, etc.) but adds no semantics that is not already present in *N*, other than the semantic role of *V*'s subject with respect to *N*, as in (24);

(24) *Gydytojai* ***padarė išvadą****, kad gijimo procesas vyksta*
Doctors made conclusion, that recovery process happens
sėkmingai. (LT)
successfully.
The doctors made the conclusion that the recovery process is successful. 'The doctors came to the conclusion that the recovery process is successful.'

Test 12 [V-REDUC]: An NP headed by *N* can be formed containing all of *V*'s syntactic arguments, and denoting the same event or state as the LVC, e.g. (EN) *Paul* ***had*** *a nice* ***walk*** denotes the same event as (EN) *the nice walk of Paul.*

Test 13 [N-PROHIBIT-ARG]: A semantic argument of the same type cannot be syntactically realised twice – both for *N* and for *V*, e.g. (EN) _*Paul **made** the **decision** of the committee_ is meaningless, while (EN) *Paul leads the discussion of the committee* is acceptable. Therefore, *to lead a discussion* is not an LVC.

Tests 12 and 13 are syntactic tests approximating the property that one of *V*'s syntactic arguments (generally its subject) is *N*'s semantic argument.

Note that our definition of an LVC does not fully overlap with the state of the art. On the one hand, we are more restrictive than some approaches in that we do not include cases in which the verb does add some (even bleached) semantics to the noun. For instance, inchoative verbs combined with non-inchoative nouns such as (PL) *objąć patronat* 'to embrace patronage' ⇒ 'to take on patronage' fail Test 11 and are therefore not classified as LVCs, although their fully bleached counterparts are, as (PL) ***sprawować patronat*** 'to perform patronage' ⇒ 'to dispense patronage'. On the other hand, we include in LVCs those combinations in which a semantically void verb selects a large class of action/state nouns so that its lexical non-compositionality is hard to establish, e.g. (FR) ***commettre un crime/délit/meurtre****/…* ‘to commit a crime/offence/murder/…’.

The latter reason makes LVCs belong to the grey area of (non-)compositionality. They are mostly morphologically and syntactically regular. They can also be seen as semantically compositional in the sense that the semantically void light verb is simply omitted in the semantic calculus. However, this omission may itself be seen as an irregular property. This confirms the observation of Kracht (2007) that compositionality is a property of linguistic analyses rather than of language items.

4.2.3 Idioms

A verbal idiomatic expression (ID) comprises a head verb *V* (possibly phrasal) and at least one of its arguments. Following the decision tree from Figure 1, a VMWE is classified as an ID in one of the 3 cases:

1. *V* has more than one lexicalised argument, as in (25) and (26)

(25) ***Srce*** *mu je* ***padlo v hlače****.* (SL)
heart him is fallen in pants
His heart fell into his pants. ‘He lost courage.’

(26) جانم به لبم رسید. (FA)
resid **lab**am **be** **jan**am
arrived lips-my to soul-my
My soul arrived at my lips. ‘I am frustrated.’

2. *V*’s single lexicalised argument is of any category other than a reflexive clitic, a particle or a nominal phrase (possibly introduced by a preposition), as in (27), (28) and (29);

(27) *Platforma* ***dopięła swego****.* (PL)
Platform PART-buttoned own
The Platform buttoned PART her own. ‘The Platform fulfilled its plans.’

(28) ***Es gibt*** *kein Zurück.* (DE)
it gives no back
It gives no retreat. ‘There is no retreat.’

(29) *Ele* ***sabe onde pisar****.* (PT)
he knows where step
He knows where to step. ‘He knows how to succeed.’

3. *V*’s single lexicalised argument is a nominal phrase (possibly introduced by a preposition), at least one of the LVC-specific Tests 9–13 fails but at least one of the identification Tests 1–5 applies, as in (30).

(30) *Artık kimsenin* ***aklına gelmeyecek****.* (TR)
anymore of-anyone to-his-mind it-will-not-come
It will not come to the mind of anyone anymore. ‘No one will remember it anymore.’

Distinguishing an ID from an LVC in case 3 is one of the hardest and most frequent annotation challenges. In case 1, care must be taken to identify and also annotate nested VMWEs (if any), e.g. the VMWE in (31) contains a nested ID (RO) ***dă pe față*** ‘gives on face’ ⇒ ‘reveals’.

(31) *El* ***dă cărțile pe față****.* (RO)
he gives cards on face
He gives the cards on the face. ‘He reveals his intentions.’

Idioms whose head verb is the copula (*to be*) pose special challenges because their complements may be (nominal, adjectival, etc.) MWEs themselves. In this task, we consider constructions with a copula to be VMWEs only if the complement does not retain the idiomatic meaning when used without the verb. For instance, (PL) *on* ***jest jedną nogą na tamtym świecie*** ‘he is with one leg in the other world’ ⇒ ‘he is close to death’ is an ID because (PL) *jedna noga na tamtym świecie* ‘one leg in the other world’ loses the idiomatic meaning, while (PL) to stwierdzenie *jest do rzeczy* ‘this statement is to the thing’ ⇒ ‘this statement is relevant’ is not a VMWE since (PL) *do rzeczy* ‘to the thing’ ⇒ ‘relevant’ keeps the idiomatic reading.

4.2.4 Inherently reflexive verbs

Pronominal verbs, sometimes also called reflexive verbs, are formed by a verb combined with a reflexive clitic (REFL). They are very common in Romance and Slavic languages, and occur in some Germanic languages such as German and Swedish. Clitics can be highly polysemous and sometimes have an idiomatic rather than a reflexive meaning, in which case we call them inherently reflexive verbs (IReflVs). To distinguish regular from idiomatic uses of reflexive clitics, we rely on an IReflV-specific decision tree[7] containing 8 tests, which are meant to capture an idiosyncratic relation between a verb with a reflexive clitic and the same verb alone. The first 3 of these tests are sufficient to identify most of the actual IReflVs:

Test 14 [INHERENT]: *V* never occurs without *C*, as in (32);

(32) *Jonas har **försovit sig** idag.* (SV)
Jonas has overslept REFL.3.SG today
Jonas overslept REFL today. 'Jonas overslept today.'

Test 15 [DIFF-SENSE]: *C* markedly changes the meaning of *V*, as in (33);

(33) *kar **se tiče** Kosova* (SL)
what REFL touches Kosovo
what REFL touches Kosovo 'as far as Kosovo is concerned'

Test 16 [DIFF-SUBCAT]: *C* changes the subcategorisation frame of *V*, as in (34) vs. (PT) *você me esqueceu* 'you forgot me'.

(34) *Você **se esqueceu** de mim.* (PT)
you REFL.3.SG forgot of me
You forgot REFL about me. 'You forgot about me.'

IReflVs are hard to annotate because pronominal clitics have several different uses. For example, (IT) *si* 'REFL' can occur not only in IReflVs such as (IT) ***riferirsi*** 'to report.REFL' ⇒ 'to refer', but also in the following non-idiomatic cases: reflexive (IT) *lavarsi* 'to wash.REFL', possessive reflexive (IT) *grattarsi la testa* 'to scratch.REFL head' ⇒ 'to scratch one's head', reciprocal (IT) *baciarsi* 'to

[7] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/?page=ireflv

kiss.REFL' ⇒'to kiss each other', impersonal (IT) *si dorme molto* 'REFL sleeps much' ⇒ 'people sleep a lot', middle alternation (IT) *si affittano case* 'REFL rent houses' ⇒ 'houses are rented' or inchoative (IT) *la porta si apre* 'the door REFL opens' ⇒ 'the door opens'. The IReflV category was reported as the most challenging to annotate by some teams, notably the Spanish and the Romanian ones.

4.2.5 Verb-particle constructions

Verb-particle constructions (VPCs) are pervasive notably in Germanic languages and Hungarian, but virtually non-existent in Romance or Slavic languages. They are formed by a lexicalised head verb *V* and a lexicalised particle *P* dependent on *V*, whose joint meaning is non-compositional. The latter property is approximated by a unique syntactic test:

> **Test 22** [V+PART-DIFF-SENSE] A sentence without *P* does not refer to the same event/state as the sentence with *P*. For example, the sentence in (35) does not imply (HU) *nekem jött ez a koktél* 'this cocktail bumped into me', while (DE) *er legt das Buch auf dem Tisch ab* 'he puts the book on the table PART' implies (DE) *er legt das Buch auf dem Tisch* 'he puts the book on the table'.

(35) ***Be-jött*** *ez a koktél nekem.* (HU)
PART-bumped this the cocktail for.me
This cocktail bumped PART into me. 'I like this cocktail.'

The first challenge in identifying a VPC is to distinguish a particle, as in (EN) *to **get** __**up**__ a party*, from a homographic preposition, as in (EN) *to get __up__ the hill.* Language-specific tests were designed for German and English to this aim.

In some Germanic languages and also in Hungarian, verb-particle constructions can be spelled either as one (multiword) token, as in (36), or separated, as in (37). Both types of occurrences are to be annotated.

(36) *ő* ***be-rúgott***. (HU)
he PART-kicked
He kicked PART. 'He got drunk.'

(37) *Nem ő* ***rúgott be***. (HU)
not he kicked PART
He did not kick PART. 'He did not get drunk.'

Special care must be taken with polysemous constructions having both a compositional and a non-compositional reading, as in (DE) *ein Schild aufstellen* ‘to put up a sign’ vs. (DE) *einen Plan* ***aufstellen*** ‘to put up a plan’ ⇒ ‘to draw up a plan’.

4.2.6 Other VMWEs

This category gathers the VMWEs which do not have a single verbal head (cf. Test 6 in Figure 1 and §4.2.1). Those include:

- Coordinations like in example (14) p. 101, or (38)

(38) בריטניה ***נשאה ונתנה*** עם מצרים. (HE)
micrayim ’im **ve-natna nas’a** britanya
Egypt with and-gave carried Britain
Britain carried and gave with Egypt. ‘Britain negotiated with Egypt.’

- Compound verbs, resulting usually from conversion of nominal compounds, and therefore having no regular verbal structure, as in (39) or in (EN) *to* ***pretty-print***.

(39) *On* ***court-circuite*** *le réseau terrestre.* (FR)
one short-circuits the network terrestrial
One short-circuits the terrestrial network. ‘One bypasses the terrestrial network.’

4.3 Language-specific interpretation of the guidelines

Despite huge efforts put into setting up generic terminologies and methodologies, as well as into the pilot annotations and the project coordination, language-specific interpretation of the final guidelines could not be avoided. This was mainly due to different linguistic sensitivities and traditions, language-specific challenges and incompleteness or imprecision of the guidelines.

The most notable deviation occurred in Farsi, where no categorisation was performed, and the OTH label was used for all identified VMWEs instead. The main reason is the particularly challenging nature of the VMWE phenomenon in this language. There are less than 200 actively used simple (single-word) verbs, and a large majority of events and processes are expressed by multiword combinations, many of which are potential VMWEs. The implications on our annotation process are at least threefold. Firstly, verbs are extremely polysemous, so Test 11 (§4.2.2) is very difficult to apply. In particular, the highly frequent light verb کردن */kardan/* 'to do/make' is ambiguous in its passive form شدن*/šodan/* 'done/made' with the semi-copula equivalent roughly to 'become'. Only the former interpretation should yield a VMWE annotation but the difference is hard to capture. Secondly, rephrasing an LVC by a single verb, often used to approximate Test 9 in other languages (*to **make** a **decision** = to decide*), is rarely feasible in Farsi. Thirdly, VMWEs are extremely pervasive, which is easily visible in Table 3: the number of annotated VMWEs is roughly the same as the number of sentences, i.e. almost every main verb is the head of a VMWE. As a result, the VMWE phenomenon is particularly hard to capture in Farsi since it can rarely be contrasted with verbal constructions deemed compositional.

Another notable deviation occurred in Slovene, where the VPC category, as defined by the generic guidelines, hardly or never occurs, however it was used instead to annotate idiomatic verb-preposition combinations, such as (SL) ***prišlo*** *je **do** nesreče* 'it came to an accident' ⇒ 'an accident occurred'.

The status of VPCs in Italian is interesting. As a Romance language, Italian was expected not to exhibit VPCs, but several dozens of VPC annotations do occur in the Italian corpus, e.g. (IT) ***volata via*** 'flew PART' ⇒ 'slipped away', ***tira fuori*** 'pulls PART' ⇒ 'shows', or ***va avanti*** 'goes PART' ⇒ 'goes on'. This shows the possibly ambiguous status of *via* 'by/away', *avanti* 'on/forward', *fuori* 'out/outside', etc. as either adverbs or particles, triggering the ID or the VPC category, respectively. The semantic compositionality of some of these constructions might also be examined more closely.

In Bulgarian and Czech, the auxiliaries accompanying the head verbs were annotated as VMWE components, e.g. in (CS) *on **se** **bude** **bavit*** 'he REFL will play' ⇒ 'he will play', in (BG) *te ne **sa** **dali saglasie*** 'they not are given consent' ⇒'they have not given consent'. This is in contrast with the guidelines, which stipulate that only the lexicalised components should be annotated. The motivation for this deviation was to always include a finite verb in the annotated expression, so as to e.g. easily study the tense and mood restrictions in VMWEs. Since such studies are enabled by the accompanying morpho-syntactic data (currently existent in Czech and to be provided in Bulgarian in the future), these divergences should be eliminated in new editions of the corpus.

In German, a deviation was observed with respect to VMWEs containing both a reflexive clitic and a particle such as (DE) *sie **bringen sich ein*** 'they bring REFL

PART' ⇒ 'they contribute'. Such cases were annotated as IReflVs with nested VPCs, which does not conform to Test 7 (§4.2.1) stipulating that, whenever the VMWE has more than one lexicalised dependent of the head verb, it should be classified as an ID (here: with a nested VPC). Good reasons exist for each of these strategies and more discussion is needed to arbitrate for future releases of the guidelines.

Lithuanian seems to have a surprisingly low number of LVCs, despite the large size of the annotated corpus. It would be worthwhile to study in more detail if this phenomenon is inherent to the language or results from a more restrictive understanding of the LVC scope.

In Hebrew, a relatively large number of VMWEs of type OTH was observed (cf. Table 3), and a necessity of defining a new category (specific to non-Indo-European languages) was hypothesised. A more detailed study revealed that most OTH annotations were spurious: they concerned statistical collocations or VMWEs of the ID or LVC types. Some idiomatic verb-preposition combinations were also annotated in Hebrew, despite the fact that we had abandoned the IPrepV category in the earlier stages of the project (§3). There, the annotators faced a particular challenge from prepositions which often attach to the governed noun and annotating them as separate lexicalised tokens was mostly impossible. Thus, in the following sequence: (HE) *sovel me.achuz avtala* 'suffers from.a.percentage of.unemployment' the free complement *achuz* 'percentage' had to be annotated as lexicalised together with its governing preposition *me* 'from'. This problem will be dealt with in the future, when inherently adpositional verbs will be addressed (§10).

In Turkish, the LVC and OTH types also had their language-specific interpretation. Namely, the Turkish PARSEME corpus resulted from adapting a pre-existing MWE typology and dataset (Adalı et al. 2016). There, the definition of a light verb, based on Turkish linguistic works (Siemieniec-Gołaś 2010), was context-independent, i.e. restricted to a closed list of 6 verbs: *olmak* 'to be', *etmek* 'to do', *yapmak* 'to make', *kılmak* 'to render', *eylemek* 'to make' and *buyurmak* 'to order'. Verb-noun combinations with other operator verbs, such as ***söz vermek*** 'promise to give' ⇒ 'to promise', were then classified as OTH. A closer look at the existing OTH annotations reveals, indeed, that most of them can be re-classified as LVC in future releases of the corpus.

Czech is another language in which a pre-existing MWE-annotated corpus (Hajič et al. 2017) was adapted to the needs of the PARSEME initiative. There, complex identification and conversion procedures had to be designed (Bejček et al. 2017). The resulting mapping procedure could be fully automatic, which suggests that the understanding of the VMWE phenomenon is similar in both annotation projects. It would still be interesting to compare both annotation guidelines more thoroughly and look for possible divergences.

5 Annotation methodology and tools

Mathet et al. (2015) mention several challenging features of linguistic annotation, some of which are relevant to the VMWE annotation task:

- *Unitising*, i.e. identifying the boundaries of a VMWE in the text;
- *Categorisation*, i.e. assigning each identified VMWE to one of the pre-defined categories (§3);
- *Sporadicity*, i.e. the fact that not all text tokens are subject to annotation (unlike in part-of-speech annotation, for instance);
- *Free overlap*, e.g. in (CS) ***ukládal*** *různé* ***sankce*** *a* ***penále*** 'put various sanctions and penalties', where two LVCs share a light verb;
- *Nesting*,
  - at the syntactic level, as in (40), where an IReflV (PL) ***skarżyć się*** 'to complain REFL' ⇒ 'to complain' occurs in a relative clause modifying the predicative noun of the LVC (PL) ***popełnić oszustwo*** 'to commit a fraud'.

    (40) ***Oszustwa****, na jakie* ***skarżą*** ***się*** *Cyganie,* ***popełniły***
    frauds, on which complain REFL Gypsies, committed
    grupy zorganizowane. (PL)
    groups organised
    Organised groups committed frauds about which the Gypsies REFL complain. 'Frauds which Gipsies complain about were committed by organised groups.'

  - at the level of lexicalised components, as in (41), where the ID (PT) ***fazer justiça*** 'to make justice' ⇒ 'to do justice' is nested within a larger ID.

    (41) *Ales* ***fizeram justiça com as próprias mãos****.* (PT)
    they made justice with their own hands
    They made justice with their own hands. 'They took the law into their own hands.'

Two other specific challenges are:

- *Discontinuities*, e.g. (CS) *on* ***ukládal*** *různé* ***sankce*** ‘he put various sanctions’;
- *Multiword token* VMWEs, e.g. separable IReflVs or VPCs:[8]
  (ES) ***abstener****.se* ‘to abstain.REFL’ ⇒ ‘to abstain’,
  (HU) ***át****.**ruház*** ‘to PART.dress’ ⇒ ‘to transfer’.

This complexity is largely increased by the multilingual nature of the task, and calls for efficient project management and powerful annotation tools.

5.1 Project management

The list of language teams having initially expressed their interest in this initiative included those mentioned in p. 91, as well as English, Croatian and Yiddish, for which no corpus release could be achieved due to the lack of sufficiently available native annotators. All languages were divided into four language groups (LGs) - Balto-Slavic, Germanic, Romance and others - as also described in p. 91. The coordination of this large project included the definition of roles – project leaders, technical experts, language group leaders (LGLs), language leaders (LLs) and annotators – and their tasks.

The biggest challenge in the initial phase of the project was the development of the annotation guidelines[9] which would be as unified as possible but which would still allow for language-specific categories and tests. To this end, a two-phase pilot annotation in most of the participating languages was carried out. Some corpora were annotated at this stage not only by native but also by near-native speakers, so as to promote cross-language convergences. Each pilot annotation phase provided feedback from annotators, triggered discussions among language (group) leaders and organisers, and led to enhancements of the guidelines, corpus format and tools.

We also defined strategies for selecting the final corpora. They should: (i) be written in the original, in order to avoid MWE-related translationese issues; (ii)

[8] Note that annotating separate syntactic words within such tokens would be linguistically more appropriate, and would avoid bias in inter-annotator agreement and evaluation measures – cf. §6.2 and (Savary et al. 2017). However, we preferred to avoid token-to-word homogenising mainly for the reasons of compatibility. Namely, for many languages, pre-existing corpora were used, and we would like VMWE annotations to rely on the same tokenisation as the other annotation layers.

[9] Their final version, with examples in many participating languages, is available under the CC BY 4.0 license at http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/.

correspond to the same genre: newspaper texts or Wikipedia articles;[10] (iii) consist of longer text fragments (rather than isolated sentences), so as to enable disambiguation and coreference resolution; (iv) not be automatically pre-selected in view of a higher density of VMWEs (so as to provide both positive and negative examples); (v) be free from copyright issues, i.e. compatible with open licenses.

5.2 Annotation platform

For this large-scale corpus construction, we needed a centralised web-based annotation tool. Its choice was based on the following criteria: (i) handling different alphabets; (ii) accounting for right-to-left scripts; and (iii) allowing for discontinuous, nested and overlapping annotations. We chose FLAT,[11] a web platform which, in addition to the required criteria, enables token-based selection of text spans, including cases in which adjacent tokens are not separated by spaces. It is possible to authenticate and manage annotators, define roles and fine-grained access rights, as well as customise specific settings for different languages.

FLAT is implemented as a web-based frontend with support for multiple users, user groups, and with configurable access rights. The frontend communicates with the FoLiA document server backend,[12] which loads and holds documents in memory as they are being edited, writes them to disk again at convenient times, and unloads them when they are not used anymore. The document server has Git version control support,[13] allowing changes to be tracked. In addition, for each individual FoLiA annotation, e.g. each VMWE, information such as who made the annotation, and when, is automatically registered.

FLAT is document-centric, i.e. it supports annotation of full documents together with their structure (headers, bulleted lists, figures, etc.). This contrasts with tools which take a more corpus-based approach with keyword-in-context visualisation. FLAT does allow for various other *perspectives* on the document; for the PARSEME annotation task a sentence-based perspective was chosen, presenting users with one or more pages of clearly delimited sentences to annotate. An example is shown in Figure 2.

FLAT is based on FoLiA,[14] a rich XML-based format for linguistic annotation (van Gompel & Reynaert 2013), and is compatible with a wide variety of linguis-

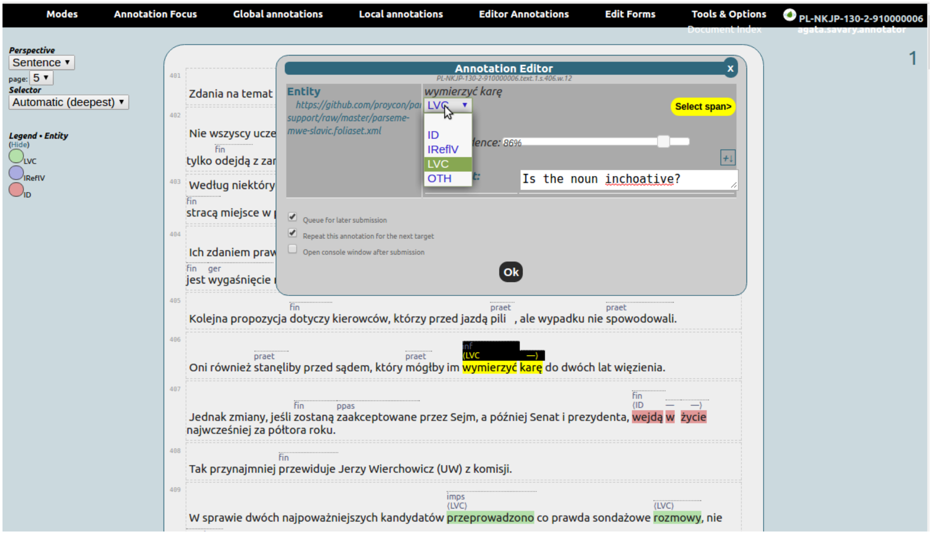
[10] Deviations from this rule occurred in some languages due to the choice of pre-existing corpora, e.g. in Hungarian legal texts were used.

[11] https://github.com/proycon/flat

[12] https://github.com/foliadocserve

[13] https://git-scm.com/

[14] https://proycon.github.io/folia



wymierzyć karę

...dence: 86%

...t: Is the noun inchoative?

☑ Queue for later submission

☑ Repeat this annotation for the next target

□ Open console window after submission

401 Zdania na temat

402 Nie wszyscy ucze
fin
tylko odejdą z zar

403 Według niektóry
fin
stracą miejsce w

404 Ich zdaniem praw
fin ger
jest wygaśnięcie

405 fin praet praet
Kolejna propozycja dotyczy kierowców, którzy przed jazdą pili , ale wypadku nie spowodowali.

406 praet praet inf (LVC —)
Oni również stanęliby przed sądem, który mógłby im wymierzyć karę do dwóch lat więzienia.

407 fin ppas fin (ID — —)
Jednak zmiany, jeśli zostaną zaakceptowane przez Sejm, a później Senat i prezydenta, wejdą w życie najwcześniej za półtora roku.

408 fin
Tak przynajmniej przewiduje Jerzy Wierchowicz (UW) z komisji.

409 imps (LVC) (LVC)
W sprawie dwóch najpoważniejszych kandydatów przeprowadzono co prawda sondażowe rozmowy, nie

Figure 2: FLAT annotation interface with a Polish text. The VMWEs are coloured according to their categories. POS tags (*fin*, *ger*, *imps*, *ppas*, and *praet*) are displayed above all verbal tokens. Some attributes (VMWE category, confidence level and a comment) of the highlighted VMWE (PL) ***wymierzyć karę*** 'to PART.measure a punishment' ⟹ 'to mete out a punishment' are edited in the annotation editor.

tic annotation types. VMWEs, or entities as they are called more generically in FoLiA, constitute the most important annotation type for PARSEME. Still, certain language teams worked on documents enriched with more linguistic annotations, such as part-of-speech tags, to aid the annotation process, as shown in Figure 2. The underlying aspiration of both FoLiA and FLAT is to provide a single unified solution for multiple annotation needs, with respect to the encoding format and the annotation environment, respectively.

While the FoLiA format specifies possible linguistic annotation types and structural types, it does not commit to any particular tagset/vocabulary nor language. Instead, tagsets are defined externally in *FoLiA set definitions*, which can be published anywhere online by anyone and are deliberately separate from the annotation format itself. A dozen of set definitions for PARSEME, based on the VMWE categories relevant to different languages or language groups (§3) are likewise published in a public repository.[15] All FoLiA documents declare which particular set definitions to use for which annotation types. FLAT uses these set definitions to populate various selection boxes, as shown in Figure 2.

[15] https://github.com/proycon/parseme-support

All software discussed here is available under an open-source license.[16] It is part of a wider and growing infrastructure of FoLiA-capable NLP tools (van Gompel et al. 2017), developed and funded in the scope of the CLARIAH[17] project and its predecessor CLARIN-NL.

Although FLAT has been in use for various other annotation projects, the PARSEME initiative, currently with over 80 active FLAT users, is the biggest use case to date, and as such has had a very positive influence in terms of the maturity of the software, fixing bugs, attaining improved performance and scalability, and compiling appropriate documentation. Various features were added to accommodate PARSEME specifically: (i) uploading documents in non-FoLiA formats, needed for the parseme-tsv format (6.1); (ii) right-to-left support necessary for Farsi and Hebrew; (iii) a metadata editor; (iv) enhanced file and user management; (v) confidence level and free-text comments as part of the editable attributes (Figure 2).

Out of 18 language teams which achieved a corpus release, 13 used FLAT as their main annotation environment. The 5 remaining teams either used other (generic or in-house) annotation tools, or converted existing VMWE-annotated corpora.

5.3 Automatic VMWE pre-annotation

Automatic pre-annotation of corpora is a current practice in many annotation tasks. In the PARSEME corpus project, it was applied by the Bulgarian and Hungarian teams, on the basis of manually compiled lists of VMWEs. All texts were then manually checked and corrected.

More precisely, pre-annotation in Bulgarian included automatic annotation of: (a) verb forms (triggers for VMWEs), (b) IReflV candidates consisting of a verb and a reflexive particle, and (c) VMWEs from a large dictionary of Bulgarian MWEs (Koeva et al. 2016). Cases of false positives included: (i) literal uses of existing VMWEs, (ii) false IReflVs which are true reflexive or passive constructions instead (§4.2.4), or (iii) coincidental co-occurrence of VMWE components. All annotations were manually verified and such cases were eliminated. False negatives could also be efficiently tracked thanks to the highlighted verb forms.

Automatic pre-annotation is known to introduce a task-dependent bias (Marcus et al. 1993; Fort & Sagot 2010) which may be both positive (simple repetitive tasks are handled uniformly and speeded up) and negative (annotators may tend

[16] GNU Public License v3

[17] https://www.clariah.nl

to rely too much on the automatic pre-annotation and fail to detect false negatives). We are not aware of any studies about biases related to VMWE annotation. We expect a minor risk of bias to stem from a possibly unbalanced VMWE dictionary: if one category (e.g. LVCs) is better represented than others, annotators may become more attentive to it. A bias might also be introduced by relatively productive constructions, when a large majority, but not all, of their occurrences belong to a unique category. For instance, the verb (BG) *davam* 'to give' occurs often and in many different LVCs, e.g. with *saglasie* 'consent', *razreshenie* 'permission' *obyasnenie* 'explanation', etc. The annotators could, therefore, tend to wrongly assign the LVC category to other expressions containing the same verb, such as ***davam duma*** 'to give word' (ID), or *davam prizovka* 'to give subpoena' (non-VMWE or borderline case).

5.4 Consistency checks and homogenisation

Even though the guidelines heavily evolved during the two-stage pilot annotation, there were still questions from annotators at the beginning of the final annotation phase. We used an issue tracker (on Gitlab)[18] in which language leaders and annotators could discuss issues with other language teams.

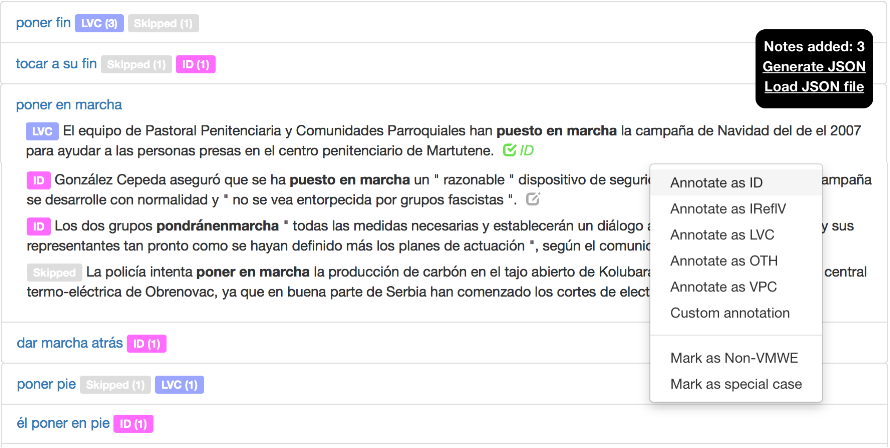
High-quality annotation standards require independent double annotation of a corpus followed by adjudication, which we could not systematically apply due to time and resource constraints. For most languages, each text was handled by one annotator only (except for a small corpus subset used to compute inter-annotator agreement, see §6.2). This practice is known to yield inattention errors and inconsistencies between annotators, and since the number of annotators per language varies from 1 to 10, we used consistency support tools.

Firstly, some language teams (Bulgarian, French, Hungarian, Italian, Polish, and Portuguese) kept a list of VMWEs and their classification, agreed upon by all annotators and updated collaboratively over time.[19] Secondly, for some languages (German, French, Hebrew, Italian, Polish, Portuguese, Romanian and Spanish) the annotation was followed by homogenisation. An in-house tool extracted the annotated VMWEs from a given corpus and rescanned the corpus to find all potential occurrences of the same VMWEs, whether already annotated or not. It then generated an HTML page where all positive and negative examples of a given VMWE were grouped, and could be accepted or rejected manually. En-

[18] https://gitlab.com/parseme/sharedtask-guidelines/issues

[19] Like automatic pre-annotation, this practice increases the consistency and speed of the annotator's work, but it also introduces a risk of bias, since collective decisions may override linguistic intuition. Therefore, such instruments should always be used with special care.

tries were sorted so that similar VMWEs, such as (EN) ***payed** a **visit*** and ***received** a **visit***, appeared next to each other. In this way, noise and silence errors could easily be spotted and manually corrected. The tool was mostly used by language leaders and/or highly committed annotators. The resulting gain in precision and recall was substantial. For instance, in Spanish the number of the annotated MWEs increased by 40% (from 742 to 1248), most notably in the IReflV category. Figure 3 shows the interface used to correct consistency problems.



LVC El equipo de Pastoral Penitenciaria y Comunidades Parroquiales han **puesto en marcha** la campaña de Navidad del de el 2007 para ayudar a las personas presas en el centro penitenciario de Martutene. ✓ *ID*

ID González Cepeda aseguró que se ha **puesto en marcha** un " razonable " dispositivo de seguri… ampaña se desarrolle con normalidad y " no se vea entorpecida por grupos fascistas ".

ID Los dos grupos **pondránenmarcha** " todas las medidas necesarias y establecerán un diálogo … y sus representantes tan pronto como se hayan definido más los planes de actuación ", según el comunic…

Skipped La policía intenta **poner en marcha** la producción de carbón en el tajo abierto de Kolubar… central termo-eléctrica de Obrenovac, ya que en buena parte de Serbia han comenzado los cortes de elect…

Figure 3: Consistency-check tool at work. Here, (ES) ***poner en marcha*** 'to put in march' ⇒ 'to start' was annotated once as LVC, twice as ID and once skipped. The clickable icon next to each example allows the user to add, correct or delete an annotation. VMWEs with the same noun, e.g. (ES) ***poner fin*** 'to put end' ⇒ 'to terminate' and ***tocar a su fin*** 'to touch to its end' ⇒ 'to come to its end' on the top of the screen, are gathered so as to enhance annotation consistency, especially for LVCs.

6 Properties of the annotated corpus

Table 3 provides overall statistics of the corpus annotated for the shared task.[20] In total, it contains almost 5,5 million tokens, 274 thousand sentences and 62 thousand VMWE annotations. The amount and distribution of VMWEs over categories varies considerably across languages.

No category was used in all languages, but the two universal categories, ID and LVC, were used in almost all languages. In Hungarian, no ID was annotated

[20] The split into training and test corpora is indicated in Savary et al. (2017).

Table 3: Overview of the annotated corpora in terms of the number of sentences, of tokens (whether belonging to the annotated VMWEs or not), and of the annotated VMWEs occurrences (overall and per category).

Language	Sentences	Tokens	VMWE occurrences					
			All	IDs	IReflVs	LVCs	OTHs	VPCs
BG	8,860	200,128	2,406	517	1,376	511	2	0
CS	49,431	833,193	14,536	1,611	10,000	2,923	2	0
DE	7,500	144,856	2,947	1,219	131	218	10	1,369
EL	8,811	226,265	2,018	642	0	1,291	37	48
ES	4,634	159,807	1,248	362	556	320	10	0
FA	3,226	55,207	3,207	0	0	0	3,207	0
FR	19,547	486,005	4,962	1,905	1,418	1,633	6	0
HE	7,000	147,361	1,782	116	0	380	693	593
HU	4,311	108,175	3,499	0	0	730	0	2,769
IT	17,000	427,848	2,454	1,163	730	482	6	73
LT	14,863	256,235	502	287	0	215	0	0
MT	10,600	152,285	1,272	446	0	693	133	0
PL	13,606	220,934	3,649	383	1,813	1,453	0	0
PT	22,240	414,020	3,947	910	596	2,439	2	0
RO	51,500	879,427	4,540	599	2,786	1,154	1	0
SL	11,411	235,864	2,287	375	1,198	231	4	479
SV	1,800	29,517	292	60	17	27	2	186
TR	18,036	362,077	6,670	3,160	0	2,823	687	0
Total	274,376	5,439,204	62,218	13,755	20,621	17,523	4,802	5,517

due to the genre of the corpus, mainly composed of legal texts. In Farsi, no categorisation was performed (§4.3), and all annotated VMWEs are marked as OTH instead.

The most frequent category is IReflV, in spite of it being quasi-universal, mainly due to its prevalence in Czech. IReflVs were annotated in all Romance and Slavic languages, and in German and Swedish. VPCs were annotated in German, Swedish, Greek, Hungarian, Hebrew, Italian, and Slovene. In the three last languages this category had a language-specific interpretation, as was the case of OTH in Hebrew and Turkish (§4.3). No language-specific categories have been defined.

All the corpora are freely available on the LINDAT/CLARIN platform.[21] The VMWE annotations are released under Creative Commons licenses, with constraints on commercial use and sharing for some languages. Some languages use data from other corpora (notably from the UD project), including additional annotations. These are released under the terms of the original licenses.

6.1 Format

The official format of the annotated data is the parseme-tsv format,[22] exemplified in Figure 4. It is adapted from the CoNLL format, with one token per line and an empty line indicating the end of a sentence. Each token is represented by 4 tab-separated columns featuring (i) the position of the token in the sentence, or a range of positions (e.g. `1–2`) in case of MWTs such as contractions; (ii) the token surface form; (iii) an optional `nsp` (no space) flag indicating that the current token is adjacent to the next one; and (iv) an optional VMWE code composed of the VMWE's consecutive number in the sentence and – for the initial token in a VMWE – its category, for example, `2:ID` if a token is the first one in an idiom which is the second VMWE in the current sentence. In case of nested, coordinated or overlapping VMWEs, multiple codes are separated with a semicolon.

Formatting of the final corpus required a language-specific tokenisation procedure, which can be particularly tedious in languages presenting contractions. For instance, (FR) *du* 'of-the' is a contraction of the preposition (FR) *de* 'of' and the article (FR) *le* 'the.MASC'.

Some language teams resorted to previously annotated corpora which have been converted to the parseme-tsv format automatically (or semi-automatically if some tokenisation rules were revisited). Finally, scripts for converting the parseme-tsv format into the FoLiA format and back were developed to ensure corpus compatibility with FLAT (5.2).

6.2 Inter-annotator agreement

Inter-annotator agreement (IAA) measures are meant to assess the hardness of the annotation task, as well as the quality of the annotation guidelines, of the annotation methodology, and of the resulting annotations. Defining such measures is not always straightforward due to the challenges listed in §5.

To assess unitising, two annotators double-annotated an extract of the corpus in each language. We then calculated the MWE-based F-score ($F1_{unit}$) of one

[21] http://hdl.handle.net/11372/LRT-2282

[22] http://typo.uni-konstanz.de/parseme/index.php/2-general/184-parseme-shared-task-format-of-the-final-annotation

1-2	Wouldn’t		
1	Would		
2	not		
3	questioning		
4	colonial		
5	boundaries		
6	open		1:ID
7	a		
8	dangerous		
9	Pandora	nsp	1
10	’	nsp	1
11	s		1
12	box	nsp	1
13	?		

1	They		
2	were		
3	letting		1:VPC;2:VPC
4	him		
5	in		1
6	and		
7	out		2
8	.	nsp	

Figure 4: Annotation of two sample sentences containing a contraction (*wouldn’t*), a verbal idiom, and two overlapping VPCs.

annotator with respect to the other.[23] MWE-based F-score is defined in Savary et al. (2017) and was used to evaluate the systems submitted to the shared task.

We also report an estimated Cohen’s κ (κ_{unit}). Measuring IAA, particularly κ, for unitising is not straightforward due to the absence of negative examples, that is, spans for which both annotators agreed that they are not VMWEs. From an extreme perspective, any combination of a verb with other tokens (of any length) in a sentence is a potential VMWE.[24] Consequently, as the density of VMWEs in most languages is rather low, one can argue that the probability of chance agreement approaches 0, and IAA can be measured simply using the observed agreement $F1_{unit}$. However, in order to provide a possibly less biased measure

[23] That is, we suppose that one annotator represents the system, and the other one represents the gold standard. Note that F-score is symmetrical (depending on the order, recall and precision are inverted), so none of the two annotators is prioritised.

[24] Also note that annotated segments can overlap.

to the reported F-scores, we assume that the total number of stimuli in the annotated corpora is approximately equivalent to the number of verbs, which is slightly higher than the number of sentences. We roughly estimate this quantity as the number of sentences plus the number of VMWEs annotated by at least one annotator.[25] Finally, to assess categorisation, we apply the standard κ (κ_{cat}) to the VMWEs for which annotators agree on the span.

Due to time and resource constraints, the majority of the corpus for most languages was annotated by a single annotator. Only small fractions were double-annotated for the purpose of the IAA calculation. All available IAA results are presented in Table 4. For some languages the IAA in unitising is rather low. We believe that this results from particular annotation conditions. In Spanish, the annotated corpus is small (Table 3), so the annotators did not become sufficiently accustomed to the task. A similar effect occurs in Polish and Farsi, where the first annotator performed the whole annotation of the train and test corpora, while the second annotator only worked on the IAA-dedicated corpus. The cases of Hebrew, and especially of Italian, should be studied more thoroughly in the future. Note also that in some languages the numbers from Table 4 are a lower bound for the quality of the final corpus, due to post-annotation homogenisation (§5.4).

A novel proposal of the holistic γ measure (Mathet et al. 2015) combines unitising and categorisation agreement in one IAA score, because both annotation subtasks are interdependent. In our case, however, separate IAA measures seem preferable both due to the nature of VMWEs and to our annotation methodology. Firstly, VMWEs are known for their variable degree of non-compositionality. In other words, their idiomaticity is a matter of scale. But this fact is not accounted for in current corpus annotation standards and identification tools, which usually rely on binary decisions, i.e. a candidate is seen as a VMWE or a non-VMWE, with no gradation of this status. Such a binary model is largely sub-optimal for a large number of grey-zone VMWE candidates. However, once a VMWE has been considered valid, its categorisation appears to be significantly simpler, as shown in the last 2 columns of Table 4 (except for Romanian and Hebrew). Secondly, as described in §4.1 – §4.2, our annotation guidelines are structured in two main decision trees – an identification and a categorisation tree – to be applied mostly sequentially. Therefore, separate evaluation of these two stages may be helpful in enhancing the guidelines.

[25] In other words, the number of items on which both annotators agree as being no VMWEs is estimated as the number of sentences. This assumption ignores the fact that some verbs may be part of more than one VMWE, since this is rare.

Table 4: IAA scores: #S, and #T show the the number of sentences and tokens in the double-annotated sample used to measure IAA, respectively. #A$_1$ and #A$_2$ refer to the number of VMWE instances annotated by each of the annotators.

	#S	#T	#A$_1$	#A$_2$	F1$_{unit}$	κ_{unit}	κ_{cat}
BG	608	27491	298	261	81.6	0.738	0.925
EL	1383	33964	217	299	68.6	0.632	0.745
ES	524	10059	54	61	38.3	0.319	0.672
FA	200	5076	302	251	73.9	0.479	n/a
FR	1000	24666	220	205	81.9	0.782	0.93
HE	1000	20938	196	206	52.2	0.435	0.587
HU	308	8359	229	248	89.9	0.827	1.0
IT	2000	52639	336	316	41.7	0.331	0.78
PL	1175	19533	336	220	52.9	0.434	0.939
PT	2000	41636	411	448	77.1	0.724	0.964
RO	2500	43728	183	243	70.9	0.685	0.592
TR	6000	107734	3093	3241	71.1	0.578	0.871

6.3 Cross-language analysis

The common terminology and annotation methodology achieved in this endeavour enable cross-language observations. In this section we offer a comparative quantitative analysis of several phenomena relevant to the challenges VMWEs pose in NLP, as discussed in §1. Namely, we analyse the lengths, discontinuities, coverage, overlapping and nesting of VMWEs across languages and VMWE types.

Table 5 provides statistics about the length and discontinuities of annotated VMWEs in terms of the number of tokens.[26] The average lengths range between 1.27 (in Hungarian) and 2.71 (in Hebrew) tokens, but the dispersion varies across languages: the mean absolute deviation (MAD) is 0.75 for Hebrew, while it is 0.11 for Turkish. Single-token VMWEs (length=1) are frequent in Hungarian and German (63% and 24% of all VMWEs, respectively) but rare or non-existent in other languages. The right part of Table 5 shows the lengths of discontinuities (gaps). This factor is measured in terms of the total number of tokens not belonging to

[26] Since the version published in Savary et al. (2017), we corrected a bug in the length average and MAD calculation, which impacted the results for languages containing VMWEs with one token only (especially DE and HU).

Table 5: Length and discontinuities of VMWE occurrences in number of tokens in the training corpora. Col. 2–3: average and mean absolute deviation (MAD) for length. Col. 4: number of single-token VMWEs. Col. 5–6: average and MAD for the length of discontinuities. Col. 7–8: number and percentage of continuous VMWEs. Col. 9–11: number of VMWEs with discontinuities of length 1, 2 and 3. Col. 12–13: number and percentage of VMWEs discontinuities of length > 3.

	Length of VMWE			Length of discontinuities (excl. VMWEs of length 1)								
Lang.	Avg	MAD	=1	Avg	MAD	0	%0	1	2	3	>3	%>3
BG	2.45	0.63	1	0.64	1.05	1586	82.1	206	33	25	82	(4.2%)
CS	2.30	0.46	0	1.35	1.53	6625	51.5	2357	1465	944	1461	(11.4%)
DE	2.02	0.61	715	2.96	2.94	619	35.7	283	159	142	529	(30.5%)
EL	2.45	0.61	3	0.94	1.08	870	57.4	389	124	50	82	(5.4%)
ES	2.24	0.39	0	0.47	0.66	523	69.9	162	33	14	16	(2.1%)
FA	2.16	0.27	0	0.42	0.70	2243	82.9	202	103	60	99	(3.7%)
FR	2.29	0.44	1	0.65	0.80	2761	61.9	1116	336	125	123	(2.8%)
HE	2.71	0.75	0	0.47	0.74	1011	78.9	129	54	43	45	(3.5%)
HU	1.27	0.39	2205	1.01	1.29	506	63.7	178	34	15	61	(7.7%)
IT	2.58	0.64	2	0.28	0.46	1580	80.9	278	56	22	16	(0.8%)
LT	2.35	0.53	0	0.72	0.94	261	64.9	79	36	9	17	(4.2%)
MT	2.64	0.68	7	0.34	0.53	589	77.0	123	33	12	8	(1.0%)
PL	2.11	0.20	0	0.53	0.77	2307	73.3	470	195	90	87	(2.8%)
PT	2.19	0.37	76	0.67	0.78	1964	58.3	1016	223	82	86	(2.6%)
RO	2.15	0.25	1	0.55	0.72	2612	64.7	689	693	32	13	(0.3%)
SL	2.27	0.43	14	1.47	1.54	787	44.4	445	221	118	202	(11.4%)
SV	2.14	0.25	0	0.38	0.59	44	78.6	7	3	1	1	(1.8%)
TR	2.06	0.11	3	0.57	0.57	3043	49.4	2900	162	33	28	(0.5%)

a VMWE but appearing between its left- and right-most lexicalised components. For instance, a gap of length 3 is counted in (DE) *jetzt* ***bin*** *ich bestimmt* ***aus*** *dem* ***Alter heraus*** 'now am I certainly out-of the age PART' ⇒ 'now I am too old'. The discontinuities vary greatly across languages. While for Bulgarian, Farsi and Italian more than 80% of VMWEs are continuous, only 35.7% of German VMWEs do not have any gaps, and 30.5% of them contain discontinuities of 4 or more tokens.

Figure 5 and Figure 6 show a breakdown of the length and discontinuity scores per VMWE category (Farsi, where categorisation was not performed, is not included). Not surprisingly, IDs are longer on average than all other categories (OTHs are omitted due to their rarity), and the average ID length ranges roughly between 2.5 and 3 components. The average lengths for the other categories are closer to 2, which is expected given their definitions. Note though that VPCs are

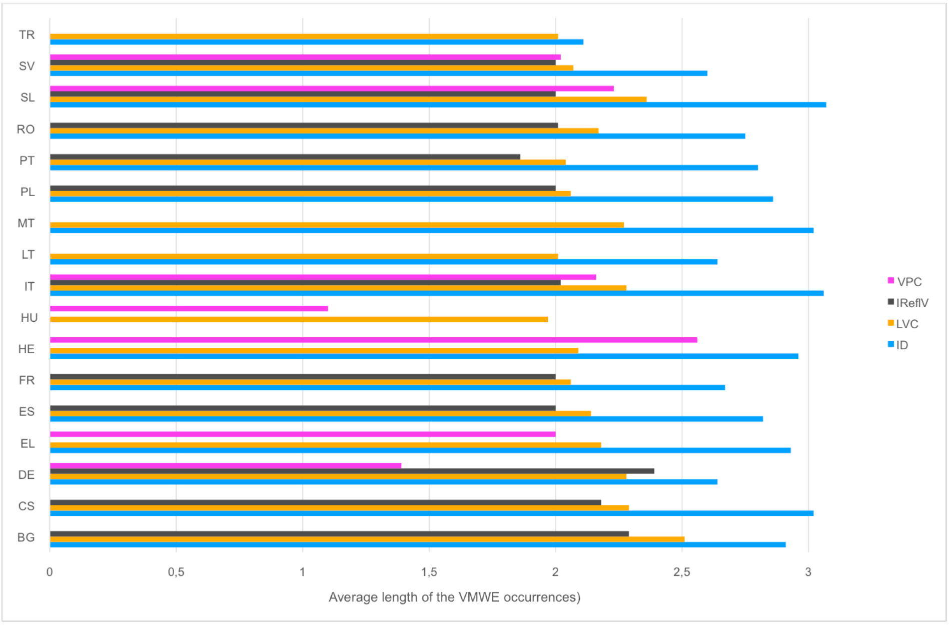


Figure 5: Average lengths of VMWE occurrences per category, in number of components. Single-token VMWEs (frequent for Hungarian and German) are included.

more contrasted across languages, with a low average length for German and Hungarian, due to the massive presence of single-token VMWEs. As far as IReflVs are concerned, a similar effect can be observed for some languages depending on morphological and tokenisation rules, due to the presence of IReflVs of length 1, for instance (ES) ***referir*.*se*** 'to refer.REFL' ⟹ 'to refer'. IReflVs of length greater than 2 in Czech, Bulgarian and German result from language-specific interpretations of the guidelines (§4.3).

When comparing the lengths of discontinuities across languages (Figure 6), German stands clearly out in all categories and so does Slovene to a smaller extent (probably due to the language-specific interpretation of the VPC category, §4.3), whereas Italian, Hebrew or Maltese show very few discontinuities. Note the difference for LVCs within Romance languages, which should be studied in more detail. LVCs are clearly the category showing the longest discontinuities overall, mainly due to the presence of non-lexicalised determiners and pre-modifiers of the predicative nouns, although extraction of the nouns also comes into play.

While regularities do exist in the formation of MWEs, it essentially remains an idiosyncratic and lexical phenomenon. Hence, it is very likely that the annotated

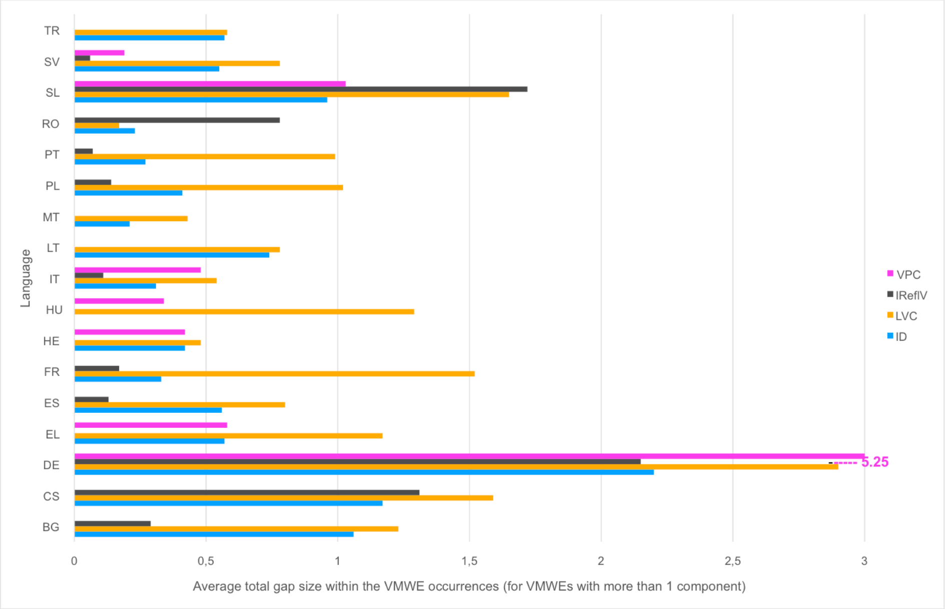


Figure 6: Size of discontinuities in VMWEs. The gap size is the total number of tokens not belonging to a VMWE but appearing between its left- and right-most lexicalised components. VMWEs of length 1 are not considered. For German the VPC average gap size is 5.25.

datasets cover only a small fraction of all the VMWEs existing in each of the 18 languages. In order to evaluate this coverage, we propose to measure the ratio of unknown VMWEs considering a corpus split into training and test sets, similar to the split used in the shared task (Savary et al. 2017). In other words, we arbitrarily split the corpus into a training and a test set, and study the proportion of VMWEs present in the test but absent in the training set.[27]

Ideally, we should perform this estimation on an intra- and inter-domain basis. Unfortunately, we do not know the domain of the source text for each annotated sentence.[28] To circumvent this limitation, we can still provide a lower bound of the unknown VMWE ratios by considering different splits that use continuous portions of the corpus, as shown in Figure 7. For each language for which the morphological companion files were provided, we show the average rate of un-

[27] See also Maldonado & QasemiZadeh (2018 [this volume]) and Taslimipoor et al. (2018 [this volume]) for in-depth considerations on how the training vs. test corpus split influences the results of automatic VMWE identification.

[28] For instance the French dataset contains the UD corpus, whose sentences come from various untraced sources and are mixed.

known VMWEs[29] computed over 5 cross-validation splits, plotted against the total number of VMWE occurrences. For instance for Italian we get an average unknown rate of 66.2%, with roughly 2,000 annotated VMWE tokens, which means that, on average, in a fraction of 400 VMWEs, two thirds are not present in the remaining 1,600 VMWEs. The ratios are rather high, except for Hungarian and Romanian. Although we would expect these scores to have negative correlation with the size of the annotated data, the plot shows great differences even among languages with comparable numbers of annotated VMWEs. We can hypothesise that other factors come into play, such as cross-language variability of domains, text genres and annotation quality.

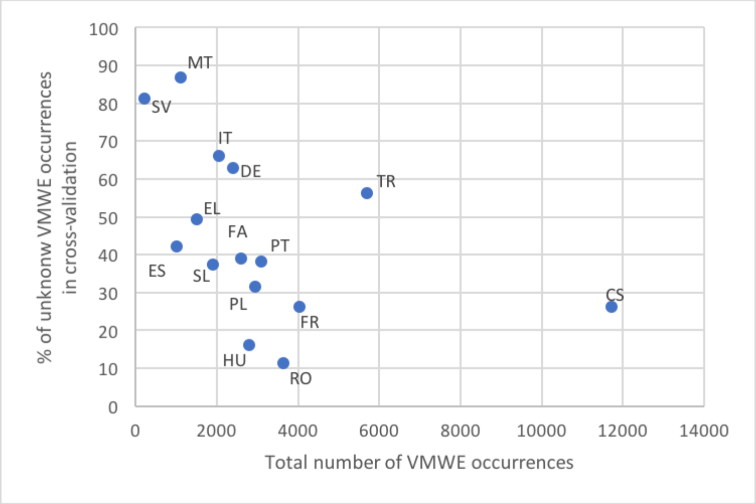


Figure 7: Ratios of unknown VMWEs in the different language datasets. X-axis: the total number of VMWEs tokens in the train+test corpus. Y-axis: average proportion of unknown VMWEs (present in the test but not in the train set) when performing cross-validation with 5 different train/test splits.

We also investigated two other challenging phenomena: overlapping and nesting of VMWEs. The former was measured in terms of the frequency of tokens belonging to at least 2 VMWEs. It occurs – most often due to ellipsis in coordinated VMWEs – in most of the languages but rarely concerns more than two VMWEs at a time, as shown in Table 6. The highest number of overlapping VMWEs was

[29] Matching of VMWEs in train and test sets is performed on lemmatised forms, and with limited normalisation of the order of components (in particular verb-noun for LVCs, and clitic-verb for IReflVs). Note that better normalisation should be performed in order to match multitoken VMWEs against their single-token variants.

five, as seen in (42), where the light verb (PL) *wykonywać* 'perform' is shared by five LVCs.

(42) *Piloci* ***wykonywali*** *podstawowe* ***manewry*** *i serie* ***wznoszeń****,*
pilots performed basic maneuvers and series climbs.GEN,
nurkowań*,* ***pętli*** *i* ***zwrotów****.* (PL)
dives.GEN, rolls.GEN and turns.GEN
'The pilots performed basic maneuvers and series of climbs, dives, rolls and turns.'

As far as nesting is concerned, measuring this phenomenon precisely, as defined in §5, would require the availability of syntactic annotations for all languages. Since this is not the case, we approximated nesting at the syntactic level by pairs of VMWEs E_1 and E_2 such that all lexicalised components of E_2 are placed between the left- and right-most lexicalised components of E_1. Single-token VMWEs were disregarded. As the last line of Table 6 shows, such configurations occur very rarely in the data. This might be due to the fact that large gaps introduced within the outer-most VMWEs by the nested structure are harder to process for the human mind.

Table 6: Overlapping and nested VMWEs. Overlap >=2 and >2: the token belongs to at least 2 or more than 2 VMWEs, respectively. Only percentages above 0.49% are indicated. They are counted wrt. all tokens belonging to VMWEs.

	BG	CS	DE	EL	ES	FA	FR	HE	HU	IT	LT	MT	PL	PT	RO	SL	SV	TR
Overlap >= 2	0	520	122	5	22	1	60	235	30	73	0	1	44	65	53	0	1	19
		(1.6%)	(2%)					(5%)		(1.2%)			(0.6%)	(0.5%)	(0.5%)			
Overlap > 2	0	11	0	1	0	0	5	9	0	0	0	0	1	6	0	0	0	0
Nested VMWEs	4	29	1	0	0	0	1	0	1	3	0	0	4	1	0	2	0	0

7 Language-specific studies based on the corpus

Since its publication in January 2017, the PARSEME VMWE-annotated corpus has enabled studies in corpus linguistics in several languages.

The French corpus was addressed by Pasquer (2017), who focuses on the variability of the most frequent VMWEs. Three aspects are studied: (i) morphological variability of VMWE components, (ii) length and nature of discontinuities between the VMWE components, (iii) syntactic dependencies between the VMWE components and their dependents/governors. The results show a distinctly higher variability in LVCs than in IDs. Namely, nouns inflect and govern external modifiers, respectively, 8 and 1.7 times more often in LVCs (*il **rend** les derniers **hommages*** 'he pays the last tributes') than in IDs. IDs include a lexicalised determiner (*elle **tourne la page*** 'she turns the page' vs. *elle **joue** un **role*** 'she plays a role') and a compulsory negation (*ça **ne paye pas de mine*** 'it does not pay a face' ⇒ 'it is not much to look at'), 20 and 10 times more often than LVC, respectively. LVCs exhibit discontinuities and passivise 1.5 and 29 times more often than IDs, respectively. Additionally, types of syntactic variants are listed and quantified for the 3 most variable VMWEs. Interesting types of morphological variants, such as prefixations (***redonner raison*** 'to re-give reason' ⇒ 'to admit again that someone is right'), are also revealed.

In Maltese, investigations on LVCs were also carried out in the PARSEME corpus extended with the Maltese UD corpus. The annotated LVCs were extracted and proofread, and the 20 most frequent light verbs (LVs) were listed. Those were used to find other candidate LVCs in a larger raw corpus (not annotated for VMWEs). For each LV the number of unique predicative nouns they combine with could be established. The results show that some LVs are inherently light (e.g. *ta* 'to give', *ħa* 'to take' and *għamel* 'to make/do') and combine with large numbers of nouns (here: 60, 48, and 46, respectively), while others are light only when combined with a few nouns (e.g. *ġarr* 'to carry', *laħaq* 'to reach/achieve', *talab* 'to request/ask'). An analogous experiment, performed for nouns, shows that most of them occur with two LVs (*ta* 'to give' and *ħa* 'to take'), while only few (*appoġġ* 'support', *kura* 'care/treatment' and *kenn* 'shelter') combine with many LVs. Other interesting findings are of etymological nature. Maltese is a language with influences from Semitic and Romance languages, as well as English. The inspected LVCs were mostly of Romance origin (70%), some of Semitic (25%) and some of English (5%). Interestingly, some LVCs accommodate borrowings and Semitic elements that are no longer productive, for example, ***ħa nifs*** 'to take a breath' is ten times more frequent than the Semitic *niffes* 'to breathe'.

LVC-specific analyses were also performed in Lithuanian. Two groups of verbs were identified based on their frequencies in LVCs: (i) 4 high-connectivity verbs i.e. those that combine with large numbers of nouns: *vykdyti* 'to carry out' connects with 19 nouns, *atlikti* 'to perform' – 14, *turėti* 'to have' – 12, *daryti* 'to do/to make' – 10; (ii) 17 low-connectivity verbs i.e. those combining with less than 10 nouns, e.g. *teikti* 'to deliver' – 6, *surengti* 'to arrange' – 4, *imtis* 'to undertake' – 3, *priimti* 'to accept' – 3, *patirti* 'to experience' – 3, *duoti* 'to give' – 3, *sudaryti* 'to make' – 3, etc. The numbers of the LVCs containing the verbs from (i) and (ii) are comparable – 55 and 38, respectively – but the diversity of the verbs is significantly higher in (ii) than in (i). The LVCs containing the verbs from group (i) seem to be the most prototypical ones, e.g. ***vykdyti patikrinimus*** 'to carry out inspections', ***atlikti analizę*** 'to perform an analysis', ***daryti spaudimą*** 'to put pressure', etc. These findings pave the way towards developing a comprehensive list of light verbs for Lithuanian.

8 Interesting problems

The considerable collective PARSEME corpus effort led us to confront various phenomena across different language families, various linguistic traditions, and annotation practices. As a result, some interesting findings allow us to view the VMWE phenomenon more globally, which should enable further cross-language generalisations.

Since semantic non-compositionality is the most pervasive property of MWEs, it should possibly be captured by generic definitions and tests in a multilingual endeavour like ours. However, semantic properties show up in different languages via different morphological, syntactic and semantic means. As a result, some semantic non-compositionality phenomena cross word boundaries in some languages, and are therefore relevant to MWEs, and others do not. This distinction can also vary from language to language for the same phenomenon.

For instance, particles in Germanic and Finno-Ugric VPCs, like (EN) *to* ***turn off***, have similar roles as prefixes in Slavic verbs, like (PL) *wy.łączyć* 'to PART.connect' ⇒'to turn off'. The former are traditionally considered separate lexemes, and can therefore form VMWEs with their governing verbs. The latter, conversely, are considered inherent components of verbs, and therefore cannot trigger MWE-related considerations.

Similarly, aspect can be realised by various lexical, morphological and syntactic means, and can therefore be seen as either a semantic or a morphological feature (or both). For instance, perfective or continuous aspect can be introduced by inflection and analytical tenses: (EN) *is doing*, *has done.* Starting, continuation, completion and perfective aspect can also be expressed by specific verbs modifying other verbs: (EN) *to start/continue/stop/complete the action.* Finally, in Slavic languages each verbal lexeme (i.e. independently of its inflected form), has inherent aspect, either perfective or imperfective, and is marked as a morphological feature (recognisable either by a prefix or by an ending): (PL) *robić* 'to do.IMPERF' vs. *z.robić* 'to PART.do.PERF'; *wy.łączać* 'to PART.connect.IMPERF' ⇒ 'to turn off' vs. *wy.łączyć* 'to PART.connect.PERF' ⇒ 'to turn off'. Therefore, in Slavic languages the verb in an LVC necessarily adds aspect to the predicate, so its status in Test 11 (§4.2.2) should be examined along slightly different lines than in Romance and Germanic languages. Additionally, if adding any aspectual semantics to the predicate should necessarily block the LVC classification in Test 11, then (EN) *to **take** a **decision*** should be annotated as an LVC, while (EN) ***taking** a **decision*** might not. These observations led us to revise the LVC tests for future editions of the guidelines.

Another finding concerns productivity. Some verbs admit arguments from large semantic classes, and, conversely, some nouns select various verbal operators. More precisely, we observed the hardness of delimiting productive from non-productive cases in VMWE categories: (i) whose semantic non-compositionality is weak, or (ii) whose components are not content words. The former mainly concerns LVCs. We found no effective and reproducible way to distinguish lexical selection from selection of large semantic classes. For instance, (EN) *to deliver* is often used with the class of nouns expressing formal speech acts such as *speech*, *lecture*, *verdict*, etc. However, we can also use the verb *to give* instead of *to deliver* with the same class of nouns, which likely shows a productive rather than a strict lexical selection. Problem (ii) concerns VPCs, IReflVs and prepositional verbs. Namely, as the semantics of particles is hard to establish, we could come up with only one VPC-related test (§4.2.5), which should clearly evolve in future work. Also, the ambiguity of various uses of the reflexive clitic, and the resulting hardness of the IReflV annotation, was stressed by many language teams. Finally, the non-compositionality of prepositional verbs was so hard to establish in the pilot annotation that we abandoned them in the final annotation.

We also underestimated the importance of modelling not only the semantic non-compositionality of idioms but their conventionalisation as well. As a result, we currently have no efficient way to distinguish MWEs from metaphors. The resemblance is strong since many idioms are metaphors, e.g. (PT) *ele **abre mão*** 'he opens hand' ⇒ 'he gives up', but non-idiomatic metaphors, created for the need of a particular text, do occur, e.g. (PL) *podpisanie tej umowy to stryczek założony na szyję Polski* 'signing this treaty is a noose put around Poland's neck'. The difference is hard to tackle, and especially to test, since it seems to lie precisely in the fact that MWEs are conventionalised while metaphors are not necessarily so. A partial solution to this problem may probably stem from statistical estimations, although the "long tail" of conventionalised and still infrequent MWEs may largely resemble non-conventionalised metaphors. We put forward the MWE vs. metaphor distinction as a future research issue.

9 Related work

In this section we contextualise our work with respect to existing MWE typologies, annotation methodologies and annotated corpora.

9.1 MWE typologies

In previous approaches to modelling MWEs, various classifications of MWEs were put forward. Here, we focus on several proposals, summarised in Table 7, which seem relevant to our work in that they: (i) have been particularly influential in the NLP community (Sag et al. 2002; Baldwin & Kim 2010; Mel'čuk 2010) (ii) were tested against a representative data set (Mel'čuk 2010), notably in corpus annotation (Schneider et al. 2014), (iii) use MWE flexibility, which is a pervasive feature of verbal MWEs, as a major classification criterion (Sag et al. 2002), (iv) focus exclusively on verbal MWEs (Sheinfux et al. forthcoming), (v) put a verbal component in the heart of the classification criterion (Laporte 2018).

Sag et al. (2002) is a highly influential seminal work whose MWE classification implements the hypothesis put forward by Nunberg et al. (1994) about the correlation between the semantic decomposability of an idiom and its syntactic flexibility. According to this theory, it is because *pull* can be rephrased as *use* and *strings* as *one's influence* that the idiom *to **pull strings*** admits variations like *to **pull** all the (political) **strings***, *the **strings** he **pulled***, etc. The hypothesis has been criticised, e.g. by Sheinfux et al. (forthcoming) and Laporte (2018), notably by demonstrating non-decomposable MWEs which still exhibit flexibility. The Sag et al. (2002) classification also calls for adjustments in inflectionally rich and free-word-order languages. Still, it remains widely used, notably due to its usefulness for NLP applications. Namely, MWE flexibility is a major obstacle in MWE identification since it prohibits seeing a MWE as a "word with spaces" and using sequence labelling approaches.

Baldwin & Kim (2010) assume the flexibility-driven classification by Sag et al. (2002) and they additionally introduce an orthogonal typology based on purely syntactic criteria, that is, on the syntactic structure of the MWE. There, verbal subcategories are both English-specific and non-exhaustive since verb-noun idioms are considered, but not, for example, verb-adjective ones.

The typology of Mel'čuk (2010) is based, conversely, on mainly semantic criteria. Different types of semantic compositionality are defined, and non-compositional subtypes are those where the semantic head is missing. The latter further subdivide into: (i) *quasi-locutions* in which the meanings of the components are combined, as in (FR) ***donner le sein*** 'to give the breast' ⇒ 'to breastfeed', (ii)

Table 7: Various MWE classifications compared.

Reference	Language	Scope	Classes	# classified expressions	Defining criteria
Sag et al. (2002)	EN	MWEs and collocations	I. Lexicalised: 1. Fixed (*by and large*); 2. Semi-fixed: non-decomposable idioms (*shoot the breeze* ‘chat’), compound nominals (*part of speech*), proper names (*San Francisco 49ers*); 3. Syntactically-flexible: VPCs (*break up*), decomposable idioms (*spill the beans*), LVCs (*make a decision*); II. Institutionalised (*traffic lights*)	unknown	lexicalisation, morphological and syntactic flexibility, semantic decomposability
Baldwin & Kim (2010)	EN	MWEs and collocations	I. Nominal (*golf club*, *connecting flight*); II. Verbal: 1. VPCs (*take off*, *cut short*, *let go*); 2. Prepositional verbs (*come accross*); 3. LVCs (*take a walk*); 4. Verb-noun idioms (*shoot the breeze*); III. Prepositional: 1. Determinerless prepositional phrases (*on top*, *by car*); 2. Complex prepositions (*on top of*, *in addition to*)	unknown	syntactic structure
Mel’čuk (2010)	FR	MWEs and collocations	I. Pragmatic (*emphasis mine*); II. Semantic: 1. Semantically compositional: clichés (*in other words*), collocations (*busy as a bee*, *award a prize*); 2. Semantically non-compositional: quasi-locutions ((FR) *donner le sein* ‘give the breast’ ⇒ ‘breastfeed’), 2. Semi-locutions ((FR) *fruits de mer* ‘sea fruit’ ⇒ ‘seafood’), 3. Complete locutions ((FR) *en tenue d’Adam et Eve* ‘in Adam’s and Eve’s dress’ ⇒ ‘naked’)	4,400 collocations, 3,200 locutions (Pausé 2017)	selection constraints, semantic non-compositionality
Schneider et al. (2014)	EN	all MWEs	I. Strong (*close call*); II. Weak (*narrow escape*)	3,500 occurrences	strength of association between words
Sheinfux et al. (forthcoming)	HE	verbal idioms	I. Transparent figurative (*saw logs* ‘snore’); II. Opaque figurative (*shoot the breeze* ‘chat’); III. Opaque non-figurative (*take umbrage* ‘feel offended’)	15 VMWEs, 400 occurrences	transparency, figuration
Laporte (2018)	FR	MWEs and collocations	I. Lexicalised: 1. MWEs without support verbs: verbal (*take stock*), nominal (*traffic lights*), adverbial (*for instance*); 2. Support-verb constr.: a. Vsup is not copula (*have an aim*, *get loose*), b. Vsup in copula (*be a genius*, *be angry*, *be on time*); II. Non-lexicalised (*salt and pepper*)	dozens of thousands of (lexicalised) MWEs	lexicalisation, presence of a support verb
This chapter	BG,CS,DE,EL, ES,FA,FR,HE, HU,IT,LT,MT, PL,PT,RO,SL, SV,TR	verbal MWEs	I. Universal: LVCs (*make a decision*), IDs (*spill the beans*); II. Quasi-universal: IReflVs ((FR) *s’avérer* ‘REFL’reveal’ ⇒ ‘prove (to be)’), VPCs (*take off*); III. OTH (*drink and drive*, *to voice act*)	62,000 occurrences	universalism, syntactic structure, lexical, syntactic and semantic idiosyncrasy

semi-locutions which include the meaning of only a part of their components, as in (FR) *fruits de mer* 'sea fruit' ⇒ 'seafood', (iii) *complete locutions*, which include the meaning of none of their components, as in (FR) *en tenue d'Adam et Eve* 'in Adam's and Eve's dress' ⇒ 'naked'.

Schneider et al. (2014) propose a rather shallow typology with only two types based on the strength of association between component words. Strong MWEs are those whose meaning is not readily predictable from component words, as in (EN) *close call* 'a situation in which something bad almost happened but could be avoided'. Weak MWEs are those with more transparent semantics and more flexibility, like (EN) *narrow escape* 'a situation in which something bad almost happened but could be avoided'. This typology was applied to annotate a large publicly available corpus, underlying the DiMSUM[30] shared task on identification of minimal semantic units and their supersenses.

In Sheinfux et al. (forthcoming) the hypothesis of Nunberg et al. (1994) is questioned on a sample of verbal Hebrew idioms, and a novel classification is put forward which relies on figuration (the degree to which the idiom can be assigned a literal meaning) and transparency (the relationship between the literal and idiomatic reading). In *transparent figurative* idioms the relationship between the literal and the idiomatic reading is easy to recover (*to* ***saw logs*** 'snore'). In *opaque figurative* idioms the literal picture is easy to imagine but its relationship to the idiomatic reading is unclear (*to* ***shoot the breeze*** 'chat'). Finally, in *opaque non-figurative* idioms no comprehensible literal meaning is available, notably due to cranberry words which have no status as individual lexical units (*to* ***take umbrage*** 'to feel offended'). The study further tests VMWEs of the 3 categories against 4 types of lexical and syntactic flexibility, and stresses the fact that flexibility is a matter of scale rather than a binary property.

Laporte (2018) formalises a MWE classification emerging from the lexicon-grammar theory and encoding practice (Gross 1986; 1994). Its specificity is to put the notion of support verb (roughly equivalent to light verb) in the heart of the classification, and push the MWE frontier far beyond what is admitted in other approaches. Namely, with the copula support verb *to be*, large classes of nouns, adjectives and PPs are seen as predicates of support-verb constructions, which should, thus, be lexically described.

Comparing our classification (§3) to the above ones (Table 7), several facts are striking: (i) we restrict ourselves to verbal MWEs only, (ii) we perform a large-scale multilingual evaluation and enhancement of the classification via corpus annotation in 18 languages, (iii) we assess semantic non-compositionality via

[30] https://dimsum16.github.io/

mostly syntactic tests, (iv) we define a novel VMWE category of IReflVs and linguistic tests delimiting its borders, we also display the quantitative importance of this category, mainly in Romance and Slavic languages, (v) we give access to detailed annotation guidelines organised as decision trees, with linguistic tests illustrated in many languages. As far as the scope of the MWE-related phenomena are concerned, recall that we exclude statistical collocations and retain only lexically, syntactically or semantically idiosyncratic expressions. This fact seemingly contrasts with other approaches shown in Table 7. Note, however, that some of these authors understand collocations differently, as discussed in §2.

9.2 MWE annotation practices

Modelling the behaviour of MWEs in annotated corpora, and prominently in treebanks, has been undertaken in various languages and linguistic frameworks. Rosén et al. (2015) offer a survey of MWE annotation in 17 treebanks for 15 languages, collaboratively documented according to common guidelines.[31] According to this survey, multiword named entities constitute by far the most frequently annotated category (Erjavec et al. 2010), sometimes with elaborate annotation schemes accounting for nesting and coordination (Savary et al. 2010). Continuous MWEs such as compound nouns, adverbs, prepositions and conjunctions are also covered in some corpora (Abeillé et al. 2003; Laporte et al. 2008; Branco et al. 2010). Verbal MWEs have been addressed for fewer languages. The survey also shows the heterogeneity of MWE annotation practices. For instance, VPCs are represented in dependency treebanks by dedicated relations between head verbs and particles. In constituency treebanks, particles constitute separate daughter nodes of sentential or verbal phrases and are assigned categories explicitly indicating their status of selected particles. Additionally, in an LFG (Lexical Functional Grammar) treebank, verbs and their particles are merged into single predicates appearing in functional structures.

Similar conclusions about the heterogeneity of MWE annotation were drawn concerning UD (McDonald et al. 2013), an initiative towards developing syntactically full-fledged and cross-linguistically consistent treebank annotation for many languages. Nivre & Vincze (2015) show that LVCs annotation in UD treebanks is threefold: (i) some treebanks lack or do not distinguish LVCs from regular verb-object pairs, (ii) some distinguish them by their structure (the direct object is dependent on the light verb rather than on the predicative noun), (iii) some account for them explicitly by the dependency labels between the noun

[31] http://clarino.uib.no/iness/page?page-id=MWEs_in_Parseme

and the verb. Furthermore, De Smedt et al. (2015) point out that 3 different dependency relations in UD[32] can be used to describe MWEs - `compound`, `mwe` and `name` (with possible sub-relations, e.g. `compound:prt` for verb-particle constructions) - and that these are used across different UD treebanks in a largely inconsistent way. More recent efforts (Adalı et al. 2016), while addressing VMWEs in a comprehensive way, still suffer from missing annotation standards.

As compared to this state of the art, the PARSEME effort aims at developing annotation guidelines and practices which would be universal but would leave room for language-dependent specificities. Our scope covers all types of VMWEs.

9.3 Corpora and datasets with VMWEs

As seen in the previous section, most efforts towards anotating MWEs were either language- or MWE category-specific. The same holds for verbal MWEs in particular. In this section we mention some outcomes of the previous VMWE annotation initiatives.

The Wiki50 (Vincze et al. 2011) corpus contains 50 English Wikipedia articles annotated for MWEs, including several VMWEs types. The dataset of Tu & Roth (2011) consists of 2,162 sentences from the British National Corpus in which verb-object pairs formed with *do*, *get*, *give*, *have*, *make*, and *take* are marked as positive and negative examples of LVCs. Tu & Roth (2012) built a crowdsourced corpus in which VPCs are manually distinguished from compositional verb-preposition combinations, again for six selected verbs. Baldwin (2005) presents another dataset of English VPCs. Finally, SZPFX (Vincze 2012) is an English-Hungarian parallel corpus with LVC annotations in both languages. For German, idiomatic combinations of verbs and prepositional phrases were described in a database by Krenn (2008) and annotated in the TIGER corpus by Brants et al. (2005).

In Slavic languages, a notable effort was made with the Prague Dependency Treebank of Czech (Hajič et al. 2017), annotated at 3 layers: morphological, analytical (accounting for syntax) and tectogrammatical (accounting for functional relations). MWEs, including some VMWEs, are annotated by identifying monosemic subtrees in the 3rd layer and replacing them by single nodes (Bejček & Straňák 2010), which unifies different morphosyntactic variants of the same MWE (Bejček et al. 2011). Each MWE occurrence is linked to its entry in an associated MWE lexicon. It is also argued that elements elided in MWEs (e.g. due to coordination) should be restored in deep syntactic trees. The Czech PARSEME corpus results from a mostly automatic (although challenging) transformation of the PDT annotations into the parseme-tsv format (Bejček et al. 2017).

[32] This analysis concerns UD v1 - these labels evolved in UD v2.

Kaalep & Muischnek (2006; 2008) and Vincze & Csirik (2010) present databases and corpora of VMWEs for Estonian particle verbs and Hungarian LVCs, respectively. VMWE annotations are available in several Turkish treebanks. In Eryiğit et al. (2015) various MWEs are labeled with a unique dependency label independently of their category, while in Adalı et al. (2016) they are classified as either strong or weak, similarly to Schneider et al. (2014). Finally, QasemiZadeh & Rahimi (2006) provide annotations for Farsi LVCs in the framework of the MULTEXT-East initiative, and in the Uppsala Persian Dependency Treebank (Seraji et al. 2014) the *lvc* dependency relationship is used for annotating non-verbal component of Farsi LVCs that are not in any other type of syntactic relationship.

The PARSEME corpus initiative builds upon these previous efforts by incorporating and extending some pre-existing datasets and annotation experiences. In some languages it is novel in that: (i) it constitutes the first attempt to annotate and analyse VMWEs in running text, e.g. in Greek and Maltese, (ii) it pays special attention, for the first time, to certain VMWE categories, e.g. to VPCs in Greek, to LVCs in Lithuanian, to IReflVs in most Slavic and Romance languages, and to distinguishing VMWEs from semi-copula-based expressions in Farsi (§4.3). But the most notable achievement going beyond the state of the art is to offer the first large highly multilingual VMWE corpus annotated according to unified guidelines and methodologies.

10 Conclusions and future work

We described the results of a considerable collective effort towards setting up a common framework for annotating VMWEs in 18 languages from 9 different language families. Unlike McDonald et al. (2013), our methodology is not English-centred. We draft the guidelines and test them on many languages in parallel, without giving priority to any of them (except for communication purposes). We offer a classification of VMWEs where properties hypothesised as universal or quasi-universal are treated in a homogeneous way, while leaving room to language-specific categories and features at the same time. Additionally to its importance for language modelling, and contrastive linguistic studies, this typology may be useful for various language technology tasks, notably because different VMWE types show different degrees of semantic decomposability, which influences their interpretation and translation. For instance, in LVCs nouns may translate literally and verbs may be omitted in the semantic calculus, but the same usually does not hold for IDs. Our annotation guidelines are organised in decision trees, so as to maximise the replicability of the annotators' decisions.

Our efforts also pave the way towards unified terminology and notation conventions. In particular, we stress the relations between words and tokens, which are crucial for defining the scope of the MWE phenomenon. We formalise the notion of a canonical form of a VMWE. Moreover, the notational conventions used in this volume for citing, glossing and translating multilingual examples of VMWEs largely result from our documentation work.

The PARSEME VMWE corpus[33] and its annotation guidelines,[34] both available under open licenses, are meant as dynamic resources, subject to continuous enhancements and updates. The size of the corpus is still modest for many languages and should be progressively increased. Adopting higher annotation standards, including a double annotation and adjudication, would lead to more reliable guidelines, increase the quality of the data, and strengthen our claims and findings. Since the publication of version 1.0 of the corpus, rich feedback was gathered from language teams, several dozens of issues were formulated and were discussed in a dedicated Gitlab space[35] and version 1.1[36] of the guidelines was elaborated. The most important evolutions include:

- Abandoning the category-neutral identification stage, since the annotation practice showed that VMWE identification is virtually always done in a category-specific way. The previous identification tests become ID-specific tests.

- Abandoning the OTH category due to its very restricted use. VMWEs classified previously as OTH now enter the ID category (except when the interpretation of the OTH category was language-specific).

- Introducing the multiverb construction (MVC) category to account for idiomatic serial verbs in Asian languages such as Hindi, Indonesian, Japanese and Chinese.

- Redesigning the tests and the decision trees for the LVC and VPC category, so as to increase the determinism in the annotation of these two categories.

- Introducing – optionally and experimentally – the category of inherently adpositional verbs (IAVs), roughly equivalent to the previously abandoned

[33] http://hdl.handle.net/11372/LRT-2282

[34] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/

[35] https://gitlab.com/parseme/sharedtask-guidelines/issues (restricted access, new users are welcome upon registration with the project leaders)

[36] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/

inherently prepositional verbs (IPrepVs). The IAV should be addressed in the post-annotation step, i.e. once the VMWEs of all other categories have been identified.

- Renaming the IReflV category by IRV, for an easier pronunciation.
- Renaming the ID category to VID (verbal idiom), to explicitly account for the verbal-only scope.

Adjustments of the previously annotated corpus to the guidelines version 1.1 are ongoing. The corpus should also significantly grow, as new portions of data are being annotated and new language teams (Arabic, Basque, Croatian, English and Hindi) are joining the project. Edition 1.1 of the PARSEME shared task (cf. Savary et al. 2017 for edition 1.0), based on the enhanced guidelines and corpus, is taking place as this volume is being edited.

In the long run, we intend to include other categories of MWEs (nominal, adjectival, adverbial, prepositional, named entities, etc.) under the annotation scope, as well as pave the way towards consistent representation and processing of both MWEs and syntax.

Acknowledgments

The work described in this chapter was supported by: (i) the IC1207 PARSEME COST action,[37]; (ii) national funded projects: LD-PARSEME[38] (LD14117) in the Czech Republic, PARSEME-FR[39] (ANR-14-CERA-0001) in France, and PASTOVU[40] in Lithuania; (iii) European Union’s Horizon 2020 research and innovation programme (Marie Skłodowska-Curie grant No 713567); (iv) Science Foundation Ireland in the ADAPT Centre[41] (Grant 13/RC/2106) at Dublin City University.

We are grateful to all language teams for their contributions to preparing the annotation guidelines and the annotated corpora. The full composition of the annotation team is the following.

Balto-Slavic languages:

- (BG) Ivelina Stoyanova (LGL, LL), Tsvetana Dimitrova, Svetla Koeva, Svetlozara Leseva, Valentina Stefanova, Maria Todorova;

[37] http://www.parseme.eu

[38] https://ufal.mff.cuni.cz/grants/ld-parseme

[39] http://parsemefr.lif.univ-mrs.fr/

[40] http://mwe.lt/en_US/

[41] www.adaptcentre.ie

- (CS) Eduard Bejček (LL), Zdeňka Urešová, Milena Hnátková;
- (LT) Jolanta Kovalevskaitė (LL), Loic Boizou, Erika Rimkutė, Ieva Bumbulienė;
- (SL) Simon Krek (LL), Polona Gantar, Taja Kuzman;
- (PL) Agata Savary (LL), Monika Czerepowicka.

Germanic languages:

- (DE) Fabienne Cap (LGL, LL), Glorianna Jagfeld, Agata Savary;
- (EN) Ismail El Maarouf (LL), Teresa Lynn, Michael Oakes, Jamie Findlay, John McCrae, Veronika Vincze;
- (SV) Fabienne Cap (LL), Joakim Nivre, Sara Stymne.

Romance languages:

- (ES) Carla Parra Escartín (LL), Cristina Aceta, Itziar Aduriz, Uxoa Iñurrieta, Carlos Herrero, Héctor Martínez Alonso, Belem Priego Sanchez;
- (FR) Marie Candito (LGL, LL), Matthieu Constant, Ismail El Maarouf, Carlos Ramisch (LGL), Caroline Pasquer, Yannick Parmentier, Jean-Yves Antoine;
- (IT) Johanna Monti (LL), Valeria Caruso, Manuela Cherchi, Anna De Santis, Maria Pia di Buono, Annalisa Raffone;
- (RO) Verginica Barbu Mititelu (LL), Monica-Mihaela Rizea, Mihaela Ionescu, Mihaela Onofrei;
- (PT) Silvio Ricardo Cordeiro (LL), Aline Villavicencio, Carlos Ramisch, Leonardo Zilio, Helena de Medeiros Caseli, Renata Ramisch;

Other languages:

- (EL) Voula Giouli (LGL,LL), Vassiliki Foufi, Aggeliki Fotopoulou, Sevi Louisou;
- (FA) Behrang QasemiZadeh (LL);
- (HE) Chaya Liebeskind (LL), Yaakov Ha-Cohen Kerner (LL), Hevi Elyovich, Ruth Malka;
- (HU) Veronika Vincze (LL), Katalin Simkó, Viktória Kovács;
- (MT) Lonneke van der Plas (LL), Luke Galea (LL), Greta Attard, Kirsty Azzopardi, Janice Bonnici, Jael Busuttil, Ray Fabri, Alison Farrugia, Sara Anne Galea, Albert Gatt, Anabelle Gatt, Amanda Muscat, Michael Spagnol, Nicole Tabone, Marc Tanti;
- (TR) Kübra Adalı (LL), Gülşen Eryiğit (LL), Tutkum Dinç, Ayşenur Miral, Mert Boz, Umut Sulubacak.

We also thank Mozhgan Neisani from University of Isfahan and Mojgan Seraji from the Uppsala Universitet for their contribution to the inter-annotator agreement calculation.

Abbreviations

FUT	future	MTW	multitoken word
GEN	genitive	MWT	multiword token
IAA	inter-annotator-agreement	NLP	natural language processing
ID	idiom	OTH	other VMWEs
IReflV	inherently reflexive verb	PART	particle
LGL	language group leader	REFL	reflexive clitic
LL	language leader	SG	singular
LV	light verb	UD	Universal Dependencies
LVC	light-verb construction	VID	verbal idiom
MAD	mean absolute deviation	VMWE	verbal multiword expression
MASC	masculine	VPC	verb-particle construction
MWE	multiword expression	1, 2, 3	first, second, third person

References

Abeillé, Anne, Lionel Clément & François Toussenel. 2003. Building a treebank for French. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 165–187. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Adalı, Kübra, Tutkum Dinç, Memduh Gokirmak & Gülşen Eryiğit. 2016. Comprehensive annotation of multiword expressions for Turkish. In *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*.

Al Saied, Hazem, Marie Candito & Matthieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 209–226. Berlin: Language Science Press. DOI:10.5281/zenodo.1469561

Baggio, Giosuè, Michiel van Lambalgen & Peter Hagoort. 2012. The processing consequences of compositionality. In *The Oxford handbook of compositionality*, 655–672. New York: Oxford University Press.

Baldwin, Timothy. 2005. Deep lexical acquisition of verb-particle constructions. *Computer Speech and Language* 19(4). 398–414. DOI:10.1016/j.csl.2005.02.004

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

Bauer, Laurie. 1983. *English word-formation* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press. https://books.google.pl/books?id=yGfUHs6FCvIC.

Bejček, Eduard, Jan Hajič, Pavel Straňák & Zděnska Urěnsová. 2017. Extracting verbal multiword data from rich treebank annotation. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories* (TLT 15), 13–24.

Bejček, Eduard & Pavel Straňák. 2010. Annotation of multiword expressions in the prague dependency treebank. *Language Resources and Evaluation* 44(1–2). 7–21.

Bejček, Eduard, Pavel Straňák & Daniel Zeman. 2011. Influence of treebank design on representation of multiword expressions. In Alexander F. Gelbukh (ed.), *Computational Linguistics and intelligent text processing - 12th International Conference,* (CICLing 2011), Tokyo, Japan, February 20-26, 2011. *Proceedings, Part I,* vol. 6608 (Lecture Notes in Computer Science), 1–14. Springer. DOI:10.1007/978-3-642-19400-9_1

Branco, António, Francisco Costa, João Silva, Sara Silveira, Sérgio Castro, Mariana Avelãs, Clara Pinto & João Graça. 2010. Developing a deep linguistic databank supporting a collection of treebanks: The CINTIL DeepGramBank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th conference on international language resources and evaluation* (LREC 2010). European Language Resources Association (ELRA).

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith & Hans Uszkoreit. 2005. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation* 2(4). 597–620. DOI:10.1007/s11168-004-7431-3

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), 4585–4592. European Language Resources Association (ELRA).

De Smedt, Koenraad, Victoria Rosén & Paul Meurer. 2015. *MWEs in universal dependency treebanks.* IC1207 COST PARSEME 5th general meeting. Iaşi, Romania. http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015.

Erjavec, Tomaz, Darja Fiser, Simon Krek & Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th International Conference on Language Resources and Evaluation* (LREC 2010), 1806–1809. European Language Resources Association (ELRA).

Eryiğit, Gülşen, Kübra Adali, Dilara Torunoğlu-Selamet, Umut Sulubacak & Tuğba Pamay. 2015. Annotation and extraction of multiword expressions in Turkish treebanks. In *Proceedings of the 11th Workshop on Multiword Expressions* (MWE '15), 70–76. Association for Computational Linguistics. http://www.aclweb.org/anthology/W15-0912.

Fort, Karën & Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the 4th Linguistic Annotation Workshop* (LAW IV '10), 56–63. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1868720.1868727.

Gross, Gaston. 1988. Degré de figement des noms composés. *Langages* 90. 57–71.

Gross, Maurice. 1986. Lexicon-grammar: The representation of compound words. In *Proceedings of the 11th coference on computational linguistics* (COLING '86), 1–6. Association for Computational Linguistics. DOI:10.3115/991365.991367

Gross, Maurice. 1994. The lexicon-grammar of a language: Application to French. In Ashley R. E. (ed.), *The encyclopedia of language and linguistics*, 2195–2205. Oxford: Oxford/NewYork/Seoul/Tokyo: Pergamon. https://hal-upec-upem.archives-ouvertes.fr/hal-00621380.

Hajič, Jan, Eva Hajičová, Marie Mikulová & Jiří Mírovský. 2017. Prague Dependency Treebank. In *Handbook on Linguistic Annotation* (Springer Handbooks), 555–594. Berlin, Germany: Springer Verlag.

Janssen, Theo M. V. 2001. Frege, contextuality and compositionality. *Journal of Logic, Language and Information* 10(1). 115–136. DOI:10.1023/A:1026542332224

Jespersen, Otto. 1965. *A Modern English grammar on historical principles, Part VI, Morphology*. London: Allen & Unwin.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2006. Multi-word verbs in a flective language: The case of Estonian. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Context* (MWE '06), 57–64. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W06/W06-2400.pdf.

Kaalep, Heiki-Jaan & Kadri Muischnek. 2008. Multi-word verbs of Estonian: A database and a corpus. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions* (MWE '08), 23–26. Association for Computational Linguistics. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W20_Proceedings.pdf.

Kim, Su Nam. 2008. *Statistical modeling of multiword expressions*. Melbourne: University of Melbourne dissertation.

Koeva, Svetla, Ivelina Stoyanova, Maria Todorova & Svetlozara Leseva. 2016. Semi-automatic compilation of the dictionary of Bulgarian multiword expressions. In *Proceedings of the Workshop on Lexicographic Resources for Human Language Technology* (GLOBALEX 2016), 86–95.

Kracht, Marcus. 2007. Compositionality: The very idea. *Research on Language and Computation* 5(3). 287–308. DOI:10.1007/s11168-007-9031-5

Krenn, Brigitte. 2008. Description of evaluation resource – German PP-verb data. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions* (MWE '08), 7–10. Association for Computational Linguistics.

Laporte, Éric. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective* (Phraseology and Multiword Expressions). Language Science Press.

Laporte, Éric, Takuya Nakamura & Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Proceedings of the 2nd Linguistic Annotation Workshop*, 48–51. https://halshs.archives-ouvertes.fr/halshs-00286541.

Lipka, Leonhard, Susanne Handl & Wolfgang Falkner. 2004. Lexicalization & institutionalization. The state of the art in 2004. *SKASE Journal of Theoretical Linguistics* 1(1). 2–19. http://www.skase.sk/Volumes/JTL01/lipka.pdf.

Maldonado, Alfredo & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–175. Berlin: Language Science Press. DOI:10.5281/zenodo.1469557

Marcus, Mitchell P., Mary Ann Marcinkiewicz & Beatrice Santorini. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics* 19(2). 313–330. http://dl.acm.org/citation.cfm?id=972470.972475.

Mathet, Yann, Antoine Widlöcher & Jean-Philippe Métivier. 2015. The unified and holistic method Gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics* 41(3). 437–479. DOI:10.1162/COLI_a_00227

McDonald, Ryan, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló & Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 2: Short papers)*, 92–97. Association for Computational Linguistics. http://www.aclweb.org/anthology/P13-2017.

Mel'čuk, Igor A. 2010. La phraséologie en langue, en dictionnaire et en TALN. In *Actes de la 17ème conférence sur le traitement automatique des langues naturelles 2010*.

Moreau, Erwan, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel & Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 177–207. Berlin: Language Science Press. DOI:10.5281/zenodo.1469559

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Nivre, Joakim & Veronika Vincze. 2015. *Light verb constructions in universal dependencies*. IC1207 COST PARSEME 5th general meeting. Iaşi, Romania. http://typo.uni-konstanz.de/parseme/index.php/2-general/138-admitted-posters-iasi-23-24-september-2015.

Nunberg, Geoffrey, Ivan A. Sag & Thomas Wasow. 1994. Idioms. *Language* 70(3). 491–538.

Pagin, Peter & Dag Westerståhl. 2001. Compositionality II: Arguments and problems. *Philosophy Compass* 5. 250–264. DOI:10.1111/j.1747-9991.2009.00228.x

Partee, Barbara H., Alice ter Meulen & Robert E. Wall. 1990. *Mathematical methods in linguistics* (Studies in Linguistics and Philosophy 30). Dordrecht: Kluwer.

Pasquer, Caroline. 2017. Expressions polylexicales verbales : étude de la variabilité en corpus. In *Actes de la 18e rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues* (TALN-RÉCITAL 2017).

Pausé, Marie-Sophie. 2017. *Structure lexico-sentaxique des locutions du français et incidence sur leur combinatoire*. Nancy, France: Université de Lorraine dissertation.

QasemiZadeh, Behrang & Saeed Rahimi. 2006. Persian in MULTEXT-East framework. In *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006, Turku, Finland, August 23-25, 2006, Proceedings*, 541–551. DOI:10.1007/11816508_54

Rosén, Victoria, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova & Verginica Barbu Mitetelu. 2015. A survey of multiword expressions in treebanks. In *Proceedings of the 14th International Workshop on Treebanks & Linguistic Theories Cconference*. https://hal.archives-ouvertes.fr/hal-01226001.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Savary, Agata, Jakub Waszczuk & Adam Przepiórkowski. 2010. Towards the annotation of named entities in the National Corpus of Polish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th conference on international language resources and evaluation* (LREC 2010). European Language Resources Association (ELRA).

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Nora Emily Danchik, Michael T. Mordowanec, Henrietta Conrad & Smith Noah A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evalua-*

tion (LREC 2014), 455–461. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf.

Seraji, Mojgan, Carina Jahani, Beáta Megyesi & Joakim Nivre. 2014. A Persian treebank with Stanford typed dependencies. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014). European Language Resources Association (ELRA).

Sheinfux, Livnat Herzig, Tali Arad Greshler, Nurit Melnik & Shuly Wintner. Forthcoming. Verbal MWEs: Idiomaticity and flexibility. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions* (Phraseology and Multiword Expressions), 5–38. Berlin: Language Science Press.

Siemieniec-Gołaś, Ewa. 2010. On some Turkish auxiliary verbs in giovanni molino's dittionario della lingua italiana, turchesca (1641). *Studia Linguistica Universitatis Iagellonicae Cracoviensis* 127(1). 57–77.

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2018. Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 227–243. Berlin: Language Science Press. DOI:10.5281/zenodo.1469563

Szarvas, György, Veronika Vincze, Richárd Farkas, György Móra & Iryna Gurevych. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics – Special Issue on Modality and Negation* 38(2). 335–367.

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 299–317. Berlin: Language Science Press. DOI:10.5281/zenodo.1469569

Tu, Yuancheng & Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE '11), 31–39. Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0807.

Tu, Yuancheng & Dan Roth. 2012. Sorting out the most confusing English phrasal verbs. In *Proceedings of the First Joint Conference on Lexical and Computational*

Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval '12), 65–69. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2387636.2387648.

van Gompel, Maarten & Martin Reynaert. 2013. FoLia: A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal* 3. 63–81.

van Gompel, Maarten, Ko van der Sloot, Martin Reynaert & Antal van den Bosch. 2017. FoLiA in practice: The infrastructure of a linguistic annotation format. In, (To appear). Ubiquity Press.

Vincze, Veronika. 2012. Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC-2012), 2381–2388. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/177_Paper.pdf.

Vincze, Veronika & János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING '10), 1110–1118. Association for Computational Linguistics. http://www.aclweb.org/anthology/C10-1125.

Vincze, Veronika, István Nagy T. & Gábor Berend. 2011. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 289–295. RANLP 2011 Organising Committee. http://aclweb.org/anthology/R11-1040.

Chapter 5

Analysis and Insights from the PARSEME Shared Task dataset

Alfredo Maldonado
ADAPT Centre, Trinity College Dublin

Behrang QasemiZadeh
University of Düsseldorf

The PARSEME Shared Task on the automatic identification of verbal multiword expressions (VMWEs) was the first collaborative study on the subject to cover a wide and diverse range of languages. One observation that emerged from the official results is that participating systems performed similarly on each language but differently across languages. That is, intra-language evaluation scores are relatively similar whereas inter-language scores are quite different. We hypothesise that this pattern cannot be attributed solely to the intrinsic linguistic properties in each language corpus, but also to more practical aspects such as the evaluation framework, characteristics of the test and training sets as well as metrics used for measuring performance. This chapter takes a close look at the shared task dataset and the systems' output to explain this pattern. In this process, we produce evaluation results for the systems on VMWEs that only appear in the test set and contrast them with the official evaluation results, which include VMWEs that also occur in the training set. Additionally, we conduct an analysis aimed at estimating the relative difficulty of VMWE detection for each language. This analysis consists of a) assessing the impact on performance of the ability, or lack-thereof, of systems to handle discontinuous and overlapped VMWEs, b) measuring the relative sparsity of sentences with at least one VMWE, and c) interpreting the performance of each system with respect to two baseline systems: a system that simply tags every verb as a VMWE, and a dictionary lookup system. Based on our data analysis, we assess the suitability of the official evaluation methods, specifically the token-based method, and propose to use Cohen's kappa score as an additional evaluation method.

Alfredo Maldonado & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–176. Berlin: Language Science Press. DOI:10.5281/zenodo.1469557

1 Introduction

Multiword expressions (MWEs) have been studied extensively due to the fact that many natural language processing (NLP) pipelines depend on their correct identification and processing (Sag et al. 2002). However, there has been relatively little work on *Verbal* MWEs (VMWEs). The PARSEME[1] Shared Task on VMWEs (Savary et al. 2017) was the first initiative focusing on the problem of identifying VMWEs for a relatively large number of languages, 18 in total. This initiative produced an array of annotated training and test sets for each language. Using these training sets, shared task participants developed and trained VMWE-identification systems, which were then evaluated on separate test sets also produced by PARSEME.

Several patterns have emerged from the evaluation results in this pioneering shared task. One is that individual systems tend to perform very differently across languages (inter-language performance) and yet different systems performed similarly in most languages (intra-language performance). In particular, participating systems scored highest on Farsi, Romanian, Czech and Polish, and lowest on Swedish, Hebrew, Lithuanian and Maltese, whilst ranging somewhere in between for the rest of the languages. It has been observed that the inter-language performance is positively correlated with the proportion of VMWEs shared by the training and test sets in each language (Maldonado et al. 2017). This observation suggests that the reported systems' performance and ranking could potentially be dependent on the proportion of shared VMWEs across languages. At the very least, it is clear that inter-language performance differences cannot be attributed to linguistic differences among languages alone, but to particularities of the dataset that interplay with these linguistic differences.

This chapter conducts a detailed data analysis of the PARSEME dataset and the official systems' submissions in order to try to understand how these particularities impact systems' performance and to propose possible modifications to the dataset in order to balance out said particularities among the language corpora.

To this end, we start our discussion in §2 by computing statistics for each language to get a sense of their differences. We then measure the relative difficulty in identifying VMWEs in each language corpus by focusing on three factors that could potentially pose challenges to the systems: 1) the relative sparsity of VMWEs in each language corpus (by measuring the proportion of sentences with and without VMWEs); 2) the prevalence and significance of discontinuous VMWEs and embedded (or overlapped) VMWEs; and 3) corpus similarity and

[1] http://parseme.eu

homogeneity measures between the training and test portions for each language section. We observe that the importance of these factors varies across languages: while some are inherent to each language's linguistic properties (e.g., proportion of continuous vs discontinuous VMWEs or the dominant category of VMWEs in a language), others (e.g., relative sparsity of VMWEs) can be controlled by altering the size of the training and test sets, the proportion of shared VMWEs between these two sets, and, in general, the homogeneity of the distribution of VMWEs in these sets for each of the languages.

We then turn our attention to the shared task official evaluation scores on the participating systems in §3 and §4. In §3, we focus on the effect of the proportion of shared VMWEs between the training and test sets in each language corpus. We evaluate the systems on shared VMWEs and on VMWEs occurring exclusively in the test set. We also introduce two baseline systems (a system that simply tags every verb as a VMWE and a simple dictionary look-up system) and observe that the performance of the participating systems follows trends that the performance of these baselines shows.

In §4, we concentrate on the evaluation metrics used in the shared task: one that measures the ability of retrieving full VMWEs (MWE-based evaluation) and another that gives credit to systems on partially identified VMWEs (Token-based evaluation). We observe that the Token-based evaluation measure gives more weight to long VMWEs and, in addition, can be exploited by a system that simply detects verbs. Lastly, we use Cohen's κ inter-annotator agreement measure as an evaluation metric based on the intuition that it provides a 'chance-corrected' degree of similarity between a system output and a gold standard.

In §5, we conclude that the PARSEME VMWE dataset is a valuable resource for evaluating VMWE identification systems as long as certain variables are controlled for and purpose-specific evaluation frameworks are considered. We also propose avenues for future work.

Before we delve into the analysis and discussion, it should be mentioned that systems were considered to be participating in one of two separate tracks under the original shared task rules: a) an open track in which participants were free to use any external data (other than the training data provided) to train and develop their systems, and b) a closed track, where participants were allowed to use the provided training data only. Given that only one system (LATL) participated in the open track and for only one language (French), this chapter completely ignores the open/closed distinction and compares all systems on the same evaluation scores.

2 Shared task dataset

This section explores several numerical properties of the dataset developed for the shared task in order to gain an insight into differences among languages and to identify potential *difficulty factors* in the corpora. We consider difficulty factors to be corpus-specific characteristics (such as corpus size, sparsity of VMWEs or corpus heterogeneity) that could potentially hinder an algorithm's ability to identify VMWEs. We assess a factor's degree of difficulty by observing the overall systems' performance on languages that present the factor in question, in comparison to languages that do not present that factor. The performance of the systems is measured by the official shared task evaluation F1 scores, shown in Table 1. That table also contains the averages all systems' scores for a given language (*avg* column) and the ranks of the languages according to these averages (*rnk* column). Recall that two evaluation modalities were measured in the shared task: MWE-based evaluation, which counts as a success the matching of a full VMWE, and Token-based evaluation, which gives partial credit to partially matched VMWEs. Figure 1 summarises these scores per language as box plots.

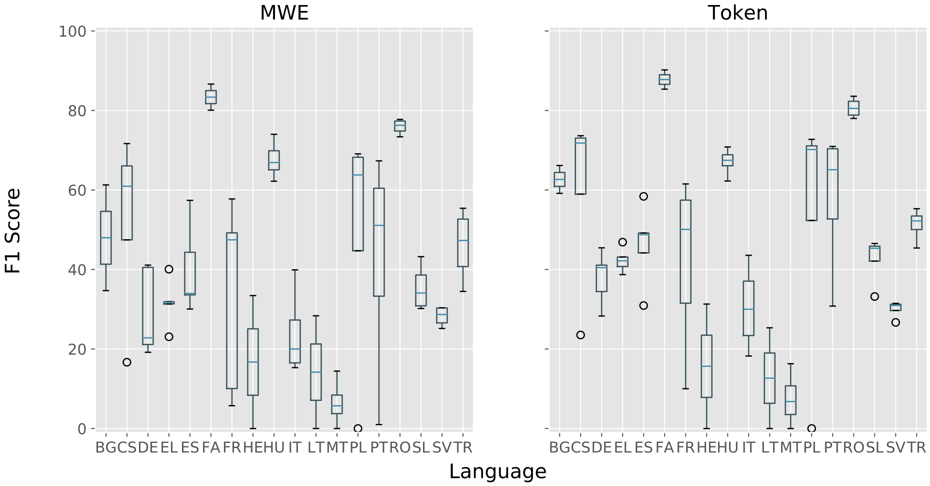


Figure 1: Box plots summarising F1 scores achieved by all systems on each language, using the MWE-based and Token-based evaluation modalities

Table 1: F1 evaluation scores by language and system with averages (avg), rank (rnk) in Token-based, MWE-based and Cohen's κ evaluations. Baselines: dictionary look-up (BD) and verb detection (BV).

		ADAPT	LATL	LIF	MUMULS	RACAI	SZEGED	TRANSITION	*avg*	*rnk*	BD	BV
BG	**Token-based**				59.16			66.15	*62.66*	*4*	47.44	
	MWE-based				34.68			61.27	*47.98*	*6*	34.67	
	Cohen's κ				21.36			53.57	37.47	6	21.27	
CS	**Token-based**	72.86			23.52	70.76		73.65	*60.20*	*5*	64.34	20.41
	MWE-based	57.72			16.67	64.18		71.67	*52.56*	*4*	51.91	0
	Cohen's κ	46.49			8.82	55.04		64.36	43.68	5	37.87	-18.54
DE	**Token-based**	40.48			34.45	28.30	45.45	41.09	*37.95*	*13*	40.70	28.52
	MWE-based	22.80			21.14	19.17	40.53	41.10	*28.95*	*13*	41.34	9.22
	Cohen's κ	5.86			5.01	5.82	24.97	26.44	13.62	16	26.42	-15.49
EL	**Token-based**	43.14			42.17	38.71	40.75	46.88	*42.33*	*12*	34.16	9.14
	MWE-based	31.34			23.08	31.74	31.88	40.07	*31.62*	*12*	21.81	0.02
	Cohen's κ	23.28			12.46	25.2	22.9	31.57	23.08	11	9.46	-7.41
ES	**Token-based**	49.17			48.75	30.93	44.18	58.39	*46.28*	*9*	50.97	15.56
	MWE-based	44.33			33.62	30.06	33.99	57.39	*39.88*	*9*	44.22	0
	Cohen's κ	35.84			21.18	23.41	17.81	48.83	30.91	9	32.7	-13.41
FA	**Token-based**	85.36						90.20	*87.78*	*1*	65.75	47.73
	MWE-based	80.08						86.64	*83.36*	*1*	55.92	0
	Cohen's κ	63.13						74.77	68.95	2	22.71	-50.01
FR	**Token-based**	61.52	54.61	10.00	29.40	50.09	33.64	60.28	*42.79*	*10*	45.73	18.28
	MWE-based	50.88	47.46	10.82	9.29	47.55	5.73	57.74	*32.78*	*11*	38.42	0.21
	Cohen's κ	40.12	33.77	7.98	-4.75	38.74	-14.42	48.98	21.49	13	24.19	-15.99
HE	**Token-based**				0.00			31.30	*15.65*	*16*	33.80	
	MWE-based				0.00			33.44	*16.72*	*16*	37.44	
	Cohen's κ				0.00			27.74	13.87	14	32.69	
HU	**Token-based**	66.10			68.86	62.26	70.81	67.47	*67.10*	*3*	68.13	12.49
	MWE-based	66.89			62.21	65.08	74.01	69.87	*67.61*	*3*	68.09	2.44
	Cohen's κ	50.6			42.13	49.45	60.04	52.03	50.85	3	49.01	-35.81
IT	**Token-based**	25.11				18.24	34.90	43.57	*30.46*	*14*	37.85	14.4
	MWE-based	23.09				16.90	15.31	39.90	*23.80*	*15*	29.03	0
	Cohen's κ	14.26				10.01	-10.01	25.33	9.9	17	8.27	-14.44
LT	**Token-based**				0.00			25.33	*12.67*	*17*	28.85	
	MWE-based				0.00			28.35	*14.18*	*17*	30.08	
	Cohen's κ				0.00			27.25	13.62	15	28.82	
MT	**Token-based**	8.87			0.00	4.69		16.29	*7.46*	*18*	11.42	6.79
	MWE-based	6.41			0.00	5.00		14.44	*6.46*	*18*	6.75	0.02
	Cohen's κ	3.5			0.00	2.99		6.6	3.27	18	-5.25	-5.74
PL	**Token-based**	72.74			69.77		0.00	70.56	*53.27*	*7*	74.40	18.33
	MWE-based	67.95			59.61		0.00	69.09	*49.16*	*5*	69.98	0
	Cohen's κ	61.53			51.33		0.00	62.72	43.9	4	63.46	-15.01
PT	**Token-based**	70.18			60.01		30.79	70.94	*57.98*	*6*	59.97	14.32
	MWE-based	58.14			44.05		0.99	67.33	*42.63*	*8*	54.49	0
	Cohen's κ	51.35			35.98		-11.52	62.03	34.46	7	46.35	-11.86
RO	**Token-based**	81.90			83.58	77.99		79.12	*80.65*	*2*	63.76	11.51
	MWE-based	73.38			77.21	77.75		75.31	*75.91*	*2*	57.74	0
	Cohen's κ	71.28			75.35	76.12		73.18	73.98	1	53.75	-7.32
SL	**Token-based**	45.06			45.62	33.20		46.55	*42.61*	*11*	28.47	0.08
	MWE-based	37.08			31.08	30.19		43.22	*35.39*	*10*	21.65	0
	Cohen's κ	29			20.49	23.45		33.17	26.53	10	5.23	-0.07
SV	**Token-based**	31.49				26.69	31.19	30.70	*30.02*	*15*	8.94	13.23
	MWE-based	30.32				25.17	27.03	30.36	*28.22*	*14*	7.32	0
	Cohen's κ	24.44				20.78	16.56	24.75	21.63	12	-5.62	-10.29
TR	**Token-based**	52.85			45.40	51.59		55.28	*51.28*	*8*	16.60	10.45
	MWE-based	42.83			34.49	51.76		55.40	*46.12*	*7*	5.95	0
	Cohen's κ	25.88			19.05	38.88		42.14	31.49	8	-8.57	-17.81
avg	***Token-based***	*53.79*	*54.61*	*10.00*	*40.71*	*41.12*	*36.86*	*54.10*			*43.40*	*16.08*
	MWE-based	*46.22*	*47.46*	*10.82*	*29.81*	*38.71*	*25.50*	*52.37*			*37.60*	*0.79*
	Cohen's κ	*36.44*	*33.77*	*7.98*	*20.56*	*30.82*	*11.82*	43.64			24.6	-16.92

2.1 Corpora sizes, VMWE sparsity and frequency distributions

We start by discussing the sizes of the training and test portions in each language corpus, depicted in Figure 2. Sizes are measured in terms of the total number of sentences. Traditionally, corpora sizes are discussed in terms of number of words, rather than number of sentences. We use number of sentences instead for a variety of reasons: 1) Each language corpus in the dataset consists of a collection of individual sentences. So the sentence is a natural unit to describe the dataset. 2) A sentence is expected to have a single main verb. On average, we can expect to have a little more than one verb per sentence. However, we would like to know what this average is for the case of *verbal* MWEs (VMWEs). That is, we would like to know how sparse VMWEs are in a given language corpus, and what impact this sparsity may have. 3) Measures such as the rate of VMWEs per *n* tokens could also be used, but are less linguistically motivated. Finally, 4) the training-to-test size ratios in terms of number of words are largely the same in this dataset as in terms of number of sentences.

Notice that Romanian and Czech have by far the largest training sets, dwarfing corpora of all other languages. This seems to work in favour of these two languages as, on average, Romanian ranked 2nd place in both evaluation modalities and Czech ranked at 4th and 5th places in the MWE-based and Token-based modalities, respectively. Swedish is the language with the smallest training set (only 200 sentences). The average F1 score of systems participating in Swedish is around 30% for both evaluation modalities. Indeed, the size of the training set is somewhat positively correlated with the average system evaluation scores for each language. The Pearson correlation coefficients for MWE-based and Token-based evaluations are 0.33 and 0.35, respectively.

The size of the test set relative to its corresponding training set varies widely across languages. The test-to-training proportions vary from 8% to 60% for most languages, except for Maltese (79%), Spanish (85%) and most notably, Swedish, with a test set about 8 times larger than its training set.[2] Although both Maltese and Swedish performed rather poorly (Maltese actually ranked last), there is no clear pattern between the test-to-training proportion of a language corpus and the performance of systems. In fact, Spanish ranked exactly in the middle at 9th place. These proportions were found to be mildly negatively correlated against MWE-based and Token-based evaluations: -0.20 and -0.23, respectively (Pearson correlation coefficients).

[2] 200 training sentences vs. 1600 test sentences, making the proportion of the training set almost invisible in Figure 2

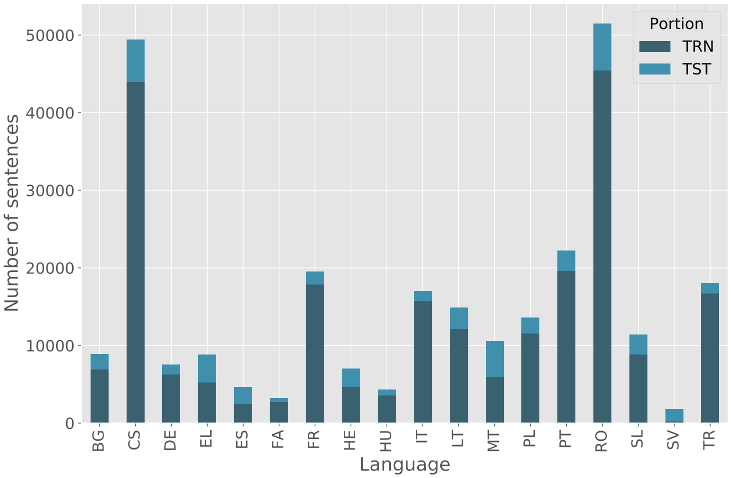


Figure 2: Relative sizes (in sentences) of the training and test portions of each language corpus.

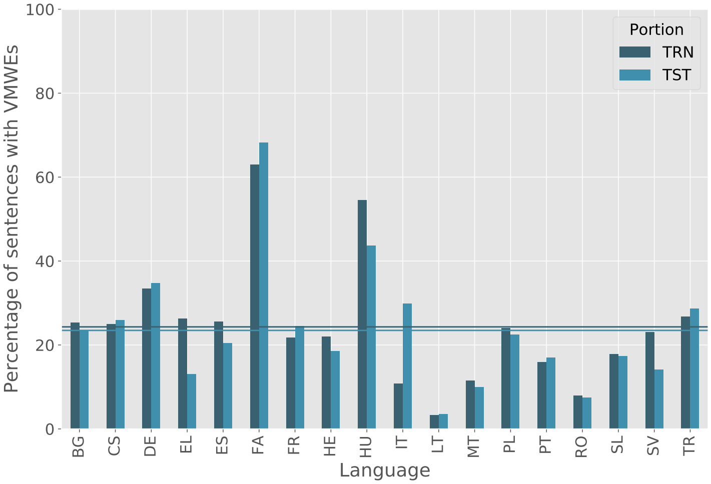


Figure 3: VMWE Sparsity – Percentage of sentences with VMWEs; horizontal lines depict average percentages across languages for training (TRN) and test (TST) sets, respectively.

Figure 3 shows how sparse VMWEs are in the language corpora. VMWE sparsity can be understood as the inverse of the proportion of sentences that have at least one VMWE. The figure shows the proportion of VMWEs within each set (training and test) using percentages. The graphs show that language corpora differ widely in their VWME sparsity. The overall proportion average (depicted by the two horizontal lines in the figure) is 24% and 23% for the training and test sets, respectively. Only Farsi and Hungarian are well above this average, and German is slightly above. For most languages, the vast majority of sentences do not contain a single VMWE. Whilst sentences without VMWE examples are indeed needed by machine learning algorithms, too few examples could hinder learning processes due to class imbalance. Indeed, there is a strong positive correlation between the proportion of sentences with VMWEs and the average system evaluation scores: 0.58 Pearson correlation coefficient against MWE-based evaluation and 0.56 against Token-based evaluation. Lithuanian and Maltese are the two lowest scoring languages in both evaluation modalities (see Table 1 and Figure 1). They are two of the three languages with the highest VMWE sparsity. The third language is Romanian, which turns out to be the second highest scoring language. Romanian is, as previously mentioned, the language with the largest amount of training data. The Romanian corpus' large volume seems to outweigh its high VMWE sparsity in systems' performance.

Another feature which seems to help systems perform well in the Romanian corpus is the frequency distribution of its VMWEs, as shown in Figure 4. This figure shows how many VMWE types occur at each VMWE frequency and how many of those VMWEs are successfully retrieved by the systems on the test portion of each language corpus. The grey bars on each chart show the total number of VMWE types occurring at each frequency inscribed on the *x* axis. The coloured bars count the number of VMWE types at each frequency that were fully detected by each system. This figure shows that Romanian VMWEs are *well distributed*: whilst Romanian hapax legomena (VMWEs occurring only once) dominate with 208 instances, there are many VMWEs with higher frequencies. The total number of VMWEs that occur more than once is 292, with frequencies up to 31 well represented. By contrast, 88 Lithuanian VMWEs appear only once and the rest, 12 of them, just twice! For Maltese, 82.57% of its VMWEs are hapax legomena. The remaining 17.43% have frequencies between 2 and 9. In short, VMWEs in the Lithuanian and Maltese corpora are not as well distributed by frequency as those in the Romanian corpus. The less frequent a VMWE is, the less opportunities a system has to learn it. So if the majority of VMWEs in a corpus are of low frequency (as in Lithuanian and Maltese), it will be harder for a system to learn them, which will lead to potentially low performance scores for the system.

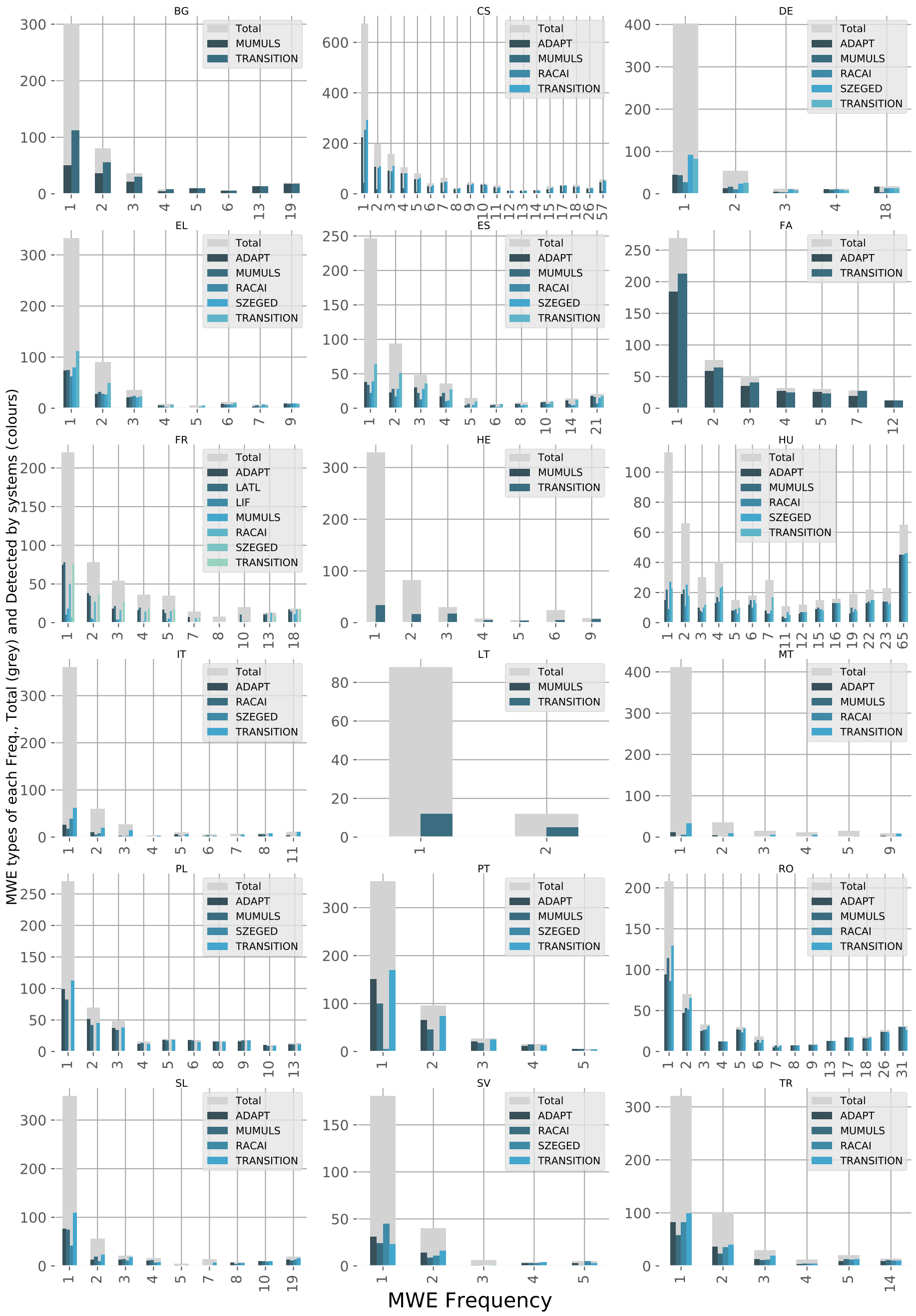


Figure 4: Distribution of VMWEs of different frequencies on the test set (grey bars) and the proportion of such VMWEs detected by systems (coloured bars) based on full MWE-based detection.

As an aside, the grey bars in Figure 4 show, for most languages, that the majority of VMWEs are hapax legomena and that the number of VMWEs occurring more frequently decreases dramatically as their frequency increases. This is the hallmark of the Zipfian distribution, which is something to be expected with lexical phenomena (Manning & Schütze 1999: pp. 22–6). This is not the usual way in which this distribution is traditionally plotted from data. However, it can be seen that most charts follow it approximately.

The issue of *frequency distribution* is important. Hungarian and Spanish are modest in size in comparison with Lithuanian and Maltese (see Figure 2), and yet the systems perform better in the former languages (especially in Hungarian) than in the latter languages. Figure 4 reveals that both Hungarian and Spanish are well distributed by frequency. Hungarian, despite having a smaller test set, is in fact even better distributed by frequency and has a lower VMWE sparsity (Figure 3) than Spanish. It obtains a 67 average F1 score whereas Spanish gets an F1 score average of 40–46, in both evaluation modalities (see *avg* column in Table 1).

From these observations, we can point out that language corpora with small amounts of training data, especially when combined with high VMWE sparsity and a poor frequency distribution, tend to obtain low scores in most systems. So increasing the size of training and test data is definitely a recommendation to follow. VMWE sparsity can be reduced by simply trying to balance out sentences with VMWEs against sentences without VMWEs. However, corpus designers should be cautious of doing this, as it could lead to a corpus that does not reflect the real distribution of VMWEs in the language and/or domain in question. Perhaps, it should be the task of system developers to design systems capable of coping with the natural VMWE imbalance/sparsity in a language corpus.[3] Improving the VMWE frequency distribution in language corpora could also help systems. Ensuring that several examples of each VMWE type are included in the training data will be a challenge, however, due to the natural Zipfian tendency of a majority of VMWEs to appear only once in any given corpus. We propose offsetting this tendency by aiming to compile a corpus where the total frequency of VMWE types that occur *frequently enough* outnumber the total frequency of VMWE types that occur *less frequently*. That is, if θ is the minimum frequency a VMWE needs to have in order to be considered to have *enough frequency*,[4] then we could ensure that the language corpus satisfies the condition:

[3] Systems could, for example, run a classifier to distinguish sentences that contain VMWEs from sentences that do not, and train/run their VMWE extractors only on sentences that do.

[4] θ, a minimum desirable frequency, is a parameter to be set empirically, with $\theta = 2$ a reasonable default value.

(1) $$\sum_{v_i \in \{f(v_j) \geq \theta\}} f(v_i) > \sum_{v_k \in \{f(v_j) < \theta\}} f(v_k)$$

where $f(v)$ is the frequency of VMWE v in the corpus. Note that a corpus with a good VMWE frequency distribution cannot be created by simply increasing the size of the corpus, but by better selecting sentences that are good examples of as many VMWEs as possible.

2.2 VMWEs shared between the training and test sets

Maldonado et al. (2017) noticed that the proportion of VMWEs shared between the training set and the test set of a language corpus is strongly positively correlated with the performance scores achieved by participating systems on that language test set (see also Savary et al. 2018 [this volume] §6.3). The most likely explanation is that when evaluated on the test set, machine learning systems would tend to perform better on VMWE examples they encountered in the training set (i.e. exact VMWEs that systems have already *seen* during training) than on VMWE examples that systems encounter for the first time in testing. The higher the proportion of shared/seen VMWEs is in one language, the higher a machine learning system can be expected to perform on that language. Figure 5 depicts this relationship by plotting the score achieved by each system on each language against the proportion of shared/seen VMWEs in that language. The languages on the x axis are sorted and labelled by this proportion. Notice the near-linear relationship between this proportion and the system scores.

It is of interest to evaluate systems on non-shared/unseen VMWEs only. This can be done by using the official systems' outputs, which were kindly provided to us by the shared task organisers. In order to evaluate unseen VMWEs only, the labels for seen VMWEs in the systems' outputs and the gold standards were cleared (i.e. changed to the underscore '_' flag) so that they would be ignored by the official evaluation scripts. Figure 6 shows the systems' performance scores when evaluated in this manner on unseen VMWEs only. Notice that the x axis was kept from Figure 5 to enable an easy visual comparison between both figures.

The first thing to notice is that all systems' scores go down dramatically for all languages. Notice however that for Farsi, the TRANSITION and ADAPT scores do not fall as dramatically as in the other languages. At first glance, this can be associated with the density of annotated instances of VMWEs in the Farsi corpus, i.e., Farsi has the lowest VMWE sparsity in the dataset (as discussed in §2.1). On the other hand, the second least VMWE-sparse language, Hungarian,

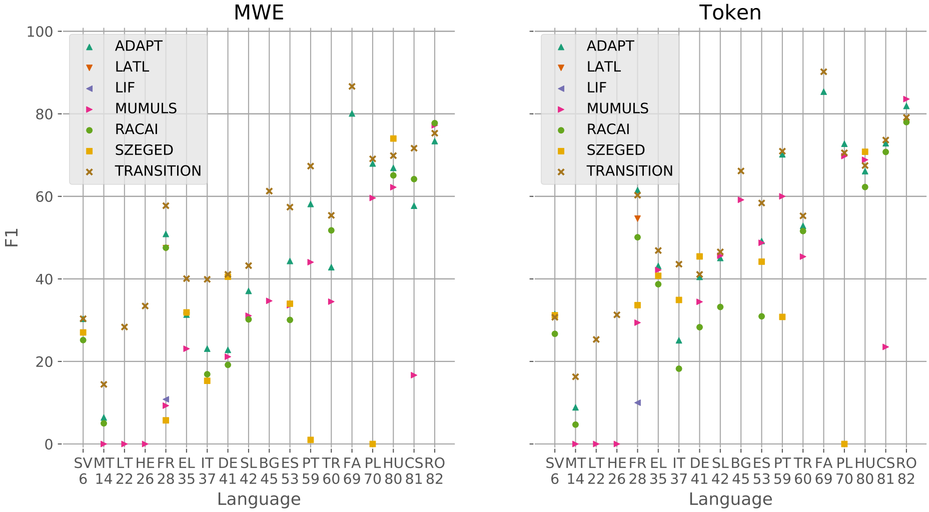


Figure 5: System evaluation scores (MWE-based, left; Token-based, right) for each language against the proportion (percentage) of test VMWEs seen during training

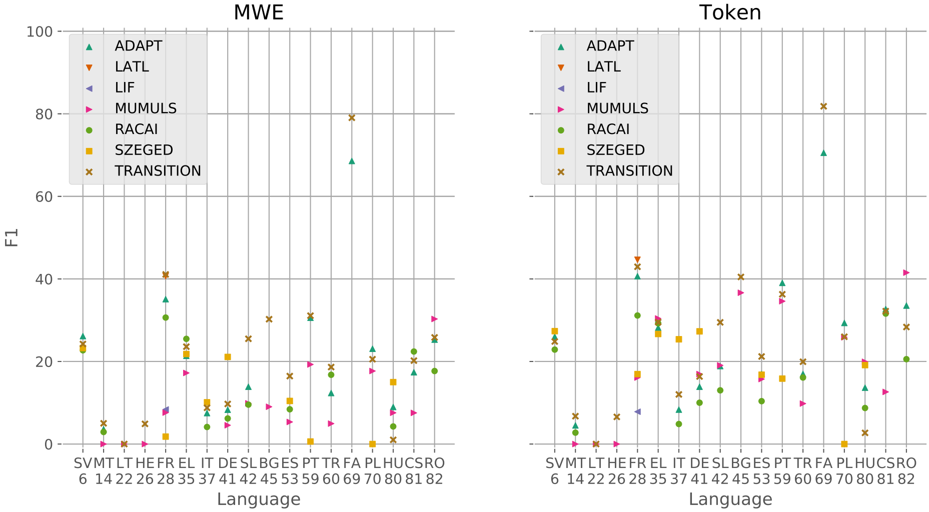


Figure 6: System evaluation scores (MWE-based, left; Token-based, right) on non-shared/unseen VMWEs only

did not fare nearly as well in this unseen VMWE evaluation. Taking a closer look at Farsi VMWEs, we observe that they show a higher level of *collostructional regularity*[5] compared to VMWEs in other languages. We observe that 86% of Farsi VMWEs are of length 2 and the last token in all Farsi VMWEs is always a verbs, while this is not the case for other languages such as Hungarian. In addition, verbs constitute a relatively small vocabulary in Farsi and as a consequence, the same set of verbs are used repeatedly in various VMWEs. For example, the 2,707 annotated VMWEs in the Farsi training set end with verbs of 46 different lemmas, and the 500 annotated instances in the test set end with 34 lemmas. Among these 34 different lemmas, only 4 do not appear in the training set. Last but not least, most of these verb lemmas are strong indicators of the presence of VMWEs, too. The overall occurrences of these lemmas in the Farsi corpus is 6,969, from which 3,207 are part of a VMWE, i.e., nearly half of them (46%). More precisely, 16 of these lemmas (with 29 occurrences) appear only as constituents of VMWEs; most importantly, for the most frequent lemma in VMWEs (the past and present forms of the infinitive کردن /kærdæn/ 'to make/to do', a light verb, which appears as the verb in 1,096 VMWEs) this proportion is 97% (i.e., out of 1,128 occurrences of this verb, only 32 do not surface as VMWE). To this, we can add observations concerning syntactic patterns in which VMWEs are used, e.g., the light verb کردن /kærdæn/ usually forms a transitive VMWE in which the non-verbal component of the VMWEs appear right after the adposition را /ra/ (i.e., which signals the presence of the syntactic object). We maintain that these exemplified regularities can justify the obtained results over the Farsi corpus.

In general, however, it is fair to expect that systems will tend to perform worse on VMWEs they did not see in training.

2.3 Discontinuous VMWEs and embedded/overlapped VMWEs

Two innovations in the PARSEME shared task were discontinuous VMWEs and embedded or overlapped VMWEs (see Savary et al. 2018 [this volume] §6.3).

Figure 7 shows that for most languages, the majority of VMWEs are continuous. For Czech and Turkish, there is about a 50–50 proportion between continuous and discontinuous VMWEs. For many other languages, the proportion of discontinuous VMWEs is considerable (German, Greek, French, Polish, Portuguese, Romanian, Slovenian). There is therefore a clear advantage in designing systems capable of detecting discontinuous VMWEs.

[5] Degree to which words tend to form (appear with) grammatical constructions (Stefanowitsch & Gries 2003).

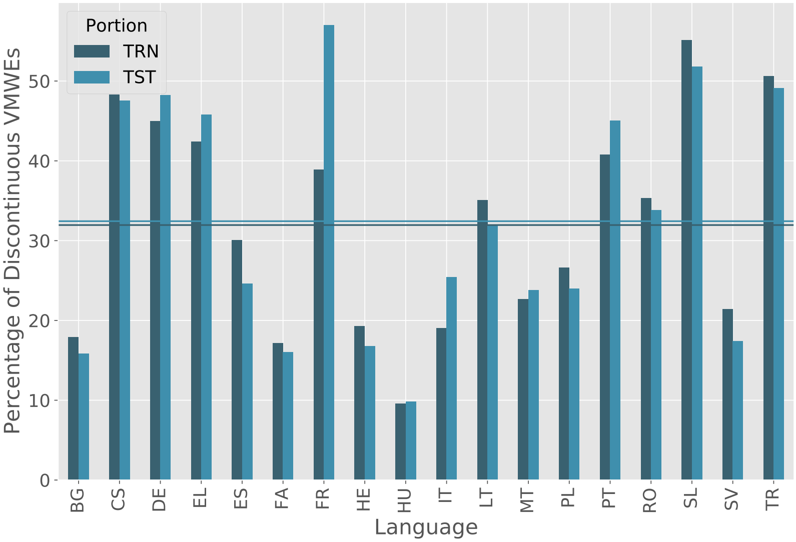


Figure 7: Percentage of discontinuous VMWEs across language corpora.

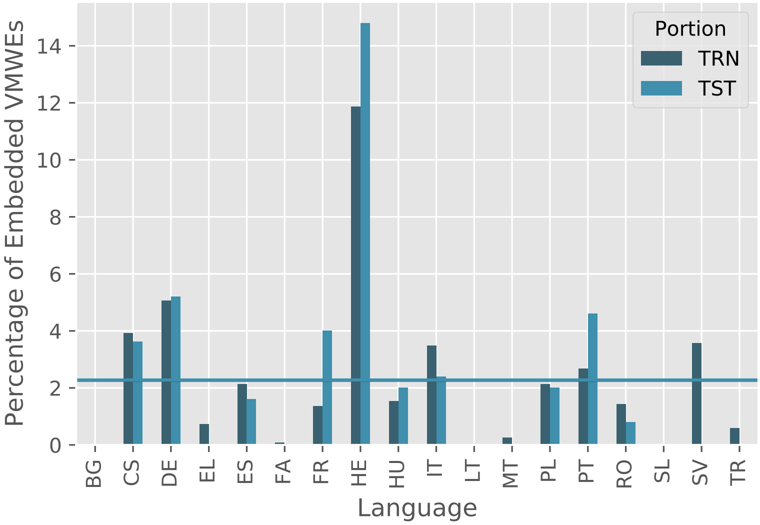


Figure 8: Proportion of Embedded/Overlapped VMWEs across language corpora.

The proportion of embedded/overlapped VMWEs, shown in Figure 8, is very low across languages, with an average of around 2.3% in both training and test portions. Hebrew is the language with the highest rate of embedded VMWEs at only 12–14.5%. Some languages do not even register a single embedded VMWE. Because of these low numbers, a system not designed to deal with embedded VMWEs will not be severely penalised. We therefore do not consider embedded VMWEs to be a difficulty factor in this dataset, with the exception of Hebrew.

2.4 Relative training-test corpus heterogeneity

The evaluation paradigm followed in the PARSEME shared task dictates that systems must be evaluated on a strictly unseen test set, guaranteeing fairness to all participating system developers. However, a valid expectation is that the data that systems will be tested on should be roughly of the same kind as the data they were trained on. The training and test portions of a language corpus should be fairly homogeneous.

Kilgarriff & Rose (1998) introduced a statistical metric to estimate the similarity of two corpora of similar size by computing the χ^2 score of the n most frequent words in the corpora. The lower this score, the less variability between the corpora and thus the more similar they are. They also adapted this similarity score to measure the homogeneity of a single corpus by computing χ^2 scores on pairs of similarly sized partitions of the corpus and averaging the individual χ^2 scores. The lower this averaged score is, the more homogeneous the corpus is deemed to be. Here, we adapt this homogeneity score in order to estimate the homogeneity between the training and test sets of a language corpus. This is done by computing similarity scores of training set partitions against similarly-sized test set partitions and averaging them together to obtain a single cross-set homogeneity score. The higher this score is, the more heterogeneous the training and test sets are. In order to allow comparisons across languages, this cross-set homogeneity score is normalised by dividing it by the average of the within-training set and within-test set homogeneity scores, calculated from the training and test sets separately. We call the result of this division, the *heterogeneity ratio of a language corpus*. Table 2 sorts the languages by their heterogeneity ratio. The detailed algorithm used is listed in Algorithm 1.

French comes out on top. Its heterogeneity ratio of 4.31 can be interpreted as the number of times that the training-test sets are more heterogeneous than the training and test sets on their own. This suggests that the French test was not derived from the same sources as the training set, or at least not in the same proportions.

Table 2: Heterogeneity ratios between training and test sets

FR	TR	IT	PT	RO	CS	PL	HU	LT	DE	FA	BG	SL	ES	SV	HE	EL	MT
4.31	2.89	2.53	2.03	2.03	1.92	1.77	1.73	1.62	1.59	1.56	1.51	1.39	1.28	1.25	1.18	1.15	1.03

Algorithm 1 Computing a language heterogeneity ratio

```
$R \longleftarrow$ number of repetitions
$n \longleftarrow$ number of words in a partition
$hr\_sum \longleftarrow 0$
$r \longleftarrow 0$
while $r < R$ do
    $trn \longleftarrow$ partition_set($n$, shuffle_sentences(training_set))
    $tst \longleftarrow$ partition_set($n$, shuffle_sentences(test_set))

    ▸ Cross homogeneity:
    $s \longleftarrow 0$
    $c \longleftarrow 0$
    for $i = 1$ to $|trn|$ do
        for $j = 1$ to $|tst|$ do
            $s \longleftarrow s+$ corpus_similarity($partition_i$, $partition_j$)
            $c \longleftarrow c + 1$
    $cross \longleftarrow s/c$

    ▸ Within-training homogeneity:
    $s \longleftarrow 0$
    $c \longleftarrow 0$
    for $i = 1$ to $|trn|$ do
        for $j = i + 1$ to $|trn|$ do
            $s \longleftarrow s+$ corpus_similarity($partition_i$, $partition_j$)
            $c \longleftarrow c + 1$
    $within\_trn \longleftarrow s/c$

    ▸ Within-test homogeneity:
    $s \longleftarrow 0$
    $c \longleftarrow 0$
    for $i = 1$ to $|tst|$ do
        for $j = 1 + 1$ to $|tst|$ do
            $s \longleftarrow s+$ corpus_similarity($partition_i$, $partition_j$)
            $c \longleftarrow c + 1$
    $within\_tst \longleftarrow s/c$

    ▸ Heterogeneity ratio:
    $hr \longleftarrow cross/((within\_trn + within\_tst)/2)$
    $hr\_sum \longleftarrow hr\_sum + hr$

    $r \longleftarrow r + 1$
return $hr\_sum/R$
```

French is followed by Turkish, Italian, Portuguese and Romanian, with ratios around 2. The rest of the languages are closer to 1, reflecting a more balanced/homogeneous partitioning between the training and the test corpora. Notice however that systems participating in French, Turkish, Italian, Portuguese and Romanian did relatively well despite their heterogeneity. Nonetheless, adopting a similar corpus selection and balancing policy across languages, like mixing the corpora before splitting them into training and test portions in comparable proportions, could be a way to put all languages on a similar footing.

3 Participating systems and baselines

This section focuses on the actual systems in the competition and introduces two baseline systems: (i) a dictionary lookup-based system that attempts to match known VMWEs against the test set, (ii) a system that flags every verb in the test set as a VMWE.

3.1 Overview of participating systems

Seven systems participated in the PARSEME shared task. Their performance was presented and discussed in §2, although not individually. The techniques employed by the different systems can be summarised as follows:

- ADAPT (Maldonado et al. 2017) uses a Conditional Random Fields (CRF) sequence labelling approach to identify the tokens of VMWEs. The features that helped most were dependency-based: the token's head, dependency relation with the head and the head's part of speech (POS) tag, along with standard bigram and trigram features commonly used in named-entity recognisers. The ADAPT system did not attempt to classify VMWEs by category. An extended version of this system is described in Moreau et al. (2018 [this volume]).

- RACAI (Boroş et al. 2017) also employs a CRF sequence labelling approach using lemma and POS tag features. However, this system conducts the VMWE identification task in two steps: head labelling (identifying the verb) and tail labelling (identifying the words linked to the head). The RACAI system does attempt to classify the VMWEs by their category.

- MUMULS (Klyueva et al. 2017) also models the VMWE identification problem as a sequence labelling task, but using a recurrent neural network via the TensorFlow package. As input features, they build embeddings of 100 dimensions from the concatenation of a token's surface form, lemma and POS tag.

- TRANSITION (Al Saied et al. 2017) is a greedy transition-based system of the kind typically used in parsing. This system does not have a syntax prediction module, however, and focuses on the lexical analysis phase of the parsing mechanism. An extended version of this system is described in Al Saied et al. (2018 [this volume]).
- LIF (Savary et al. 2017) also employs a probabilistic transition-based technique. The team focused on French light-verb constructions.
- SZEGED (Simkó et al. 2017) trains a dependency parser on a modified training set in which the dependency relation label of tokens belonging to a VMWE were relabelled with the corresponding VMWE category label. Simkó et al. (2018 [this volume]) describes an extended version of this system.
- LATL (Nerima et al. 2017) uses a rule-based constituent parser that prioritises parsing alternatives of known collocations, and uses its parsing features to detect known collocations even if they are in a different word order or if they are discontinuous.

Not all systems participated in all languages. French was the language covered by most systems. The languages least covered were Bulgarian, Hebrew, Lithuanian (covered only by MUMULS and TRANSITION) and Farsi (covered by ADAPT and TRANSITION). Since only raw surface tokens and no syntactic dependency information or POS tags were provided for Bulgarian, Hebrew and Lithuanian, most system developers decided not to cover them. The systems that covered most languages were TRANSITION (all 18 languages), ADAPT (15), MUMULS (15), RACAI (12) and SZEGED (9). LATL and LIF focused on French only.

In Token-based evaluation, ADAPT ranked first on two languages (French and Polish), while MUMULS and SZEGED ranked first on Romanian and Hungarian, respectively. In MWE-based evaluation, TRANSITION beat all systems in all languages, except Hungarian (won by SZEGED) and Romanian (won by RACAI and very closely followed by MUMULS).

The ADAPT and the RACAI systems are clearly related, as are the TRANSITION and the LIF systems. These four systems, along with the MUMULS system, are all probabilistic sequence labelling methods, although quite different in their implementation details. It is interesting to see that, on average (see bottom row in Table 1), ADAPT and TRANSITION performed very similarly in the Token-based evaluation, while MUMULS and RACAI also performed very similarly in the same average evaluation.

3.2 Baseline systems

This section proposes two types of baseline systems that put into perspective the participating systems' performance. One such baseline system is a simple dictionary lookup, which collects all VMWEs encountered during training and simply attempts to match collected VMWEs in the test set. The other is a baseline system which flags every verb as a VMWE. More details on these two baselines and their results are described in what follows.

Dictionary lookup baseline The implemented system is very simplistic: it attempts to match VMWE lemmas from the training file in the test file sequentially. If lemmas are not available, then the token's surface form is used. Discontinuous VMWEs are matched in the test file as long as they appear in the same order as in the training file: intervening words are ignored when collecting VMWEs from the training file and when matching VMWEs in the test file. If one VMWE appears in more than one word order in the training file, each word order will be considered to be a separate VMWE. Tokens are marked as belonging to a VMWE only if a full match is detected; partial matches are not flagged. This is to avoid making too many, potentially spurious, partial matches. Embedded/overlapped VMWEs are attempted by using separate VMWE matching automata.

Notice that the maximum performance that can be achieved by this lookup system is determined by the proportion of shared VMWEs between the training and the test set in a language corpus. This proportion of shared VMWEs, indicated as percentages under the language labels in Figure 5 and Figure 6, is thus the maximum recall such a system can achieve.

The actual F1 score for the dictionary lookup system described here appears in the BD column in Table 1. It is evident from this table that this simple baseline is quite competitive, beating some of the participating systems in several languages. In fact, it beat all systems on both evaluation modalities in Hebrew, Lithuanian and Polish, and on MWE-based evaluation in German.

Verb baseline As mentioned earlier, this system simply flags each verb in the test set as a VMWE. Column BV in Table 1 shows the F1 scores for the verb baseline. Notice that no scores are supplied for Bulgarian, Hebrew and Lithuanian.

This is because no POS tag was provided in these languages' datasets. So we omit them from this discussion.

For BV, notice that the Token-based F1 scores range between 10 to 47 for most languages. This is a relatively high score range. Table 3 provides precision and recall details for these Token-based scores.

Table 3: Token-based scores for the Verb baseline

Language	CS	DE	EL	ES	FA	FR	HU	IT	MT	PL	PT	RO	SL	SV	TR
P-token	13.57	20.87	5.14	9.58	48.64	11.52	9.29	8.87	3.74	11.42	8.55	6.61	4.17	7.8	6.54
R-token	41.13	45.02	40.85	41.49	46.86	44.13	19.08	38.26	36.81	46.31	44.1	44.3	44.59	43.59	25.97
F1-token	20.41	28.52	9.14	15.56	47.73	18.28	12.49	14.4	6.79	18.33	14.32	11.51	7.63	13.23	10.45

Notice that this baseline's recall directly depends on each language's proportion of sentences with VMWEs (see Figure 3). Recall is particularly high with most languages scoring around the 40-point mark. We interpret this result as indicating that Token-based scores tend to overestimate systems' performance. We elaborate on this issue in §4. The recall values in Hungarian and Turkish are considerably lower than in the rest of the languages. This is because there is a large proportion of VMWEs in these languages that are not tagged with a *verb* POS tag (this baseline exploits that tag): 74% of VMWEs in Hungarian and 50% of VMWEs in Turkish do not have a single token with a *verb* POS tag. Different teams make different decisions as to what MWEs constitute *verbal* MWEs. For example, the Hungarian team informed us that they flag nominalised verbs as VMWEs, even if they are not functioning as verbs anymore.

Given that the verb baseline only labels a single word (a verb) and that VMWEs are made up of at least two words (the verb plus at least another word), the reader might find it puzzling that, in Table 1, the verb baseline (BV) has non-zero MWE-based scores on a few languages. The MWE-based evaluation modality only rewards full MWE matches, not partial matches. How is it possible to get non-zero scores on full MWE matches for single-word labels which surely will never form a full match, given that the minimum length of a full VMWE is two words? It turns out that there are VMWEs of one-word length in some languages. This is usually due to linguistic reasons specific to each language in which a single word is consdered composed of more than one unit. In Spanish, for example, reflexives can sometimes appear separated from the verb and sometimes postfixed to the verb: *ella* ***se levanta*** *temprano* 'she gets up early' vs. *es difícil* ***levantarse*** *temprano* 'getting up early is hard'. Both, *se levanta* and *levantarse*, are considered to be VMWEs.

4 Evaluation methods

As previously mentioned, system performance was measured on two modalities: MWE-based evaluation and Token-based evaluation. Whilst the MWE-based evaluation is an all-or-nothing measure, which might unfairly penalise systems that partially identify correct VMWEs, the Token-based evaluation is intended to compensate for this coarse penalisation by giving partial credit for every word of the identified VMWE. Thus, it is reasonable to expect systems to perform better on Token-based evaluation than on MWE-based evaluation. Indeed, Table 1 shows that for the most part, Token-based scores are higher than MWE-based scores within every system-language combination, including baseline systems.

By definition, every single VMWE will involve a verb. So, the verb baseline system is able to make gains on the Token-based F1 score by increasing recall, at the expense of reducing precision. However, if the dataset were less unbalanced (i.e. if it had less VMWE sparsity), the verb baseline would also increase its precision. In addition, the Token-based evaluation gives more weight to longer VMWEs than shorter ones. Matching one VMWE of say four tokens gets the same credit as matching two VMWEs of two tokens each. More credit should perhaps be given for matching more (even if partially) VMWEs than for matching fewer, longer VMWEs.

Even though Token-based scores are expected to be higher than MWE-based scores, the system rankings differ across modalities. Because of these issues, we cannot categorically say that system *A*, which scored higher than system *B* in Token-based evaluation, is better at detecting partial VMWEs. It could well be that system *A* is good at identifying simple verbs and/or long and formulaic VMWEs but not necessarily at detecting partial VMWEs. One solution would be giving a fraction of a point corresponding to the proportion of a matched VMWE, as well as subtracting a fraction of a point proportional to matched non-VMWE tokens.

On a slightly different note, we would like to propose an alternative evaluation metric: Cohen's κ measure, which is commonly used to measure inter-annotator agreement. We use it here to measure the degree to which systems agree with gold standards. The obtained κ score is similar to the MWE-based F1 score, but with a correction that removes the possible bias from chance agreement.

We compare the similarity between systems' rankings given by the averaged results per language per performance measure, by reporting their Spearman's rank correlation ρ and Pearson's moment correlation. As shown in Table 4, the rankings and assigned scores to systems remain very similar across performance measures. However, overall, the Token-based and MWE-based measures show the highest correlation (both in terms of ranking, ρ, and the relative magnitude of the assigned scores, r). With respect to Cohen's κ, while it yields a ranking more similar to the MWE-based measure, the distribution of the assigned Cohen's κ scores are more similar to the token-based method (i.e., their linear relationship signified by r).

Table 4: Similarity of systems' ranking per performance measure: Spearman's ρ and Pearson's r are reported to show similarity between systems' ranking per performance measure.

Measure	Measure	ρ	r
Token-based	MWE-based	98.14	97.48
Token-based	Cohen's κ	94.06	93.28
MWE-based	Cohen's κ	96.75	97.18

4.1 On Using the Cohen's κ as an evaluation score

The use of the F1 score, i.e., the harmonic mean of precision and recall, for evaluation can be biased unless certain criteria are met, e.g. that the distribution of annotated instances in the test and training data are identical. Since in the PARSEME shared task, the VMWE identification task is reduced to a binary classification problem, Cohen's κ can be used reliably to obtain a measure of performance that can, at least, cancel out the influence of certain sources of bias. In particular, it penalises the overall score of the systems by the expected chance agreement (as done in the computation of inter-annotator agreement) and takes into account a notion of true negative rate in the overall evaluation of systems (Powers 2012; 2015).

The count of true negative outputs and subsequently true negative rate, however, cannot be computed directly from the evaluation setup and the test set. Simply put, we do not know how many "*is this a VMWE?*" questions are answered by a system[6] (or human annotators) in order to perform the identification task on a test set (or to manually annotate a corpus). Hence, further assumptions about the problem setting are required to devise the number of true negatives in the respective evaluation contingency table. Here, likewise (Savary et al. 2017), we assume

[6] This discussion also implies a way to justify the better performance of transition-based systems, i.e., the total number of classification problems in these systems is often less than in non-transition-based systems.

that the total number of stimuli, i.e., the total number of "*is this a VMWE?*" questions to complete a VMWE identification problem, is approximately equivalent to the number of verbs in the test set (or the corpus which must be annotated).

Given the abovementioned assumption for a test set, let v be the number of verbs in the set that are not part of a VMWE. For a system, we define tp and fp as being the number of correctly and incorrectly identified VMWEs, respectively, and fn as the number of VMWEs in the test set that are not identified by the system. If

(2)

$$t = tp + fp + fn + v$$

we compute

(3)

$$p_o = \frac{tp + v}{t}$$
$$p_e = p_0 + p_1$$

in which

(4)

$$p_0 = \frac{(tp + fp) \times (tp + fn)}{t^2}$$
$$p_1 = \frac{(fn + v) \times (fp + v)}{t^2}$$

Finally, we compute Cohen's κ:

(5)

$$\kappa = \frac{1 - p_o}{1 - p_e}$$

and report it as an additional performance measure. Evidently, the suggested method can be refined and improved, e.g., by taking the partial matches between VMWEs (particularly the verbal part) into account.

5 Conclusions

This chapter analysed different statistical properties of the language corpora used in the PARSEME shared task. We found that having large training sets allows systems to better learn to identify VMWEs. But size is not the whole story. High VMWE sparsity can hinder a system's performance. However, it can be offset by a large training corpus and, even better, by ensuring that the corpus has many examples of a majority of VMWEs, a property we call good VMWE frequency distribution. Romanian seems to be the language corpus that hits the sweet spot: it is large in size (training and test portions) and it has a good frequency distribution, even if it suffers from high VMWE sparsity.

This chapter also showed that the higher the proportion of VMWEs shared between training and test sets, the better the systems will perform. We also saw that it is advisable to design systems capable of detecting discontinuous VMWEs, but we observed that systems would not be significantly penalised for ignoring embedded VMWEs. There was no clear pattern on the effect of the training-to-test proportions on systems' performance. Shuffling corpora before splitting into training and test portions will also reduce its heterogeneity ratio and help put all languages on a similar footing.

On the evaluation front, we found the token-based evaluation method to overestimate the performance of systems. As future work, the authors will investigate alternative partial-matching measures, especially those that favour number of the detected VMWEs over their lengths. And finally, this chapter described the use of Cohen's κ metric to produce less biased estimations of systems' performance.

We would also like to recommend shared task organisers to consider application scenarios of the VMWE identification task. Different application scenarios will dictate different evaluation criteria, corpus selection and priorities. For example, if VMWEs are being identified to compile a dictionary, perhaps recall should be favoured over precision. If the application is to identify a few but good VMWEs examples for a language learning system, then precision should be favoured. Evaluation could also be done *in vivo* in actual parsing or machine translation systems, which is something the authors will seek to investigate as future work.

The quality of the analysis presented here depends directly on the quality of the annotated data. Whilst the annotation guidelines try to be as universal as possible, we have found that significant differences in annotation approach remain. For example, at least one language team annotated MWEs derived from verbs that do not function as verbs (e.g., nominalised verbs). So we hope that this work can spark a discussion in the community as to what constitutes a VMWE more precisely. Is it simply a MWE that involves a word of verbal origin (even if it does not function as a verb anymore) or must it be a MWE involving a verb that still functions as a verb?

The authors hope that the insights and recommendations included in this chapter inform future editions of the shared task. At the same time, the authors plan, as future work, to repeat the analysis presented here on the second edition of this dataset, which is being prepared at the time of writing. This will help us determine to what extent our observations generalise to new datasets.

Acknowledgements

We would like to thank the shared task organisers for providing systems' outputs which were instrumental in some of the analysis presented in this chapter. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. This work is also supported by the DFG (Deutsche Forschungsgemeinschaft) through "Collaborative Research Centre 991 (CRC 991): The Structure of Representations in Language, Cognition, and Science".

Abbreviations

AVG	average
BD	baseline: dictionary lookup
BV	baseline: verb detection
CRF	conditional random fields
F1	F1 score aka F-measure
κ	Cohen's inter-annotation agreement measure
MWE	multiword expression
POS	part of speech
r	Pearson's correlation coefficient
ρ	Spearman's rank correlation coefficient
RNK	rank
TRN	training
TST	test
VMWE	verbal multiword expression

References

Al Saied, Hazem, Marie Candito & Matthieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions* *at length and in depth: Extended papers from the MWE 2017 workshop*, 209–226. Berlin: Language Science Press. DOI:10.5281/zenodo.1469561

Al Saied, Hazem, Matthieu Constant & Marie Candito. 2017. The ATILF-LLF system for parseme shared task: A transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 127–132. Association for Computational Linguistics. DOI:10.18653/v1/W17-1717

Boroş, Tiberiu, Sonia Pipa, Verginica Barbu Mititelu & Dan Tufiş. 2017. A data-driven approach to verbal multiword expression detection. PARSEME shared task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 121–126. Association for Computational Linguistics. DOI:10.18653/v1/W17-1716

Kilgarriff, Adam & Tony Rose. 1998. Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, 46–52.

Klyueva, Natalia, Antoine Doucet & Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 60–65. Association for Computational Linguistics. April 4, 2017. DOI:10.18653/v1/W17-1707

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 114–120. Association for Computational Linguistics. DOI:10.18653/v1/W17-1715

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Moreau, Erwan, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel & Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 177–207. Berlin: Language Science Press. DOI:10.5281/zenodo.1469559

Nerima, Luka, Vasiliki Foufi & Eric Wehrli. 2017. Parsing and MWE detection: Fips at the PARSEME shared task. In *Proceedings of The 13th Workshop on Multiword Expressions* (MWE '17), 54–59. Association for Computational Linguistics. DOI:10.18653/v1/W17-1706

Powers, David M. W. 2012. The problem with Kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL '12), 345–355. Avignon, France: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2380816.2380859.

Powers, David M. W. 2015. What the f-measure doesn't measure: Features, flaws, fallacies and fixes. *CoRR* abs/1503.06410. http://arxiv.org/abs/1503.06410.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1471591

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2017. USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. In *Proceedings of The 13th Workshop on Multiword Expressions* (MWE '17), 48–53. Association for Computational Linguistics.

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2018. Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 227–243. Berlin: Language Science Press. DOI:10.5281/zenodo.1469563

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. DOI:10.1075/ijcl.8.2.03ste

Chapter 6

Semantic reranking of CRF label sequences for verbal multiword expression identification

Erwan Moreau
ADAPT Centre, Trinity College Dublin

Ashjan Alsulaimani
Trinity College Dublin

Alfredo Maldonado
ADAPT Centre, Trinity College Dublin

Lifeng Han
ADAPT Centre, Dublin City University

Carl Vogel
Trinity Centre for Computing and Language Studies, Trinity College Dublin

Koel Dutta Chowdhury
ADAPT Centre, Dublin City University

Verbal multiword Expressions (VMWE) identification can be addressed successfully as a sequence labelling problem via conditional random fields (CRFs) by returning the one label sequence with maximal probability. This work describes a system that reranks the top 10 most likely CRF candidate VMWE sequences using a decision tree regression model. The reranker aims to operationalise the intuition that a non-compositional MWE can have a different distributional behaviour than

Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado, Lifeng Han, Carl Vogel & Koel Dutta Chowdhury. 2018. Semantic reranking of CRF label sequences for verbal multiword expression identification. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 177–207. Berlin: Language Science Press. DOI:10.5281/zenodo.1469559

that of its constituent words. This is why it uses semantic features based on comparing the context vector of a candidate expression against those of its constituent words. However, not all VMWE are non-compostional, and analysis shows that non-semantic features also play an important role in the behaviour of the reranker. In fact, the analysis shows that the combination of the sequential approach of the CRF component with the context-based approach of the reranker is the main factor of improvement: our reranker achieves a 12% macro-average F1-score improvement on the basic CRF method, as measured using data from PARSEME shared task on VMWE identification.

1 Introduction

The automatic identification of multiword expressions (MWEs) is an important but challenging task in natural language processing (NLP) (Sinclair 1991; Sag et al. 2002). An effort in response to this challenge is the shared task on detecting multiword verbal constructions (Savary et al. 2017) organised by the PARSing and Multiword Expressions (PARSEME) European COST Action.[1] The shared task consisted of two tracks: a closed one, restricted to the data provided by the organisers, and an open track that permitted participants to employ additional external data.

The ADAPT team participated in the closed track with a system that exploited syntactic dependency features in a Conditional Random Fields (CRF) sequence model (Lafferty et al. 2001) and ranked 2nd in the detection of full MWEs in most languages (Maldonado et al. 2017).[2] In addition to extending the description of our CRF-based solution in §3, this chapter focuses on a second component aimed at reranking the top 10 sequences predicted by the CRF decoder, using a regression model. This component, called a *semantic reranker* and described in §4, increases the performance of the system by an average 12% in F1-score over the datasets at the MWE level. Because the reranker requires a third-party corpus, the system using both components (the CRF-based and the reranker) would compete in the open track task.

The design of the semantic reranker was originally oriented towards detecting non-compositional expressions. In such expressions, the meaning of the expression cannot be obtained by combining the meanings of its individual words, i.e. the actual meaning is unrelated to the literal meaning (e.g. *to kick the bucket*). This is a distinctive feature which can be recognised by comparing their context

[1] http://www.parseme.eu.

[2] Official results: http://multiword.sourceforge.net/sharedtask2017/; system details, feature templates, code and experiment instructions: https://github.com/alfredomg/ADAPT-MWE17.

vectors (these vectors can be built from any large corpus). This idea has been used for bigram expressions (Schütze 1998; Maldonado & Emms 2011), and we adapted it to multiword expressions. Nevertheless, most verbal MWEs are actually compositional, at least to some extent (e.g. *to give somebody a break*). In light of the performance improvement obtained when adding the reranker to our system, it is clear that the reranker gives a boost in detecting MWEs across the board, and not only for a few non-compositional expressions. In order to understand how the reranker contributes to the performance, we carried out a thorough study and provide a detailed analysis of the results in §5.

2 Related work

MWEs have long been discussed in NLP research and a myriad of processing techniques have been developed, such as combining statistical and symbolic methods (Sag et al. 2002), single and multi-prototype word embeddings (Salehi et al. 2015), and integrating MWE identification within larger NLP tasks, such as parsing (Green et al. 2011; 2013; Constant et al. 2012) and machine translation (Tsvetkov & Wintner 2010; Salehi et al. 2014a,b).

More directly related to our closed-track approach are works such as that of Venkatapathy & Joshi (2006), who showed that information about the degree of compositionality of MWEs helps the word alignment of verbs, and of Boukobza & Rappoport (2009) who used sentence surface features based on the canonical form of VMWEs. In addition, Sun et al. (2013) applied a hidden semi-CRF model to capture latent semantics from Chinese microblogging posts; Hosseini et al. (2016) used double-chained CRF for minimal semantic units detection in a SemEval task. Bar et al. (2014) discussed that syntactic construction classes are helpful for verb-noun and verb-particle MWE identification. Schneider et al. (2014) also used a sequence tagger to annotate MWEs, including VMWEs, while Blunsom & Baldwin (2006) and Vincze et al. (2011) used CRF taggers for identifying continuous MWEs.

In relation to our open-track approach, Attia et al. (2010) demonstrated that large corpora can be exploited to identify MWEs, whilst Legrand & Collobert (2016) showed that fixed-size continuous vector representations for phrases of various lengths can have a performance comparable to CRF-based methods in the same task. Finally, Constant et al. (2012) used a reranker for MWEs in an *n*-best parser. We combine these ideas by reranking the *n* best CRF VMWE predictions for each sentence using regression scores computed from vectors that represent different combinations of VMWE candidates. The vectors are computed from a large corpus, namely EUROPARL's individual language subcorpora.

3 VMWE identification via CRF

We decided to model the problem of VMWE identification as a sequence labelling and classification problem. We operationalise our solution through CRFs (Lafferty et al. 2001), implemented using the CRF++ system.[3] CRFs have been successfully applied to such sequence-sensitive NLP tasks such as segmentation, named-entity recognition (Han et al. 2013; 2015) and part-of-speech (POS) tagging. Our team attempted 15 out of the 18 languages involved in the shared task. It should be noted that of these 15 languages, four (Czech, Farsi, Maltese and Romanian) were provided without syntactic dependency information, although morphological information (i.e. tokens' lemmas and POS) was indeed supplied. The data for the languages we did not attempt (Bulgarian, Hebrew and Lithuanian) lacked even morpho-syntactic information, leaving the CRF with only tokens as features; so we felt that we were unlikely to obtain good results with them, and chose to focus on the richer datasets.

3.1 Features

We assume that features based on the relationships between the different types of morpho-syntactic information provided by the organisers will help identify VMWEs. Ideally, one feature set (or *feature template* in the terminology of CRF++) per language should be developed. Due to time constraints, we developed a feature set for three languages (German, French and Polish), then for every language the feature template that performed best in cross-validation among these three was selected.

For each token in the corpus, the direct linguistic features available are its word surface (W), word lemma (L) and POS (P). In the languages where syntactic dependency information is provided, each token also has its head's word surface (HW), its head's word lemma (HL), its head's POS (HP) and the dependency relation between the token and its head (DR). It is possible to create CRF++ feature templates that combine these features. In addition, it is also possible to use the predicted output label of the previous token (B).

The three final feature templates, which we call FT3, FT4 and FT5,[4] are shown in Table 1. Whilst the feature templates in this table are expressed in the CRF++ format, a comment (starting with #) at each feature (line) expresses the type of feature and its relative position to the current token. For instance, `L-2` refers to

[3] https://taku910.github.io/crfpp/. Release 0.58, Last verified 2017-12-29.

[4] The feature template numbering starts at 3 for consistency with their original description in Maldonado et al. (2017).

Table 1: CRF++ Feature Templates developed. Example: template `U32:%x[0,3]` indicates current token (row 0, i.e. current row) and lemma (column 3) while template `U41:%x[-1,2]` refers to the previous token (row -1 from current row) and POS tag (column 2), etc.

FT3	FT4	FT5
# L-2	# L-2	# L-2
U30:%x[-2,3]	U30:%x[-2,3]	U30:%x[-2,3]
# L-1	# L-1	# L-1
U31:%x[-1,3]	U31:%x[-1,3]	U31:%x[-1,3]
# L	# L	# L
U32:%x[0,3]	U32:%x[0,3]	U32:%x[0,3]
# L+1	# L+1	# L+1
U33:%x[1,3]	U33:%x[1,3]	U33:%x[1,3]
# L+2	# L+2	# L+2
U34:%x[2,3]	U34:%x[2,3]	U34:%x[2,3]
# L-2/L-1	# L-2/L-1	# L-2/L-1
U35:%x[-2,3]/%x[-1,3]	U35:%x[-2,3]/%x[-1,3]	U35:%x[-2,3]/%x[-1,3]
# L-1/L	# L-1/L	# L-1/L
U36:%x[-1,3]/%x[0,3]	U36:%x[-1,3]/%x[0,3]	U36:%x[-1,3]/%x[0,3]
# L/L+1	# L/L+1	# L/L+1
U37:%x[0,3]/%x[1,3]	U37:%x[0,3]/%x[1,3]	U37:%x[0,3]/%x[1,3]
# L+1/L+2	# L+1/L+2	# L+1/L+2
U38:%x[1,3]/%x[2,3]	U38:%x[1,3]/%x[2,3]	U38:%x[1,3]/%x[2,3]
# P	# HL/DR	# HL/DR
U00:%x[0,2]	U01:%x[0,4]/%x[0,6]	U01:%x[0,4]/%x[0,6]
# HL/DR	# P/DR	# P/DR
U01:%x[0,4]/%x[0,6]	U02:%x[0,2]/%x[0,6]	U02:%x[0,2]/%x[0,6]
# P/DR	# HP/DR	# HP/DR
U02:%x[0,2]/%x[0,6]	U03:%x[0,5]/%x[0,6]	U03:%x[0,5]/%x[0,6]
# HP/DR		
U03:%x[0,5]/%x[0,6]	# Previous token’s label	# Previous token’s label
	B	B
# Previous token’s label		
B	# P-2	# P-2
	U40:%x[-2,2]	U40:%x[-2,2]
	# P-1	# P-1
	U41:%x[-1,2]	U41:%x[-1,2]
	# P	# P
	U42:%x[0,2]	U42:%x[0,2]
	# P+1	# P+1
	U43:%x[1,2]	U43:%x[1,2]
	# P+2	# P+2
	U44:%x[2,2]	U44:%x[2,2]
	# P-1/P	# P-1/P
	U45:%x[-1,2]/%x[0,2]	U45:%x[-1,2]/%x[0,2]
	# P/P+1	# P/P+1
	U46:%x[0,2]/%x[1,2]	U46:%x[0,2]/%x[1,2]
		# L/HP
		U52:%x[0,3]/%x[0,5]

the lemma of the token at position $i-2$ relative to the current token at position i, `P+1` refers to the part of speech of the token at position $i+1$, and `HL/DR` refers to the combination of head's lemma of the current token and the dependency relation between the current token and its head.[5]

FT5 is based on FT4, which in turn, is based on FT3. FT3 is itself based on the CRF++ example feature template, commonly used in NER experiments. The difference between FT3 and this example feature template is that FT3 exploits syntactic dependency information.

We conducted preliminary 5-fold cross validation experiments on German, French and Polish training data using the FT3 template. We then started tweaking the template independently for each of these three languages based on successive 5-fold cross validation results. This exercise resulted in the three final templates: FT3 for French and FT5 for German and Polish. Given that FT4's performance was very similar to that of FT5, we decided not to discard it.

During this preliminary experimentation, we also observed that templates exploiting token word surface features (W) performed unsurprisingly worse than those based on token lemmas (L) and POS (P). Templates using head features (HL, HP, DR) in addition to token features (L, P) fared better than those relying on token features only.

We also attempted to test the assumption that these feature templates would perform similarly in other languages of the same language family. That is, that FT3 would also perform better than FT4 and FT5 in other Romance languages and that FT5 would score higher than FT3 and FT4 in the rest of the languages. So we conducted a final set of 5-fold cross validation experiments on all 15 languages, this time trying each feature template (FT3, FT4 and FT5) independently on each language. The results are shown in Table 2. The F1 scores in bold italic are the maximum scores per language. For each given language, the results of the three templates are very similar. Therefore we are not able to comfortably confirm or refute our language family assumption. Nonetheless, we decided to choose for the final challenge the template that maximised the MWE-based F1 score for each language.

In order to use these templates with the provided data, we combined the supplied PARSEMETSV (VMWE annotations) and CONLLU files (linguistic features) into a single file. The training and blind test files were combined separately. The resulting file is also columnar in format, with column 0 representing the token ID as per the original PARSEMETSV file, column 1 the token's surface form, col-

[5] CRF++ feature template format described in https://taku910.github.io/crfpp/#templ Last verified 2017-12-19.

Table 2: F1 scores from cross validation experiments on 15 languages using feature templates FT3, FT4 and FT5.

Lang	CS		DE		EL		ES		FA	
Eval. Type	MWE	Token	MWE	Token	MWE	Token	MWE	Token	MWE	Token
FT3	54.23	70.79	23.84	39.02	***50.41***	***62.03***	***56.04***	60.74	76.09	83.52
FT4	55.91	71.81	24.41	39.62	50.16	61.76	55.72	60.77	77.88	84.75
FT5	***57.12***	***72.57***	***25.23***	***40.53***	49.72	62.02	55.70	***61.00***	***78.61***	***85.24***

Lang	FR		HU		IT		MT		PL	
Eval. Type	MWE	Token	MWE	Token	MWE	Token	MWE	Token	MWE	Token
FT3	3.99	6.71	66.03	70.24	***65.85***	75.81	81.44	80.96	28.31	31.12
FT4	4.20	6.92	65.70	70.56	65.62	76.07	81.30	81.00	28.19	30.80
FT5	***5.35***	***7.96***	***66.28***	***71.21***	65.6	***76.08***	***81.86***	***81.76***	***28.68***	***31.51***

Lang	PT		RO		SL		SV		TR	
Eval. Type	MWE	Token	MWE	Token	MWE	Token	MWE	Token	MWE	Token
FT3	56.56	64.81	75.87	***83.76***	***37.06***	***48.90***	19.68	20.09	***49.71***	59.42
FT4	***56.73***	65.32	75.91	83.47	34.04	46.70	***24.47***	***24.56***	49.49	***59.43***
FT5	56.64	***65.52***	***76.00***	83.69	34.65	47.57	22.09	22.33	49.46	59.38

umn 2 the token's POS tag (P), column 3 the lemma (L), column 4 the head's lemma (HL), column 5 the head's POS tag (HP), column 6 the dependency relationship between the token and its head (DR), and column 7 the VMWE label for the token.

The VMWE label was changed from the numerical values in the PARSEMETSV file to "B" for the tokens that start a VMWE and "I" for subsequent tokens that belong to a VMWE. This labelling scheme (usually called BIO, for *Begin, Inside, Outside*) is common in CRF-based implementations of NER systems. The BIO scheme can represent several consecutive VMWEs but cannot represent embedded or overlapping VMWEs, so these were ignored and a single B or I label was used for overlapping tokens.[6] The proportion of overlapping VMWEs is between 2 and 6% for Czech, German, Spanish, French, Italian, Polish, Portuguese and

[6] Remark: Schneider et al. (2014) proposes several tagging schemes, some using special "o" labels for discontinuous expressions; since we use the most simple scheme (BIO), the words which appear between the lexicalized components of the expressions are labeled with a regular "O".

Swedish, and it is even less for the rest of the languages we studied (see Maldonado & QasemiZadeh 2018 [this volume] for further details). Because of these low proportions, we consider embedded/overlapping VMWEs to have only a small negative impact on our system's performance. Therefore, we decided to ignore them.

Our system does not distinguish among the different categories of VMWEs, treating them all equally. The templates in Table 1 make reference to each feature based on the position of the current token and the column in which they appear in the input file.

3.2 CRF results

Table 3 shows, under the "CRF only" category, the Precision (P), Recall (R) and F1 (F) scores on the test set based on the shared task two evaluation modalities, MWE-based and token-based.[7] On token-based evaluation, our system was ranked in first place in French, Polish and Swedish, second place in eight languages (Czech, Greek, Farsi, Maltese, Portuguese, Romanian, Spanish and Turkish) and third in three (German, Italian and Slovenian). For MWE-based scores, our system ranked second place in nine languages (Farsi, French, Italian, Maltese, Polish, Portuguese, Slovenian, Spanish and Swedish) and third in four languages (Czech, German, Hungarian and Turkish). If all languages' scores are averaged per system, our system ranks in third place on both token-based and MWE-based evaluation (see Maldonado & QasemiZadeh 2018 [this volume] for average scores per system).

4 Semantic reranker

4.1 Motivations

The semantic reranker is the second component of the VMWE detection system. While the first component (CRF) offers a decent level of performance in its predictions, the reranker is intended to fix as many mistaken predictions as possible, by exploiting features that CRF are poorly equipped to deal with. These features are based on a distributional semantics approach (Schütze 1998; Maldonado & Emms 2011): they rely on comparing context vectors which are extracted from a *reference corpus* (usually a large third-party corpus). As it is often the case with

[7] We did not include the languages for which there is no Europarl data in this table.

Table 3: Performance by language according to official evaluation measures for the CRF component alone and the two components together (CRF and semantic reranker) P/R/F stands for precision/recall/f-score. All the values are expressed as percentages. The last two rows show the macro-average performance over the 12 languages.

Lang.	Eval.	CRF only			CRF + Reranker			Improvement (%)		
	Type	P	R	F	P	R	F	P	R	F
CS	MWE	59.3	56.2	57.7	79.8	63.4	70.6	+34.5	+12.8	+22.4
	Token	81.9	65.6	72.9	86.4	65.7	74.6	+5.4	+0.2	+2.4
DE	MWE	33.1	17.4	22.8	53.5	19.6	28.7	+61.9	+12.6	+25.9
	Token	70.6	28.4	40.5	72.2	25.6	37.8	+2.3	-9.7	-6.6
EL	MWE	34.4	28.8	31.3	45.1	32.2	37.6	+31.2	+11.8	+19.9
	Token	53.8	36.0	43.1	54.8	36.3	43.6	+1.8	+0.7	+1.1
ES	MWE	61.1	34.8	44.3	66.7	38.8	49.0	+9.2	+11.5	+10.6
	Token	74.5	36.7	49.2	74.2	38.9	51.1	-0.3	+6.0	+3.9
FR	MWE	61.5	43.4	50.9	75.6	47.6	58.4	+22.9	+9.7	+14.8
	Token	80.9	49.6	61.5	82.6	50.0	62.3	+2.1	+0.7	+1.2
HU	MWE	75.7	59.9	66.9	74.5	63.3	68.5	-1.5	+5.7	+2.4
	Token	78.5	57.1	66.1	77.3	61.5	68.5	-1.5	+7.8	+3.7
IT	MWE	61.7	14.2	23.1	70.8	9.2	16.3	+14.6	-35.2	-29.5
	Token	69.6	15.3	25.1	76.1	9.7	17.3	+9.2	-36.4	-31.2
PL	MWE	78.0	60.2	68.0	83.7	63.8	72.4	+7.4	+6.0	+6.6
	Token	87.4	62.3	72.7	87.0	63.9	73.7	-0.5	+2.6	+1.3
PT	MWE	64.1	53.2	58.1	75.9	57.2	65.2	+18.3	+7.5	+12.2
	Token	83.5	60.5	70.2	82.2	60.1	69.4	-1.5	-0.8	-1.1
RO	MWE	75.5	71.4	73.4	90.4	77.6	83.5	+19.8	+8.7	+13.8
	Token	88.3	76.4	81.9	91.8	77.9	84.3	+4.0	+2.1	+3.0
SL	MWE	51.4	29.0	37.1	68.6	32.4	44.0	+33.5	+11.7	+18.7
	Token	72.9	32.6	45.1	75.4	32.4	45.3	+3.5	-0.8	+0.5
SV	MWE	48.6	22.0	30.3	48.7	23.7	31.9	+0.2	+7.7	+5.2
	Token	52.5	22.5	31.5	52.1	24.1	32.9	-0.7	+7.0	+4.6
macro-	MWE	58.7	40.9	48.2	69.4	44.1	53.9	+18.3	+7.8	+11.9
average	Token	74.5	45.3	56.3	76.0	45.5	56.9	+2.0	+0.6	+1.1

complex machine learning problems, the orthogonality of the information, obtained on the one hand from the sequential CRF model and computed on the other hand from an independent semantic vector space, proves fruitful; this will be demonstrated in §5.

4.2 Design

Intuitively, the goal is to estimate whether a candidate expression should be considered a MWE or not. Thus, the core part of the reranker is to generate features which are relevant to assess the likeliness of a given candidate MWE being an actual MWE. These expression-level features are then combined to produce a set of sentence-level features, which in turn are used to train a decision tree regression model. Later, this model is applied to the candidate sequences provided by the CRF component, so that their predicted scores can be compared; finally, the sequence with the highest score (among a set of *N* candidate sequences) is selected as the final answer. This is why the reranker receives the output produced by CRF++ in the form of the 10 most likely predictions for every sentence.

4.3 A distributional semantics approach

Distibutional semantics is a well known method to represent the meaning of a single word as a context vector (Schütze 1998). However, our algorithm must calculate a context vector for the multiple words in an expression whether they are continuous or discontinuous (see §5.7). This raises new questions about the optimal way to take the co-occurences into account in the algorithm. To the authors' knowledge, such questions have not been previously studied.

In order to calculate the context vector, we count the words co-occurring with the MWE in a *reference corpus* (see §4.4). Given a candidate MWE identified by the CRF, the reference corpus is searched for every occurrence of this MWE. This consists in matching the words which compose the expression; because MWEs are not continuous in general, the matching only requires that the words appear in the same order in a sentence, i.e. allows other words to appear between the words of the expression.[8] As a consequence, false positive matches may happen, in particular if the words of the MWE appear in a sentence by chance, without any direct relation between them (neither syntactic or semantic).

[8] This implies that ambiguities can arise if the same word is used several times in a sentence. In such cases, the matching always selects the shortest possible distance between the first and the last word if the MWE.

Once an occurrence of the MWE expression is identified, the words which appear within a fixed-size window around its lexicalized components are counted as its co-occurrences. The number of times a given word co-occurs with the MWE across all the sentences in the reference corpus is recorded; by doing this for every word which co-occurs with a component of an MWE, we obtain a vector which represents the meaning of the MWE. However this method is traditionally used with single words, and its adaptation to sequences of words raises new questions. This is why we propose several options, as detailed below, that determine how the co-occurences are extracted and counted. The combinations of these options offer multiple possibilities that we analyse in §5.8.

- *IncludeWordsInExpression (WIE):* This option determines whether to add the actual expression words to the context vector as contexts of other components or to exclude them, when they fall within the scope of the window.
- *MultipleOccurencesPosition (MO):* This option determines whether or not to count multiple occurrences of a word within the window scope. Such a word can be part of the actual expression or not, depending on the first option.
- *ContextNormalization (CN):* This option determines whether the frequencies of the co-occurrences are normalized for every occurrence of the expression, i.e. divided by the number of co-ccurrences found.[9] This is meant to account for the differences in the number of context words across different sentences or expressions (since a longer expression generally has more words in its context window).

Table 4 illustrates the impact of these options when applied to the following example, in which the words of the expressions are in bold:

(1) French (Indo-European; FR training data)
Les gens ne ***se rendent*** *pas bien* ***compte*** *du coût énorme de*
The people NEG self give not well account of.the cost enormous of
l' opération.
the operation.
'People do not fully **realize** the enormous cost of the operation.'

[9] Otherwise absolute frequencies are used at the level of a single expression. In both cases a different stage of normalization is carried out once all the occurrences of the expression have been collected, where the values are divided the number of occurrences.

Table 4: Example of how context words are counted (for one occurence of the expression). A context window of size 2 is assumed on both sides. IWIE and MO represent the options *IncludeWordsInExpression*, *MultipleOccurrences*, respectively. The values are represented in the case where *ContextNormalization* is false, i.e. they are not normalized; in the case where *ContextNormalization* is true, every value is divided by the sum of the values in the row.

IWIE	MO	gens	ne	**se**	**rendent**	pas	bien	**compte**	du	coût
False	False	1	1	0	0	1	1	0	1	1
True	True	1	2	1	1	3	2	0	1	1
False	True	1	2	0	0	3	2	0	1	1
True	False	1	1	1	1	1	1	0	1	1

4.4 Third-party reference corpus: Europarl

We use Europarl (Koehn 2005) as reference corpus, because it is large and conveniently contains most of the languages addressed in the shared task data. However there is no Europarl data for Farsi, Maltese and Turkish. This is why these languages are excluded from this part of the process. For each of the 12 remaining languages, we use only the monolingual Europarl corpus, and we tokenise it using the generic tokeniser provided by the shared task organisers. However this tokeniser was not necessarily used for the shared task data, because each language team was free to use their own tools or to use existing pre-tokenised data.[10] Therefore, discrepancies are to be expected between the tokenisation of the shared task corpus and the one performed on Europarl. Additionally, Europarl consists of raw text, so the reranker cannot use the morpho-syntactic information (POS tags and lemmas) provided with the input data.

4.5 Features

The core component of the reranker computes a set of feature values for every sequence proposed by the CRF; these features are later used for training or applying the decision tree model. First, these features include a few simple values which are either available directly from the CRF output or easily computed from the reference corpus:

- The confidence score assigned by the CRF component to the sequence (i.e. the probability of the sequence according to the CRF);

[10] For instance, the French dataset originates from several existing corpora; their tokenisation follows language-specific patterns which cannot be obtained with a generic tokeniser.

- The number of expressions labeled by the CRF in the sequence;
- The minimum/mean/maximum number of words in the expression, over the expressions of the sequence;
- The minimum/mean/maximum frequency of the expression in the reference corpus, over the expressions of the sequence.

The original intuition behind the semantic reranker is to calculate features which give indications about the level of compositionality between the words in the expression. The underlying assumption is that such features might help detect at least non-compositional expressions. In the past this idea has been used successfully to detect collocations among two-words sequences: in Maldonado & Emms (2011), for every two-words collocation *xy* word vectors are computed for the word *x*, the word *y* and the full sequence *xy*; the cosine similarity is used to compare (1) the vector for *x* with the vector for *xy* and (2) the vector for *y* with the vector for *xy* . In a compositional expression both scores are expected to be high, because the semantic overlap between the full expression and each of its words is high, as opposed to a non-compositional expression. Hence the average similarity score between (1) and (2) can be used as a measure of the compositionality of the pair.

This idea is generalized to the case of VMWEs of any length by comparing different parts of the expression against the full candidate expression; we call *pseudo-expressions* these "parts of the expression". Every pseudo-expression extracted from the candidate expression is analyzed as if it was a regular candidate expression: the reference corpus is searched to identify all its occurrences, then a context vector is built, as described in §4.3. Examples of pseudo-expressions based on the expression *avoir sa place* ('have their place') are provided below.

Four different similiarity measures have been implemented for comparing pairs of context vectors: Jaccard index, min/max similarity, and cosine similarity with or without inverse document frequency (IDF) weighting. Additionally to the semantic context vectors comparison, the frequencies of the pseudo-expressions are compared with respect to the frequency of the full MWE: this feature is the ratio of the full expression frequency divided by the pseudo-expression frequency. Finally every candidate MWE is compared against every other candidate MWE in the 10 predicted sequences. Thus, for each of the three groups of features below, the frequency ratio and the similarity score obtained between the context vectors of the pseudo-MWEs and the full MWE are added as features.

- Features comparing each pseudo-MWE consisting of a single word of the MWE against the full MWE. Example: for the candidate expression *avoir sa* *place* ('have their place'), three comparisons are performed: *avoir* vs. *avoir sa place*, *sa* vs. *avoir sa place*, *place* vs. *avoir sa place*.

- Features comparing each pseudo-MWE consisting of the MWE minus one word against the full MWE. Example: for the candidate expression *avoir sa place*, three comparisons are performed: *sa place* vs. *avoir sa place*, *avoir place* vs. *avoir sa place*, *avoir sa* vs. *avoir sa place*.

- Features comparing one of the other MWEs found in the 10 predicted sequences[11] against the current MWE.[12]

The main difficulty in representing a predicted sequence as a fixed set of features is that each sentence can contain any number of MWEs, and each MWE can contain any number of words. We opted for a simple method which consists in "summarising" any non-fixed number of features with three statistics: minimum, mean and maximum. For instance, the similarity scores between each individual word and the corresponding MWE (n scores, where n is the number of words) are represented with these three statistics, each computed across these n scores. Similarly, if the sequence contains m expressions, a feature f has m values $f_1, \ldots, f_m$, with each value f_i corresponding to one expression; here again the minimum, mean and maximum are calculated across these m values, i.e. every expression-level feature f is converted into three features f_{min}, f_{mean} and f_{max} in the final set of sequence-level features.

4.6 Supervised regression and cross-validation process

As explained above, the reranker has to assign a score to each of the 10 sequences provided by the CRF component, in order to select the highest one. We use regression, rather than classification, because a categorical answer would cause some sentences to have either no positive answer or multiple positive answers in its set of predicted sequences, thus making the decision impossible.

[11] This includes the 9 other sequences as well as the other candidate expressions in the current one, if any.

[12] The last group is not meant to measure compositionality of the expression. The rationale is that such features might help eliminate candidate expressions which are very different from the other candidates, under the hypothesis that likely candidates tend to share words together (such features are unlikely to help with sentences which contain several expressions). As explained in §5.5, this group of features turned out to be the least useful.

In training mode, an instance (i.e. sequence) is assigned the score 1 if it corresponds exactly to the sequence in the gold standard, or 0 otherwise.[13] Since the 10 candidate sequences are all different, there should always be only one correct sequence. This way, the regression model assigns scores in $[0, 1]$ to every instance, with the highest values expected to be more likely correct answers. As regression model, we use the Weka (Hall et al. 2009) implementation of decision trees regression (Quinlan 1992): the final score is determined by the decision tree rules, then by a linear combination of the features values. This choice has the advantage of simplicity, but also of the interpretability of decision trees models. Of course, other regression models could be considered as well.

Because of the two-components approach (CRF and reranker), the training process requires special care. In order to train the reranker with the kind of data that it will receive in testing mode, predictions from the CRF are needed. The testing data cannot be used for this, which is why we use 5 fold cross-validation on the training data: on each of the five 20% subsets of the data, a model trained from the 80% data left is applied. This way the reranker can be fed with real predictions for the full training data, including the classification errors of the CRF. If necessary, the cross-validation process can be repeated, for instance to tune the parameters of the reranker. Otherwise, the reranker can simply be trained with the full training set of predictions, then applied to the test set (after the CRF predictions have been predicted on this test set).

5 Results and analyses

In this section we present detailed results of our system and analyze how the reranker helps improving performance with respect to various factors. All the experiments presented in this section have been carried out using the official PARSEME shared task 2017 training and test data. Unless otherwise stated, we use the same “standard” configuration of options for the reranker throughout the experiments (e.g. context window size, minimum frequency, etc.).

5.1 Results

Table 3 shows the performance of both the CRF component and the semantic reranker by language, as well as the improvement brought by the reranker. Despite differences by language, all but one language (Italian) show a significant

[13] It might happen that none of the 10 sequences corresponds to the gold sequence; in such cases all the instances are left as negative cases.

increase in MWE-level F-score, with a macro-average F-score improvement of +11.9%. The increase in token-level F-score is much smaller, with even a decrease in a few languages; the macro-average token-level F-score improvement is only +1.1%. This means that the reranker does not drastically change the expressions predicted by the CRF (hence the little improvement at the token level), but instead tends to fix the proposed expressions by finding their exact boundaries, thus bringing the MWE F-score closer to the token F-score. This is because the top 10 CRF predictions tend to be variants of one another, rather than drastically different labelings; they frequently focus on the same part(s) of the sentence, varying only by labeling one or two words differently; thus the reranker seldom introduces a new expression that the top prediction would have missed, but can select a more likely variant among the remaining 9 predictions. This hypothesis is also backed by the observation that the increase in precision is larger in general than the increase in recall: the reranker mostly follows the top CRF prediction and possibly fixes it, hence turning a false positive instance into a true positive. These observations are consistent with the design of the system and validate the reranking approach in general.

5.2 Error analysis methodology

The performance of the reranker can be evaluated straightforwardly using the official evaluation measures, as presented above; these measures are useful to compare against other systems or between datasets. However, in order to get a clear understanding of how the system works, we also look at the different combinations of error status between the CRF component and the semantic reranker: the CRF is said to be right if and only if it ranks the actual (gold standard) answer as its top prediction; similarly, the reranker is right if and only if it assigns the top score to the actual answer. Thus the four following categories are defined:

- **RR** stands for right-right, which means that the CRF component ranked the right answer as first sequence (first **R**) and the reranker kept it as the final answer (second **R**).
- **WR** stands for wrong-right: the CRF answer was wrong, but the reranker successfully selected an alternative answer.
- **RW** stands for right-wrong: in this case the reranker mistakenly changed the CRF answer, which was correct in the first place.
- **WW** stands for wrong-wrong: the CRF answer was wrong, and the reranker either kept it or changed it to another wrong answer.

These four categories cover all the cases, except when the correct answer is not present in the 10 most likely sequences that the CRF component provides. For this case we use the special label **GoldNotIn10**.

It is worth noticing that the reranker works at the sentence level (as opposed to the expression level or the token level). This is why these categories apply to complete sentences, in accordance with the design of the two-components system. In particular, the number of expressions in a sentence is not taken into account in this categorization. As a consequence, sentences which contain no expression at all are considered as important as sentences which contain one or multiple expressions.

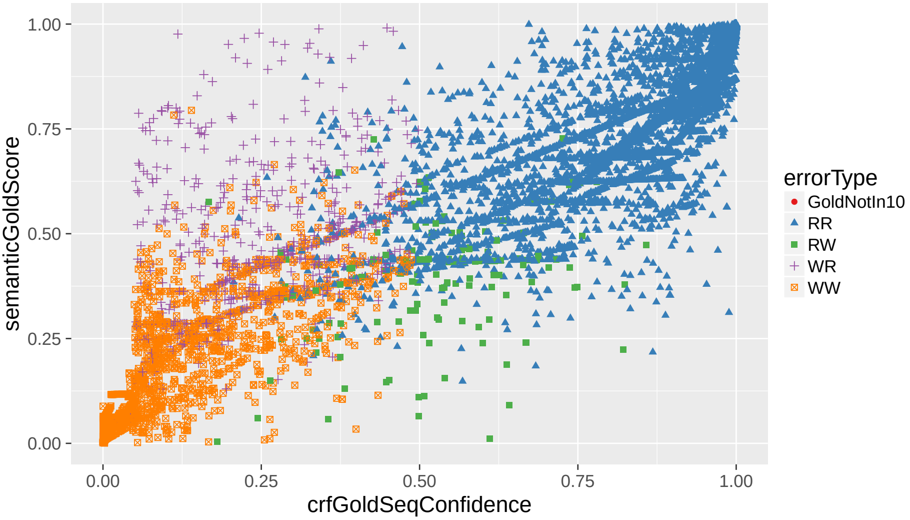


Figure 1: Reranker score w.r.t CRF confidence for the gold sequence in every sentence, by error type (all languages together). A sentence is represented in the right half (resp. left half) if the CRF assigned a high confidence (resp. low) to its gold sequence, i.e. the CRF answer is correct (resp. incorrect). Similarly, a sentence appears in the top half (resp. bottom half) if the reranker assigned a high score (resp. low) to its gold sequence, i.e. the reranker answer is correct (resp. incorrect).[14]

Figure 1 gives an overview of how the reranker improves performance over the CRF predictions. Every point in this graph represents a sentence, positioned according to the CRF confidence (X axis) and reranker final score (Y axis) for its gold sequence. This way, if the CRF finds the right answer for a sentence,

[14] Remark: Category *GoldNotIn10* is not visible on this graph, since in such cases the gold sequence cannot be assigned a CRF confidence nor a reranker score.

i.e. the gold sequence obtains the highest confidence among the 10 sequences, it is represented in the right half of the graph, and conversely for wrong answers. If the reranker finds the right answer, it assigns a high score to the gold sequence, so the sentence appears in the top half, and conversely. This explains why the four error types appear mostly clustered each in its own quadrant: the top right quadrant contains sentences for which the correct sequence is recognized by both the CRF and the reranker (hence in the **RR** error category), and the bottom left contains sentences for which neither finds the right answer (**WW**). The last two quadrants are the interesting ones, since this is where the reranker changes the CRF prediction to another sequence: in the top left quadrant, the **WR** points correspond to successful changes, whereas the **RW** cases, in the bottom right quadrant, correspond to mistakes introduced by the reranker.[15] It can be observed that the former category outnumbers the latter, thus confirming the positive contribution of the reranker.

5.3 Insight: what the reranker actually does

The vast majority of the sentences (85.3% of a dataset in average) fall into the *RR* category, i.e. the reranker simply confirms the correct CRF answer. The *WW* cases account for 7.6% of the sentences, and the *GoldNotIn10* cases for 4.4%. The reranker actually changes only 2.7% of the answers, and when it does, it does it correctly 81.5% of the time (2.2% *WR*, 0.5% *RW*).

Figure 2 shows how the positions of the sequences selected by the reranker are distributed. The reranker strongly favors top positions for its selected sequence; more precisely, as the position of the sequence decreases, the number of sentences for which this position is selected decreases exponentially (this is why a logarithmic scale is used in Figure 2). This trend is regular from the top position, which is selected 92.5% of the time, down to the 9th position, selected in only two cases.[16] This shows that increasing the number of candidate sequences supplied by the CRF (10 in all our experiments) would not improve the performance of the reranker, since it seldom selects a sequence associated with a low CRF confidence (the importance of the CRF confidence as a feature is shown more clearly

[15] This graph can give the impression that one could easily prevent **RW** mistakes (bottom right quadrant) by accepting any CRF answer with high confidence, but this is due to the fact that only the gold sequence is represented here. Thus, for cases where the gold sequence has low confidence, some other (wrong) sequence has high confidence. Therefore selecting the highest confidence sequence would simply prevent any reranking to happen, in particular for the CRF mistakes which can be fixed.

[16] The small rebound in position 10 is not significant, as it represents only 3 cases.

in §5.5 below). Finally the fact that the reranker makes more correct changes than mistakes is confirmed in this graph again, by observing that the number of *WR* cases is higher or equal than the number of *RW* cases at every position.

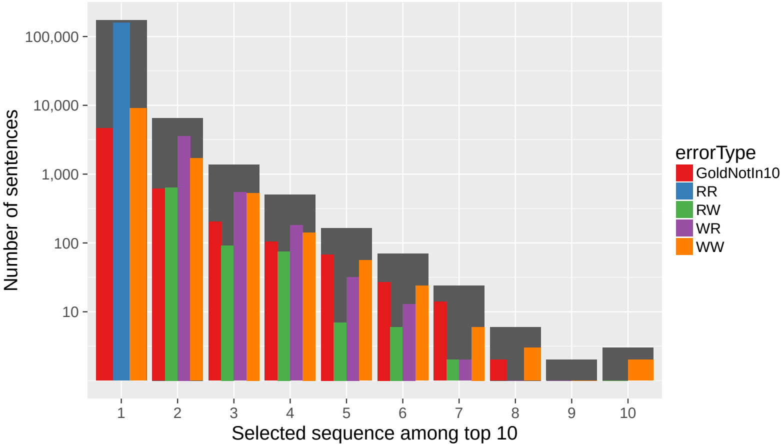


Figure 2: Distribution of the position selected by the reranker (logarithmic scale, all languages together). For every position, the large dark grey bar shows the total number of sentences, while the coloured bars show the number of sentences for every possible error type.[17]

The distribution of errors is not uniform over the data, and the number of expressions in a sentence is one of the most obvious factors: For example, Figure 3 shows that 96% of the sentences with no expression in the gold sequence are correctly identified by both the CRF and the reranker (*RR*), whereas only 9% of the sentences with three expressions are. The difference is mostly due to the proportion of sentences for which the CRF does not propose the right answer in the top 10 candidate sequences (*GoldNotIn10*), which is naturally higher in the more complex cases with multiple expressions in the sentence. Figure 3 also shows that the reranker is more useful with the sentences which contain one or two expressions (with 7.4% and 7.5% of changes, respectively), because these contain more mistakes to correct compared to sentences with no expressions, and contain more possibilities to correct the mistakes compared to sentences with three (or more) expressions (since the reranker cannot correct anything in the *GoldNotIn10* cases).

[17] This means that the dark grey bar represents the sum of the coloured bars, although the logarithmic scale makes this difficult to observe. Since the sequence selected by the CRF is the top one, position 1 is the only way for both the CRF and the reranker answers to be correct, thus it contains all the *RR* cases. Similarly, it cannot contain any *WR* or *RW* case, by definition.

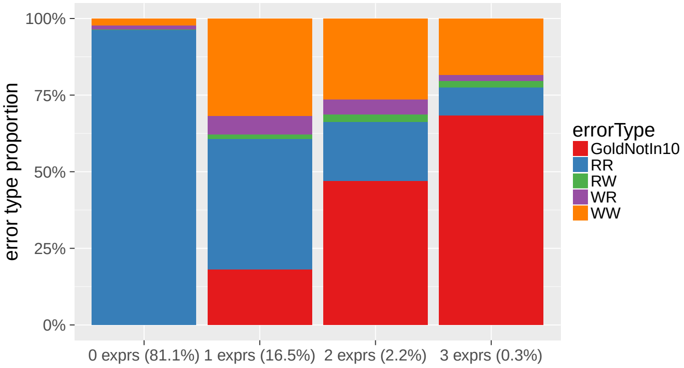


Figure 3: Proportion of error type by number of expressions in the gold sequence (all languages). Sentences with more than 3 expressions were discarded (30 cases, 0.1% of the data). Example: 16.5% of the sentences contain exactly one expression; among these, 18% belong to the *GodNotIn10* category, and 41%, 1%, 6%, 32% belong respectively to categories *RR, RW, WR, WW*.

5.4 Reranker-specific evaluation

Based on these error types, a new reranker-centered evaluation method can be defined. Indeed, using this categorization, the reranker can be seen as a binary classifier: from this perspective, the job of the reranker is to detect the sentences for which the CRF answer was wrong, and leave the right ones as they are. Thus, for every sentence, either the answer is changed (positive instances) or not (negative instances). With this idea in mind, the four main categories can be translated to the standard true/false positive/negative categories in a straightforward way: if the reranker changes the answer correctly (**WR**), the instance is a true positive; if it changes the answer incorrectly (**RW**), the instance is a false positive, and similarly for the last two cases: **RR** and **WW** correspond respectively to true negative and false negative instances (the former was rightly not changed, and the latter should have been changed). Thanks to this interpretation, performance measures like precision, recall and F-score can be calculated for the semantic reranker, independently from the performance of the CRF component.[18] An example of using such performance scores is given in Table 5.

[18] The **GoldNotIn10** category is ignored when calculating these performance measures for the reranker, consistently with the idea of evaluating the reranker on its own: since these cases are impossible to solve, they should not be taken into account.

Table 5: Reranker-specific performance by number of expressions in the gold sequence (all languages). P/R/F stands for precision/recall/f-score; the macro-average performance is the average over languages (datasets with NaN F-scores are ignored).

Nb exprs	Macro-average			Micro-average		
	P	R	F	P	R	F
0	82.1	40.6	53.8	83.9	38.3	52.6
1	79.5	14.5	23.5	81.1	16.1	26.8
2	64.3	20.7	30.4	66.0	15.9	25.6
3	75.0	20.8	31.0	50.0	10.0	16.7
all	74.9	18.9	29.1	81.4	22.4	35.2

In the rest of this section we do not detail results by dataset, since the large number of languages and the dataset particularities would make it harder to recognize the general patterns related to the reranker. Nevertheless, it is important to keep in mind that such dataset-specific factors exist even though they are not shown. Additionally, the inequal size of the datasets clearly favors large datasets over small ones when grouping all the sentences together. This is why we present both the micro-average and macro-average performance whenever relevant, like in Table 5.

5.5 Feature analysis

As explained in §4, the reranker uses various kinds of features to determine the likelihood of the expressions in a sequence. Table 6 gives a brief overview of the impact of the different groups of features on performance, expressed as the F-score, computed from the micro-average precision and recall of the reranker alone (column 1, see §5.4), macro-average of the same over languages (column 2) and expression-level F-score (column 3, official evaluation on the full system).

First, it should be observed that the reranker relies heavily on the CRF confidence to make its decisions: without this feature, the performance drops to a ridiculously low level. Nevertheless, the reranker needs additional features in order to improve over the CRF alone (since otherwise the best it can do is to always agree with the CRF top prediction). A few simple features allow a large gain in performance (*SF* in Table 6: number of candidate expressions in the sequence, min./mean/max. number of words by expression and frequency in the reference corpus). Adding more complex features based on frequency and semantic similarity of the candidate expression allows the reranker to make even better decisions: the micro F-score reaches 35.6 with the best combination, compared to 32.0 with only *SF*. Among these features, frequency and semantic similarity features seem to be equally useful, and combining both of them achieves the best performance; the only group of features which performs poorly (and is apparently even counter-productive) is the one where the candidate expression is compared against all other candidates in the sentence (group III in Table 6).

Table 6: Performance of the reranker using various subsets of features (percentages). Simple features *(SF)* represents the number of expressions and words as well as the frequency in the reference corpus; Groups *I*, *II* and *III* represent respectively *single word*, *expression minus one word* and *alternative expressions features*; Groups *a* and *b* represent respectively frequency features and semantic similarity features (see §4); NaN values correspond to cases where the precision and/or recall is zero.

Features	micro F-score reranker	macro F-score reranker	macro F-score MWE-level
baseline: CRF answer	NaN	NaN	48.2
all but confidence	00.6	NaN	09.6
confidence + SF (*)	32.0	NaN	53.1
(*) + Ia, Ib	34.9	29.7	53.4
(*) + IIa, IIb	34.3	29.0	53.4
(*) + IIIa, IIIb	33.2	27.8	53.6
(*) + IIa, IIb, IIIa, IIIb	34.4	29.1	53.7
(*) + Ia, Ib, IIIa, IIIb	34.2	29.0	53.4
(*) + Ia, Ib, IIa, IIb	**35.6**	**30.2**	53.7
freq. only: (*) + Ia, IIa, IIIa	33.9	28.7	53.7
sem. sim. only: (*) + Ib, IIb, IIIb	34.0	28.4	53.7
all features, with mean only	34.8	29.3	53.6
all features, with min/mean/max	35.2	30.1	**53.9**

5.6 Analysis: impact of the coverage in the reference corpus

Some candidate expressions might not be found in the reference corpus, either because they are simply rare or because of tokenization/lemmatization issues (see §4.4). In fact, the coverage rate of the expressions in Europarl is quite low: for 36.2% of the sentences containing at least one expression, the expression(s) they contain are not found at all in the reference corpus. Figure 4 shows that the error type depends greatly on whether the expression appears in the reference corpus or not. First, the CRF finds the right answer much more often when the expression is covered (*RR* + *RW* = 73%) than when it is not (*RR* + *RW* = 30%). This can be explained by the fact that the least frequent expressions are hard to identify by the CRF, and they also tend not to appear in the reference corpus. While this implies that the reranker has potentially more mistakes to fix in the zero-coverage cases, it actually changes fewer sentences (4.4% against 12.5% for covered expressions), resulting in a very low recall (3.5% against 37.7%); the precision is also lower, with 67% against 81%.

As explained in §5.5, the reranker can work with only a small set of "simple features", which is why its performance in the zero-coverage case is lower but positive. Clearly, the more advanced features which rely on the reference corpus increase performance. This means that the coverage in the reference corpus is critical for the reranker to give its best results, but our current implementation of the system is probably not optimized from this point of view; in particular, the tokenization process might not be identical between the input data (where tokenization is provided) and the reference data (for which we apply a generic tokenizer), and the reference corpus is not lemmatized (see §4.4). This might explain why the recall is low with the current implementation. Ideally, a larger corpus would also help by covering a broader range of expressions; but there are very few such large datasets available for multiple languages.

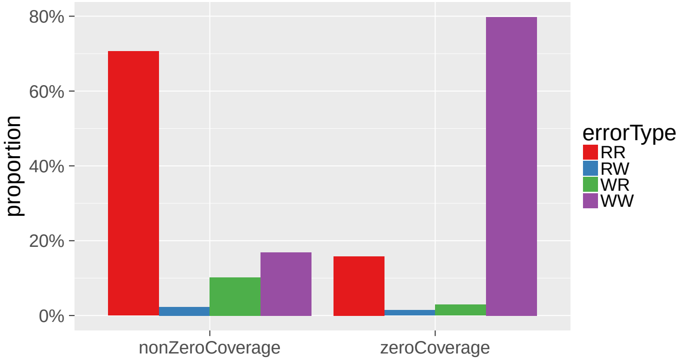


Figure 4: Sentences error types by coverage/non-coverage of their expressions in the reference corpus (all languages). Example: 70% of the sentences containing expressions which appear in the reference corpus belong to the *RR* category, whereas only 16% of the sentences with expressions not covered in the reference corpus belong to this category.

5.7 Analysis: continuous vs. discontinuous expressions

Verbal multiword expressions can be classified as either continuous or discontinuous: in the former case, the expression appears as a sequence of contiguous words, as in the following idiomatic expression:

(2) French (Indo-European; FR training data)
Celles-ci peuvent à tout moment jeter l' éponge.
they.3.fem.pl can at any time throw the sponge
'They.3.fem.pl can give up at any time.'

In the latter case, the expression appears with words inserted between its lexicalized components, e.g.:

(3) French (Indo-European; FR training data)
J'ai ***obtenu*** *de Jean-Marie Molitor [...]* ***la permission*** *de publier.*
I have obtained from Jean-Marie Molitor [...] the permission to publish
'Jean-Marie Molitor gave me the permission to publish.'

It is worth noticing the same lexicalized components might appear sometimes as a continuous expression and other times as a discontinuous expression.

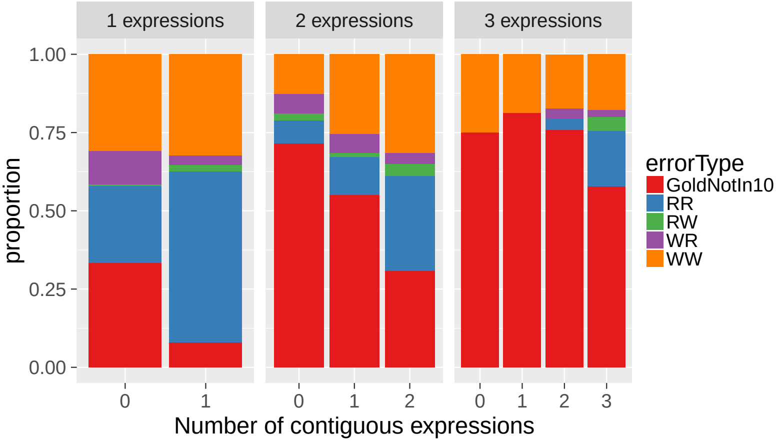


Figure 5: Proportion of error type by number of continuous expressions in the sentence, for sentences containing 1, 2 or 3 expressions (all languages). Example: among sentences which contain two expressions, the proportion of *RR* cases is 7% (respectively 12%, 30%), when there are no (resp. 1, 2) continuous expressions among the two.

Figure 5 shows the impact of continuity in expressions: *RR+RW* cases increase with the number of continuous expressions for any number of expressions in the sentence, which means that the more there are continuous expressions, the better the performance of the CRF (there are also much less *GoldNotIn10* cases). Interestingly, however, the semantic reranker follows an opposite trend: the less there are continuous expressions (i.e. the more there are discontinuous expressions), the better its performance: not only it fixes more mistakes from the CRF (better recall), it also fixes them better (better precision).[19] The most likely explanation for these observations is that the CRF suffers from a "sequential bias" which makes it less good with discontinuous expressions, whereas such cases are not any harder for the semantic reranker, which is "sequence-agnostic". In our opinion, this point clearly illustrates the complementarity of the two components.

5.8 Analysis: context vectors options

In §4.3, we presented several options which modify the way words which cooccur with the expression are taken into account in the MWE context vector. Table 7 shows the impact on performance of these options. Although there is no decisive pattern in these results, the absence of context-level normalization (CN)

[19] Except in the case of three expressions with zero or one continuous; this is probably due to the low number of cases.

Table 7: Performance of the reranker depending on context vector options. Left: Overall micro-average performance; right: F-score for continuous/discontinuous cases, for sentences with one expression exactly. CN, IWIE and MO represent the options presented in §4.3, respectively: *ContextNormalization, IncludeWordsInExpression, MultipleOccurrences.* P/R/F stands for precision/recall/f-score.

Options			Micro-average		
CN	IWIE	MO	P	R	F
0	0	0	82.9	21.9	34.6
0	0	1	81.7	**22.4**	**35.2**
0	1	0	82.5	22.1	34.8
0	1	1	81.6	21.7	34.3
1	0	0	**83.3**	21.8	34.5
1	0	1	82.8	21.7	34.4
1	1	0	80.9	21.8	34.4
1	1	1	81.4	**22.4**	**35.2**

Options			F-score	F-score
CN	IWIE	MO	Continuous	Discontinuous
0	0	0	14.7	**41.4**
0	0	1	**16.8**	40.3
0	1	0	14.7	40.9
0	1	1	15.9	39.6
1	0	0	15.2	39.8
1	0	1	15.2	39.6
1	1	0	14.0	41.0
1	1	1	14.8	40.8

as well as allowing multiple occurrences of the same word to be counted multiple times (MO) obtain slightly higher performance in general. Looking at the effect of these options on continuous/discontinuous expressions, including expressions words in the context window (IWIE) has a negative effect on all the cases, except if CN is selected but only in the discontinuous case. In fact, an interesting pattern can be observed in the continuous/discontinuous table: the combinations of options which make the F-score increase for continuous expressions tend to make the F-score in the discontinuous case decrease, and conversely. This is confirmed by a moderate negative Pearson's correlation coefficient of -0.56 and a high negative Spearman's rank correlation coefficient of -0.79. Here again the differences in performance are too moderate to conclude decisively; however this point suggests that there is a trade-off between the continuous and discontinuous cases, and this trade-off might be controlled through these options to some extent. This means that the system could potentially be tuned to favor one or the other case.

6 Conclusion and future work

In this chapter we described a two stages approach for identifying VMWEs, based on sequence labeling with CRF followed by reranking of the CRF candidates. We showed experimentally that the reranker significantly improves the performance of the system, with in average a 12% F1-score improvement over using the CRF component alone. Then we proceeded to analyze how the reranker works.

We found that the reranker follows the CRF quite closely, rarely selecting a candidate with a low CRF confidence, and selecting the CRF top prediction in 92.5% of the cases. Consistently with this observation, when the reranker diverges from the top CRF prediction, it does so correctly with high confidence (81.5% of correct answers among the changed predictions).

The contribution of the reranker is more important with the sentences which contain one or two expressions: sentences with no expressions are almost always correctly detected by the CRF alone, whereas the cases with 3 or more expressions are so complex that the CRF does not usually provide the right candidate among its top 10 predictions, leaving the reranker unable to fix these errors. The coverage of the MWE in the reference corpus is another major factor of performance for the reranker, with the recall dropping 10 times for expressions which do not appear in the reference corpus. Finally the last important finding of our study is that the reranker seems to compensate the CRF sequential bias: while the latter performs better with continuous MWEs, the reranker performs comparatively better with discontinuous cases.

The semantic reranker presented in this chapter is a proof-of-concept version, and new perspectives emerge from the fact that the combination of a CRF component with this reranker proves fruitful for detecting MWEs. There are a few obvious areas in which the reranker could be improved, especially in the tokenization/lemmatization part, and it is likely that choosing a more adequate reference corpus would help. But there are also deeper questions which are worth studying:

- Computing a context vector for a MWE is not as trivial as for a single word (especially if the words in the expression are not continuous), and the authors are not aware of any standard approach for this. While several options were tested, this question deserves to be studied on its own.
- In the same line of thought, the current state of the art in distributional semantics is based on word embeddings (Legrand & Collobert 2016). Here again, the authors are not aware of any software able to retrieve word embeddings for multiple (possibly discontinuous) words.
- Would it be possible for the reranker to work at the expression level instead of the sentence level? Indeed, the current method used to "merge" multiple expressions in a sentence is likely to lose some information in the process. One could also ask whether some of the information currently computed by the reranker could be fed directly into the CRF. An iterative process could even be considered, perhaps allowing to refine the quality of the predicted expressions over iterations.

Acknowledgements

We are grateful to the editors and anonymous reviewers for their valuable comments. Lifeng Han also thanks Qun Liu for his continued support.

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

The graphics in this chapter were created with R (R Core Team 2012), using the ggplot2 library (Wickham 2009).

Abbreviations

CRF	conditional random fields	RW	right-wrong
IDF	inverse document frequency	VMWE	verbal mutiword expression
MWE	multiword expression	WR	wrong-right
NLP	Natural Language Processing	WW	wrong-wrong
RR	right-right		

References

Attia, Mohammed, Lamia Tounsi, Pavel Pecina, Josef van Genabith & Antonio Toral. 2010. Automatic extraction of Arabic multiword expressions. In *Proceedings of the COLING Workshop on Multiword Expressions: From Theory to Applications* (MWE '10), 19–27. Association for Computational Linguistics.

Bar, Kfir, Mona Diab & Abdelati Hawwari. 2014. Arabic multiword expressions. In Nachum Dershowitz & Ephraim Nissan (eds.), *Language, culture, computation. Computational Linguistics and Linguistics: Essays dedicated to Yaacov Choueka on the occasion of his 75th birthday, Part III*, 64–81. Berlin: Springer. DOI:10.1007/978-3-642-45327-4_5

Blunsom, Phil & Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 164–171. Sydney, Australia: Association for Computational Linguistics. DOI:10.3115/1610075.1610101

Boukobza, Ram & Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 468–477. August 6-7, 2009.

Constant, Matthieu, Anthony Sigogne & Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 204–212. Association for Computational Linguistics.

Green, Spence, Marie-Catherine de Marneffe, John Bauer & Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 725–735. Association for Computational Linguistics. http://www.aclweb.org/anthology/D11-1067.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39(1). 195–227. DOI:10.1162/COLI_a_00139

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann & Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11(1). 10–18.

Han, Aaron Li-Feng, Derek F. Wong & Lidia S. Chao. 2013. Chinese named entity recognition with conditional random fields in the light of Chinese characteristics. In Mieczysław A. Kłopotek, Jacek Koronacki, Małgorzata Marciniak, Agnieszka Mykowiecka & Sławomir T. Wierzchoń (eds.), *Language processing and intelligent information systems. iis 2013*, vol. 7912 (Language Processing and Intelligent Information Systems), 57–68. Berlin Heidelberg: Springer.

Han, Aaron Li-Feng, Xiaodong Zeng, Derek F. Wong & Lidia S. Chao. 2015. Chinese named entity recognition with Graph-based semi-supervised learning model. In *Proceedings of the 8th SIGHAN Workshop on Chinese Language Processing* (SIGHAN-8), 15–20. Association for Computational Linguistics & Asian Federation of Natural Language Processing. July 30-31, 2015.

Hosseini, Mohammad Javad, Noah A. Smith & Su-In Lee. 2016. UW-CSE at SemEval-2016 task 10: Detecting multiword expressions and supersenses using double-chained conditional random fields. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, 931–936. http://aclweb.org/anthology/S/S16/S16-1143.pdf. June 16-17, 2016.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, 79–86. Phuket, Thailand.

Lafferty, John D., Andrew McCallum & Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (ICML '01), 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=645530.655813.

Legrand, Joël & Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions* (MWE '16), 67–71. Association for Computational Linguistics. http://anthology.aclweb.org/W16-1810.

Maldonado, Alfredo & Martin Emms. 2011. Measuring the compositionality of collocations via word co-occurrence vectors: Shared task system description. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 48–53. http://dl.acm.org/citation.cfm?id=2043121.2043130.

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 114–120. Association for Computational Linguistics. DOI:10.18653/v1/W17-1715

Maldonado, Alfredo & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–175. Berlin: Language Science Press. DOI:10.5281/zenodo.1469557

Quinlan, J. Ross. 1992. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, 343–348.

R Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/. ISBN 3-900051-07-0.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014a. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014), 1792–1797. Doha, Qatar. http://aclweb.org/anthology/D/D14/D14-1189.pdf.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014b. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2014), 472–481. Gothenburg. http://aclweb.org/anthology/E/E14/E14-1050.pdf.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies* (NAACL '15), 977–983. http://dblp.uni-trier.de/db/conf/naacl/naacl2015.html#SalehiCB15.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* 2(1). 193–206. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/281.

Schütze, Hinrich. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1). 97–123. http://dl.acm.org/citation.cfm?id=972719.972724.

Sinclair, John. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sun, Xiao, Chengcheng Li, Chenyi Tang & Fuji Ren. 2013. Mining semantic orientation of multiword expression from Chinese microblogging with Discriminative Latent Model. In *Proceedings of the 2013 International Conference on Asian Language Processing*, 117–120. http://dblp.uni-trier.de/db/conf/ialp/ialp2013.html#SunLTR13.

Tsvetkov, Yulia & Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING-10), 1256–1264. http://dl.acm.org/citation.cfm?id=1944566.1944710.

Venkatapathy, Sriram & Aravind K. Joshi. 2006. Using information about multiword expressions for the word-alignment task. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (MWE '06), 20–27. http://dl.acm.org/citation.cfm?id=1613692.1613697.

Vincze, Veronika, István Nagy T. & Gábor Berend. 2011. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 289–295. RANLP 2011 Organising Committee. http://aclweb.org/anthology/R11-1040.

Wickham, Hadley. 2009. *Ggplot2. Elegant graphics for data analysis.* New York: Springer-Verlag.

Chapter 7

A transition-based verbal multiword expression analyzer

Hazem Al Saied
ATILF UMR 7118, Université de Lorraine/CNRS

Marie Candito
LLF UMR 7110, Université Paris Diderot/CNRS

Matthieu Constant
ATILF UMR 7118, Université de Lorraine/CNRS

We describe a robust transition-based analyzer for identifying and categorizing Verbal Multiword Expressions (VMWEs). The system was developed and evaluated using the datasets of the PARSEME shared task on VMWE identification (Savary et al. 2017).

We accommodate the variety of linguistic resources provided for each language, in terms of accompanying morphological and syntactic information. Our system produces very competitive scores, for both VMWE identification and categorization, with respect to the shared task results.

1 Introduction

We present a generic system for the identification and categorization of verbal multiword expressions (hereafter VMWEs). With respect to grammatical or nominal multiword expressions, VMWEs tend to exhibit more morphological and syntactic variation than other MWEs, if only because in general the verb is inflected and can receive adverbial modifiers. Furthermore, some VMWE types, in particular light-verb constructions, allow for the full range of syntactic variation

Hazem Al Saied, Marie Candito & Matthieu Constant. 2018. A transition-based verbal multiword expression analyzer. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 209–226. Berlin: Language Science Press. DOI:10.5281/zenodo.1469561

(extraction, coordination etc...). This renders the VMWE identification task even more challenging than general MWE identification, in which fully frozen and continuous expressions contribute to an increase in the overall performance.

Our objective was to design a data-driven system applicable to several languages, with limited language-specific tuning. We took advantage of the datasets provided within the shared task on automatic identification of VMWEs (Savary et al. 2017) to train and test our system. These datasets concern 18 languages, and consist of tokenized sentences in which VMWEs are annotated. One VMWE instance is a set made either of several tokens, potentially non-continuous, or of one single token (i.e. a multiword token, hereafter MWT, such as amalgamated verb-particle in German).[1] A VMWE may be embedded in another longer one, and two VMWEs can overlap. Each annotated VMWE is tagged with a category among Light-Verb Constructions (LVC), IDioms (ID), Inherently REFLexive Verbs (IReflV), Verb-Particle Constructions (VPC) and OTHer verbal MWEs (OTH). The datasets are quite heterogeneous, both in terms of size and of accompanying resources: 4 languages have none (Bulgarian, Spanish, Hebrew, Lithuanian), for 4 languages, morphological information such as lemmas and POS is provided (Czech, Maltese, Romanian, Slovene), and for the 10 remaining languages, full dependency parses are provided.

The system we describe in the current paper is an extension of the ATILF-LLF system (Al Saied et al. 2017), a one-pass greedy transition-based system which participated in the shared task, obtaining very competitive results (hereafter ATILF-LLF 1). The new system (ATILF-LLF 2) categorizes the VMWEs on top of identifying them, and has an extended expressive power, handling some cases of VMWE embedded in another one. Both for ATILF-LLF 1 and 2, we tuned a set of feature template for each language, relying exclusively on training data, accompanying CoNLL-U files when available, and basic feature engineering.

The remainder of the article is organized as follows: we describe our system in Section 2, and comment its expressive power as opposed to ATILF-LLF 1. We then describe the experimental setup in Section 3, and comment the results in Section 4. Section 5 is devoted to related work. We conclude in Section 6 and give perspectives for future work.

[1] The majority of annotated VMWEs are multi-token. The prevalence of MWTs varies greatly among languages. While absent from seven languages and very rare for nine other languages, they are very frequent in German and Hungarian.

2 System description

The analyzers we developed (ATILF-LLF 1 and 2) are simplified versions of the system proposed by Constant & Nivre (2016). Building on the classic arc-standard dependency parser (Nivre 2004), Constant & Nivre (2016) designed a parsing algorithm that jointly predicts a syntactic dependency tree and a forest of multiword lexical units. Their system uses both a syntactic and a lexical stack and specific transitions merge tokens to create MWEs, as proposed by Nivre (2014). We have simplified this formal apparatus keeping only the lexical stack, and predicting MWEs only.

A transition-based system applies a sequence of actions (usually called *transitions*) to incrementally build the expected output structure in a bottom-up manner. Each transition is usually predicted by a classifier given the current state of the system (namely a *configuration*).

A configuration in our system consists of a triplet $c = (\sigma, \beta, L)$, where σ is a stack containing units under processing, β is a buffer containing the remaining input tokens, and L is a set of output MWEs.

The *initial configuration* for a sentence $x = x_1, ..., x_n$, i.e. a sequence of n tokens, is represented by c_s as: $c_s(x) = ([], [x_1, \dots, x_n], \emptyset)$ and the set of *terminal configurations* C_t contains any configuration of the form $c_t = ([], [], L)$. At the end of the analysis, the identified VMWEs are simply extracted from L.

Single tokens are brought to the stack by the Shift transition, and are potentially marked as (mono-token) VMWE using the Mark as C transition, whereas trees are formed using merge transitions (cf. §2.1).

The output VMWEs are the units added to L, either by the Mark as C or Merge as C transitions. When one VMWE is embedded in another one, both VMWEs appear separately in L (which is thus redundant).[2]

2.1 Transition set

Our system uses the following transitions:

1. the Shift transition moves the first element of the buffer to the stack

   Precondition: the buffer is not empty.

2. the Reduce removes the top element of the stack

   Precondition: the stack is not empty.

[2] For instance, if we represent the binary tree in bracketing format and the categorization with a subscript, $((a, b)_{IReflV}, c)_{ID}$ represents an IReflV $a + b$ embedded within an ID $a + b + c$. Both $((a, b)_{IReflV}, c)_{ID}$ and $(a, b)_{IReflV}$ will appear in L.

3. the **White Merge** transition combines the two top elements of the stack into a single element;

   Precondition: the stack contains at least two elements.

4. five **Merge As C** transitions (where **C** stands for a VMWE category) perform a white merge, mark the resulting unit as a VMWE of category C, and add it to *L*.

   Precondition: the stack contains at least two elements.

5. In order to cope with MWTs, we added five **Mark as *C*** transitions, which mark the top stack element as a VMWE, assign to it the category C, and add it to *L*.

   Precondition: The stack is not empty and its head is a non-marked single token.[3]

Figure 1 shows the analysis of a German sentence containing a multiword token VPC embedded within an IReflV.

In the input, each token is associated with linguistic attributes (form, and depending on the data sets, lemma and POS). When a merge transition is applied, the newly created element gets its attributes using basic concatenation over forms, and over lemmas and POS tags when available.[4]

2.2 Parsing algorithm and training oracle

In all the following, at parsing time we use a greedy algorithm, starting with an initial configuration c_s, and applying in sequence the best-scoring legal transition until a terminal configuration is reached.[5]

The training phase learns the transition-scoring classifier. This is done through supervised learning, by converting the training sentences into sequences of [con-

[3] Because mono- and multi-tokens have very different linguistic properties, we preferred to distinguish transitions coping with both kinds of VMWEs. Without this restriction, as noted by a reviewer, Merge as C would be equivalent to White Merge + Mark as C. The effectiveness of this alternative solution remains to be tested.

[4] This would deserve to be improved in future experiments, with finer procedures to predict the lemmatized form and more importantly to predict the POS tag of the merged node, although in the special case of VMWE prediction the POS is verbal.

[5] Sentence analysis is composed of exactly $2n + r$ transitions, with n being the number of tokens and r the number of MWTs. Every single token not entering a VMWE requires a Shift and a Reduce, every multi-token VMWE of length m requires $2m$ transitions (m Shifts, $m-2$ White Merges, one Merge As C and one Reduce), while every MWT requires three transitions: a Shift, a Mark As C and a Reduce.

Transition		Configuration
		[ ], [Damit, müsste, man, sich, nun, herumschlagen], [ ]
Shift	⇒	[Damit], [müsste, man, sich, nun, herumschlagen], [ ]
Reduce	⇒	[ ], [müsste, man, sich, nun, herumschlagen], [ ]
Shift	⇒	[müsste], [man, sich, nun, herumschlagen], [ ]
Reduce	⇒	[ ], [man, sich, nun, herumschlagen], [ ]
Shift	⇒	[man], [sich, nun, herumschlagen], [ ]
Reduce	⇒	[ ], [sich, nun, herumschlagen], [ ]
Shift	⇒	[sich], [nun, herumschlagen], [ ]
Shift	⇒	[sich, nun], [herumschlagen], [ ]
Reduce	⇒	[sich], [herumschlagen], [ ]
Shift	⇒	[sich, herumschlagen], [ ], [ ]
Mark as VPC	⇒	[sich, herumschlagen$_{VPC}$], [ ], [herumschlagen$_{VPC}$]
Merge as IreflV	⇒	[(sich, herumschlagen$_{VPC}$)$_{IReflV}$], [ ], [herumschlagen$_{VPC}$, (sich, herumschlagen$_{VPC}$)$_{IReflV}$]
Reduce	⇒	[ ], [ ], [herumschlagen$_{VPC}$, (sich, herumschlagen$_{VPC}$)$_{IReflV}$]

Figure 1: Transition sequence for tagging the German sentence *Damit müsste man* ***sich*** *nun* ***herumschlagen*** 'With-that must-SUBJUNCTIVE one REFLEXIVE now around-struggle' ⇒ 'One would have to struggle with that' , containing two VMWEs: sich herumschlagen tagged as inherently reflexive verb (IReflV), in which herumschlagen is itself a multiword token tagged as verb-particle combination (VPC).

figuration, gold transition to apply] pairs, by using a static oracle.[6] Our static oracle returns for a given configuration the first applicable transition using the following priority order: Mark as C, Merge As C, White Merge, Reduce and Shift. Applicability here means not only the standard preconditions for transitions, but also that the output configuration is compatible with the gold annotations. We added the constraint that the White Merge is only applicable to the right suffix of a gold VMWE. For instance, suppose we have a gold continuous VMWE *kick the bucket*, when the first two elements are on top of the stack, the White Merge is not applicable yet, it will be applied when the right suffix *the* and *bucket* is on top of the stack.

To produce our training examples, we start by generating the initial configuration for each sentence, and apply in sequence the transition predicted by the static oracle, until a terminal configuration is reached. The analysis in Figure 1 corresponds to the oracle transition sequence for the example sentence.

[6] Using (Goldberg & Nivre 2013)'s terminology, a static oracle is both incomplete (defined for configurations obtained from previous oracle transitions only) and deterministic (at each such configuration, there is a single oracle transition to apply).

2.3 Expressive power

As far as expressive power is concerned, ATILF-LLF 2 is slightly more powerful than ATILF-LLF 1. ATILF-LLF 2 now performs VMWE categorization and not just identification. Both systems cannot analyze interleaving MWEs, but while ATILF-LLF 1 could cope with no overlapping at all, ATILF-LLF 2 can cope with some cases of embeddings, i.e. some cases of VMWE fully contained in another one.[7]

In effect, ATILF-LLF 1 contained a Shift transition, a White merge, a Merge as C+Reduce for identifying multi-token VMWEs, and a Mark as C+Reduce for MWTs.[8] Because the Merge as C+Reduce transition identifies a MWE and removes it from the stack, no cases of embeddings were covered.

In ATILF-LLF 2, some cases of embeddings are now covered (e.g. the example in Figure 1). More precisely, the covered cases are those where one can form a projective tree by attaching to a fictitious root all the binary trees representing the VMWEs of a sentence, ignoring tokens not belonging to any VMWE. An alternative formulation is that given any VMWE composed of the $t_1 t_2 ... t_m$ tokens and any gap $g_1 g_2 ... g_n$ appearing between a pair of components $t_i t_{i+1}$, the condition is that the g_i tokens cannot belong to a MWE having components outside the set $g_1 ... g_n$. So a non-covered case is found for instance in **Let**$_{1,2}$ the$_1$ cat$_1$ **out**$_{1,2}$ of$_1$ the$_1$ bag$_1$: the VMWE **Let out** has a gap containing tokens the$_1$ cat$_1$, which belong to a VMWE with tokens outside the gap.

3 Experimental setup

For a given language, and a given train/dev split, we use the oracle-based resulting transition sequences to train a multi-class SVM classifier.[9] We describe in the next subsections the feature templates we used (§3.1) and how we tuned them (§3.2).

[7] It is worth noting that embedded VMWEs are very rare in the datasets: there are twenty to thirty embedded VMWEs in German, Hebrew and Hungarian and about 150 in Czech.

[8] ATILF-LLF 1 used hard-coded procedures for matching MWTs (if seen in the training set), which we replaced by features used by the classifier.

[9] The whole system was developed using Python 2.7, with 4,739 lines of code, using the Scikit-learn 0.19. We used the Error-correcting output codes framework for the multi-class SVM classifier. The code is available on Github: https://goo.gl/1j8mVu under the MIT license.

Table 1: System setting code descriptions.

Code	Setting description
A	use of POS and lemmas
A’	use of suffixes
B	use of syntactic dependencies
C	use of bigrams S_1S_0, S_0B_0, S_1B_0 and S_0B_1
D	use of the trigram $S_1S_0B_0$
E	use of the S_0B_2 bigram
F	use of transition history (length 1)
G	use of transition history (length 2)
H	use of transition history (length 3)
I	use of distance between S_0 and S_1
J	use of distance between S_0 and B_0
K	use of B_1
L	use of training corpus VMWE lexicon
M	use of stack length
N	use of MWT dictionary

3.1 Feature templates

A key point in a classical transition-based system is feature engineering, known to have great impact on performance. We have gathered feature templates into groups, for which we provide short descriptions in Table 1, along with code letters that we use in §4 to describe which feature groups were used for each language in the final shared task results. We describe in this section each feature group. We hereafter use symbol B_i to indicate the ɪth element in the buffer. S_0 and S_1 stand for the top and the second elements of the stack. For every unit X in the stack or buffer, we denote X_w its word form, X_l its lemma and X_p its POS tag. The concatenation of two elements X and Y is noted XY.

Basic linguistic features

For each language, we used a precise set of stack or buffer elements, hereafter the *focused elements*, to derive unigram, bigram and trigram features. By default, the focused elements are S_0, S_1 and B_0. For some languages, B_1 was also used (code K in Table 1). If bigrams are on (code C in Table 1) features are generated for the element pairs S_1S_0, S_0B_0, S_1B_0, plus S_0B_1 if K is on, and plus S_0B_2 for a few languages (code E). For trigrams, we only used the features of the $S_1 S_0 B_0$ triple (code D).

For any resulting unigram, bigram or trigram, we use by default the word form (e.g. S_{0w}). For languages whose datasets comprise morphological information, we further use the lemmas and POS tags (code A in Table 1), i.e. X_l and X_p. The precise features for a bigram XY are $X_w Y_w$, $X_p Y_p$, $X_l Y_l$, $X_p Y_l$ and $X_l Y_p$. Those for a trigram XYZ are $X_w Y_w Z_w$, $X_l Y_l Z_l$, $X_p Y_p Z_p$, $X_l Y_p Z_p$, $X_p Y_l Z_p$, $X_p Y_p Z_l$, $X_l Y_l Z_p$, $X_l Y_p Z_l$, $X_p Y_l Z_l$.

For the languages lacking companion morphological information, we tried to mimic that information using suffixes (code A' in Table 1), more precisely the last two and last three letters, which we used for unigram elements only.

Syntax-based features

After integrating classical linguistic attributes, we investigated using more linguistically sophisticated features. First of all, syntactic structure is known to help MWE identification (Fazly et al. 2009; Seretan 2011; Nagy T. & Vincze 2014). So for datasets comprising syntactic information, we introduced features capturing the existence of syntactic dependencies between elements of the buffer and of the stack (code B in Table 1). More precisely, provided that S_0 is a single token, we generate (i) the features **RightDep (S_0, B_i)** = **True** and **RightDepLab (S_0, B_i)** = **l** for each buffer token B_i that is a syntactic dependent of S_0 with label l, and (ii) the features **LeftDep (S_0, B_i)** = **True** and **LeftDepLab (S_0, B_i)** = **l** when a buffer element B_i is S_0's syntactic governor.

Other syntax-based features aim at modeling the direction and label of a syntactic relation between the top two tokens of the stack (feature **syntacticRelation (S_0, S_1)** = ± **l** is used for S_0 governing/governed by S_1, provided S_0 and S_1 are single tokens).[10] All these syntactic features try to capture syntactic regularities between the tokens composing a VMWE.

Miscellaneous features

We found that other traditional features, used in transition-based systems, were sometimes useful like (local) transition history of the system. We thus added **History-based features** to represent the sequence of previous transitions (of length one, two or three, cf. codes F, G and H in Table 1).

[10] For ATILF-LLF 1, we used gold syntactic features for the languages accompanied with gold dependency companion files, as authorized in the closed track. Performance when using predicted syntax will be evaluated in future work.

We also added **Distance-based features**, known to help transition-based dependency parsing (Zhang & Nivre 2011), more precisely the distance between S_0 and S_1 and between S_0 and B_0 (respectively codes I and J in Table 1). We also extracted **Stack-length-based features** (code M in Table 1).

The VMWE identification task is highly lexical so we found it useful to use **dictionary-based features**, which use "dictionaries" extracted from the training set, both for multi-token VMWEs and MWTs. The dictionaries are lemma-based when lemmas are available, and form-based otherwise. These dictionary-based features include (i) a boolean feature set to true when S_0 belongs to the MWT dictionary (code N in Table 1), and (ii) boolean features firing when S_0, S_1, B_0, B_1 or B_2 belong to an entry of the extracted VMWE dictionary (code L in Table 1).

3.2 Feature tuning

We first divided the data sets into three groups, based on the availability of CoNLL-U files: (a) for **Bulgarian**, **Hebrew** and **Lithuanian** neither morphological nor syntactic information is available on top of tokens and VMWE annotation; (b) **Czech**, **Spanish**, **Farsi**, **Maltese** and **Romanian** are accompanied by CoNLL-U files with morphological information only, and (c) **the other languages**[11] are accompanied by a fully annotated CoNLL-U file.

In the first tuning phase, we used one "pilot" language for each group (Bulgarian, Czech and French). Then, German was added as pilot language to investigate features for languages with high percentage of MWTs and embedded VMWEs. We tuned the features using both development sets extracted from the provided training sets, and using cross-validation.

Finally, we used the discovered feature groups as a source of inspiration for producing specialized feature groups for all other languages. Note that given the combinatorial explosion of feature combinations, we could not apply a full grid-search for the pilot languages, and a fortiori for all languages.

4 Results

We provide the identification results in Table 2, in which the performance of ATILF-LLF 2, both in the shared task test sets and in cross-validation, can be compared with (i) a baseline system, (ii) the best performing system of the shared

[11] These languages are German, Greek, French, Hungarian, Italian, Polish, Portuguese, Slovene, Swedish, and Turkish.

Table 2: **VMWE identification**: The first column provides the language, it is shown whether the companion file contains morpho and syntax, morpho only (*) or nothing (**). The last column lists the feature groups used for that language (using the codes of Table 1). Columns 2, 3, 4: VMWE-based F-scores on test sets, for the **baseline** system, **ATILF-LLF 2**, and the best performing Shared task systems (**Best of ST**). Columns 5, 6, 7: same as 2, 3, 4 but token-based. Columns 8, 9, 10: VMWE-based results in 5-fold cross-validation over training sets, for the baseline system, ATILF-LLF 1, and ATILF-LLF 2. The last row (Avg) provides the average results weighted by the size of the test sets (or train sets for cross-validation results). The stars in columns Best of ST are those for which ATILF-LLF 1 did not rank first.

	Test dataset						Cross validation			
	VMWE-based F_1			Token-based F_1			VMWE-based F_1			
Language	Baseline	ATILF-LLF 2	Best of ST	Baseline	ATILF-LLF 2	Best of ST	Baseline	ATILF-LLF 1	ATILF-LLF 2	Feature Settings
BG**	47.6	55.8	**61.3**	50.7	60.2	**66.2**	48.3	**57.1**	53.0	A' C D F G I L
CS*	61.6	70.9	**71.7**	66.7	**73.9**	73.7	60.1	**71.4**	68.9	A C F G H I J K L M
DE	37.9	**45.8**	41,1	33.3	44.8	***45.5**	39.9	27.9	**47.6**	A B C D E J L N
EL	35.6	**42.8**	40.1	39.9	46.3	**46.9**	48.0	56.2	**57.3**	A B C E J K L
ES*	56.9	**58.9**	57.4	56.7	**60.3**	58.4	61.2	63.5	**66.0**	A C D F G H I J K L
FA*	72.2	84.3	**86.6**	72.5	84.8	**90.2**	67.3	**87.7**	81.1	A C I J K
FR	44.6	**60.6**	57.7	49.3	**62.6**	*61.5	66.0	71.1	**73.8**	A B C E I J K L
HE**	**33.4**	29.9	**33.4**	29.6	30.5	**31.3**	**30.0**	17.0	26.8	A' C E F G H K L
HU	68.3	**74.8**	*74.0	64.9	**72.1**	*70.8	73.7	23.5	**83.7**	A B C D F G H K L N
IT	39.2	28.2	**39,9**	39.5	29.8	**43.6**	**33.7**	27.4	27.2	A B C H J L
LT**	30.5	**34.1**	28.4	27.3	**31.7**	25.3	20.7	8.6	**21.5**	A' C D E F G H I J K L M
MT*	8.2	6.9	**14.4**	12.3	9.4	**16.3**	7.7	**8.1**	7.2	A C F G H J L M
PL	72.6	**75.1**	69.1	71.8	**75.5**	*72.7	70.0	70.4	**73.6**	A B C H L
PT	65.5	**69.6**	67.3	67.4	**71.4**	70.9	65.2	64.7	**67.5**	All features
RO*	55.0	**86.3**	*77.8	65.4	**87.0**	*83.6	61.8	**86.3**	86.0	A C D E F G H I J K
SL	13.9	42.9	**43.2**	17.6	45.7	**46.6**	17.3	**47.7**	40.8	A B C F G H I K N
SV	10.4	30.1	**30.4**	10.1	**34.3**	*31.5	6.9	**25.0**	24.7	All features except N
TR	11.3	53.8	**55.4**	18.1	**53.9**	55.3	19.3	58.1	**60.1**	A B C F G H I K
AVG	46.2	56.5	**56.7**	48.5	58.1	**59.2**	52.0	60.3	**64.5**	

task, and with (iii) ATILF-LLF 1.[12] The table shows that results are very heterogeneous across languages. We can hypothesize that multiple factors come into play, such as the size of corpora, the availability and the quality of annotations, the most common VMWE categories in train and test sets, the percentage of unknown VMWEs in test sets. For example, Figure 2 illustrates the impact of this last trait, showing an approximative linear negative correlation between VMWE-based F-score and the proportion of unknown VMWE occurrences in test sets.[13]

Because the datasets have very varying sizes across languages, we provide in the last row of the table the weighted average F-scores, with each language F-score weighted by the size of the test set (or of the training set in cross-validation).

Comparison with the best results at the shared task: Although the ATILF-LLF 2 benefited from more design time, it is interesting to compare its results to the best results obtained at the shared task for each language. When considering the weighted average results (last row of Table 2), it can be seen that the VMWE-based results are almost as high for ATILF-LLF 2 as for the Best of ST (56.5 versus 56.7), and are ahead for 9 languages out of 18. For token-based results, our system is a bit less effective: while still ahead for 10 languages out of 18, it is on average 1.1 point lower (58.1 versus 59.2). This can be viewed as a particularity of our system: while the token-based results are generally higher than the VMWE-based ones (for the baseline, or for other participating systems, cf. Savary et al. 2017), the gap is less pronounced in our case.

Comparison with the baseline: The baseline system is a string matching-based system that uses a lemma-based VMWE dictionary extracted from the training set and identifies as VMWEs all matching strings in the test set.

The matching procedure is very simple: a VMWE is identified inside a sentence if all of its components (lemmas if available, otherwise word forms) occur in the sentence, provided that the order of the components corresponds to an order observed in the dictionary and that the distances between them do not exceed the maximal observed distances in the training dataset.

[12] More precisely, for the results on the test sets (columns 2 to 7), the **Best of ST** columns reflect the performance of ATILF-LLF 1 for the non starred values (cf. no star means we ranked first). F-scores of ATILF-LLF 1 for starred values are as following: Hungarian=70%, Romanian=75% for VMWE-based and German=41%, French=60%, Hungarian=67.4%, Polish=70.5%, Romanian=79.1% and Swedish=30% for token-based.

[13] We also checked for the correlation between the F-score and the training set size, and obtained a positive correlation, but less marked, in particular some languages like Czech and Turkish reach relatively low scores given the size of training data, which is better explained considering the unknown VMWE rate.

Regarding VMWE-based evaluation, ATILF-LLF 2 outperforms the baseline in all experimental settings but four (VMWE-based evaluation on test set for Hebrew and VMWE-based cross-validation for Hebrew and Italian): on average, we obtain about 10- and 12.5-point F-score difference when evaluating on the test set and in cross-validation respectively. Yet, the baseline consistently beats our system on Hebrew. This might be explained by several characteristics of this language preventing the system to generalize well to morpho-syntactic variants: (i) small training set and (ii) no companion linguistic information (no POS, no lemmas, no syntactic parses).

Comparison between ATILF-LLF 2 and ATILF-LLF 1: The ATILF-LLF 1 system participated in the shared task and reached the best VMWE-based scores for almost all languages (cf. the two starred results out of 18 in column Best of ST, for Hungarian and Romanian). It can be seen that ATILF-LLF 2 shows comparable performance on the same test sets (see in particular the weighted average performance shown in the last row: 56.5 versus 56.7). It is worth noticing though that there is great variation between results on test sets and results in cross-validation. As the latter are more representative, let us focus on them (columns 8 to 10). Despite a few languages showing a drop in performance (in particular Bulgarian, Farsi and Slovene), ATILF-LLF 2 beats ATILF-LLF 1 for 10 languages out of 18, and the average result (last row of Table 2) has improved (4.2-point gain). Again, even though ATILF-LLF 2 benefited from more design time, this is a good result considering that (i) ATILF-LLF 1 did identification only, and the introduction of the categorization task led us to multiply the number of transitions (e.g. 5 Merge as C transitions instead of 1), (ii) the expressive power was increased to some cases of embeddings and (iii) the overall architecture is more elegant since hard-coded procedures included in the rush of the shared task have been replaced by features.[14]

Categorization results: Table 3 details the categorization results for the basic categories over all languages but Farsi (cf. the Farsi dataset does not comprise VMWE category information).[15] The table allows us to compare the performance of our system with best-performing shared task systems (for the systems having the optional categorization predictions, note that our former system ATILF-LLF 1 is excluded given that it does not categorize VMWEs). It can be seen

[14] It is worth noting that feature groups for each language were very close for both systems (ATILF-LLF 1, 2). However, we transformed the dictionary-based hard-coded feature groups into dynamic ones.

[15] We do not include the category OTH because of its marginal presence in test and train datasets for all languages but Turkish (for which our F-score is 51.5 and the Shared task best F-score is 54.6).

Table 3: **VMWE categorization**: detailed results for the four basic categories over all the languages except Farsi. For each category, we display the proportion of the given category **in test set**, the F-scores $\mathbf{F_1}$ for **ATILF-LLF 2**, and the best performing shared task systems (**Best of ST**), among systems having provided categorization information.

	LVC			IReflV			VPC			ID		
	%	F_1	F_1	%	F_1	F_1	%	F_1	F_1	%	F_1	F_1
Languages	In test set	ATILF-LLF 2	Best of ST	In test set	ATILF-LLF 2	Best of ST	In test set	ATILF-LLF 2	Best of ST	In test set	ATILF-LLF 2	Best of ST
BG	16	**28.6**		63	**68.3**	46.6				21	**19.4**	
CS	20	**53.3**	40.1	68	**80.7**	73.3				11	**33.6**	22.0
DE	8	**4.7**	2.3	4	8.9	**16.0**	45	**58.3**	43.3	43	**29.1**	16.4
EL	67	**44.4**	33.2				3	**52.2**	36.4	25	**27.1**	15.4
ES	21	**49.0**	35.1	44	**72.0**	40.4				33	**39.0**	13.8
FR	54	40.0	**42.8**	21	**78.6**	68.3				24	**75.0**	60.8
HE	25	**31.4**					37	**27.2**		6	**16.2**	
HU	29	**50.9**	41.5				71	**80.3**	77.2			
IT	17	**17.5**	12.9	30	**30.3**	9.3	2	**50.0**	14.3	50	**25.7**	20.3
LT	42	**56.3**								58	**14.9**	
MT	52	**6.6**	5.8							52	7.7	2.1
PL	34	**62.9**	39.1	53	**87.3**	80.2				13	**51.5**	
PT	66	**68.0**		16	**68.9**					18	**65.3**	
RO	27	**87.4**	86.3	58	**86.3**	79.1				15	**76.7**	65.6
SL	9	7.4	**8.3**	51	**45.7**	40.8	22	**47.5**	34.5	18	**09.1**	3.9
SV	6	16.7	**21.1**	6	**08.7**		66	**33.6**	30.2	21	**13.3**	3.8
TR	40	57.6	**59.1**							11	**49.3**	49.8

that our system reaches high performance on categorization too, although performance varies greatly across categories. Although the general trend is higher performance for IReflV, then LVC, then ID, Figure 3 shows that this pattern is not systematic. For instance, results are relatively low for Czech given its high IReflV proportion. On the contrary, results for Portuguese are high despite a high LVC ratio.

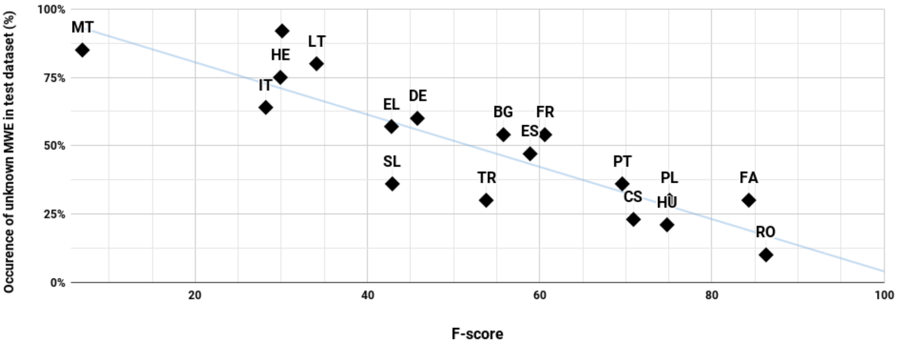


Figure 2: Correlation between the ATILF-LLF 2 identification results for each language (F-score, on the x axis) and the percentage of **occurrences of test VMWEs unknown in the train set** (y axis).

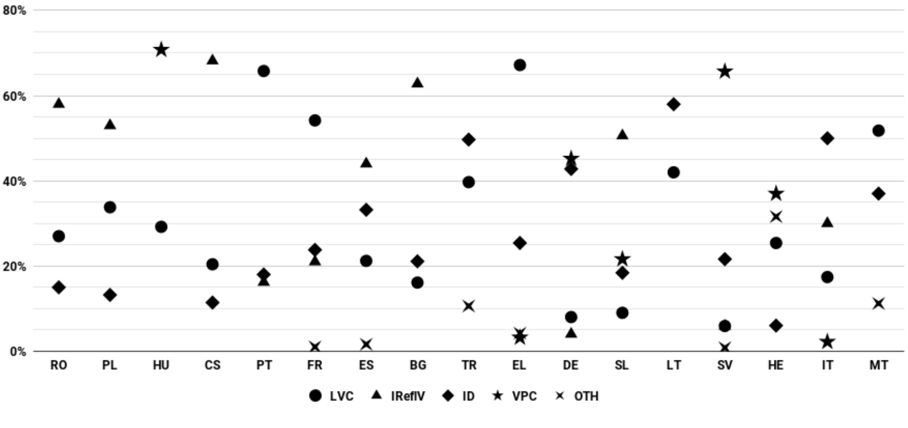


Figure 3: A graph ranking all languages except Farsi according to their F-scores. In each bar, the proportions of VMWE categories in the test set are shown using symbols.

5 Related work

A popular VMWE identification method is to use a sequence labeling approach, with IOB-based tagsets. For instance, Diab & Bhutada (2009) apply a sequential SVM to identify verb-noun idiomatic combinations in English. Note also that three (out of seven) systems participating in the PARSEME shared task used such approach (Boroş et al. 2017; Maldonado et al. 2017; Klyueva et al. 2017). Such an approach was also investigated for MWE identification in general (including verbal expressions) ranging from continuous expressions (Blunsom & Baldwin 2006) to gappy ones (Schneider et al. 2014). Recently, neural networks have been successfully integrated into this framework (Legrand & Collobert 2016; Klyueva et al. 2017).

VMWE identification can naturally take advantage of previously predicted syntactic parses. Some systems use them as soft constraints. For instance, the sequence labeling systems of the shared task and our system use them as source of features in their statistical tagging models. There also exist approaches using syntactic parses as hard constraints. For example, Baptista et al. (2015) apply hand-crafted identification rules on them. Fazly et al. (2009) and Nagy T. & Vincze (2014) propose a two-pass identification process consisting of candidate extraction followed by binary classification. In particular, candidate extraction takes advantage of predicted syntactic parses, through the use of linguistic patterns.

A joint syntactic analysis and VMWE identification approach using off-the-shelf parsers is another interesting alternative that has shown to help VMWE identification such as light-verb constructions (Eryiğit et al. 2011; Vincze & Csirik 2010). Some parsers integrate mechanisms into the parsing algorithm to identify MWEs on top of predicting the syntactic structure, like in Wehrli (2014) and Constant & Nivre (2016), our system being a simplified version of the latter.

6 Conclusion and future work

This article presents a simple transition-based system devoted to VMWE identification and categorization. In particular, it offers a simple mechanism to handle discontinuity and embedding, which is a crucial point for VMWEs. Results on the PARSEME Shared Task datasets show that our system is quite robust across languages, and achieves very competitive results. Its linear time complexity is also an asset.

As future work, we would like to apply more sophisticated syntax-based features, as well as more advanced machine-learning techniques like neural networks. We also plan to investigate the use of a dynamic oracle (Goldberg & Nivre 2012).

Acknowledgements

This work was partially funded by the French National Research Agency (PARSEME-FR ANR-14-CERA-0001).

Abbreviations

ID	Idiom	MWT	multiword token
IOB	inside outside beginning	OTH	other verbal MWE
IReflV	inherently reflexive verb	VMWE	verbal multiword expression
LVC	light-verb construction	VPC	verb-particle construction
MWE	multiword expression		

References

Al Saied, Hazem, Matthieu Constant & Marie Candito. 2017. The ATILF-LLF system for parseme shared task: A transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 127–132. Association for Computational Linguistics. DOI:10.18653/v1/W17-1717

Baptista, Jorge, Graça Fernandes, Rui Talhadas, Francisco Dias & Nuno Mamede. 2015. Implementing European Portuguese verbal idioms in a natural language processing system. In Gloria Corpas Pastor (ed.), *Proceedings of europhras 2015*, 102–115.

Blunsom, Phil & Timothy Baldwin. 2006. Multilingual deep lexical acquisition for HPSGs via supertagging. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 164–171. Sydney, Australia: Association for Computational Linguistics. DOI:10.3115/1610075.1610101

Boroş, Tiberiu, Sonia Pipa, Verginica Barbu Mititelu & Dan Tufiş. 2017. A data-driven approach to verbal multiword expression detection. PARSEME shared task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 121–126. Association for Computational Linguistics. DOI:10.18653/v1/W17-1716

Constant, Matthieu & Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 161–171. Association for Computational Linguistics. http://www.aclweb.org/anthology/P16-1016.

Diab, Mona & Pravin Bhutada. 2009. Verb noun construction MWE token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 17–22. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-2903.

Eryiğit, Gülşen, Tugay İlbay & Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of IWPT Workshop on Statistical Parsing of Morphologically-Rich Languages* (SPMRL 2011), 45–55. http://dl.acm.org/citation.cfm?id=2206359.2206365. October 6, 2011.

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103. http://aclweb.org/anthology/J09-1005.

Goldberg, Yoav & Joakim Nivre. 2012. A dynamic oracle for arc-eager dependency parsing. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, 959–976.

Goldberg, Yoav & Joakim Nivre. 2013. Training deterministic parsers with non-deterministic oracles. *Transactions of the Association for Computational Linguistics* 1. 403–414.

Klyueva, Natalia, Antoine Doucet & Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 60–65. Association for Computational Linguistics. April 4, 2017. DOI:10.18653/v1/W17-1707

Legrand, Joël & Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions* (MWE '16), 67–71. Association for Computational Linguistics. http://anthology.aclweb.org/W16-1810.

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 114–120. Association for Computational Linguistics. DOI:10.18653/v1/W17-1715

Nagy T., István & Veronika Vincze. 2014. VPCTagger: Detecting verb-particle constructions with syntax-based methods. In *Proceedings of the 10th Workshop* *on Multiword Expressions* (MWE '14), 17–25. Association for Computational Linguistics. http://www.aclweb.org/anthology/W14-0803.

Nivre, Joakim. 2004. Incrementality in deterministic dependency parsing. In Frank Keller, Stephen Clark, Matthew Crocker & Mark Steedman (eds.), *Proceedings of the ACL Workshop on Incremental Parsing: Bringing Engineering and Cognition together*, 50–57. Association for Computational Linguistics.

Nivre, Joakim. 2014. *Transition-based parsing with multiword expressions*. IC1207 COST PARSEME 2nd general meeting. Athens, Greece. https://typo.uni-konstanz.de/PARSEME/images/Meeting/2014-03-11-Athens-meeting/PosterAbstracts/WG3-Nivre-athens-poster.pdf.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* 2(1). 193–206. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/281.

Seretan, Violeta. 2011. *Syntax-based collocation extraction* (Text, Speech and Language Technology). Dordrecht, Heidelberg, London, New York: Springer.

Vincze, Veronika & János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING '10), 1110–1118. Association for Computational Linguistics. http://www.aclweb.org/anthology/C10-1125.

Wehrli, Eric. 2014. The relevance of collocations for parsing. In *Proceedings of the 10th Workshop on Multiword Expressions* (MWE '14), 26–32. Association for Computational Linguistics. 26-27 April, 2014.

Zhang, Yue & Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 188–193. Association for Computational Linguistics. http://www.aclweb.org/anthology/P11-2033.

Chapter 8

Identifying verbal multiword expressions with POS tagging and parsing techniques

Katalin Ilona Simkó
University of Szeged

Viktória Kovács
University of Szeged

Veronika Vincze
University of Szeged
MTA-SZTE Research Group on Artificial Intelligence

The chapter describes an extended version (USzeged+) of our previous system (USzeged) submitted to PARSEME's Shared Task on automatic identification of verbal multiword expressions. USzeged+ exploits POS tagging and dependency parsing to identify single- and multi-token verbal MWEs in text. USzeged competed on nine of the eighteen languages, where USzeged+ aims to identify the VMWEs in all eighteen languages of the shared task and contains fixes for deficiencies of the previously submitted system. Our chapter describes how our system works and gives a detailed error analysis.

1 Introduction

Multiword expressions (MWEs) are frequent elements of all natural languages. They are made up of more than one lexeme, but their meaning is not predictable from the meaning of their components. There are different types of MWEs such

Katalin Ilona Simkó, Viktória Kovács & Veronika Vincze. 2018. Identifying verbal multiword expressions with POS tagging and parsing techniques. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 227–243. Berlin: Language Science Press. DOI:10.5281/zenodo.1469563

as stereotyped similes (*as white as snow*), collocations (*strong tea*), or idioms (*to kick the bucket*). This chapter deals with verbal MWEs (VMWEs) where the head element of the MWE is a verb, for example verb-particle constructions (*look after*), or light-verb constructions (*take a shower*).

This chapter describes our system for verbal MWE recognition. It was built for the PARSEME Shared Task 1.0 (Savary et al. 2017), USzeged and its extension, USzeged+. Both systems use POS tagging and dependency parsing and are capable of identifying single- and multi-token verbal MWEs. They are language-independent: USzeged was submitted for nine of the eighteen languages of the Shared Task, while for this extended version, USzeged+, we present results for all eighteen languages.

In this chapter, we first describe the original USzeged system and give our results submitted to the Shared Task with detailed error analysis. This part of the chapter builds heavily on our workshop paper (Simkó et al. 2017). Then, we describe the details of the updated, USzeged+ version and give the results we achieved with this new system. Last, we give a comparison of the results achieved using our approach in the original, USzeged system and the new USzeged+ one in an experiment using the available Hungarian data.

2 USzeged - The original system

The USzeged system was built for the shared task on automatic identification of verbal multiword expressions organized as part of the 2017 MWE workshop (Savary et al. 2017).[1] The shared task's aim is to identify verbal MWEs in multiple languages. In total, eighteen languages are covered that were annotated using guidelines taking universal and language-specific phenomena into account.

The guidelines identify five different types of verbal MWEs: idioms (ID), light-verb constructions (LVC), verb-particle constructions (VPC), inherently reflexive verbs (IReflV) and "other" (OTH). Their identification in natural language processing is difficult because they are often discontinuous and non-compositional, the categories are heterogeneous and the structures show high syntactic variability.

The precise definitions of MWE, VMWE and the VMWE types can be found in Savary et al. (2018 [this volume]), as well as details on the different languages' databases used.

Our team created the Hungarian shared task database and VMWE annotation. Our system is mostly based on our experiences with the Hungarian data in this

[1] http://multiword.sourceforge.net/sharedtask2017

annotation phase. Our goal was to create a simple system capable of handling MWE identification in multiple languages.

2.1 System description

The USzeged system exploits the syntactic relations within MWEs, i.e. it directly connects MWEs and parsing, an approach described in many sources (Constant & Nivre 2016; Nasr et al. 2015; Candito & Constant 2014; Green et al. 2011; 2013; Wehrli et al. 2010; Waszczuk et al. 2016) and one of the basic ideas behind the work done by the PARSEME group.[2] The core of our system is directly based on the work described in Vincze et al. (2013): using dependency parsing to identify MWEs. That system uses complex dependency relations specific to the given syntactic relation and MWE type. We note that a high number of the languages of the shared task are morphologically rich and have free word order, which entails that syntactically flexible MWEs might not be adjacent. Hence, a syntax-based approach seems a better fit for the task than sequence labeling or similar strategies.

The USzeged system uses only the MWE type as a merged dependency label, i.e. no clue is encoded to the syntactic relation between two parts of the MWE. Moreover, it also treats single-token MWEs. As multiple languages had single-token MWEs as well as multi-token ones that are dealt with in dependency parsing, we expanded the approach using POS tagging. Frequent single-token MWEs are, for example, German and Hungarian VPCs: when the particle directly preceeds the verb, German and Hungarian spelling rules require that they are spelled as one word, however, it still remains a construction made up of two lexemes with non-compositional meaning (e.g. (HU) **kinyír** (ki+nyír) 'out+cut' ⇒ 'kill' or (DE) **aufmachen** (auf+machen) 'up+do' ⇒ 'open').

MWEs have specific morphological, syntactic and semantic properties. Our approach treats multi-token MWEs on the level of syntax – similarly to the `mwe` dependency relation in the Universal Dependency grammar (Nivre 2015) – and single-token MWEs on the level of morphology.

The USzeged system works in four steps, and the main MWE identification happens during POS tagging and dependency parsing of the text. Our system relies on the POS tagging and dependency annotations provided by the organizers of the shared task in the companion CoNLL files and the verbal MWE annotation of the texts and is completely language-independent given those inputs.

[2] http://typo.uni-konstanz.de/parseme/

In the first step, we prepared the training file from the above mentioned inputs. We merged the training MWE annotation into its morphological and dependency annotation for single- and multi-token MWEs, respectively. The POS tag of single-token MWEs got replaced with their MWE type, while for the multi-token MWEs the dependency graphs' label changed: the label of the dependent node in the tree was replaced with a label denoting the MWE type.

Figure 1, Figure 2 and Figure 3 show the single-token MWE's change in POS tag and multi-token MWE dependency relabeling for VPCs and LVCs in a Hungarian example.

	original label	relabeled	(HU)
bekezdés in+starting, ‘paragraph’	NOUN	VPC	(HU)
határozathozatal decision+bringing, ‘decision-making’	NOUN	LVC	(HU)

Figure 1: Adding the VPC and LVC single-token MWE POS tags to (HU) **bekezdés** (be+kezdés) ‘in+starting’ ⇒ ‘paragraph’ and (HU) **határozathozatal** (határozat hozatal) ‘decision+bringing’ ⇒ ‘decision-making’.

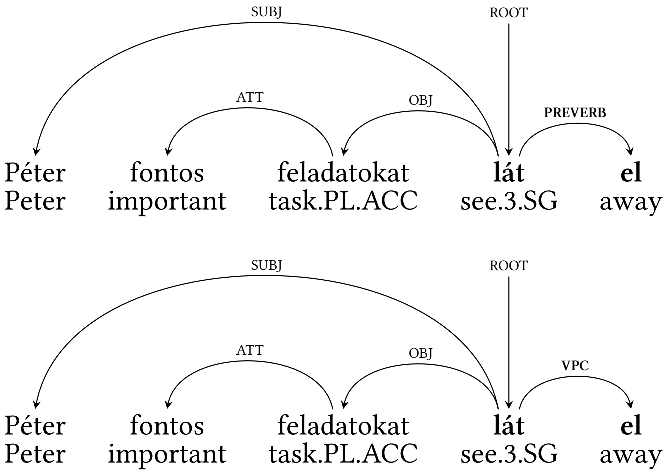

SUBJ
ROOT
ATT
OBJ
PREVERB
Péter fontos feladatokat **lát** **el**
Peter important task.PL.ACC see.3.SG away
SUBJ
ROOT
ATT
OBJ
VPC
Péter fontos feladatokat **lát** **el**
Peter important task.PL.ACC see.3.SG away

Figure 2: Adding the VPC multi-token MWEs label to the dependency graph in (HU) *Péter fontos feladatokat* ***lát el***. ‘Peter important tasks sees away’ ⇒ ‘Peter takes care of important tasks’.

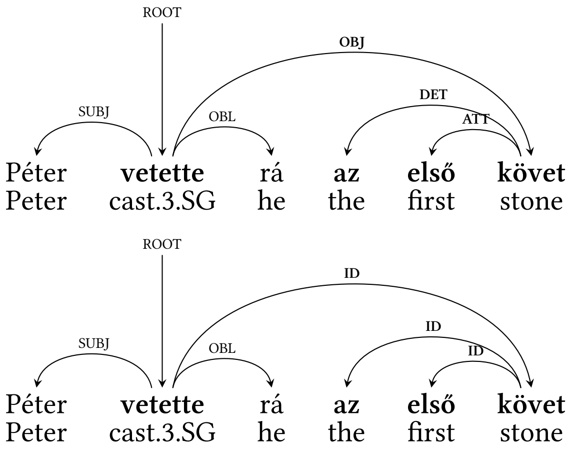

ROOT
OBJ
DET
SUBJ
OBL
ATT
Péter vetette rá az első követ
Peter cast.3.SG he the first stone
ROOT
ID
ID
SUBJ
OBL
ID
Péter vetette rá az első követ
Peter cast.3.SG he the first stone

Figure 3: Adding the ID multi-token MWE label to the dependency graph in (HU) Péter **vetette** rá **az első követ**. 'Peter cast the first stone on him'.

For multi-token MWEs our approach is based on our hypothesis that the dependent MWE elements will be directly connected to the other MWE element(s). We do not change the structure of the dependency relations in the tree, but change the dependency label of the dependent MWE element to the MWE type, therefore making the MWE element retraceable from the dependency annotation of the sentence. For example *lát* and *el* in Figure 2 make up a VPC (**ellát** 'take care'), so the dependency relation label of the dependent element, *el* changes from the general syntactic label **PREVERB** to the MWE label **VPC**, with this **VPC** label now connecting the two elements of the MWE.

For MWEs of more than two tokens, the conversion replaces the dependency labels of all MWE elements that depend on the head. In Figure 3, the head of the idiom (**az első követ veti** 'casts the first stone') is the verb, *vetette* (cast.Sg3.Past). All other elements' dependency labels are changed to **ID**.

The second step is training the parser: we used the Bohnet parser (Bohnet 2010) for both POS tagging and dependency parsing. For the single-token MWEs, we trained the Bohnet parser's POS tagger module on the MWE-merged corpora and its dependency parser for the multi-token MWEs. The parser would treat the MWE POS tags and dependency labels as any other POS tag and dependency label.

We did the same for each language and created POS tagging and dependency parsing models capable of identifying MWEs for them. For some languages in the shared task, we had to omit sentences from the training data that were overly long (spanning over 500 tokens in some cases) and therefore caused errors in training due to lack of memory. This affected one French, one Polish, two Italian, five Romanian and nine Turkish sentences.

Third, we ran the POS tagging and dependency parsing models of each language on their respective test corpora. The output contains the MWE POS tags and dependency labels used in that language as well as the standard POS and syntactic ones.

The fourth and last step is to extract the MWE tags and labels from the output of the POS tagger and the dependency parser. The MWE POS tagged words are annotated as single-token MWEs of the type of their POS tag. From the MWE dependency labels, we annotate the words connected by MWE labels of the same type as making up a multi-token MWE of that type (see Figure 4).

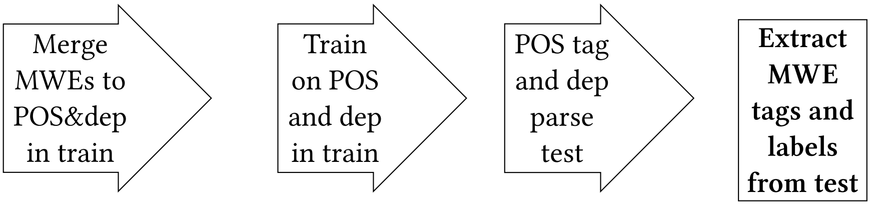

Merge MWEs to POS&dep in train
Train on POS and dep in train
POS tag and dep parse test
Extract MWE tags and labels from test

Figure 4: Steps of the USzeged system.

There are arguments for and against our approach. The system cannot handle multi-token MWEs where the elements are not connected in the tree and replacing the POS tags and dependency labels can have a negative effect on the accuracy of POS tagging and parsing. However, as our end goal is not the POS tagging or dependency parse of the data, we believe that this side effect is negligible since higher-level applications (e.g. machine translation) can profit from more accurate MWE identification. On the other hand, the approach has low technical requirements and it is very easily adaptable to other languages.

2.2 Results

We submitted the USzeged system for all languages in the shared task with provided dependency analysis and POS tagging. We attempted to use just the POS tagging component of our system on the languages that only had POS tagging available to give partial results (i.e. identifying only single-token MWEs), but we found that these languages incidentally had no or very few single-token MWEs (Farsi 0, Maltese 4, Romanian 44, Slovene 3, Turkish 22), therefore we had no access to adequate training data and did not submit results for these languages.

Our results on the nine languages are reported in Simkó et al. (2017). Our system was submitted for German, Greek, Spanish, French, Hungarian, Italian, Polish, Portuguese, and Swedish. For the evaluation, we employed the metrics used for the evaluation of the shared task (Savary et al. 2017).

The F-scores show great differences between languages, but so did they for the other systems submitted. Compared to the other, mostly closed-track systems, the USzeged system ranked close to or at the top on German, Hungarian, and Swedish. For the other languages (except for Polish and Portuguese, where ours is the worst performing system), we ranked in the mid-range.

2.3 Error analysis

After receiving the gold annotation for the test corpora, we investigated the strengths and weaknesses of our system.

Our error analysis showed that the USzeged system performs by far best on single-token MWEs, which in this dataset are mostly made up of the verb-particle construction category, correctly identifying around 60% of VPCs, but only about 40% of other types on average. It is probably due to the fact that single-token MWEs are identified by POS tagging techniques, which are known to obtain more accurate results in most languages than dependency parsing.

German, Hungarian, and Swedish were also the languages with the highest proportions of the VPC type of verbal MWEs in the shared task, which also correlates with why our system performed best on them. Romance languages contain almost no VPCs and the remaining ones have much less also. In this way, the frequency of VPCs strongly influences our results on the given language.

For French and Italian, our system also performed worse on IReflVs. In general, we had some trouble identifying longer IDs and LVCs and MWEs including prepositions. A further source of error was when there was no syntactic edge in between members of a specific MWE, for instance, in German, the copula **sein** 'be' was often indirectly connected to the other words of the MWE (e.g. ***im Rennen sein*** 'in race be' ⇒ 'to compete'), hence our method was not able to recognize it as part of the MWE. As our system does not restructure the syntactic trees, if the elements of a multi-token VMWE are not connected (i.e. they do not form a graph) in their dependency annotation, we cannot identify the full MWE, however, we can still identify tokens of it correctly if at least two tokens within the MWE are attached.

3 The extended system - USzeged+

The primary aim of our extension was to be able to use our system for the languages in the shared task without any available POS and dependency data. We achieved this by parsing the annotated set in a preprocessing step. For the languages with gold POS and dependency data already available, we did not use this extra step (see Figure 5).

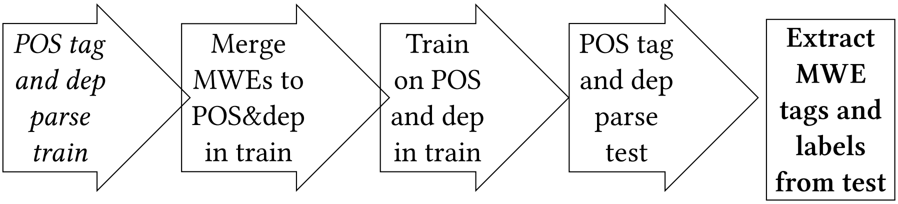

POS tag and dep parse train
Merge MWEs to POS&dep in train
Train on POS and dep in train
POS tag and dep parse test
Extract MWE tags and labels from test

Figure 5: Steps of the USzeged+ system.

We used data for the remaining languages from the Universal Dependencies Project release 2.0 (Nivre et al. 2016) to train the Bohnet parser for POS-tagging and dependency parsing and parsed the VMWE annotated shared task's training sets. We should note that for some languages, the VMWE corpus and the Universal Dependencies corpus are overlapping. This influences our dependency parse to some degree as the training data might partially include the test data, but as our end goal here is not the full dependency parse of the texts (moreover, we already use gold dependency annotations for the languages which have it directly available), we feel that this factor is negligible. Henceforward, we exploited the very same processes as before: we merged the parsed data with the VMWE annotations and once again, trained the Bohnet parser on the VMWE merged data. We then parsed the test sets for the shared task and extracted the MWE POS-tagged and MWE dependency labeled words and phrases.

3.1 Results

Table 1 shows the USzeged+ results for all shared task languages. The languages covered by USzeged can be found in the upper part of the table, and the ones covered by USzeged+ are in the lower part. The "upper" languages of this table show differences to the results presented in (Simkó et al. 2017). This is due to two main factors: the Bohnet parser was updated between our USzeged and USzeged+

versions of the system and we also corrected some bugs in our conversion tool. The basic working principles of our system are the same as described above.

Table 1: USzeged+ results: Languages covered by the previous system also are on top.

	P-MWE	R-MWE	F1-MWE	P-token	R-token	F1-token
DE	31.16	40.20	35.11	40.65	43.05	41.82
EL	37.01	30.20	33.26	49.14	32.65	39.23
ES	25.67	52.00	34.37	32.13	55.20	40.62
FR	31.23	31.60	31.41	43.57	39.44	41.40
HU	62.02	71.34	66.36	58.45	69.08	63.32
IT	9.21	6.80	7.83	33.29	18.70	23.94
PL	35.96	59.40	44.80	41.33	63.68	50.13
PT	33.29	52.80	40.84	40.76	58.96	48.20
SV	14.96	22.88	18.09	20.55	28.21	23.77
BG	53.26	43.13	47.66	77.79	49.25	60.32
CS	44.95	57.93	50.62	57.60	64.46	60.84
FA	69.58	46.20	55.53	85.78	53.14	65.63
HE	41.18	8.40	13.95	55.22	8.63	14.93
LT	33.33	7.00	11.57	40.48	6.97	11.89
MT	0.00	0.00	0.00	4.65	0.31	0.59
RO	46.29	67.40	54.89	53.01	71.68	60.95
SL	59.49	18.80	28.57	66.46	18.77	29.27
TR	39.34	37.92	38.62	42.07	39.49	40.74

Using gold or parsed POS and dependency data as the starting phase does not have a significant impact on the results (as we will show in another experiment in §4), with the exception of Maltese. As Maltese currently has no available Universal Dependencies treebank, we used cross-language training to train our parser. As a Semitic language, Maltese is basically related to Arabic but spelt with Latin characters and about half of its vocabulary originates from Italian. Thus, we selected the available Italian Universal Dependencies treebank to train the parser and parse the VMWE train data. This had a very bad effect on our results: no full MWE could be correctly identified in the VMWE test set. Hence, for Maltese, a more suitable solution is still to be found for our approach. For all other languages – where the parser for the VMWE train data was trained on the same language – the final results are much more comparable to those of the languages with gold trees.

Besides Maltese, one of the languages where our system performed poorly is Italian. We investigated the Italian training corpus and found that its annotation has different underlying principles than most of the other corpora. Namely, it allows sentences to have multiple roots (which is prohibited in other dependency theories), hence it confuses the parser's training to a high degree and therefore very few valuable results (i.e. MWE annotations) can be converted from the parsed sentences. Finally, Swedish results are probably due to the small size of the training corpora.

Table 2 and Table 3 show our results in F-score for the different MWE types; crossed out cells indicate that the type was not present for the given language.

Overall, the USzeged+ system performs best on inherently reflexive verbs (IReflV). IReflVs contain irreflexive pronouns, which show little variability, thus they can be relatively easily recognized by the system. However, the system performs worst on idioms and the "other" category due to their bigger variability and the longer MWEs in these types. Light-verb constructions and verb-particle constructions show varied results depending on the variability of the category in the given language. VPCs could be easily recognized in Hungarian (an F-score over 80), while LVC identification was most successful in Romanian (an F-score of 70).

There are also differences in the annotations of the languages: for instance, Farsi contains only VMWEs for the "other" category, which makes it hard to make any comparisons with the other languages on the effective identification of VMWE categories.

The results also show that while our system has very similar average results on the other language group of the shared task (interestingly even the "other" category, which is probably due to Farsi), results are much lower on Romance languages on average. This is most probably due to the issues on the Italian dependency data (see above), which resulted in poor performance for almost all of the VMWE categories in Italian.

We did some error analysis based on the languages we can speak. This revealed that as for LVCs, our system usually marks as false positives those verb-object pairs where the verb is an otherwise frequent light verb in the given language (e.g. (PT) **ter** 'to have'). Also, participle forms of LVCs were often missed in (FR) **études menées** 'studies conducted'. As for VPCs, many compositional instances of verbs with particles were falsely marked in German and in Hungarian, like (DE) **anheben** 'hang up'. The same is true for IReflVs: compositional ones like (PT) **encantar-se** 'enchant' were sometimes falsely identified as VMWEs. A fur-

Table 2: USzeged+ MWE-level F-score results for the different MWE types.

	all	VPC	LVC	ID	IReflV	OTH
BG	47.66	-	20.44	18.79	60.70	-
CS	50.62	-	26.96	2.51	61.71	0.00
DE	35.11	54.55	10.39	16.16	17.50	-
EL	33.26	56.00	36.68	9.52	-	8.00
ES	34.37	-	32.40	9.70	43.04	0.00
FA	55.53	-	-	-	-	55.53
FR	31.41	-	27.67	16.48	52.94	0.00
HE	13.95	9.35	24.20	0.00	-	13.76
HU	66.36	80.59	35.47	-	-	-
IT	7.83	22.22	5.13	5.05	0.00	0.00
LT	11.57	-	23.53	0.00	-	-
MT	0.00	-	0.00	0.00	-	0.00
PL	44.80	-	25.38	0.00	65.63	-
PT	40.84	-	45.89	11.57	40.99	-
RO	54.89	-	70.59	20.51	54.92	-
SL	28.57	0.00	14.29	1.90	45.71	0.00
SV	18.09	22.11	8.70	0.00	2.90	0.00
TR	38.62	-	39.77	32.71	-	34.34
Average	34.08	34.97	26.32	9.06	40.54	10.15

ther source of errors could also be some inconsistencies in the data: in a few cases, annotators missed to mark some clear examples of VMWEs in the test data, which resulted again in false positives. Finally, the German corpus contained some English sentences, e.g. *[...] if Proporz were to be taken out of the Austrian economy, actual unemployment would be ... higher?* In this sentence, *be* and *higher* are marked as an instance of VPC. The word *be* is a typical particle in German, while last words of the sentences are often verbs in German due to word order reasons. Probably this is the reason why the system gave this analysis.

As our system uses different methods to assign single- and multi-token MWE labels, we also investigated our results for these separately. We found that most languages only contain no or very few single-token MWEs, with the exception of German and Hungarian. Approximately 12% of VMWEs are single-token in the

Table 3: USzeged+ system's token-level F-score results for the different MWE types.

	all	VPC	LVC	ID	IReflV	OTH
BG	60.32	-	32.63	24.30	74.05	-
CS	60.84	-	31.78	17.87	72.44	0.00
DE	41.82	56.14	12.58	32.41	24.85	-
EL	39.23	56.00	40.65	19.97	-	4.04
ES	40.62	-	35.86	22.47	44.27	0.00
FA	65.63	-	-	-	-	65.63
FR	41.40	-	30.17	36.11	52.94	0.00
HE	14.93	13.82	23.21	10.62	-	10.83
HU	63.32	80.59	40.82	-	-	-
IT	23.94	27.78	12.60	23.78	0.00	0.00
LT	11.89	-	23.30	3.28	-	-
MT	0.59	-	0.71	0.33	-	0.00
PL	50.13	-	28.18	13.19	69.44	-
PT	48.20	-	50.73	27.80	42.07	-
RO	60.95	-	73.03	48.88	55.75	-
SL	29.27	1.70	15.87	9.36	46.29	0.00
SV	23.77	26.80	8.79	2.26	2.90	0.00
TR	40.74	-	41.32	34.43	-	36.36
Average	39.87	37.55	29.54	20.44	44.09	10.62

German data, while they make up of 40% of VMWEs in the Hungarian data. Table 4 shows our system's accuracy on single- and multi-token MWEs for German and Hungarian.

Table 4: Accuracy on single- and multi-token MWEs for German and Hungarian.

	single-token	multi-token	overall
DE	64	36	46
HU	83	43.3	68.9

These results confirm that our system achieves better results on single-token MWEs than on multi-token ones.

4 Gold or parsed?

In this last section, we describe a small experiment comparing our new addition to the system: the parsed POS and dependency data. We compare our results on Hungarian using the gold POS and dependency data with an experimental setup mirroring that of the languages without this gold data. We used the Hungarian Universal Dependencies treebank to train the Bohnet parser for POS tagging and dependency parsing and exploited these trained models to parse the VMWE train sentences.

Table 5 shows the results of this experiment; the results for HU-GOLD are the same as the ones for Hungarian in the above tables. The results show that gold and parsed methods in our system can provide very comparable results. Interestingly, in both MWE-level and token-level results, the parsed method provides much higher precision but lower recall than the gold method.

For MWE types, LVCs are causing the main difference in the two systems. Many VPCs in the Hungarian data are single-token, so our system deals with them on the level of POS tagging, which is not affected by gold or parsed dependency trees.

Overall, both options achieved approximately the same results in the automatic VMWE recognition task.

Table 5: Gold and parsed results for Hungarian.

	HU-GOLD	HU-PARSED
P-MWE	62.02	74.94
R-MWE	71.34	62.93
F-MWE	66.36	68.41
P-token	58.45	74.74
R-token	69.08	53.85
F-token	63.32	62.60
F-VPC-MWE	80.59	78.59
F-LVC-MWE	35.47	25.41
F-VPC-token	79.05	77.66
F-LVC-token	40.82	25.71

5 Conclusions

In our chapter, we presented our system for verbal MWE recognition. The system uses POS tagging and dependency parsing as a means of finding verbal MWEs in multiple languages.

Apart from parsing-based solutions (Al Saied et al. (2017), Nerima et al. (2017) and our system), the shared task hosted a number of other approaches, like neural networks (Klyueva et al. (2017)) or sequence labeling based models (Boroş et al. (2017), Maldonado et al. (2017)). In the final results, parsing-based systems achieved the best results for almost all languages, showing that this approach works very well for language independent MWE identification.

Our chapter further shows that it is possible to build a highly language independent MWE detection methodology that makes use of a limited amount of language-specific data and achieve reasonable results.

Acknowledgements

We would like to thank the anonymous reviewers and the volume editors for their invaluable comments on our work.

Veronika Vincze was supported by the UNKP-17-4 New National Excellence Program of the Ministry of Human Capacities, Hungary.

Abbreviations

ID	idiom	OTH	other
IReflV	inherently reflexive verb	POS	part of speech
LVC	light-verb construction	VPC	verb-particle construction
MWE	multiword expression	VMWE	verbal multiword expression

References

Al Saied, Hazem, Matthieu Constant & Marie Candito. 2017. The ATILF-LLF system for parseme shared task: A transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 127–132. Association for Computational Linguistics. DOI:10.18653/v1/W17-1717

Bohnet, Bernd. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics* (COLING 2010), 89–97. http://www.aclweb.org/anthology/C10-1011.

Boroş, Tiberiu, Sonia Pipa, Verginica Barbu Mititelu & Dan Tufiş. 2017. A data-driven approach to verbal multiword expression detection. PARSEME shared task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 121–126. Association for Computational Linguistics. DOI:10.18653/v1/W17-1716

Candito, Marie & Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 743–753. Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-1070.

Constant, Matthieu & Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 161–171. Association for Computational Linguistics. http://www.aclweb.org/anthology/P16-1016.

Green, Spence, Marie-Catherine de Marneffe, John Bauer & Christopher D. Manning. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 725–735. Association for Computational Linguistics. http://www.aclweb.org/anthology/D11-1067.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39(1). 195–227. DOI:10.1162/COLI_a_00139

Klyueva, Natalia, Antoine Doucet & Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 60–65. Association for Computational Linguistics. April 4, 2017. DOI:10.18653/v1/W17-1707

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 114–120. Association for Computational Linguistics. DOI:10.18653/v1/W17-1715

Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1116–1126. Association for Computational Linguistics. http://www.aclweb.org/anthology/P15-1108.

Nerima, Luka, Vasiliki Foufi & Eric Wehrli. 2017. Parsing and MWE detection: Fips at the PARSEME shared task. In *Proceedings of The 13th Workshop on Multiword Expressions* (MWE '17), 54–59. Association for Computational Linguistics. DOI:10.18653/v1/W17-1706

Nivre, Joakim. 2015. Towards a universal grammar for natural language processing. In Alexander Gelbukh (ed.), *Computational Linguistics and intelligent text processing*, 3–16. Cham: Springer. http://stp.lingfil.uu.se/~nivre/docs/nivre15cicling.pdf.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1471591

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th*

Workshop on Multiword Expressions (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Simkó, Katalin Ilona, Viktória Kovács & Veronika Vincze. 2017. USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. In *Proceedings of The 13th Workshop on Multiword Expressions* (MWE '17), 48–53. Association for Computational Linguistics.

Vincze, Veronika, János Zsibrita & István Nagy T. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, 207–215. Nagoya, Japan: Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/I13-1024.

Waszczuk, Jakub, Agata Savary & Yannick Parmentier. 2016. Promoting multiword expressions in a* TAG parsing. In Nicoletta Calzolari, Yuji Matsumoto & Rashmi Prasad (eds.), *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers* (COLING-16), 429–439. Association for Computational Linguistics. http://aclweb.org/anthology/C/C16/C16-1042.pdf. December 11-16, 2016.

Wehrli, Eric, Violeta Seretan & Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of the Workshop on Multiword Expressions: From Theory to Applications* (MWE '10), 27–35. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W10/W10-??04.

Chapter 9

Semi-automated resolution of inconsistency for a harmonized multiword-expression and dependency-parse annotation

Julian Brooke
The University of Melbourne

King Chan
The University of Melbourne

Timothy Baldwin
The University of Melbourne

This chapter presents a methodology for identifying and resolving various kinds of inconsistency in the context of merging dependency and multiword expression (MWE) annotations, to generate a dependency treebank with comprehensive MWE annotations. Candidates for correction are identified using a variety of heuristics, including an entirely novel one which identifies violations of MWE constituency in the dependency tree, and resolved by arbitration with minimal human intervention. Using this technique, we identified and corrected several hundred inconsistencies across both parse and MWE annotations, representing changes to a significant percentage (well over 10%) of the MWE instances in the joint corpus and a large difference in MWE tagging performance relative to earlier versions.

1 Introduction

The availability of gold-standard annotations is important for the training and evaluation of a wide variety of Natural Language Processing (NLP) tasks, includ-

Julian Brooke, King Chan & Timothy Baldwin. 2018. Semi-automated resolution of inconsistency for a harmonized multiword-expression and dependency-parse annotation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 245–262. Berlin: Language Science Press. DOI:10.5281/zenodo.1469565

ing the evaluation of dependency parsers (Buchholz & Marsi 2006). In recent years, there has been a focus on multi-annotation of a single corpus, such as joint syntactic, semantic role, named entity, coreference and word sense annotation in Ontonotes (Hovy et al. 2006) or constituency, semantic role, discourse, opinion, temporal, event and coreference (among others) annotation of the Manually Annotated Sub-Corpus of the American National Corpus (Ide et al. 2010). As part of this, there has been an increased focus on harmonizing and merging existing annotated data sets as a means of extending the scope of reference corpora (Ide & Suderman 2007; Declerck 2008; Simi et al. 2015). This effort sometimes presents an opportunity to address conflicts among annotations, a worthwhile endeavour since even a small number of errors in a gold-standard syntactic annotation can, for example, result in significant changes in downstream applications (Habash et al. 2007). This chapter presents the results of a harmonization effort for the overlapping STREUSLE annotation (Schneider, Onuffer, et al. 2014) of multiword expressions (MWEs: Baldwin & Kim 2010) and dependency parse structure in the English Web Treebank (EWT: Bies et al. 2012), with the long-term goal of building reliable resources for joint MWE/syntactic parsing (Constant & Nivre 2016).

As part of merging these two sets of annotations, we use analysis of cross-annotation and type-level consistency to identify instances of potential annotation inconsistency, with an eye to improving the quality of the component and combined annotations. It is important to point out that our approach to identifying and handling inconsistencies does not involve re-annotating the corpus; instead we act as arbitrators, resolving inconsistency in only those cases where human intervention is necessary. Our three methods for identifying potentially problematic annotations are:

- a cross-annotation heuristic that identifies MWE tokens whose parse structure is incompatible with the syntactic annotation of the MWE;
- a cross-type heuristic that identifies *n*-grams with inconsistent token-level MWE annotations; and
- a cross-type, cross-annotation heuristic that identifies MWE types whose parse structure is inconsistent across its token occurrences.

The first of these is specific to this harmonization process, and as far as we are aware, entirely novel. The other two are adaptions of an approach to improving syntactic annotations proposed by Dickinson & Meurers (2003). After applying these heuristics and reviewing the candidates, we identified hundreds of errors in MWE annotation and about a hundred errors in the original syntactic annotations. We make available a tool that applies these fixes in the process of joining the two annotations into a single harmonized, corrected annotation, and release the harmonized annotations in the form of HAMSTER (the HArmonized Multiword and Syntactic TreE Resource): https://github.com/eltimster/HAMSTER. This chapter goes beyond the MWE2017 paper that first introduced HAMSTER (Chan et al. 2017) to show that the application of these and other corpus fixes has a major effect on MWE identification performance: we find that almost a quarter of the error originally assumed to be tagger error is actually attributable to errors in the corpus.

2 Related work

Our long-term goal is building reliable resources for joint MWE/syntactic parsing. Explicit modelling of MWEs has been shown to improve parser accuracy (Nivre 2004; Seretan & Wehrli 2006; Finkel & Manning 2009; Korkontzelos & Manandhar 2010; Green et al. 2013; Vincze et al. 2013; Wehrli 2014; Candito & Constant 2014; Constant & Nivre 2016). Treatment of MWEs has typically involved parsing MWEs as single lexical units (Nivre 2004; Eryiğit et al. 2011; Fotopoulou et al. 2014), but this flattened, "words with spaces" (Sag et al. 2002) approach is inflexible in its coverage of MWEs where components have some level of flexibility.

The English Web Treebank (Bies et al. 2012) represents a gold-standard annotation effort over informal web text. The original syntactic constituency annotation of the corpus was based on hand-correcting the output of the Stanford Parser (Manning et al. 2014); for our purposes we have converted this into a dependency parse using the Stanford Typed Dependency converter (de Marneffe et al. 2006). We considered the use of the Universal Dependencies representation (Nivre et al. 2016), but we noted that several aspects of that annotation (in particular the treatment of all prepositions as case markers dependent on their noun) make it inappropriate for joint MWE/syntactic parsing since it results in large numbers of MWEs that are non-continuous in their syntactic structure (despite being continuous at the token level).[1] As such, the Stanford Typed Dependencies is the representation which has the greatest currency for joint MWE/syntactic parsing work (Constant & Nivre 2016).

The STREUSLE corpus (Schneider, Onuffer, et al. 2014) is based entirely on the Reviews subset of the EWT, and comprises 3,812 sentences representing 55,579 tokens. The annotation was completed by six linguists who are native English

[1] An example of this would be a phrase such as ***think of*** *home*, where we consider ***think of*** to be an MWE, but the Universal Dependencies framework would treat *of* as a syntactic dependent of *home*, not *think*.

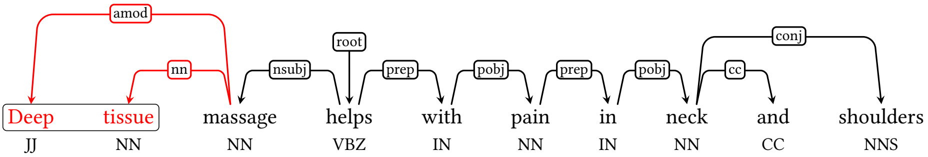

amod
nn
nsubj
root
prep
pobj
prep
pobj
conj
cc
Deep tissue massage helps with pain in neck and shoulders
JJ NN NN VBZ IN NN IN NN CC NNS

Figure 1: An example where the arc count heuristic is breached. ***Deep tissue*** has been labeled in the sentence here as an MWE in STREUSLE. *Deep* and *tissue* act as modifiers to *massage*, a term that has not been included as part of the MWE.

speakers. Every sentence was assessed by at least two annotators, which resulted in an average inter-annotator F1 agreement of 0.7. The idiosyncratic nature of MWEs lends itself to challenges associated with their interpretation, and this was readily acknowledged by those involved in the development of the STREUSLE corpus (Hollenstein et al. 2016). Two important aspects of the MWE annotation are that it includes both continuous and non-continuous MWEs (e.g., ***check* * *out***), and that it supports both WEAK and STRONG annotation. With regards to the latter, a variety of cues are employed to determine associative strength. The primary factor relates to the degree in which the expression is semantically opaque and/or morphosyntactically idiosyncratic. An example of a strong MWE would be ***top notch***, as used in the sentence: *We stayed at a **top notch** hotel.* The semantics of this expression are not immediately predictable from the meanings of *top* and *notch*. On the other hand, the expression ***highly recommend*** is considered to be a weak expression as it is largely compositional – one can ***highly recommend*** a product – as indicated by the presence of alternatives such as ***greatly recommend*** which are also acceptable though less idiomatic. A total of 3,626 MWE instances were identified in STREUSLE, across 2,334 MWE types.

Other MWE-aware dependency treebanks include the various UD treebanks (Nivre et al. 2016), the Prague Dependency Treebank (Bejček et al. 2013), the Redwoods Treebank (Oepen et al. 2002), and others (Nivre 2004; Eryiğit et al. 2011; Candito & Constant 2014). The representation of MWEs, and the scope of types covered by these treebanks, can vary significantly. For example, the internal syntactic structure may be flattened (Nivre 2004), or in the case of Candito & Constant (2014), allow for distinctions in the granularity of syntactic representation for regular vs. irregular MWE types.

The identification of inconsistencies in annotation requires comparisons to be made between similar instances that are labeled differently. Boyd et al. (2007) employed an alignment-based approach to assess differences in the annotation of *n*-gram word sequences in order to establish the likelihood of error occurrence. Other work in the syntactic inconsistency detection domain includes those related to POS tagging (Loftsson 2009; Eskin 2000; Ma et al. 2001) and parse structure (Ule & Simov 2004; Kato & Matsubara 2010). Dickinson & Meurers (2003) outline various approaches for detecting inconsistencies in parse structure within treebanks.

In general, inconsistencies associated with MWE annotation fall under two categories: (1) annotator error (i.e. false positives and false negatives); and (2) ambiguity associated with the assessment of hard cases. While annotation errors apply to situations where a correct label can be applied but is not, hard cases are those where the correct label is inherently difficult to assign. We address both these categories in this work.

3 Error candidate identification

3.1 MWE syntactic constituency conflicts

The hypothesis that drives our first analysis is that for nearly all MWE types, the component words of the MWE should be syntactically connected, which is to say that every word is a dependent of another word in the MWE, except one word which connects the MWE to the rest of the sentence (or the root of the sentence). We can realise this intuition by using an arc-count heuristic: for each labeled MWE instance we count the number of incoming dependency arcs that are headed by a term outside the MWE, and if the count is greater than one, we flag it for manual analysis. Figure 1 gives an example where the arc count heuristic is breached since both terms of the MWE ***deep tissue*** act as modifiers to the head noun that sits outside the MWE.

3.2 MWE type inconsistency

Our second analysis involves first collecting a list of all MWE types in the STREUSLE corpus, corresponding to lemmatized *n*-grams, possibly with gaps. We then match these *n*-grams across the same corpus, and flag any MWE type which has at least one inconsistency with regards to the annotation. That is, we extract as candidates any MWE types where there were at least two occurrences of the corresponding *n*-gram in the corpus that were incompatible with respect to their annotation in STREUSLE, including discrepancies in weak/STRONGISH designation. For non-continuous MWE types, matches containing up to 4 words of intervening context between the two parts of the MWE type were included as candidates for further assessment. Some examples of many *n*-gram types which showed inconsistency in their MWE annotation include ***interested in***, ***high quality***, ***ask for***, ***in town***, ***pizza place***, ***even though***, and ***easy to work with***.

3.3 MWE type parse inconsistency

The hypothesis that drives our third analysis is that we would generally expect the internal syntax of an MWE type to be consistent across all its instances.[2] For each MWE type, we extracted the internal dependency structure of all its labeled instances, and flagged for further assessment any type for which the parse structure (including typed dependency label) varied between at least two of those instances. Note that although this analysis is aimed at fixing parse errors, it makes direct use of the MWE annotation provided by STREUSLE to greatly limit the scope of error candidates to those which are most relevant to our interest. Some MWE types which showed syntactic inconsistency include ***years old***, ***up front***, ***set up***, ***check out***, ***other than***, and ***get in touch with***.

4 Error arbitration

Error arbitration was carried out by the authors (all native English speakers with experience in MWE identification), with at least two authors looking at each error candidate in most instances, and for certain difficult cases, the final annotation being based on discussion among all three authors. One advantage of our arbitration approach over a traditional token-based annotation was that we could enforce consistency across similar error candidates (e.g., *disappointed with* and *happy with*) and also investigate non-candidates to arrive at a consensus; where at all possible, our changes relied on precedents that already existed in the relevant annotation.

Arbitration for the MWE syntax conflicts usually involved identifying an error in one of the two annotations, and in most cases this was relatively obvious. For instance, in the candidate *… the usual lady called in sick hours earlier*, ***called in sick*** was correctly labeled as an MWE, but the parse incorrectly includes *sick* as a dependent of *hours*, rather than *called in*. An example of the opposite case is *… just to make the appointment …*, where *make the* had been labeled as an MWE, an obvious error which was caught by our arc count heuristic. There were cases where our arc count heuristic was breached due to what we would view as a

[2] Noting that we would not expect this to occur between MWE instances of a given combination of words, and non-MWE combinations of those same words.

general inadequacy in the syntactic annotation, but we decided not to effect a change because the impact would be too far reaching; examples of this were certain discourse markers (e.g., *as soon as*), and infinitives (e.g., *have to complete* where the *to* is considered a dependent of its verb rather than of the other term in the MWE ***have to***). The most interesting cases were a handful of non-continuous MWEs where there was truly a discontinuity in the syntax between the two parts of the MWE, for instance ***no amount of*** * ***can***. This suggests a basic limitation in our heuristic, although the vast majority of MWEs did satisfy it.

For the two type-level arbitrations, there were cases of inconsistency upheld by real usage differences (e.g., *a little house* vs. *a little tired*). We identified clear differences in usage first and divided the MWE types into sets, excluding from further analysis non-MWE usages of MWE type *n*-grams. For each consistent usage of an MWE type, the default position was to prefer the majority annotation across the set of instances, except when there were other candidates that were essentially equivalent: for instance, if we had relied on majority annotation for ***job*** * ***do*** (e.g., *the **job** that he **did***) it would have been a different annotation than ***do*** * ***job*** (e.g., ***do** a good **job***), so we considered these two together. We treated continuous and non-continuous versions of the same MWE type in the same manner.

In the MWE type consistency arbitration, for cases where majority rules did not provide a clear answer and there was no overwhelming evidence for non-compositionality, we introduced a special internal label called *hard*. These correspond to cases where the usage is consistent and the inconsistency seems to be a result of the difficulty of the annotation item (as discussed earlier in Section 2), which extended also to our arbitration. Rather than enforce a specific annotation without strong evidence or allow the inconsistency to remain when there is no usage justification for it, the corpus merging and correction tool gives the user the option to treat hard annotated MWEs in varying ways: the annotation may be kept unchanged, removed, converted to weak, or converted to hard for the purpose of excluding it from evaluation. Examples of hard cases include ***go back, go in, more than, talk to, speak to, thanks guys, not that great, pleased with, have*** * ***option, get*** * ***answer, fix*** * ***problem***. On a per capita basis, inconsistencies are more common for non-continuous MWEs relative to their continuous counterparts, and we suspect that this is partially due to their tendency to be weaker, in addition to the challenges involved in correctly discerning the non-continuous parts, which are sometimes at a significant distance from each other.

Table 1 provides a summary of changes to MWE annotation at the MWE type and token levels. *Mixed* refer to MWEs that are heterogeneous in the associative strength between terms in the MWE (between weak and strongish). Most of the changes in Table 1 (98% of the types) were the result of our type consistency analysis. Almost half of the changes involved the use of the *hard* label, but even excluding these (since only some of these annotations required actual changes in the final version of the corpus) our changes involve over 10% of the MWE tokens in the corpus, and thus represent a significant improvement to the STREUSLE annotation.

Relative to the changes to the MWE annotation, the changes to the parse annotation were more modest, but still not insignificant: for 161 MWE tokens across 72 types, we identified and corrected a dependency and/or POS annotation error. The majority of these (67%) were identified using the arc count heuristic. Note we applied the parse relevant heuristics after we fixed the MWE type consistency errors, ensuring that MWE annotations that were added were duly considered for parse errors.

Table 1: Summary of changes to MWE annotation at the MWE type and token level.

		No MWE	Weak	Strong	Mixed	Hard	TOTAL
Token	**No MWE**	—	55	136	6	151	348
	Weak	35	—	22	4	46	107
	Strong	44	42	—	9	70	165
	Mixed	2	4	3	12	2	23
	TOTAL	81	101	161	31	269	643
Type	**No MWE**	—	31	74	5	64	174
	Weak	31	—	13	4	35	83
	Strong	34	28	—	7	43	112
	Mixed	2	4	3	7	2	18
	TOTAL	67	63	90	23	144	387

5 Experiments

In this section we investigate the effect of the HAMSTER MWE inconsistency fixes on the task of MWE identification. For this we use the AMALGr MWE identification tool of Schneider, Danchik, et al. (2014), which was developed on the initial release of the STREUSLE (called then the CMWE).[3] AMALGr is a supervised structured perceptron model which makes use of external resources including 10 MWE lexicons as well as Brown cluster information. For all our experiments we use the default settings from Schneider, Danchik, et al. (2014), including the original train/test splits and automatic part-of-speech tagging provided by the ARK TweetNLP POS tagger (Owoputi et al. 2013) trained on the all non-review sections of the English Web Treebank. We note that in contrast to typical experiments in NLP, here we are holding *the approach* constant while varying the quality of the dataset, which provides a quantification of the extent to which errors in the dataset interfered with our ability to build or accurately evaluate models. Following Schneider, Danchik, et al. (2014), we report an F-score which is calculated based on links between words: a true positive occurs when two words which are supposed to appear together in an MWE do so as expected.

Table 2: AMALGr F-scores for various versions of MWE annotation of EWT Reviews.

Dataset	**F1-score (%)**
CMWE (Schneider, Danchik, et al. 2014)	59.4
STREUSLE 3.0	64.6
HAMSTER-original	69.1
HAMSTER-notMWE	68.2
HAMSTER-weak	69.4
HAMSTER-original-noeval	70.2
HAMSTER-weak-noeval	69.3
HAMSTER-original-test	67.1
HAMSTER-original-train	65.7

There are two baselines in Table 2: the first is the original performance of AMALGr as reported in Schneider, Danchik, et al. (2014) using CMWE (version 1.0 of this annotation), and the second is its performance using STREUSLE (version 3.0). Note that these involve exactly the same texts: the difference between these two numbers reflects other fixes to this dataset that have happened in the

[3] The key difference between the CMWE and STREUSLE is the inclusion of supersense tags. Though we hope to eventually include supersense information in the output of HAMSTER, supersenses are beyond the scope of the present work.

years since its initial release. The difference between the two is quite substantial, at roughly 5% F-score.

The rest of the table makes use of HAMSTERized versions of STREUSLE, which we refer to as simply HAMSTER. The options here mostly refer to our treatment of the *hard* cases, which must be removed to make use of AMALGr. *-original* indicates that we apply all fixes which result in the creation or removal of a standard STREUSLE label (i.e., weak and strongish), but leave *hard* annotations as they were in the original corpus. *-notMWE* and *-weak* create versions of the corpus where all *hard* labels have been mapped to either nothing (no MWE) or weak MWEs, respectively. Another option we consider is *-noeval*, which involved tweaking the AMALGr evaluation script to exclude particular annotations (in this case *hard*) from evaluation altogether; that is, it does not matter what the model predicted for those words which are considered *hard*. Finally, *-test* and *-train* refer to the situation where we apply our fixes to texts only in the test or training sets, respectively; this gives us a sense of whether the improved performance of the model over the HAMSTER datasets is primarily due to the removal of errors from the test set, or whether improving the consistency of the training set is playing a major role as well.

Our fixes result in roughly another 5% increase to F-score relative to STREUSLE 3.0, for a total of about 10% F-score difference relative to results using the original CMWE annotation of this corpus. With respect to options for phrases labeled as *hard*, treating them as nonMWEs seems to be a worse option than simply leaving them alone; the best explanation for this is probably that these hard cases are generally more similar to labelled MWEs. Treating them as weak appears to a better strategy. Even better, though, might be to leave *hard* inconsistencies in the training set but exclude them from consideration during testing. The results using mixed training/test datasets indicate that the fixes to the test data are clearly more important, but the consistency across the two sets also accounts for a major part of the performance increase seen here.

Our second round of experiments looks at exact match recall with respect to various subsets of the MWEs in the test set. Here we consider only the original STREUSLE and HAMSTERized version with *hard* MWEs unchanged. *N* is the number of MWEs labeled as that type in that version of the dataset. Our goal here is to get a sense of how our changes have affected the identification of specific kinds of MWE. Weak versus strongish is an obvious distinction (mixed MWE were considered strongish), but even more relevant to what we have done here is whether or not the MWE appears in both the training and test sets. We are also interested in the status of multiword named entities (identified fairly reliably using proper noun tags in the gold-standard POS tags), which occur numerously in a corpus of reviews, but often as singletons, i.e., with a frequency of one. We would expect MWEs which neither appear in our corpus nor are named entities (NEs) to be relatively unaffected by our fixes, and among the most challenging MWEs to identify in general.

Table 3: AMALGr exact recall for different MWE subsets in original and HAMSTERized STREUSLE.

	STREUSLE		HAMSTER	
MWE types	*N*	**Recall (%)**	*N*	**Recall (%)**
All	423	59.7	444	63.4
strongish	352	63.2	368	66.3
weak	71	24.0	76	35.5
In training	178	77.7	208	80.1
Not in training	247	47.4	238	49.4
Named entity (NE)	52	73.5	52	71.6
Not NE, not in training	195	40.3	186	43.9

In Table 3, AMALGr does better with the HAMSTER dataset for most of the MWE subtypes considered here. The most striking difference occurs for the weak tag, reflecting a disproportionate amount of inconsistency, enough that the model built on the earlier version was apparently hesitant to apply the tag at all. Not only are MWEs with training instances tagged better after our fixes, but the set of such MWE tokens has noticeably increased. There is a corresponding drop in those test instances without training data, which are clearly the most difficult to identify, particularly when named entities are excluded. The recall of named entities has actually dropped slightly, though since there are only 52 of these in the test set, this corresponds to a single missed example and is probably not meaningful. Though the rationale in terms of higher-level semantics is clear, we wonder whether including NER as part of MWE identification may result in a distorted view of the importance of MWE lexicons in token-level MWE identification. Here, we can see that among test-set-only MWEs, they stand out as being significantly easier than the rest, probably because in English they can be identified fairly reliably using only capitalization.

6 Discussion

Our three heuristics are useful because they identify potential errors with a high degree of precision. For the MWE type consistency analysis, 77% of candidate types were problematic, and for parse type consistency, the number was 63%. For the arc count heuristic, 54% of candidate types were ultimately changed: as mentioned earlier, many of the breaches involved systematic issues with annotation schema that we felt uncomfortable changing in isolation. By bringing these candidate instances to our attention, we were able to better focus our manual analysis effort, including in some cases looking across multiple related types, or even searching for specialist knowledge which could resolve ambiguities: for instance, in the example shown in Figure 1, though a layperson without reference material may be unsure whether it is tissue or massage which is considered to be deep, a quick online search indicates that the original EWT syntax is in error (*deep* modifies *tissue*).

However, it would be an overstatement to claim to have fixed all (or even almost all) the errors in the corpus. For instance, our type consistency heuristics only work when there are multiple instances of the same type, yet it is worth noting that 82% of the MWE types in the corpus are represented by a singleton instance. Our arc count heuristic can identify issues with singletons, but its scope is fairly limited. We cannot possibly identify missing annotations for types that were not annotated at least once. We might also miss certain kinds of systematic annotation errors, for instance those mentioned in De Smedt et al. (2015), though that work focused on the use of MWE dependency labels which are barely used in the EWT, one of the reasons a resource like STREUSLE is so useful.

Our experiments with the AMALGr tool show that our fixes result in a major improvement in MWE identification. One particularly striking result is the fact that the errors identified in the annotation since its original release account for about a quarter of all error (as measured by F-score) in the original model trained on it. This error may affect relative comparisons between systems, and we should be skeptical of results previously drawn based on relatively small differences in MWE identification in earlier versions of the corpus (e.g., Qu et al. 2015). This amount of error is also unacceptable simply in terms of the obfuscation relative to the degree of absolute progress on the task. Beyond this specific effort, we believe, for annotation efforts in general and for MWEs in particular, we should move beyond a singular focus on achieving sufficient annotator agreement in the initial annotation – the agreement in the original CWME was impressively high — and instead develop protocols for semi-automated, type-level inconsistency detection as a default step before any annotation is released.

7 Conclusion

We have proposed a methodology for merging MWE and dependency parse annotations, to generate HAMSTER: a gold-standard MWE-annotated dependency treebank with high consistency. The heuristics used to enforce consistency operate at the type- and cross-annotation level, and affected well over 10% of the MWEs in the new resource, resulting in a downstream change in MWE identification of roughly 5% F-score. More generally, we have provided here a case study in how bringing together multiple kinds of annotation done over the same corpus can facilitate rigorous error correction as part of the harmonization process.

Abbreviations

AMALGr	A Machine Analyzer of Lexical Groupings
CMWE	The Comprehensive Multiword Expression Corpus
EWT	The English Web Treebank
HAMSTER	The Harmonized Multiword and Syntactic Tree Resource
MWE	multiword expression
NE	named entity
NER	named entity recognition
NLP	Natural Language Processing
STREUSLE	Supersense-Tagged Repository of English with a Unified Semantics for Lexical Expressions

References

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

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovskỳ, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek & Šárka Zikánová. 2013. *Prague dependency treebank 3.0.* Charles University in Prague, UFAL.

Bies, Ann, Justin Mott, Colin Warner & Seth Kulick. 2012. *English web treebank.* Tech. rep. LDC2012T13. Linguistic Data Consortium. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2012T13.

Boyd, Adriane, Markus Dickinson & Detmar Meurers. 2007. Increasing the recall of corpus annotation error detection. In *Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories* (TLT 2007), 19–30. http://decca.osu.edu/publications/boyd-et-al-07b.html.

Buchholz, Sabine & Erwin Marsi. 2006. CoNLL-x shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning* (CoNLL-X '06), 149–164. http://dl.acm.org/citation.cfm?id=1596276.1596305.

Candito, Marie & Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 743–753. Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-1070.

Chan, King, Julian Brooke & Timothy Baldwin. 2017. Semi-automated resolution of inconsistency for a harmonized multiword expression and dependency parse annotation. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 187–193. Valencia, Spain. http://aclweb.org/anthology/W17-1726.

Constant, Matthieu & Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 161–171. Association for Computational Linguistics. http://www.aclweb.org/anthology/P16-1016.

de Marneffe, Marie-Catherine, Bill MacCartney & Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* (LREC 2006). European Language Resources Association (ELRA).

De Smedt, Koenraad, Victoria Rosén & Paul Meurer. 2015. Studying consistency in UD treebanks with INESS-Search. In *Proceedings of the 14th Workshop on Treebanks and Linguistic Theories* (TLT14), 258–267.

Declerck, Thierry. 2008. A framework for standardized syntactic annotation. In *Proceedings of the 6th international on language resources and evaluation* (LREC 2008), 3025–3028.

Dickinson, Markus & W. Detmar Meurers. 2003. Detecting inconsistencies in treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories* (TLT 2003), 45–56. 14-15 November, 2003.

Eryiğit, Gülşen, Tugay İlbay & Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of IWPT Workshop on Statistical*

Parsing of Morphologically-Rich Languages (SPMRL 2011), 45–55. http://dl.acm.org/citation.cfm?id=2206359.2206365. October 6, 2011.

Eskin, Eleazar. 2000. Detecting errors within a corpus using anomaly detection. In *Proceedings of the First North American Chapter of the Association for Computational Linguistics Conference*, 148–153. http://dl.acm.org/citation.cfm?id=974305.974325. April 29 - May 04, 2000.

Finkel, Jenny Rose & Christopher D. Manning. 2009. Joint parsing and named entity recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL '09), 326–334. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1620754.1620802. May 31 - June 05, 2009.

Fotopoulou, Angeliki, Stella Markantonatou & Voula Giouli. 2014. Encoding MWEs in a conceptual lexicon. In *Proceedings of the 10th Workshop on Multiword Expressions* (MWE '14), 43–47. Association for Computational Linguistics.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39(1). 195–227. DOI:10.1162/COLI_a_00139

Habash, Nizar, Ryan Gabbard, Owen Rambow, Seth Kulick & Mitchell P. Marcus. 2007. Determining case in Arabic: Learning complex linguistic behavior requires complex linguistic features. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2007* (EMNLP-CoNLL 2007), 1084–1092.

Hollenstein, Nora, Nathan Schneider & Bonnie Webber. 2016. Inconsistency detection in semantic annotation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), 3986–3990. European Language Resources Asscociation (ELRA).

Hovy, Eduard, Mitchell P. Marcus, Martha Palmer, Lance Ramshaw & Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Main Conference of Human Language Technology of the North American Chapter of the Association for Computiatonal Linguistics*, 57–60.

Ide, Nancy, Collin Baker, Christiane Fellbaum & Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus: A community resource for and by the people. In *Proceedings of the 48th annual meeting of the ACL* (ACL 2010)- *short papers*, 68–73. Uppsala, Sweden.

Ide, Nancy & Keith Suderman. 2007. GrAF: A Graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, 1–8. Prague, Czech Republic. http://dl.acm.org/citation.cfm?id=1642059.1642060.

Kato, Yoshihide & Shigeki Matsubara. 2010. Correcting errors in a treebank based on synchronous tree substitution grammar. In *Proceedings of the ACL 2010 Conference-Short Papers*, 74–79. Association for Computational Linguistics. July 11 - 16, 2010.

Korkontzelos, Ioannis & Suresh Manandhar. 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* (HLT '10), 636–644. http://dl.acm.org/citation.cfm?id=1857999.1858088.

Loftsson, Hrafn. 2009. Correcting a POS-tagged corpus using three complementary methods. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (EACL '09), 523–531. Association for Computational Linguistics.

Ma, Qing, Bao-Liang Lu, Masaki Murata, Michnori Ichikawa & Hitoshi Isahara. 2001. On-line error detection of annotated corpus using modular neural networks. In *International Conference on Artificial Neural Networks*, 1185–1192.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard & David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd annual meeting of the association for computational linguistics: System demonstrations*, 55–60.

Nivre, Joakim. 2004. Incrementality in deterministic dependency parsing. In Frank Keller, Stephen Clark, Matthew Crocker & Mark Steedman (eds.), *Proceedings of the ACL Workshop on Incremental Parsing: Bringing Engineering and Cognition together*, 50–57. Association for Computational Linguistics.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Oepen, Stephan, Dan Flickinger, Kristina Toutanova & Christoper D. Manning. 2002. LinGO Redwoods: A rich and dynamic treebank for HPSG. In *Proceedings of The First Workshop on Treebanks and Linguistic Theories* (TLT2002).

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider & Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

Qu, Lizhen, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, Nathan Schneider & Timothy Baldwin. 2015. Big Data Small Data, In Domain Out-of Domain, Known Word Unknown Word: the impact of word representations on sequence labelling tasks. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning* (CoNLL 2015), 83–93.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* 2(1). 193–206. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/281.

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Nora Emily Danchik, Michael T. Mordowanec, Henrietta Conrad & Smith Noah A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), 455–461. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf.

Seretan, Violeta & Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 953–960. 17-18 July 2006.

Simi, Maria, Simonetta Montemagni & Cristina Bosco. 2015. Harmonizing and merging Italian treebanks: Towards a merged Italian dependency treebank and beyond. In Roberto Basili, Cristina Bosco, Rodolfo Delmonte, Alessandro Moschitti & Maria Simi (eds.), *Harmonization and development of resources and tools for Italian Natural Language Processing within the PARLI project*, 3–23. Heidelberg, Germany: Springer. DOI:10.1007/978-3-319-14206-7_1

Ule, Tylman & Kiril Simov. 2004. Unexpected productions May well be errors. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, Raquel Silva, Carla Pereira, Filipa Carvalho, Milene Lopes, Mónica Catarino & Sérgio Barros (eds.), *Proceedings of the 4th international confernece on language resources and evaluation* (LREC 2004), 1795–1798.

Vincze, Veronika, János Zsibrita & István Nagy T. 2013. Dependency parsing for identifying Hungarian light verb constructions. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, 207–215. Nagoya, Japan: Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/I13-1024.

Wehrli, Eric. 2014. The relevance of collocations for parsing. In *Proceedings of the 10th Workshop on Multiword Expressions* (MWE '14), 26–32. Association for Computational Linguistics. 26-27 April, 2014.

Chapter 10

Sequence models and lexical resources for MWE identification in French

Manon Scholivet
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Carlos Ramisch
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

Silvio Cordeiro
Institute of Informatics, Federal University of Rio Grande do Sul, Brazil
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France

We present a simple and efficient sequence tagger capable of identifying continuous multiword expressions (MWEs) of several categories in French texts. It is based on conditional random fields (CRF), using as features local context information such as previous and next word lemmas and parts of speech. We show that this approach can obtain results that, in some cases, approach more sophisticated parser-based MWE identification methods without requiring syntactic trees from a treebank. Moreover, we study how well the CRF can take into account external information coming from both high-quality hand-crafted lexicons and MWE lists automatically obtained from large monolingual corpora. Results indicate that external information systematically helps improving the tagger's performance, compensating for the limited amount of training data.

1 Introduction

Identifying multiword expressions (MWEs) in running texts with the help of lexicons could be considered as a trivial search-and-replace operation. In theory, one could simply scan the text once and mark (e.g. join with an underscore)

Manon Scholivet, Carlos Ramisch & Silvio Cordeiro. 2018. Sequence models and lexical resources for MWE identification in French. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 263–297. Berlin: Language Science Press. DOI:10.5281/zenodo.1469567

all sequences of tokens that appear as headwords in the MWE lexicons. Direct matching and projection of lexical entries onto the corpus can be employed as a simple yet effective preprocessing step prior to dependency parsing (Nivre & Nilsson 2004) and machine translation (Carpuat & Diab 2010). Upon recognition, the identified member words of an MWE can be concatenated and treated as single token, that is, a “word with spaces”, as suggested by Sag et al. (2002).

However, this simple pipeline will fail when dealing with frequent categories of MWEs that present some challenging characteristics such as **variability** and **ambiguity**. For many MWE categories, **variability** due to morphological inflection may pose problems. For instance, if a lexicon contains the idiom *to make a face*, string matching will fail to identify it in *children are always **making faces*** because the verb and the noun are inflected.[1] Since lexicons usually contain canonical (lemmatised) forms, matching must take inflection into account. This can be carried out by (a) pre-analysing the text and matching lemmas instead of surface-level word forms (Finlayson & Kulkarni 2011), or by (b) looking up lexicons containing inflected MWEs (Silberztein et al. 2012).

Things get more complicated when the target MWEs are ambiguous, though. An MWE is **ambiguous** when its member words can co-occur without forming an expression. For instance, *to make a face* is an idiom meaning ‘to show a funny facial expression’, but it can also be used literally when someone is making a snowman (Fazly et al. 2009). Additionally, the words in this expression can co-occur by chance, not forming a phrase (Boukobza & Rappoport 2009; Shigeto et al. 2013). This is particularly common for multiword function words such as prepositions (e.g. *up to*), conjunctions (e.g. *now that*) and adverbials (e.g. *at all*). For example, *up to* is an MWE in *they accepted **up to** 100 candidates* but not in *you should look it up to avoid making a typo*. Similarly, *at all* is an adverbial in *they accepted no candidates **at all***, but not in *this train does not stop at all stations*. Context-dependent statistical methods (Fazly et al. 2009; Boukobza & Rappoport 2009) and syntax-based methods (Candito & Constant 2014; Nasr et al. 2015) are usually employed to deal with semantic ambiguity and accidental co-occurrence, respectively.

In addition to variability and ambiguity, an additional challenge stems from the absence or limited coverage of high-quality hand-crafted lexical resources containing MWEs for many languages. Therefore, it is not always possible to em-

[1] In addition, the determiner *a* is not mandatory. However, discontinuous expressions containing optional intervening words are out of the scope of this work because our method is based on sequence models and our corpora only contain continuous MWEs. An adaptation of sequence models to discontinuous expressions has been proposed by Schneider, Danchik, et al. (2014).

ploy purely symbolic look-up methods for MWE identification. Statistical methods are an interesting alternative, since one can learn generic models for MWE identification based on corpora where MWEs have been manually annotated. If enough evidence is provided and represented at the appropriate level of granularity, the model can make generalizations based on commonly observed patterns. It may then be able to identify MWE instances that have never occurred in annotated training data. However, annotated corpora often do not contain enough training material for robust MWE identification. Complementary evidence can be obtained with the help of unsupervised MWE discovery methods that create MWE lists from raw corpora, which are then considered as if they were hand-crafted lexicons. In short, the heterogeneous landscape in terms of available resources (annotated corpora, hand-crafted lexicons) motivates the development of statistical MWE identification models that can exploit external hand-crafted and automatically constructed lexicons as a complementary information source (Constant & Sigogne 2011; Schneider, Danchik, et al. 2014; Riedl & Biemann 2016).

We propose a simple, fast and generic sequence model for tagging continuous MWEs based on conditional random fields (CRF). It cannot deal with discontinuous expressions, but is capable of modelling variable and highly ambiguous expressions. Moreover, we propose a simple adaptation to integrate information coming from external lexicons. Another advantage of our CRF is that we do not need syntactic trees to train our model, unlike methods based on parsers (Le Roux et al. 2014; Nasr et al. 2015; Constant & Nivre 2016). Moreover, for expressions that are syntactically fixed, it is natural to ask ourself if we really need a parser for this task. Parsers are good for non-continuous MWEs, but we hypothesise that continuous expressions can be modelled by sequence models that take ambiguity into account, such as CRFs. Regardless of the syntactic nature of these ambiguities, we expect that the highly lexicalised model of the CRF compensates for its lack of structure.

The present chapter contains three significant extensions with respect to our previous publication at the MWE 2017 workshop (Scholivet & Ramisch 2017). First, we train and test our models on two complementary datasets containing nominal expressions and general MWEs in French. Second, we study the integration of automatically constructed MWE lexicons obtained with the help of MWE discovery techniques. Third, we study the performance of our system on particularly hard MWE instances such as those including variants and those that do not occur in the training corpora.

In short, we demonstrate that, in addition to being well suited to identifying highly ambiguous MWEs in French (Scholivet & Ramisch 2017), the proposed model and its corresponding free implementation[2] can also be applied to identify other MWE categories and use other types of external lexicons. We believe that this approach can be useful (a) when no treebank is available to perform parsing-based MWE identification, (b) when large monolingual corpora are available instead of hand-crafted lexical resources, and (c) as a preprocessing step to parsing, which can improve parsing quality by reducing attachment ambiguities (Candito & Constant 2014; Nivre & Nilsson 2004).

2 Related work

Token identification of MWEs in running text can be modelled as a machine learning problem, building an identification model from MWE-annotated corpora and treebanks. To date, it has been carried out using mainly two types of models: sequence taggers and parsers. Sequence taggers such as CRFs, structured support vector machines and structured perceptron allow disambiguating MWEs using local feature sets such as word affixes and surrounding word and POS *n*-grams. Parsers, on the other hand, can take into account longer-distance relations and features when building a parse tree, at the expense of using more complex models.

Sequence taggers have been proven useful in identifying MWEs. MWE identification is sometimes integrated with POS tagging in the form of special tags. Experiments have shown the feasibility of sequence tagging for general expressions and named entities in English (Vincze et al. 2011), verb-noun idioms in English (Diab & Bhutada 2009) and general expressions in French (Constant & Sigogne 2011) and in English (Schneider, Danchik, et al. 2014; Riedl & Biemann 2016). Shigeto et al. (2013) tackle specifically English function words and build a CRF from the Penn Treebank, additionally correcting incoherent annotations. We develop a similar system for French, using the MWE annotation of the French Treebank as training data and evaluating the model on a dedicated dataset.

Parsing-based MWE identification requires a treebank annotated with MWEs. Lexicalised constituency parsers model MWEs as special non-terminal nodes included in regular rules (Green et al. 2013). In dependency parsers, it is possible to employ a similar approach, using special dependency labels to identify relations between words that make up an expression (Candito & Constant 2014).

The work of Nasr et al. (2015) is a parsing-based approach evaluated on highly ambiguous grammatical MWEs in French (§5.1). In their work, they link word

[2] The CRF-MWE tagger described in this chapter is included in the mwetoolkit in the form of 2 scripts: `train_crf.py` and `annotate_crf.py`, freely available at http://mwetoolkit.sf.net/

sequences belonging to complex conjunctions such as *bien que* 'well that' ⇒ 'although' and partitive determiner such as *de la* 'of the' ⇒ 'some', using a special dependency link called `morph`, similar to Universal Dependencies' `compound` relation (Nivre et al. 2016). On the other hand, these word sequences can occur by chance, such as in *Je pense bien que je suis malade.* 'I think well that I am sick.' ⇒ 'I really think that I am sick'. Then, the adverb *well* modifies the verb *think*, which in turn has a complement introduced by *that.* Nasr et al. (2015) train a second-order graph-based dependency parser to distinguish `morph` from other syntactic relations, implicitly identifying MWEs. In addition to standard features, they extract features from a valence dictionary specifying whether a given verb licences complements introduced by *que* 'that' or *de* 'of'.

Our hypothesis is that parsing-based techniques like this are not required to obtain good performances on continuous expressions. Our paper adapts a standard CRF model inspired on the ones proposed by Constant & Sigogne (2011), Riedl & Biemann (2016) and Shigeto et al. (2013) to deal with continuous MWEs.

Concerning external lexical resources, Nasr et al. (2015) have shown that their features extracted from a valence dictionary can significantly improve identification. Moreover, most systems based on sequence taggers also integrate additional hand-crafted lexicons to obtain good results (Constant & Sigogne 2011; Schneider, Danchik, et al. 2014). Nonetheless, the integration of automatically discovered lexicons of MWEs has not been explored by many authors, with the notable exception of Riedl & Biemann (2016). We show that our CRF can naturally handle automatically and manually constructed lexicons and that, in both cases, the system benefits from the extra information present in the lexicons.

3 A CRF-based MWE tagger

Linear-chain conditional random fields (CRFs) are an instance of stochastic models that can be employed for sequence tagging (Lafferty et al. 2001). Each input sequence T is composed of $t_1 \dots t_n$ tokens considered as an observation. Each observation is tagged with a sequence $Y = y_1 \dots y_n$ of tags corresponding to the values of the hidden states that generated them. CRFs can be seen as a discriminant version of hidden Markov models, since they model the conditional probability $P(Y|T)$. This makes them particularly appealing since it is straightforward to add customised features to the model. In first-order linear-chain CRFs, the probability of a given output tag y_i for an input word x_i depends on the tag of the neighbour token y_{i-1}, and on a rich set of features of the input $\phi(T)$, that can range over any position of the input sequence, including but not limited to the current token t_i. CRF training consists in estimating individual parameters proportional to $p(y_i, y_{i-1}, \phi(T))$.

The identification of continuous MWEs is a segmentation problem. We use a tagger to perform this segmentation, employing the well-known Begin-Inside-Outside (BIO) encoding (Ramshaw & Marcus 1995). In BIO, every token t_i in the training corpus is annotated with a corresponding tag y_i with values `B`, `I` or `O`. If the tag is `B`, it means the token is the beginning of an MWE. If it is `I`, this means the token is inside an MWE. `I` tags can only be preceded by another `I` tag or by a `B`. Finally, if the token's tag is `O`, this means the token is outside the expression, and does not belong to any MWE. An example of such encoding for the 2-word expression *de la* 'some' in French is shown in Figure 1.

i	-2	-1	0	1	2	3
w_i	*Il*	*jette*	*de*	*la*	*nourriture*	*périmée*
y_i	O	O	B	I	O	O
	He	*discards*	*some*		*food*	*expired*

Figure 1: Example of BIO tagging of a French sentence containing a *De*+determiner expression, assuming that the current word (w_0) is *de*.

For our experiments, we have trained a CRF tagger with the CRFSuite toolkit[3] (Okazaki 2007). We used a modified version of the French treebank (Abeillé et al. 2003) as training, development, and test data, and the MORPH dataset[4] (Nasr et al. 2015) as development and test data. We additionally include features from an external valence lexicon, DicoValence[5] (van den Eynde & Mertens 2003), and from an automatically constructed lexicon of nominal MWEs obtained automatically from the frWaC corpus (Baroni & Bernardini 2006) with the help of the mwetoolkit (Ramisch 2014).

3.1 CRF features

Our set of features $\phi(T)$ contains 37 different combinations of values (henceforth referred to as the ALL feature set). Our features are inspired on those proposed by Constant & Sigogne (2011), and are similar to those used by Schneider, Danchik, et al. (2014) and Riedl & Biemann (2016). The feature templates described below

[3] http://www.chokkan.org/software/crfsuite/

[4] http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/morph

[5] http://bach.arts.kuleuven.be/dicovalence/

consider that the current token T_0 has surface form W_0, lemma L_0 and POS P_0. In addition to output tag bigrams (CRF's first-order assumption), we consider the following feature templates in our model, to be regarded in conjunction with the current tag to predict:

- Single-token features (T_i):[6]
  - W_0 : wordform of the current token
  - L_0 : lemma of the current token
  - P_0 : POS tag of the current token
  - W_i, L_i and P_i: wordform, lemma or POS of previous ($i \in \{-1, -2\}$) or next ($i \in \{+1, +2\}$) tokens
- *N*-gram features (bigrams T_{i-1}T_i and trigrams $\text{T}_{i-1}\text{T}_i\text{T}_{i+1}$):
  - W_{i-1}W_i, L_{i-1}L_i, P_{i-1}P_i: wordform, lemma and POS bigrams of previous-current ($i = 0$) and current-next ($i = 1$) tokens
  - $\text{W}_{i-1}\text{W}_i\text{W}_{i+1}$,$\text{L}_{i-1}\text{L}_i\text{L}_{i+1}$, $\text{P}_{i-1}\text{P}_i\text{P}_{i+1}$: wordform, lemma and POS trigrams of previous-previous-current ($i = -1$), previous-current-next ($i = 0$) and current-next-next ($i = +1$) tokens
- Orthographic features (ORTH):
  - HYPHEN and DIGITS: the current wordform W_i contains a hyphen or digits
  - F-CAPITAL: the first letter of the current wordform W_i is uppercase
  - A-CAPITAL: all letters of the current wordform W_i are uppercase
  - B-CAPITAL: the first letter of the current word W_i is uppercase, and it is at the beginning of a sentence.
- Lexicon features (LF), described in more detail in §4.3:
  - QUEV: the current wordform W_i is *que*, and the closest verb to the left licences a complement introduced by *que* according to the valence dictionary DicoValence.[7]
  - DEV: the current wordform W_i is *de*, and the closest verb to the left licences a complement introduced by *de* according to the valence dictionary DicoValence.

[6] T_i is a shortcut denoting the group of features W_i, L_i and P_i for a token T_i. In other words, each token T_i is a tuple (W_i,L_i,P_i). The same applies to *n*-grams.

[7] QUEV and DEV are sequential versions of the *subcat features* proposed by Nasr et al. (2015).

- Association measures (AM) between the current token's lemma $\textsc{l}_i$ and the previous tokens' lemmas:
    - * MLE: probability of the lemma sequence estimated using maximum likelihood estimation
    - * PMI: pointwise mutual information of the lemma sequence.
    - * DICE: Dice's coefficient of the lemma sequence
    - * T-MEAS: Student's t-score of the lemma sequence
    - * LL: log-likelihood ratio between the current lemma and the previous lemma

Our proposed feature set is similar to previous work, with only minor differences (Constant & Sigogne 2011; Schneider, Onuffer, et al. 2014; Riedl & Biemann 2016). Like all previous models, we encode output tags with BIO, and we consider as features the surface form of the current token, of surrounding tokens, and bigrams of those. Our orthographic features are practically identical to related work, but all previously proposed models include 4- to 5-character prefixes and suffixes, which we do not. The features proposed by Constant & Sigogne (2011) are only based on surface forms of words, given that their task is to jointly predict POS and MWE tags. On the other hand, the features of Schneider, Onuffer, et al. (2014) and Riedl & Biemann (2016) are based on current and surrounding lemmas and POS tags, and so are ours. Differently from these two articles, we rely on token trigram features and we do not use mixed lemma+POS features. The lemma-based features of Schneider, Onuffer, et al. (2014) are quite different from ours, because they are conditioned on particular POS tags. The main differences between previous models and ours are in the lexicon features: Constant & Sigogne (2011) and Schneider, Onuffer, et al. (2014) employ hand-crafted lexicons and extract more detailed information from them than we do. Riedl & Biemann (2016) cover both hand-crafted and automatically built lexicons. Their feature set has one feature in common with ours: Student's t-score. In short, the features are similar in nature, but present some arbitrary variation in their implementations, in addition to some variation due to the nature of the available lexicons and corpora.

Our training corpora contain syntactic dependency information. However, we decided not to include it as CRF features for two main reasons. First, we wanted to evaluate the hypothesis that sequence-based methods can perform MWE identification without resorting to treebanks, as opposed to parser-based identification. Second, representing syntactic structure in a CRF is tricky as the linear-chain model in our experiments is not adequate for representing general graphs.

Nonetheless, adding features based on simplified syntactic information (e.g. the dependency label of each word with respect to its parent) is feasible and represents an interesting idea for future work.

4 Experimental setup

In order to evaluate our systems, we test them on four MWE categories in French:

- Adverb+*que* expressions (AQ): in French, adverbs (such as *bien* 'well') and the subordinating conjunction *que* 'that' are frequently combined to build complex conjunctions such as *bien que* 'well that' ⇒ 'although'. This category was chosen because (a) these expressions present little variability,[8] and (b) they are highly ambiguous, since their components can co-occur by chance, as in *il sait bien que tu mens.* 'he knows well that you lie.' ⇒ 'he really knows that you are lying'. Thus, we can focus on ambiguity as a challenging problem to model.

- *De*+determiner expressions (DD): in French, partitive and plural determiners are formed by the word *de* 'of' followed by a definite article, for instance, *il mange de la salade, du pain et des fruits.* 'he eats of the.SG.FEM salad, of-the.SG.MASC bread and of-the.PL fruit.' ⇒ 'he is eating some salad, bread and fruit'. Similarly to AQ, these constructions present little variation[9] and are ambiguous with preposition+article combinations such as *il parle de la salade* (lit. *he talks of the salad*) 'he talks about the salad'. Their disambiguation is challenging because it relies on the argumental structure of the verb governing the noun. Moreover, these are among the most frequent tokens in a corpus of French (Nasr et al. 2015).

- Nominal expressions: at a first moment, we focus on the identification of nominal expressions for two reasons. First, they present morphological variability but are syntactically fixed, making CRFs particularly suitable to model them. Second, we test the inclusion of automatically calculated association measures as features in the identification model, and our lexicon of pre-calculated association measures contains only nominal MWEs.

- General MWEs: finally, we evaluate our model on a corpus containing several categories of continuous MWEs. These include nominal expressions,

[8] The only variability that must be taken into account is that *que* is sometimes written as *qu'* when the next word starts with a vowel.

[9] Except for contractions *de+le=du* and *de+les=des*

complex conjunctions and determiners such as AQ and DD combinations, fixed prepositional phrases, multiword named entities, some limited continuous verbal expressions such as *avoir lieu* (lit. *have place*) 'take place', and so on. Our training and test corpora do not contain any labels distinguishing these MWE categories. Therefore the only category-based analysis we perform relies on the POS tags of the component words (for nominal MWEs).

In our experiments, we used two annotated corpora: the French treebank and the MORPH dataset. Other corpora annotated with MWEs in French do exist (Laporte et al. 2008; Savary et al. 2017). However, we chose to evaluate our model on a dataset for which, at the time of writing this chapter, many studies on MWE identification methods have been reported (the French treebank) and on an in-house dataset focusing on ambiguous MWEs (MORPH). Hence, we can compare our sequence model with state-of-the art results and verify whether they are adequate to recognise ambiguous MWEs. Evaluation on other corpora is left for future work.

4.1 The French treebank

We train and test our models on the MWE-annotated French treebank (FTB), available in CONLL format and automatically transformed into the CRFsuite format. The FTB is traditionally split into three parts: train, dev and test. We train our systems systematically on the training part of the FTB, that we adapted to keep only the MWEs we are interested in. For the experiments where we considered general MWEs and nominal MWEs, we used the FTB version of the SPMRL shared task (Seddah et al. 2013). The FTB dev and test corpora were employed respectively for feature engineering and evaluation. For each word, the corpus contains its wordform, lemma, POS (15 different coarse POS tags), and syntactic dependencies (that were ignored).

In the original corpus, MWE information is represented as words with spaces. For instance, *bien_que* appears as a single token with underscores when it is a complex conjunction, whereas accidental co-occurrence is represented as two separate tokens *bien* and *que*. We argue that using such gold tokenisation is unrealistic, especially in the case of ambiguous MWEs. We thus systematically split single-token MWEs and added an extra column containing MWE tags using BIO encoding (§3). Even though this preprocessing might sound artificial, we believe that it provides a more uniform treatment to ambiguous constructions, closer to their raw-text form. This assumption is in line with the latest developments in the dependency parsing community, which has evolved from parser evaluation on gold tokenisation (Buchholz & Marsi 2006) to evaluation on raw text (Zeman et al. 2017).

The MWE-BIO tags were generated using the following transformation heuristics in the case of ambiguous AQ and DD MWEs:

- For AQ expressions:
    1. We scan the corpus looking for the lemmas *ainsi_que*, *alors_que*, *autant_que*, *bien_que*, *encore_que*, *maintenant_que* and *tant_que*.
    2. We split them into two tokens and tag the adverb as B and *que* as I.
- For DD expressions:
    1. We scan the corpus looking for the wordforms *des*, *du*, *de_la* and *de_l'*. Due to French morphology, *de* is sometimes contracted with the articles *les* (determinate plural) and *le* (determinate singular masculine). Contractions are mandatory for both partitive and preposition+determiner uses. Therefore, we systematically separate these pairs into two tokens.[10]
    2. If a sequence was tagged as a determiner (D), we split the tokens and tag *de* as B and the determiner as I.
    3. Contractions (*des*, *du*) tagged as P+D (preposition+determiner) were split in two tokens, both tagged as O.
- All other tokens are tagged as O, including all other categories of MWEs.

For the newly created tokens, we assign individual lemmas and POS tags. The word *de* is systematically tagged as P (preposition), not distinguishing partitives from prepositions at the POS level. The input to the CRF is a file containing one word per line, BIO tags as targets, and FeatureName=value pairs including *n*-grams of wordforms, lemmas and POS tags, as described in §3.1.

In the case of nominal MWEs, we applied the same procedure as for AQ pairs to the MWEs matching certain sequences of POS tags[11]. We accept that tokens can be separated by punctuation marks, as in the proper noun *Bouches-du-Rhône*.

[10] An alternative to this preprocessing would be to keep contractions untokenised, and to assign a single B tag to those representing determiners. However, this would actually move the task of MWE identification to the POS tagger, which would need to choose whether the token is a determiner or a contracted preposition before MWE identification.

[11] The exact regular-expression pattern is: `(A.N) | (N.(PONCT.)?(A |(P+D.(PONCT.)?N) | (P.(PONCT.)?(D.)?(PONCT.)?N) | N)+)`.

When the MWE starts with a noun (N), it can be followed by one or more adjectives (A), nouns (N), or nouns preceded by prepositions (P) optionally including determiners (D) between the preposition and the noun. The matched nominal MWEs include combinations composed of:

- adjective noun: *premier ministre* 'prime minister';
- noun adjective: *corps médical* 'medical community';
- noun-noun: *maison mère* 'parent company';
- noun preposition noun: *motion de censure* 'motion of censure';
- noun preposition determiner noun: *impôt sur le revenu* 'income tax';
- noun preposition+determiner noun: *ironie du sort* 'twist of fate'.

4.2 MORPH dataset

We used the MORPH dataset introduced by Nasr et al. (2015) as test and development corpora for ambiguous AQ and DD expressions. It contains a set of 1,269 example sentences, each containing one of 7 ambiguous AQ constructions and 4 ambiguous DD constructions. To build this corpus, around 100 sentences for each of the 11 target constructions were extracted from the frWaC corpus and manually annotated as to whether they contain a multiword function word (MORPH) or accidental cooccurrence (OTHER). We have preprocessed the raw sentences as follows:

1. We have automatically POS tagged and lemmatized all sentences using an off-the-shelf POS tagger and lemmatizer independently trained on the FTB.[12] This information is given to the CRF as part of its features.
2. We have located the target construction in the sentence and added BIO tags according to the annotation provided: target pairs annotated as MORPH were tagged B + I, target pairs annotated as OTHER were tagged O.
3. For each target construction, we have taken the first 25 sentences as development corpus (dev, 275 sentences) and the remaining sentences for testing (test, 994 sentences).

[12] http://macaon.lif.univ-mrs.fr/

4. We created four targeted datasets: DEV$_{AQ}$, DEV$_{DD}$, FULL$_{AQ}$ and FULL$_{DD}$, where the different construction classes are separated, in order to perform feature selection.

Table 1 summarises the corpora covered by our experiments in terms of number of tokens and MWEs. We trained all systems on the training portion of the FTB with different tokenisation choices, depending on the target MWE.[13] The density of AQ and DD being too low in FTB-dev and FTB-test, we tune and evaluate our model for AQ and DD constructions on the MORPH dataset. For nominal and general MWEs, however, we use the FTB-dev and FTB-test portions.

Table 1: Number of tokens and MWEs in each corpus of our experiments.

Corpus	Portion	Target MWEs	#tokens	#MWEs
FTB	train	AQ	285,909	216
FTB	train	DD	285,909	1,356
FTB	train	Nominal	443,115	6,413
FTB	train	General	443,115	23,522
FTB	dev	Nominal	38,820	686
FTB	dev	General	38,820	2,117
FTB	test	Nominal	75,217	1,019
FTB	test	General	75,217	4,041
MORPH	FULL$_{AQ}$	AQ	11,839	730
MORPH	FULL$_{DD}$	DD	8,319	539

4.3 External lexicons

The verbal valence dictionary DicoValence specifies the allowed types of complements per verb sense in French. For each verb, we extract two binary flags:

- QUEV: one of the senses of the verb has one object that can be introduced by *que*.[14]

[13] FTB-train for AQ/DD and for nominal/general MWEs is the same corpus, but the number of tokens differs because all MWEs other than AQ and DD were represented using words-with-spaces in FTB-train for AQ/DD.

[14] In DicoValence, an object P1, P2 or P3 licenses a complementizer QPIND

- deV: one of the senses of the verb has a locative, temporal or prepositional paradigm that can be introduced by *de*.[15]

We also use a lexicon containing nominal MWEs that were automatically extracted from the frWaC. They were obtained with the help of the mwetoolkit by first extracting all lemma sequences that match the nominal MWE pattern described above. Then, for each sequence, we calculate its number of occurrences as well as the number of occurrences of its member words, which are then used to calculate the association measures listed in §3.1.

When integrating this lexicon in the CRF as features, special treatment was required for overlapping expressions. If a given token belonged to more than one overlapping MWE, we considered the maximum value of the association measures. Moreover, since CRFs cannot deal with real-valued features, we have quantized each association score through a uniform distribution that assigned an equal number of expressions to each bin.

4.4 Evaluation measures

For general and nominal MWEs, we analyse the performance on the FTB reported by the evaluation script of PARSEME shared task (see Savary et al. 2018 [this volume]).[16] The script provides us with two different scores: one based on MWEs, and one based on MWE tokens. The MWE-based measure requires that all tokens in the MWE be predicted by the system, whereas the token-based measure is calculated based on each token individually, so that partially correct predictions are taken into account. Each variant (MWE-based and token-based) is reported in terms of precision, recall and F-measure. In this work, we will particularly focus on the F-measure.

For AQ and DD combinations, we evaluated on the MORPH dataset. We report precision (P), recall (R) and F-measure (F_1) of MWE tags. In other words, instead of calculating micro-averaged scores over all BIO tags, we only look at the proportion of correctly guessed `B` tags. Since all of our target expressions are composed of exactly 2 contiguous words, we can use this simplified score because all `B` tags are necessarily followed by exactly one `I` tag. As a consequence, the measured precision, recall and F-measure scores on `B` and `I` tags are identical.

[15] In DicoValence, the paradigm is PDL, PT or PP.

[16] http://multiword.sf.net/sharedtask2017

5 Results

We present our results for different categories of MWEs, performing feature selection on the *dev* datasets. §5.1 presents an evaluation of our approach on ambiguous AQ and DD expressions. §5.2 evaluates the broader category of nominal MWEs. §5.3 then extends the latter results to an evaluation of all MWEs. §5.4 compares the best results we obtained against the state of the art. Finally, §5.5 presents the results of a detailed analysis focusing on variable and unseen MWEs.

5.1 Experiments on AQ and DD expressions

Our first evaluation was performed on the *dev* part of the MORPH dataset. We consider a subset of around 1/4 sentences containing AQ constructions (DEV$_{AQ}$, 175 sentences) and DD constructions (DEV$_{DD}$, 100 sentences). We evaluate the results under different levels of feature selection, regarding both coarse groups and individual features.

In these experiments, the CRF is trained to predict BIO labels on the training part of the FTB, where only the target AQ and DD constructions have been annotated as MWEs, as described in §4.1. Feature selection is performed on development set of the MORPH dataset, in which each sentence contains exactly one occurrence to disambiguate (MWE or accidental co-occurrence).

5.1.1 First feature selection: coarse

As shown in the first row of Table 2, when we include all features described in §3 (ALL), we obtain an F$_1$ score of 75.47 for AQ and 69.70 for DD constructions. The following rows of the table show the results of a first ablation study, conducted to identify coarse groups of features that are not discriminant and may hurt performance.

When we ignore orthographic features (ALL - ORTH), all scores increase for DEV$_{AQ}$ and DEV$_{DD}$, suggesting that MWE occurrences are not correlated with orthographic characteristics. F$_1$ also increases when we remove all surface-level wordform features, including single words and *n*-grams (represented by W). We hypothesize that lemmas and POS are more adequate, as they can reduce sparsity by conflating variations of the same lexeme, while wordforms only seem to introduce noise.

We then evaluate the removal of lexicon features (ALL - LF). At a first intuition, one would say that this information is important to our system, as it allows assigning O tags to conjunctions and prepositions that introduce verbal

Table 2: Ablation study results on the dev portion of the MORPH dataset focusing on AQ and DD expressions - impact of the removal of coarse-grained feature sets.

Feature set	DEV$_{AQ}$			DEV$_{DD}$		
	P	**R**	**F$_1$**	**P**	**R**	**F$_1$**
ALL	89.55	65.22	75.47	92.00	56.10	69.70
ALL - ORTH	90.28	70.65	79.27	95.83	56.10	70.77
ALL - W	90.79	75.00	82.14	87.10	65.85	75.00
ALL - LF	91.18	67.39	77.50	88.89	58.54	70.59
ALL - $t_{\pm 2}$	87.67	69.57	77.58	88.00	53.66	66.67
ALL - $T_{i-1}T_iT_{i+1}$	87.84	70.65	78.31	91.67	53.66	67.69
ALL - $T_{i-1}T_i$	**93.55**	63.04	75.32	95.83	56.10	70.77
ALL - $T_{i-1}T_i$ - $T_{i-1}T_iT_{i+1}$	88.57	67.39	76.54	**96.00**	58.54	72.73
ALL - ORTH - W	90.24	**80.43**	**85.06**	87.10	65.85	75.00
ALL - ORTH - W - $t_{\pm 2}$ (REF$_1$)	89.74	76.09	82.35	85.29	**70.73**	**77.33**

complements. Surprisingly, though, the system performs better without them. We presume that this is a consequence of the sparsity of these features: since there are many features overall in the system, the CRF will naturally forgo LF features when they are present, rendering them superfluous to the system. These features will be analyzed individually later (see Table 4).

One might expect that single tokens located 2 words apart from the target token do not provide much useful information, so we evaluate the removal of the corresponding features (ALL - $t_{\pm 2}$). While this intuition may be true for DEV$_{AQ}$, it does not hold for DEV$_{DD}$. Next, we try to remove all trigrams, and then all trigam & bigram features at once. When we remove trigrams, F$_1$ decreases by 2.01 absolute points in DEV$_{DD}$ and increases by 2.84 absolute points in DEV$_{AQ}$. Bigrams are somehow included in trigrams, and their removal has little impact on the tagger's performance. When we remove bigram and trigram features altogether, scores are slightly better, even though a large amount of information is ignored. Since these results are inconclusive, we perform a more fine-grained selection considering specific *n*-grams in §5.1.2.

Finally, we try to remove several groups of features at the same time. When we remove both orthographic and wordform features, F$_1$ increases to 85.06 for DEV$_{AQ}$ and 75.00 for DEV$_{DD}$. When we also remove tokens located far away from the current one, performance increases for DEV$_{DD}$, but not for DEV$_{AQ}$. Unreported experiments have shown, however, that further feature selection also yields better results for DEV$_{AQ}$ when we ignore $t_{\pm 2}$ features. Therefore, our reference (REF$_1$) for the fine-grained feature selections experiments will be this set of features, corresponding to the last row of Table 2.

5.1.2 Second feature selection: fine

Table 3 presents the results from fine-grained feature selection. In the first row of the table, we replicate the reference (REF$_1$) feature set defined above. In the second row, we try to remove the lexicon features (LF) once again. When they were removed in previous experiments, shown in Table 2, we had a gain in performance, suggesting that these features were superfluous. When we remove them from REF$_1$, however, the precision and recall observed for DEV$_{DD}$ decrease by about 10 points. That is, the removal of LF yields a performance drop with respect to a relatively good model (REF$_1$), suggesting that these features are valuable after all. We hypothesise that LF can be better taken into account now that there are less noisy features overall in the whole system.

Table 3: Ablation study results on the dev portion of the MORPH dataset focusing on AQ and DD expressions - impact of the removal of fine-grained feature sets.

Feature set	DEV$_{AQ}$			DEV$_{DD}$		
	P	R	F$_1$	P	R	F$_1$
REF$_1$	89.74	76.09	82.35	**85.29**	70.73	77.33
REF$_1$ - LF	90.00	78.26	83.72	75.76	60.98	67.57
REF$_1$ - $T_{-1}T_0$	**90.54**	72.83	80.72	**85.29**	70.73	77.33
REF$_1$ - T_0T_{+1}	89.87	77.17	83.04	84.85	68.29	75.68
REF$_1$ - $T_0T_{+1}T_{+2}$ (BEST$_1$)	87.36	**82.61**	**84.92**	83.78	**75.61**	**79.49**

The last three rows of the table presents the results from attempts at removing individual *n*-gram features that we expected to be redundant or not highly informative. First, we consider the removal of two types of bigram features independently (towards the left and towards the right of the target word). We remove their wordforms, POS and lemmas. The results suggest that bigrams can be mildly useful, as their removal causes the most scores to drop.

In the last row of the table, we present the results from removing all trigram features of the form $T_0T_{+1}T_{+2}$. As a result, we can see that performance increases for both datasets. While trigram features could be potentially useful to recognise longer expressions, we assume that the number of all possible trigrams is actually too large, making the feature values too sparse. In other words, a much larger annotated corpus would be required for trigram features to be effective. This is the best configuration obtained on the development datasets, and we will refer to it as BEST_1 in the next experiments.

Our last feature selection experiments consider the influence of lexicon features (LF) individually, as shown in Table 4. We observe that DEV is an important feature, because when we remove it, F_1 decreases by almost 7 absolute points on the DEV_{DD} set. The feature QUEV, however, seems less important, and its absence only slightly decreases the F_1 score on the DEV_{AQ} set. This is in line with what was observed by Nasr et al. (2015) for the whole dataset. In sum, these features seem to help, but we would expect the system to benefit more from them with a more sophisticated representation.

Table 4: Ablation study results on the dev portion of the MORPH dataset focusing on AQ and DD expressions - impact of the removal of lexicon features (LF).

Dataset	Feature set	P	R	F_1
DEV_{AQ}	BEST_1	87.36	82.61	84.92
	BEST_1-QUEV	91.25	79.35	84.88
DEV_{DD}	BEST_1	83.78	75.61	79.49
	BEST_1-DEV	77.78	68.29	72.73

The results obtained in this section focus on a limited number of very frequent expressions. Since our evaluation focuses on a small sample of 11 such MWEs only, it would be tempting to train one CRF model per target expression. However, there are a few more expressions with the same characteristics in French, and many of them share similar syntactic behaviour (e.g. conjunctions formed by an adverb and a relative conjunction). An approach with a dedicated model per expression would miss such regular syntactic behaviour (e.g. the fact that the surrounding POS are similar).

The experiments reported up to here show how it is possible to identify highly ambiguous (and frequent) expressions with a CRF, but they are hard to generalise to other MWE categories. Therefore, in the next sections, we evaluate our model on broader MWE categories such as nominal MWEs and general continuous MWEs (as defined in the FTB).

5.2 Experiments on nominal MWEs

We now focus on the identification of nominal MWEs in the FTB. As above, we separate our experiments in coarse-grained and fine-grained feature selection. In these experiments, the CRF was trained on the training part of the FTB where only nominal MWEs were tagged as B-I and all other words and MWEs were tagged as O. The feature selection experiments are performed on the development set of the FTB, also transformed in the same way. For the comparison with the state of the art, we report results for the test portion of FTB.

5.2.1 First feature selection: coarse

Table 5 presents the results obtained on FTB for different levels of feature selection. In the first row (ALL), we present the evaluation of all the features described in §3, except DEV and QUEV (only relevant to the previous experiments). We obtain a baseline with MWE-based F_1 score of 71.57%, and token-based score F_1 score of 73.85%.

Table 5: Ablation study results on FTB-dev focusing on nominal MWEs - impact of the removal of coarse-grained feature sets.

Feature set	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
ALL	80.86	64.19	71.57	81.23	67.70	73.85
ALL - ORTH	81.85	64.78	72.32	82.16	68.02	74.43
ALL - W	80.41	64.78	71.75	80.95	68.44	74.17
ALL - AM	81.37	61.72	70.19	81.48	65.16	72.41
ALL - $t_{\pm 2}$	81.49	**65.84**	72.83	81.80	**69.50**	75.15
ALL - T_{+2}	80.96	65.51	72.48	81.18	69.08	74.64
ALL - $T_{i-1}T_i$	80.41	64.31	71.47	80.99	67.84	73.83
ALL - $T_{i-1}T_iT_{i+1}$ (REF$_2$)	81.61	**65.84**	**72.88**	82.05	69.40	**75.20**
ALL - $T_{i-1}T_iT_{i+1}$ - AM	81.69	63.60	71.52	82.09	67.28	73.95
ALL - ORTH - W - $t_{\pm 2}$ - $T_{i-1}T_iT_{i+1}$	79.59	63.37	70.56	81.00	67.88	73.86
ALL - ORTH - $T_{i-1}T_iT_{i+1}$	**82.51**	65.05	72.73	**82.74**	67.93	74.61

We consider the removal of the same groups of features that we removed on the AQ and DD experiments. We evaluate the independent removal of orthographic features, wordforms, association measures, $t_{\pm 2}$, t_{i+2}, bigrams and trigrams. We notice that all of these columns have better results than ALL, except for the column where we removed the bigrams and the one in which we removed association measures. In particular, we notice that the absence of the AMs significantly hurts recall, which in turn has an impact on the F_1 score (−1.38% for the MWE-based measure and −1.41% for the token-based measure). This is the first clue that indicates the importance of these features.

We then evaluate the removal of different groups of features at the same time. We begin by deleting all of the previous groups, except for AMs and bigrams, which seemed to provide useful information above. Nevertheless, we did not obtain better results. We then tried to remove only the trigrams and the orthographic features. Results were slightly higher than ALL, but still remain worse than the results with only the trigrams removed. Finally, we decided to verify if the AM features are still relevant to obtain this performance. This was confirmed, as without the AM, the MWE-based F_1 score decreased by 1.36%, and the token-based F_1 score decreased by 1.25%. Overall, the highest results were obtained by removing only trigrams from ALL, and so we take this feature set as our new reference (REF_2).

5.2.2 Second feature selection: fine

Experiments above have shown that association measures (AM) are a vital component of our system. We proceed now to evaluate the importance of individual association measures towards the identification of nominal MWEs. The results are shown in Table 6. We consider the impact of the different AMs in two baseline configurations: all features (ALL), and the features of the reference only (REF_2). We then remove individual measures and evaluate the new feature set on FTB-dev.

We consider the removal of multiple combinations of features. In most cases, we notice a slight improvement in the results against ALL, but not when compared to the reference group . The removal of the DICE measure did improve the results in both cases, ALL and REF_2. Therefore, this configuration was chosen as the $BEST_2$ set of features. We then evaluated these $BEST_2$ features on the FTB-test dataset, obtaining a MWE-based F_1 score of 71.38%, and a token-based score of 73.43%. As a sanity check, we have also evaluated the system without AMs on FTB-test (ALL - AM). The $BEST_2$ system is significantly different from both ALL and ALL - AM on the test set. Moreover, the large margin between ALL - AM and the two other systems indicates that association measures do provide useful features for this task.

Table 6: Ablation study results on FTB-dev focusing on nominal MWEs - impact of the removal of fine-grained feature sets.

Feature set	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
ALL	80.86	64.19	71.57	81.23	67.70	73.85
ALL - DICE	81.07	64.55	71.87	81.39	68.02	74.11
ALL - T-MEAS	81.07	64.55	71.87	81.40	68.07	74.14
ALL - PMI	81.26	63.84	71.50	81.51	67.33	73.74
ALL - MLE - LL	81.13	64.31	71.75	81.40	67.65	73.89
ALL - T-MEAS - DICE	80.76	64.78	71.90	81.23	68.30	74.20
ALL - MLE - LL - T-MEAS - DICE	81.72	63.72	71.61	81.58	67.05	73.61
REF_2	81.61	65.84	72.88	82.05	69.40	75.20
REF_2 - DICE ($BEST_2$)	**81.84**	65.84	72.98	82.33	**69.45**	**75.34**
REF_2 - T-MEAS	81.61	65.84	72.88	82.01	69.22	75.08
REF_2 - PMI	81.80	65.14	72.52	**82.36**	68.71	74.92
REF_2 - MLE - LL	81.67	65.61	72.76	82.03	69.08	75.00
REF_2 - T-MEAS - DICE	81.75	**65.96**	**73.01**	82.18	69.36	75.23
REF_2 - MLE - LL - T-MEAS - DICE	81.41	65.49	72.58	81.51	68.94	74.70
ALL (on FTB-test)	77.06	65.66	70.90	79.10	68.23	73.27
ALL - AM (on FTB-test)	76.96	61.81	68.56	78.94	64.91	71.24
$BEST_2$ (on FTB-test)	76.00	67.28	71.38	77.74	69.58	73.43

5.3 Experiments on general MWEs

We extend the experiments above to evaluate the feature sets against the whole FTB corpus, keeping all annotated MWEs in the training, development and test parts of the FTB. We would like to verify if our system is able to take into account the different MWE categories at the same time. This time, we only present coarse-grained feature selection (Table 7), since unreported fine-grained feature selection resulted in similar findings as in experiments focusing on nominal MWEs.

The first row in the table (ALL) presents the evaluation of all features described in §3. The prediction of general MWEs with ALL features yields a MWE-based F_1 score of 78.89% and a token-based F_1 score of 81.61%. We then consider what happens when one removes the same groups of features as in the previous sections. This time the results are quite different: all of these tests have worse results than ALL, except when we remove t_{+2} features. In some unreported experiments,

Table 7: Ablation study results on FTB-dev focusing on general MWEs - impact of the removal of feature sets.

Feature set	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
ALL	**85.60**	73.16	78.89	**87.32**	76.60	81.61
ALL - ORTH	85.09	72.97	78.57	86.97	76.56	81.44
ALL - W	83.96	72.59	77.86	86.13	76.37	80.96
ALL - AM	85.11	72.78	78.46	86.89	76.33	81.27
ALL - $t_{\pm 2}$	84.03	72.45	77.81	86.57	76.94	81.47
ALL - t_{+2}	85.50	73.68	79.15	87.19	77.21	81.90
ALL - $T_{i-1}T_i$	84.36	71.75	77.54	86.61	75.47	80.66
ALL - $T_{i-1}T_iT_{i+1}$	84.78	73.07	78.49	86.39	76.31	81.04
ALL - T_{+2} - ORTH (REF_3)	85.52	**73.82**	**79.24**	87.30	**77.35**	**82.03**
REF_3 - AM	85.37	72.69	78.52	87.08	76.33	81.35
REF_3 - T-MEAS ($BEST_3$)	85.62	**73.87**	**79.31**	87.40	**77.43**	**82.11**
ALL (on FTB-test)	83.80	74.51	78.88	86.58	78.23	82.19
ALL - AM (on FTB-test)	84.19	73.52	78.49	86.90	77.30	81.82
$BEST_3$ (on FTB-test)	84.03	74.71	79.10	86.72	78.47	82.39

we have tried to remove other groups of features along with t_{+2}. We found that removing orthographic features along with t_{+2} increased the results more than only removing t_{+2} features. This group of features will be our new reference from now on (REF_3). Once again, we tried to remove AMs from the reference to verify their impact. Here again, we notice that the removal of these features decreases the overall performance scores, even if their impact is weaker than it was in the case of nominal MWEs. Unreported experiments have shown that we obtain better results when we ignore the T-MEAS feature ($BEST_3$).

Then, we applied the feature group $BEST_3$ on the FTB-test dataset, and we obtained a MWE-based F_1 score of 79.10%, and a token-based score of 82.39%. For the feature selection experiments on the test part of the FTB (both nominal and general MWEs), we calculated the p-value of the difference between the configuration called BEST and the one called ALL, using approximate randomisation with stratified shuffling. None of the observed differences was considered statistically significant with $\alpha = 0.05$.

5.4 Comparison with state of the art

We now compare the highest-scoring reference results with the state of the art. We begin by evaluating the identification of *DD* and *AQ* constructions, and then proceed to evaluate more generally the quality of our reference system for general MWE identification. The comparisons presented here focus on MWE identification only, and our model takes gold POS and lemma information as input (except on the MORPH dataset). On the other hand, some of the works mentioned in our comparisons also predict POS and/or syntactic structure, which makes the task considerably harder. Therefore, results presented here should be taken as an indication of our position within the current landscape of MWE identification, rather than as a demonstration of our model's superiority.

5.4.1 AQ and DD constructions

We report the performance of MWE identification on the full MORPH dataset, split in two parts: sentences containing AQ constructions (FULL_{AQ}) and sentences containing DD constructions (FULL_{DD}). The use of the full datasets is not ideal, given that we performed feature selection on part of these sentences, but it allows a direct comparison with related work.

Table 8 presents a comparison between the best system score obtained after feature selection (BEST_1) and the results reported by Nasr et al. (2015). We include two versions of the latter system, since they also distinguish their results based on the presence of lexicon features (LF) coming from DicoValence.

Table 8: Comparison with baseline and state of the art of AQ and DD identification on the full MORPH dataset.

System	FULL_{AQ}			FULL_{DD}		
	P	**R**	**F_1**	**P**	**R**	**F_1**
Baseline	56.08	100.00	71.86	34.55	100.00	51.35
Nasr et al. (2015)-LF	88.71	82.03	85.24	77.00	73.09	75.00
Nasr et al. (2015)+LF	91.57	81.79	86.41	86.70	82.74	84.67
BEST_1	91.08	78.31	84.21	79.14	74.37	76.68

We additionally report results for a simple baseline:

1. We extract a list of all pairs of contiguous AQ and DD from the FTB-train.
2. We calculate the proportion of cases in which they were annotated as MWEs (B-I tags) with respect to all of their occurrences.
3. We keep in the list only those constructions which were annotated as MWE at least 50% of the time.
4. We systematically annotate these constructions as MWEs (B-I) in all sentences of the MORPH dataset, regardless of their context.

Table 8 shows that this baseline reaches 100% recall, covering all target constructions, but precision is very low, as contextual information is not taken into account during identification. Our BEST$_1$ system can identify the target ambiguous MWEs much better than the baseline for both FULL$_{AQ}$ and FULL$_{DD}$.

For some constructions, we do obtain results that are close to those obtained by the parsers (see Table 9 for more details). For FULL$_{AQ}$, our BEST$_1$ system obtains an F$_1$ score that is 1.2 absolute points lower than the best parser. For FULL$_{DD}$, however, our best system, which includes lexicon features (LF), is comparable with a parser without lexicon features. When the parser has access to the lexicon, it beats our system by a significant margin of 7.99 points, indicating that the accurate disambiguation of DD constructions indeed requires syntax-based methods rather than sequence taggers. These results contradict our hypothesis that sequence models can deal with continuous constructions with a performance equivalent to parsing-based approaches. While this may be true for non-ambiguous expressions, parsing-based methods are superior for AQ and DD constructions, given that they were trained on a full treebank, have access to more sophisticated models of a sentence's syntax, and handle long-distance relations and grammatical information.

Despite the different results obtained depending on the nature of the target constructions, the results are encouraging, as they prove the feasibility of applying sequence taggers for the identification of highly ambiguous MWEs. Our method has mainly two advantages over parsing-based MWE identification: (a) it is fast and only requires a couple of minutes on a desktop computer to be trained; and (b) it does not require the existence of a treebank annotated with MWEs.

Table 9 shows the detailed scores for each expression in the MORPH dataset. We notice that some expressions seem to be particularly difficult to identify, especially if we look at precision, whereas for others we obtain scores well above 90%. When we compare our results to those reported by Nasr et al. (2015), we can see that they are similar to ours: *ainsi* 'likewise', *alors* 'then' and *bien* 'well' have F_1 higher than 90%, while *autant* ‘as much’ and *tant* ‘while’ have a score lower than 80%. The AQ constructions with *encore* ‘still’ and *maintenant* ‘now’ are the only ones which behave differently: our system is better for *encore* ‘still’, but worse for *maintenant* ‘now’. Likewise, for DD expressions, our system obtains a performance that is close to their system without lexicon features (LF), but considerably worse than their system including LFs for three out of 4 expressions. Both approaches are more efficient in identifying the plural article *de les* ‘of the.PL’ than the partitive constructions.

Table 9: Performance of the $BEST_1$ configuration broken down by expression, along with the results for the best model of Nasr et al. (2015) (with LF).

	$BEST_1$ system			Nasr et al. (2015)		
Expression	**P**	**R**	$\mathbf{F_1}$	**P**	**R**	$\mathbf{F_1}$
ainsi que	94.44	93.15	93.79	95.94	89.87	92.81
alors que	84.00	97.67	90.32	93.81	93.81	93.81
autant que	93.48	51.81	66.67	86.66	70.65	77.84
bien que	100.00	91.43	95.52	91.66	99.18	90.41
encore que	76.19	94.12	84.21	92.85	65.00	76.47
maintenant que	97.62	64.06	77.36	90.91	74.62	81.96
tant que	100.00	60.00	75.00	82.35	70.00	75.67
de le	78.05	71.11	74.42	85.41	91.11	88.17
de la	67.74	72.41	70.00	81.25	89.65	85.24
de les	92.41	71.57	80.66	98.70	76.00	85.87
de l’	61.11	95.65	74.58	64.51	86.95	74.07

5.4.2 General MWEs

We now compare our system with two baselines and with the system proposed in Le Roux et al. (2014).[17] $Baseline_1$ consists in identifying as MWE every continuous occurrence of tokens that has been seen as an MWE in the training corpus. For example, the MWE *bien sûr* (lit. *well sure*) ‘of course’ can be seen in the training corpus, and so every occurrence of this expression was predicted as an MWE

[17] The comparison with Le Roux et al. (2014) is not ideal, since we predict MWEs with the help of gold POS and lemmas, whereas they try to predict both POS and MWEs. However, we could not find a fully comparable evaluation in the literature.

for the test corpus. Baseline$_2$ filters the list of MWEs seen in the training corpus, so that only the expressions which had been annotated more than 40% of the time are predicted as MWEs. For example, the expression *d'un côté* (lit. *of a side*) 'on the one hand' is not predicted as MWE, as it was only annotated in 38% of its occurrences in the training corpus. The baselines were directly inspired by a predictive model applied to the English language in a similar task, where a threshold of 40% was found to yield the best results (Cordeiro et al. 2016). The applied threshold in Baseline$_2$ only eliminates 6.46% of the MWEs from the list, but it contributes to an increase of 20–30 points in precision without impacting the recall.

Our system (BEST$_3$ configuration) is more accurate than the baselines, both in terms of precision and recall. It also has a higher precision than the approach proposed by Le Roux et al. (2014), but the recall is considerably worse (9.48% less than their system). This means that our system misses more expressions, even if its predictions have higher precision. This could be partly explained by the fact that they employed dictionaries, and have access to more expressions that our system has never seen and could not predict. Nonetheless, our results are sufficiently close and represent a decent alternative if high-quality external resources are not available.

Table 10: Comparison with baseline and state of the art of general MWE identification on FTB-test.

System	MWE-based			Token-based		
	P	R	F_1	P	R	F_1
Baseline$_1$	52.93	66.20	58.82	62.70	69.73	66.03
Baseline$_2$	82.76	69.36	75.47	84.80	69.62	76.46
BEST$_3$ configuration	84.03	74.71	79.10	86.72	78.47	82.39
Le Roux et al. (2014)	80.76	84.19	82.44	–	–	–

5.5 Analysis of results

The performance of our CRF identification model depends on the characteristics of the identified MWEs and of the training and test corpora. Therefore, we have performed a detailed analysis of its performance focusing on a subset of the test corpus. We focus on two phenomena: variants and unseen MWEs. We define a **variant** as an MWE whose lemmatised form occurs both in the training and in the test corpus, but whose surface form in the test corpus is different from all of its surface forms in the training corpus. We define an **unseen** MWE as one whose lemmatised form occurs in the test corpus but never (under any surface form) in the training corpus. MWEs which have identical occurrences in the training and test corpora will be referred to as **seen** MWEs.

Both variants and unseen MWEs are harder to identify than seen MWEs. Nonetheless, we hypothesise that our model is able to recognise variants correctly, since its features are based on lemmas. On the other hand, we expect that unseen MWEs cannot be easily predicted given that our system is absed on categorical features and does not have access to much information about an expression that has never been seen in the training corpus, except for its association measures in a large unannotated corpus. To verify these hypotheses, we create sub-corpora of FTB-test, where the density of variants and unseen MWEs is higher than in the full FTB-test corpus. In these experiments, the model is not newly trained, but the BEST$_2$ and BEST$_3$ models are applied to different sub-corpora with a high density of variant/unseen MWEs.

The evaluation measures reported in our experiments (§4.4) consider the best bijection between predicted and gold MWEs. Therefore, we cannot simply remove seen MWEs from the test set, since they will be predicted anyway, artificially penalising precision. Therefore, instead of completely removing seen MWEs, we remove sentences that contain only seen MWEs and keep sentences that contain (a) at least one variant MWE or (b) at least one unseen MWE.

Table 11: Results of BEST$_2$ (nominal MWEs) and BEST$_3$ (general MWEs) on FTB-test, on sub-corpus containing unseen variants of a seen MWEs, and on sub-corpus containing unseen MWEs. Columns %var and %unseen show the proportion of variants/unseen MWEs in each sub-corpus.

Feature set	%var	%unseen	**MWE-based**			**Token-based**		
			P	**R**	**F$_1$**	**P**	**R**	**F$_1$**
Nominal full	5%	28%	76.00	67.28	71.38	77.74	69.58	73.43
Nom. variants	65%	N/A	86.42	63.64	73.30	85.84	66.20	74.75
Nom. unseen	N/A	72%	82.01	42.70	56.16	87.78	46.75	61.01
General full	5%	23%	84.03	74.71	79.10	86.72	78.47	82.39
Gen. variants	32%	N/A	88.91	69.44	77.98	92.77	74.05	82.36
Gen. unseen	N/A	51%	86.94	65.22	74.25	90.40	69.14	78.35

Table 11 presents the performance of the BEST$_2$ configuration for nominal MWEs (first row) and BEST$_3$ configuration for general MWEs (fourth row) on the full FTB-test corpus. For each expression (nominal and general), we also present the results for the sub-corpus containing a higher density of variants and of unseen MWEs. The numbers in columns %var and %unseen indicate the proportion of variant/unseen MWEs in each sub-corpus. Notice that, in the case of general MWEs, these proportions are quite low (32% and 51%), indicating that sentences containing variant and unseen general MWEs often contain seen ones too. When focusing on variants (Nom. variants and Gen. variants sub-corpora), the proportion of unseen MWEs is very small and not relevant (N/A), and vice-versa.

If we focus on variants, we can observe relatively stable results with respect to the full FTB-test corpus. For nominal MWEs, precision increases by 8-10%, whereas recall decreases by about 3% for both MWE-based and token-based measures. Results for general MWEs follow a similar pattern: around 4-6% improvement in precision at the cost of around 4-5% decrease in recall. The precision of general MWE identification in the variants sub-corpus is particularly impressive, reaching 92.77%.

The variants sub-corpora contain less unseen MWEs than the full FTB-test corpus, so the predicted MWEs are more reliable (better precision), showing that our model is robust to morphological variability. On the other hand, the fact that recall drops indicates that it is indeed slightly harder to recognise variants of MWEs than those seen identically in training and test corpora. In short, we infer that variants can be correctly handled and identified by our model, provided that a good lemmatiser is available (results presented here are based on gold lemmas, their substitution by predicted lemmas should be studied in the future).

On the other hand, predicting unseen MWEs is considerably harder. Recall drops drastically by about 23-25% for nominal MWEs and by about 9% for general MWEs, and the improvements in precision do not compensate for this, yielding much lower F-measure values, specially for nominal MWEs where the concentration of unseen MWEs in the sub-corpus is higher (72%). The improvements in precision are probably due to the fact that some seen and variant MWEs are still present in the sub-corpora. AMs could also have some predictive power to identify unseen MWEs, and we intend to verify their contribution for unseen MWE identification in the future. These results show that our model is limited in the identification of unseen MWEs, and can probably only identify some of those that appear in the AM lexicons.

6 Conclusions and future work

We have described and evaluated a simple and fast CRF tagger that is able to identify several categories of continuous multiword expressions in French. We have reported feature selection studies and shown that, for AQ constructions and for general MWEs, our results are almost as good as those obtained by parsers, even though we do not rely on syntactic trees. This was not true for DD constructions, though, which seem to require parsing-based methods to be properly analysed. Based on these results, we conclude that, when treebanks are not available, sequence models such as CRFs can obtain reasonably good results in the identification of continuous MWEs. On the other hand, when MWE-annotated treebanks exist, parsing-based models seem to obtain better results, especially for expressions whose high ambiguity requires syntax to be resolved.

An interesting direction of research would be to study the interplay between automatic POS tagging and MWE identification. We recall that our results were obtained with an off-the-shelf POS tagger and lemmatizer. Potentially, performing both tasks jointly could help obtaining more precise results (Constant & Sigogne 2011). Moreover, we could explore CRFs' ability to work with lattices in order to pre-select the most plausible MWE identification (and POS tagging) solutions, and then feed them into a parser which would take the final decision.

Another idea for future work would be an investigation of the features themselves. For example, in this work, we were not fully satisfied with the quality of the representation of lexical features. We would like to investigate the reason why lexical features were not always useful for the task of MWE identification, which could be done by performing an error analysis on the current systems. Another interesting question is whether annotated corpora are at all necessary: could hand-crafted and/or automatically built lexicons be employed to identify MWEs in context in a fully unsupervised way?

While these experiments shed some light on the nature of MWEs in French, the feature selection methodology is highly empirical and cannot be easily adapted to other contexts. Therefore, we would like to experiment different techniques for generic automatic feature selection and classifier tuning (Ekbal & Saha 2012). This could be performed on a small development set, and would ease the adaptation of the tagger to other contexts.

Finally, we would like to experiment with other sequence tagging models such as recurrent neural networks. In theory, such models are very efficient to perform feature selection and can also deal with continuous word representations able to encode semantic information. Moreover, distributed word representations could be helpful in building cross-lingual MWE identification systems.

Acknowledgments

This work has been funded by projects PARSEME (Cost Action IC1207), PARSEME-FR (ANR-14-CERA-0001), and AIM-WEST (FAPERGS-INRIA 1706-2551/13-7). We also thank the anonymous reviewers for their valuable suggestions.

Abbreviations

AQ	adverb+*que*	DD	*de*+determiner
AM	association measure	FTB	French Treebank
BIO	begin-inside-outside	LF	lexicon feature
CRF	conditional random field	MWE	multiword expression

References

Abeillé, Anne, Lionel Clément & François Toussenel. 2003. Building a treebank for French. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 165–187. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Baroni, Marco & Silvia Bernardini (eds.). 2006. *Wacky! Working papers on the web as corpus*. Bologna, Italy: GEDIT.

Boukobza, Ram & Ari Rappoport. 2009. Multi-word expression identification using sentence surface features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 468–477. August 6-7, 2009.

Buchholz, Sabine & Erwin Marsi. 2006. CoNLL-x shared task on multilingual dependency parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning* (CoNLL-X '06), 149–164. http://dl.acm.org/citation.cfm?id=1596276.1596305.

Candito, Marie & Matthieu Constant. 2014. Strategies for contiguous multiword expression analysis and dependency parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, 743–753. Association for Computational Linguistics. http://www.aclweb.org/anthology/P14-1070.

Carpuat, Marine & Mona Diab. 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*, 242–245. Association for Computational Linguistics. http://www.aclweb.org/anthology/N10-1029.

Constant, Matthieu & Joakim Nivre. 2016. A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 161–171. Association for Computational Linguistics. http://www.aclweb.org/anthology/P16-1016.

Constant, Matthieu & Anthony Sigogne. 2011. MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proceedings of the ALC Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE 2011), 49–56. Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0809.

Cordeiro, Silvio, Carlos Ramisch & Aline Villavicencio. 2016. UFRGS & LIF at SemEval-2016 task 10: Rule-based MWE identification and predominant-supersense tagging. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 910–917. Association for Computational Linguistics. http://www.aclweb.org/anthology/S16-1140.

Diab, Mona & Pravin Bhutada. 2009. Verb noun construction MWE token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, 17–22. Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-2903.

Ekbal, Asif & Sriparna Saha. 2012. Multiobjective optimization for classifier ensemble and feature selection: An application to named entity recognition. *International Journal on Document Analysis and Recognition (IJDAR)* 15(2). 143–166. DOI:10.1007/s10032-011-0155-7

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103. http://aclweb.org/anthology/J09-1005.

Finlayson, Mark & Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the ACL Workshop on Multiword Expressions: From Parsing and Generation to the Real World* (MWE '11), 20–24. Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0805.

Green, Spence, Marie-Catherine de Marneffe & Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics* 39(1). 195–227. DOI:10.1162/COLI_a_00139

Lafferty, John D., Andrew McCallum & Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (ICML '01), 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=645530.655813.

Laporte, Éric, Takuya Nakamura & Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Proceedings of the 2nd Linguistic Annotation Workshop*, 48–51. https://halshs.archives-ouvertes.fr/halshs-00286541.

Le Roux, Joseph, Antoine Rozenknop & Matthieu Constant. 2014. Syntactic parsing and compound recognition via dual decomposition: Application to French. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers* (COLING 2014), 1875–1885. Dublin, Ireland: Dublin City University & Association for Computational Linguistics. http://www.aclweb.org/anthology/C14-1177.

Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1116–1126. Association for Computational Linguistics. http://www.aclweb.org/anthology/P15-1108.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Workshop on methodologies and evaluation of multiword units in real-world applications* (MEMURA 2004), 39–46. http://stp.lingfil.uu.se/~nivre/docs/mwu.pdf.

Okazaki, Naoaki. 2007. *CRFsuite: A fast implementation of conditional random fields (CRFs)*. http://www.chokkan.org/software/crfsuite/.

Ramisch, Carlos. 2014. *Multiword expressions acquisition: A generic and open framework* (Theory and Applications of Natural Language Processing XIV). Cham, Switzerland: Springer. 230. http://link.springer.com/book/10.1007%2F978-3-319-09207-2.

Ramshaw, Lance & Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *3rd Workshop on Very Large Corpora*, 82–94. http://aclweb.org/anthology/W95-0107.

Riedl, Martin & Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions* (MWE '16), 107–111. Association for Computational Linguistics. http://anthology.aclweb.org/W16-1816.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1471591

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* 2(1). 193–206. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/281.

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Nora Emily Danchik, Michael T. Mordowanec, Henrietta Conrad & Smith Noah A. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC 2014), 455–461. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/521_Paper.pdf.

Scholivet, Manon & Carlos Ramisch. 2017. Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 167–175. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1723.

Seddah, Djamé, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola Galletebeitia, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiórkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska & Eric Villemonte de la Clergerie. 2013. Overview of the SPMRL 2013 Shared Task: a cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically-Rich Languages*, 146–182. Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-4917.

Shigeto, Yutaro, Ai Azuma, Sorami Hisamoto, Shuhei Kondo, Tomoya Kouse, Keisuke Sakaguchi, Akifumi Yoshimoto, Frances Yung & Yuji Matsumoto. 2013. Construction of English MWE dictionary and its application to POS tagging. In *Proceedings of the 9th Workshop on Multiword Expressions* (MWE '13), 139–144. Association for Computational Linguistics. http://www.aclweb.org/anthology/W13-1021.

Silberztein, Max, Tamás Váradi & Marko Tadić. 2012. Open source multi-platform NooJ for NLP. In *Proceedings of the 24th International Conference on Computational Linguistics: Demonstration Papers* (COLING-12), 401–408. The Coling 2012 Organizing Committee. http://www.aclweb.org/anthology/C12-3050.

van den Eynde, Karel & Piet Mertens. 2003. La valence: L'approche pronominale et son application au lexique verbal. *Journal of French Language Studies* 13. 63–104.

Vincze, Veronika, István Nagy T. & Gábor Berend. 2011. Detecting noun compounds and light verb constructions: A contrastive study. In *Proceedings of the ALC Workshop on Multiword Expressions: From Parsing and Generation to the*

Real World (MWE '11), 116–121. Portland, OR, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W11-0817.

Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de-Paiva, Kira Droganova, Héctor Martínez Alonso, ÇağrÄ± Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj & Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, 1–19. Vancouver, Canada: Association for Computational Linguistics. http://www.aclweb.org/anthology/K17-3001.

Chapter 11

Identification of multiword expressions: A fresh look at modelling and evaluation

Shiva Taslimipoor
University of Wolverhampton

Omid Rohanian
University of Wolverhampton

Ruslan Mitkov
University of Wolverhampton

Afsaneh Fazly
Thomson Reuters

Automatic identification of multiword expressions (MWEs) in running text has recently received much attention among researchers in computational linguistics. The wide range of reported results for the task in the literature prompted us to take a closer look at the algorithms and evaluation methods. For supervised classification of Verb+Noun expressions, we propose a context-based methodology in which we find word embeddings to be appropriate features. We discuss the importance of train and test corpus splitting in validating the results and present type-aware train and test splitting. Given our specialised data, we also discuss the benefits of framing the task as classification rather than tagging.

1 Introduction

Ambiguity is a pervasive phenomenon in natural language. It starts from single words, and can propagate through larger linguistic constructs. Multiword expres-

Shiva Taslimipoor, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 299–317. Berlin: Language Science Press. DOI:10.5281/zenodo.1469569

sions (MWEs) which are idiosyncratic combinations of two or more words, behave differently in their separate usages in running text. In natural language processing (NLP) tasks such as part-of-speech tagging, parsing and machine translation, these expressions should be treated either before the task (Nivre & Nilsson 2004) or combined with the process (Constant & Tellier 2012; Kordoni et al. 2011; Nasr et al. 2015).

Examples of such expressions are: *take action*, *make sense* and *set fire*. [1] MWEs are a recurring theme in any language with some sources estimating their number to be in the same range as single words (Jackendoff 1997) or even beyond (Sag et al. 2002). Besides, new expressions come to languages on a regular basis. It is therefore not practically feasible to simply list MWEs in dictionaries or thesauri.

More importantly, most idiomatic expressions can also have literal meaning depending on context. For instance consider the expression *play games*. It is opaque with regards to its status as an MWE and depending on context could mean different things. For example in *He went to play games online* it has a literal sense but is idiomatic in *Don't play games with me as I want an honest answer*. Resolving these cases is critically important in many NLP applications (Katz 2006). Katz (2006) framed the task as sense disambiguation. Tagging corpora for MWEs or token-based identification of MWEs is an example of a task where it is necessary to differentiate between idiomatic and literal usages of each expression type.

Studies on MWEs can be divided into two main categories. One includes works regarding the canonical forms of expressions, their lexical properties and their potential to be considered as MWEs, namely type-based extraction of MWEs or MWE discovery; the other regards studies on tagging texts for the idiomatic usages of expressions, namely MWE tagging or token-based identification of MWEs. The former is a traditional approach which is of use to lexicographers as pointed in Ramisch (2014); the latter though, is more practical for NLP applications (Schneider et al. 2016).

Although discovering canonical forms of multiword expressions is still an active research area (Salehi & Cook 2013; Farahmand & Martins 2014), recently there is a significant move towards automatic tagging of corpora for MWEs (Schneider et al. 2014; Constant & Tellier 2012).

The focus of our study is token-based identification of MWEs, and we model it as a classification, rather than a sequence labelling problem. To determine the idiomaticity of each Verb+Noun occurrence, we experiment with using solely

[1] MWEs combine words from many different parts of speech. The pattern in our datasets is Verb+Noun, so all the examples in this chapter are of this kind.

context features without any sophisticated linguistic information. We do not exploit parsing, tagging or external lexicon-based information. To discriminate between idiomatic and literal Verb+Noun expression tokens, we have proposed a context-based classification approach expounded in detail in the previous publication (Taslimipoor et al. 2017). In this chapter, we build on this approach, experimenting with additional languages and several more sophisticated machine learning models. However, here we take a closer look at modelling and evaluation aiming at devising approaches that have more generalisation power and lead to less misleading results. We also conduct experiments to better demonstrate the suitability of framing our task as classification.

For token-based identification of MWEs, there is a wide range of results in the literature reported as the state-of-the art: from F-score of 64% with the DiMSUM dataset (Schneider et al. 2016) to 90% (Al Saied et al. 2017) for a dataset in the last PARSEME shared task (Savary et al. 2017). We find that in order for the performance results not to be misleadingly high, the distribution of the tokens between train and test corpus, henceforth called TRAIN AND TEST SPLITTING, should be controlled. Failure to do so will result in a kind of overfitting which could be overlooked during evaluation. For instance, an expression like *take advantage* is idiomatic consistently in all its usages in text. When different occurrences of this expression exist in both train and test corpus, the model memorises it from training data and predicts it very well in the test. Since such expressions are highly frequent, this memorisation helps the model to achieve erroneously high performance scores.

In the process of supervised identification of MWEs, we make observations with regards to the following: (1) the effect of train and test splitting of the tokens on generalisability of a model; (2) comparison between sequence labelling (tagging) and sequence classification.

1.1 Literature review

Identification of MWEs was shown to be effective in different NLP tasks, such as machine translation (Pal et al. 2011) and automatic parsing (Constant et al. 2012). There exists a considerable body of work in the literature attempting to investigate lexical and syntactic properties of expressions to account for their potential for being MWEs (Ramisch 2014; Baldwin & Kim 2010). However, recently there has been a great attention given to identifying where exactly this potential takes effect by tagging a running text for each individual occurrence (token) of an expression (Schneider et al. 2014; Constant & Tellier 2012; Gharbieh et al. 2017). Token-based identification of MWEs is effective in disambiguating between different behaviours of expressions in their individual usages. Evaluating all occurrences of expressions in the whole corpus of big size is not feasible. For this reason we have gathered a specialised dataset of concordances of particular expressions.

To the best of our knowledge, there are very few comprehensive tagged corpora for MWEs available, among which DiMSUM by Schneider et al. (2016) is very recent and well-cited. This corpus was used in the SemEval (2016) shared task in Detecting Minimal Semantic Units and their Meanings (DiMSUM). It is not particularly clear if the current methodologies applied to this corpus are capable of disambiguating between different usages of one specific canonical form.

Cook et al. (2008) prepared a dataset of English Verb+Noun constructions, categorising expressions based on their idiomaticity and how consistent they behave in their different usages. Fazly et al. (2009) have used that dataset for classifying Verb+Noun tokens into idiomatic or literal categories.

Katz (2006) used context features for identifying the idiomaticity/non-compositionality of MWEs in a different way. They represent different occurrences of an expression using LSA vectors and show that the vectors of the expressions in their idiomatic sense are very different from those of the same expressions in literal sense. Based on this observation they classify a test expression token depending on whether it is more similar to the idiomatic sense of the expression in training data or to the literal sense.

Scholivet & Ramisch (2017) recently tried to disambiguate a number of opaque French expressions using their contexts. They proposed a tagging approach using unigram and bigram features of the word forms and their POS. Qu et al. (2015) found word embedding representation of the words in context very useful for tagging a text with MWEs. We also used word vector representations of the verb and noun components of the expression and the words in a window size of two on the right of the expression as features for classifying expressions as MWE or not.

While most of the previous work on token-based identification of MWEs applied sequence tagging approaches using some kind of IOB labeling, Legrand & Collobert (2016) looked at the problem as classification. They proposed a neural network based approach that learns fixed-size representations for arbitrary sized chunks which is able to classify these representations as MWE or not. They showed better performance in MWE identification over the CRF-based approach in Constant et al. (2013).

1.2 Outline of the proposal

In almost all of the previous work on supervised modelling of MWE tokens, data is randomly split into train and test sets. In a random splitting, it is possible for occurrences of the same expression type to occur in both train and test sets. There are many instances where the expression almost always behaves idiomatically (e.g. *take part*, *make progress*) or literally (e.g. *eat food*, *give money*). In such cases a model learns every feature related to the POS and lemma form of the expression, and naturally can predict the correct tag for the expression perfectly in the test set (regardless of the expression being idiomatic or literal).

Having observed this issue, for evaluation we propose and perform type-aware train and test splitting. To this end we divide expression types into train and test folds and gather all occurrences of each type into the same fold. This makes the predication rigorous, since the model performs cross-type learning. One interesting study that considers cross-type learning of MWEs is the one by Fothergill & Baldwin (2012). However, they did not clearly explain the general advantages and effects of cross-type classification in evaluation. They used the approach in order to learn better features from specialised MWE resources.

We propose type-aware splitting of the data as a supplementary benchmark for evaluating MWE identification. We design experiments to show the effectiveness of this kind of evaluation in assessing the generalisability of models.

Recent studies on token-based identification of MWEs are heading towards using structured sequence tagging models. The choice of the model based on the data is an important issue. Our data includes occurrences of specific Verb+Noun expressions with the context around them. This makes it possible to have sizeable datasets annotated for a specific type of MWE in order to have a extensive evaluation. We observe that our data cannot benefit from sequence tagging and a regular classification approach can more reasonably model the data. We show better results from classification over a tagging model. Other than traditional machine learning classification approaches, we also propose a neural-network model by combining convolutional neural network and long short term memory models for identifying MWEs. Although some deep learning models have already been investigated for tagging MWEs by Gharbieh et al. (2017), to the best of our knowledge this is the first time this approach has been applied for classifying MWE instances.

We extensively discuss the following: 1) the division of data into train and test sets for evaluation and 2) the choice of model (classification versus tagging) based on the data.

2 Context-based identification of MWEs

In this study we use context features in a supervised environment to identify the idiomaticity of Verb+Noun expression tokens. In order to construct context features, for our first set of experiments (§4.1), given each occurrence of a Verb+Noun combination, we concatenate four different word vectors corresponding to the verb, noun, and their two following adjacent words while preserving the original order (following the previous work by Taslimipoor et al. 2017). Concatenated word vectors are fed into different classification models to be evaluated in terms of their performance.

The classification algorithms that have been used are Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), Multi Layer Perceptron (MLP) and Support Vector Machine (SVM). We also experimented with different neural network-based classification models. The best result is achieved with a combination of bidirectional Long Short-Term Memory network with a convolutional layer as a front-end (ConvNet+LSTM).

For the second set of experiments (§4.2), in which we compare Conditional Random Field (CRF) as a tagger with a simple Naive Bayes Classifier (NBC), we consider simple word forms of the verb, the noun, and the two words after as lexical context features.

We conducted our extensive experiments with Italian. The experiments are augmented by applying the approach also for smaller data in Spanish and English.

3 Experiments

3.1 Data

We first experiment with two similarly formatted datasets in Italian and Spanish and later also on a standard available dataset for English.

For Italian, our data includes a large set of concordances of Verb+Noun expressions.[2] Each item in the dataset is one concordance of a Verb+Noun expression and the whole item is annotated with 1 if the Verb+Noun inside is an MWE and with 0 otherwise. The data as explained in Taslimipoor et al. (2016) was annotated by two native speakers with Kappa agreement measure of 0.65. We resolved the disagreements by employing a third annotator who decided on most (but not all)

[2] The data as described in Taslimipoor et al. (2016) was gathered for four light verbs *fare*, *dare*, *trovare* and *prendre*. For some examples of expression instances refer to the same work.

cases of disagreements. This results in 20,030 concordances of 1,564 expression types. The Italian data is very imbalanced and almost 70% of the data is marked as MWE. To resolve this issue, we ignore the 15 most frequent expression types which are exclusively marked as MWE and also the expressions with frequency lower than 3. As a result we run the experiments on 18,540 concordances of 940 expression types.

For Spanish, we extracted concordances of Verb+Noun expressions in the same way using SketchEngine (Kilgarriff et al. 2004).[3] After ignoring the concordances for five most frequent expressions, $3,918$ usages were marked by two native speakers. The Kappa inter-annotator agreement was 0.55. Having seen the observed agreement of 0.79, we ignored all cases of disagreements and considered only the concordances on which both annotators agreed. This results in $3,090$ concordances of 747 expression types.

For English, we employ a standard dataset called VNC-tokens prepared by Cook et al. (2008).[4] The dataset is a benchmark for English verb-noun idiomatic expressions and was used for identifying MWE tokens in a number of previous studies such as Fazly et al. (2009) and Salton et al. (2016). The dataset includes sentences from the BNC corpus including occurrences of Verb+Noun expressions and is suitable for our task since it contains expressions with both skewed and balanced behaviour in being literal or idiomatic. Rather than concordances, it includes sentences from BNC containing occurrences of Verb+Noun expressions. Two English native speakers selected the expression types based on whether they have the potential for occurring in both idiomatic or literal senses. Although this dataset is slightly different from our Italian and Spanish data (which are extracted randomly), it has the same favourable pattern of different occurrences of same expression types that can be split into train and test. We find it interesting to investigate our observations on a differently gathered but standard dataset. The Verb+Nouns in this dataset are not necessarily continuous. We ignore the cases where the Verb+Noun occurs in passive form and the ones that the annotators were unsure of and this results in $2,499$ sentences consisting of Verb+Noun expressions. The statistics of the data for all three languages are reported in Table 1.

For all the three datasets, we consider the same context words as features for classification: we extract the vectors of the verb, noun and the two words after the noun.

[3] For Spanish, we focused on four light verbs *tener*, *hacer*, *dar* and *tomar*, similar to the ones we use for Italian.

[4] The dataset is available in https://sourceforge.net/projects/multiword/files/MWE_resources/20110627/

Table 1: Distribution of the data

	Italian	Spanish	English
Expression types	940	747	53
Expression tokens	18,540	3,090	2,499
MWE tokens	10,804 (58.27%)	2,094 (66.57%)	1,981 (79.27%)

3.2 Evaluation

In all cases classifier performance was measured using 10-fold cross-validation.

3.2.1 Standard splitting of data into train and test

In the standard method of performing cross-validation, the whole data is randomly divided into k folds and then the model is repeatedly trained on the data of $k - 1$ folds and tested on the data of the remaining fold. The result is averaged among k different iterations. In our task, we find the result of this evaluation misleading, since the repetition of the same expression in both train and test partitions helps the model to predict those specific types of expressions well, while the model might not work for new unseen expressions in test. Even stratified cross-validation suffers from the same kind of overfitting. In standard stratified cross-validation, imbalance is coped with by controlling the distribution of labels alone, so that all folds have the same distribution of labels. Similar to standard cross-validation, this method is not informed about the idiosyncratic distribution of types and tokens.

Therefore, these methods of evaluation cannot precisely reflect the effectiveness of the model or features and show better results for models that are more prone to overfitting. It is not particularly clear from this kind of evaluation if a good performing model could be generalised to unseen expressions and also to ambiguous expressions that have balanced distribution of occurrences as literal or idiomatic. We show the performance computed using this type of evaluation for different classifiers in Table 2.

3.2.2 Type-aware splitting and evaluation

We perform a custom cross-validation by splitting the expression occurrences into different folds considering their types/canonical forms. We split the expression types into k groups and all the occurrences of the expressions in the k^{th}

group goes into the k^{th} fold. This method ensures that the model performs cross-type learning and generalises to tokens from unseen types in the test fold. In other words, the model is learning the features and general patterns and does not overfit on highly recurrent token occurrences. The results for all classifiers evaluated using this approach is reported in Table 3.

4 Results

In this section, first a comparison of several classifiers using different train and test splitting methods is reported; then we present experiments using sequence tagging for identifying MWEs; and finally, the effectiveness of neural network-based word embeddings compared with count-based representations was analysed using one of the best classifiers.

4.1 Regular and type-aware evaluation

Evaluation performances for all classifiers using two different kinds of train and test splitting, namely regular (random) and our proposed type-aware, are reported in Table 2 and Table 3. The columns of the tables represent the results for Italian (IT), Spanish (ES) and English (EN). All traditional classifiers in this experiment use the same vectorised context features. The word vectors used in this study are available online.[5] The generated Italian and Spanish word embeddings applied Gensim's skipgram word2vec model with the window size of 10 to extract vectors of size 300. For English we use word embeddings of the same dimension trained using Glove (Pennington et al. 2014) algorithm available via spaCy.[6]

We also report the results from a more sophisticated neural network based architecture comprising of a BiLSTM with an additional convolutional layer as a front-end (ConvNet+LSTM). For this architecture the context window size is 2 (two words before and two words after the Verb+Noun expression).[7] Implementation details of these experiments can be found at https://github.com/shivaat/VN-tokens-clf.

Different classifiers show high performance with not much difference using regular cross-validation in which tokens are distributed into separate folds re-

[5] http://hlt.isti.cnr.it/wordembeddings/ for Italian and https://github.com/Kyubyong/wordvectors for Spanish

[6] https://spacy.io/docs/usage/word-vectors-similarities

[7] The difference in results were negligible when considering only the two context words on the right.

Table 2: Regular evaluation results: accuracy (standard deviation)

Classifiers	IT	ES	EN
Majority Baseline	0.5827	0.6657	0.7927
LR	0.8869 (0.007)	0.9129 (0.011)	0.8651 (0.020)
DT	0.8905 (0.008)	0.9065 (0.017)	0.8799 (0.018)
RF	0.9218 (0.005)	0.9337 (0.019)	0.9024 (0.017)
MLP	0.9069 (0.006)	0.933 (0.009)	0.9056 (0.016)
SVM	0.9116 (0.005)	0.9207 (0.009)	0.7927 (0.021)
ConvNet+LSTM	**0.9220 (0.007)**	**0.9668 (0.01)**	**0.8860 (0.024)**

Table 3: Type-aware evaluation results: accuracy (standard deviation)

Classifiers	IT	ES	EN
Majority Baseline	0.5827	0.6657	0.7927
LR	0.6909 (0.06)	0.8178 (0.074)	0.8092 (0.149)
DT	0.6048 (0.03)	0.7483 (0.078)	0.6327 (0.128)
RF	0.6337 (0.08)	0.7604 (0.097)	0.7321 (0.19)
MLP	0.7053 (0.06)	0.8319 (0.086)	0.7294 (0.169)
SVM	**0.7369 (0.07)**	0.8460 (0.093)	0.8062 (0.152)
ConvNet+LSTM	0.6601 (0.053)	**0.8681 (0.072)**	**0.8112 (0.106)**

gardless of their types (Table 2). ConvNet+LSTM, in particular, performs the best, which we believe is the result of overfitting arising from this method of train and test splitting. However, we can see notable differences between classifiers in Table 3 where we cross validate in a way that no same expression type occurs in both train and test partitions.

In the case of cross-type learning (Table 3), the SVM classifier showed the best results in identifying MWEs using vectorised context features for Italian, and close to the second best for Spanish and English data for which ConvNet+ LSTM is the best. The performance of this classifier is followed by that of MLP and LR for both Italian and Spanish. For English the results of SVM and LR are comparable. Computed performance for other classifiers like DT and RF dropped sharply when we use our type-aware cross-validation. This is also the case for ConvNet+LSTM for Italian data. This experiment determines how well a classifier can generalise among different expression types. SVM and LR in particular are shown to be fairly suitable for cross-type identification of MWEs. MLP also performs relatively well overall.

As for the English data it is worth noting that the VNC data is very imbalanced with the majority baseline of 0.7927 which is difficult to beat by classifiers.

4.2 Sequence classification versus sequence tagging

The experimental data in this study can be perfectly processed with standard classification approach, since the goal is to predict idiomaticity of an expression in a given context. However, Scholivet & Ramisch (2017) modelled such a data with sequence tagging. We believe that since not all the words in a sequences are going to be tagged, MWE identification using such a data cannot benefit from sequence labelling. We applied sequence tagging on the data to properly investigate the effects. Specifically, simple Naive Bayes Classifier (NBC) was considered as a simple sequence classification methodology and Conditional Random Field (CRF) was used as the sequence tagging approach. Both of the models use simple nominal features: the verb, the noun, and the two words after the noun.[8] The results are reported in Table 4 in terms of accuracy.

Table 4: Performance of sequence classification versus sequence tagging

	regular cross-validation			type-aware cross-validation		
	IT	ES	EN	IT	ES	EN
NBC	0.9504	0.9601	0.8560	0.7291	0.7298	0.6013
CRF	0.9165	0.9142	0.8176	0.6447	0.7199	0.6848

As can be seen in Table 4, CRF cannot even beat the simple naive bayes classifier except in the case of English data (when we apply cross-type learning). This is because our data is naturally suited for sequence classification and cannot benefit from sequence labelling models.

4.3 Effectiveness of word embedding representation

To specifically show the effectiveness of neural network-based embeddings for the classifiers to identify Verb+Noun MWEs, we performed an experiment using

[8] The features are the surface text occurrences of these words.

sparse bag-of-words count vectors with tf-idf weighting. In this case each sentence is considered as a collection of words, disregarding any word order. The sparse vector for each word is constructed based on its occurrence in different sentences. Each entry of the vector is weighted by tf (the word frequency) $*$ idf (the inverse of frequency of the sentences containing the word). Similar to previous experiments, we feed the vectors to a Multi Layer Perceptron (MLP) which works reasonably well compared to other classifiers based on the previous experiment. Note that the execution time for the best performing model, SVM, is almost 5 times that of MLP which makes it inefficient in comparison. The results of this comparison can be seen at Table 5.

Table 5: The accuracy of MLP in identifying Verb+Noun MWEs using word2vec and count-based embedding

		Accuracy (std.)	
	IT	ES	EN
MLP with count based embedding	0.6504 (0.0354)	0.7851 (0.042)	0.7002 (0.099)
MLP with word2vec	0.7053 (0.06)	0.8319 (0.086)	0.7294 (0.169)

The results in Table 5 show the improvement in performance when using word embeddings rather than the vanilla count-based vectors for all three languages (although less significant for English).

5 Discussion

In order to understand the argument behind type-aware evaluation and decide its applicability, we have to look at the distribution of data points. In the Italian data, for instance, the majority of data points belong to MWE types whose tokens occur invariably as idiomatic or literal only. In other words, if we plot the distribution of tokens with regards to the degree of idiomaticity of their corresponding types, we would see a skewed distribution (even after ignoring the 15th most frequent expressions), where only a smaller portion of tokens belong to MWE types whose usages can be fluid between literal and idiomatic. In such a scenario, a model easily overfits on the majority of the data, where labels were assigned invariably. However, this skewedness is not necessarily reflected in the distribution of MWE labels, as we might see a relatively balanced distribution of literal and idiomatic labels. This means there might be no severe class imbalance in the dataset, but within-class imbalance (Ali et al. 2015).

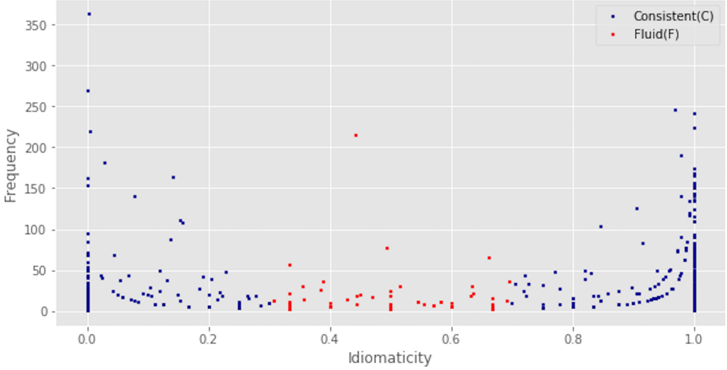


Figure 1: Distribution of expression types.

To illustrate the point, we operationalise two categories for MWE types, namely Consistent (C) and Fluid (F). Those types whose tokens occur more than 70% or less than 30% of the time as only literal or idiomatic are tagged as C, and the rest are considered F. Accordingly, Figure 1 shows the distribution of the expression types with regards to the behavior of their corresponding tokens. As can be seen, the majority of expressions with higher token frequencies are from the sub-class C. For this reason, evaluation using a vanilla cross validation or even stratified cross validation would not provide us with reliable results, since splitting of train and test disregards the within-class imbalance inherent in the data.

Since this is the case with data in real world, we propose type-aware train and test splitting as a supplementary approach for modelling the data and evaluating the results. This way, we make sure that a model has the best ability for generalisation, learns general properties for MWEs and is not merely based on memorising the words that construct MWEs.

It is worth noting that we did not used any linguistic or lexical features and we expect vector representation of context to be generalisable enough. Even with these generalisable features we observe substantial differences between regular and type-aware cross-validation. A proper method for train and test splitting is even more essential to validate the evaluation when a model trains on more exact features such as lexical ones.

With regards to previous data and models for MWEs, DiMSUM is one of the most noteworthy shared tasks. DiMSUM includes a recent tagged corpus for MWEs with a fairly small size of 4, 799 sentences in train and 1, 000 in test, including all types of MWEs. With such limited data, we observed only a few number of expressions of the form Verb+Noun occurring in both train and test. To give an example, with a selection of 6 most frequent light verbs, all their combinations with nouns are only 13 occurrences in the test data, out of which only 3 are MWEs. There are no repeated occurrences of these cases in both train and test data. Therefore, we believe that this data inherently does not lead to misleading results. In other words, a model that works well on this data could be fairly generalised.

Gharbieh et al. (2017) showed better performance when using deep neural network models compared with traditional machine learning on DiMSUM. However, in our experiment of type-aware classification, SVM performed the best, even outperforming LSTM and ConvNet and their combinations for Italian and Spanish. Since neither DiMSUM nor our data is big enough for a proper analysis with deep learning, more studies are required to find the most effective model to identify MWEs.

Another data for token-based identification of MWEs in English that we also used in this study is VNC-tokens (Cook et al. 2008). One advantage of this corpus is that the data is particularly gathered for the task of disambiguation between idiomatic and literal usages of expressions. Before the annotation, they selected only the expressions that have the potential for occurring in both idiomatic and literal senses. Although we did not follow the initial development splitting of the data for this study (i.e. we followed our proposed way of splitting the data into train and test), the development and test splitting of this data is type-aware. Therefore, an experiment with this data, is able to truly measure generalisation.

In PARSEME shared task (Savary et al. 2017), which features the most recent multi-lingual data for MWEs, Maldonado et al. (2017) presented statistics on the percentage of previously seen data in test sets of all languages (i.e. proportion of MWE instances in the test set that were seen also in the training set). The correlation between these percentages and the results stress the need for proper train and test splitting. Maldonado & QasemiZadeh (2018 [this volume]) further discuss the characteristics of the shared task data and report the performance results of the systems on seen and un-seen data separately. The experiments with the data for the Parseme shared task, which is also discussed in Savary et al. (2018 [this volume]), would definitely benefit from such type-aware train and test splitting.

6 Conclusions

In this study, we explored a context-based classification method for identification of Verb+Noun expressions. We employed word embedding to represent context features for MWEs. We evaluated the methodology using type-aware cross-validation and discussed its effectiveness compared with standard evaluation. We argue that only this proposed method properly accounts for the generalisability of a model. We also showed that our data (and similar ones) for this task cannot benefit from structured sequence tagging models.

The effectiveness of word embeddings as context features for identifying MWEs should be examined in more detail with datasets of larger size and with more sophisticated embeddings that consider linguistic features. We would also like to analyse the effect of our proposed approach on unseen and less frequent data.

Acknowledgements

The authors would like to thank Manuela Cherchi, Anna Desantis, Martina Cotella, Andrea Silvestre and Alondra Nava Zea for their help in annotations.

Abbreviations

BiLSTM	bi-directional long short-term memory
CRF	conditional random field
ConvNet	convolutional neural network
DT	decision tree
IOB	inside-outside-beginning
LSA	Latent Semantic Analysis
LR	logistic regression
LSTM	long short-term memory
MLP	multi-layer perceptron
MWE	multiword expression
NBC	naive Bayes classifier
NLP	Natural Language Processing
POS	part of speech
RF	random forest
SVM	support vector machine

References

Al Saied, Hazem, Matthieu Constant & Marie Candito. 2017. The ATILF-LLF system for parseme shared task: A transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 127–132. Association for Computational Linguistics. DOI:10.18653/v1/W17-1717

Ali, Aida, Siti Mariyam Shamsuddin & Anca L. Ralescu. 2015. Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and its Application* 7(3). 687–719.

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

Constant, Matthieu, Marie Candito & Djamé Seddah. 2013. The LIGM-alpage architecture for the SPMRL 2013 shared task: Multiword expression analysis and dependency parsing. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages*, 46–52.

Constant, Matthieu, Anthony Sigogne & Patrick Watrin. 2012. Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 204–212. Association for Computational Linguistics.

Constant, Matthieu & Isabelle Tellier. 2012. Evaluating the impact of external lexical resources into a CRF-based multiword segmenter and part-of-speech tagger. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), 646–650. European Language Resources Association (ELRA).

Cook, Paul, Afsaneh Fazly & Suzanne Stevenson. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions* (MWE '08), 19–22. Association for Computational Linguistics.

Farahmand, Meghdad & Ronaldo Martins. 2014. A supervised model for extraction of multiword expressions, based on statistical context features. In *Proceedings of the 10th Workshop on Multiword Expressions* (MWE '14), 10–16. Association for Computational Linguistics.

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103. http://aclweb.org/anthology/J09-1005.

Fothergill, Richard & Timothy Baldwin. 2012. Combining resources for MWE-token classification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation* (SemEval '12), 100–104. Association for Computational Linguistics.

Gharbieh, Waseem, Virendra Bhavsar & Paul Cook. 2017. Deep learning models for multiword expression identification. In *Proceedings of the 6th Joint Confer-*

*ence on Lexical and Computational Semantics (*SEM 2017)*, 54–64. Association for Computational Linguistics.

Jackendoff, Ray. 1997. *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Katz, Graham. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (MWE '06), 12–19. Association for Computational Linguistics.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. In Geoffrey Williams & Sandra Vessier (eds.), *Proceedings of the 11th EURALEX international congress*, 105–116. Lorient, France: Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Kordoni, Valia, Carlos Ramisch & Aline Villavicencio. 2011. Proceedings of the ACL Workshop on Multiword Expressions: From Parsing and Generation to the Real World. In (MWE '11). Association for Computational Linguistics.

Legrand, Joël & Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions* (MWE '16), 67–71. Association for Computational Linguistics. http://anthology.aclweb.org/W16-1810.

Maldonado, Alfredo, Lifeng Han, Erwan Moreau, Ashjan Alsulaimani, Koel Dutta Chowdhury, Carl Vogel & Qun Liu. 2017. Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 114–120. Association for Computational Linguistics. DOI:10.18653/v1/W17-1715

Maldonado, Alfredo & Behrang QasemiZadeh. 2018. Analysis and Insights from the PARSEME Shared Task dataset. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 149–175. Berlin: Language Science Press. DOI:10.5281/zenodo.1469557

Nasr, Alexis, Carlos Ramisch, José Deulofeu & André Valli. 2015. Joint dependency parsing and multiword expression tokenization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1116–1126. Association for Computational Linguistics. http://www.aclweb.org/anthology/P15-1108.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Workshop on methodologies and evaluation of multiword units in real-world ap-*

plications (MEMURA 2004), 39–46. http://stp.lingfil.uu.se/~nivre/docs/mwu.pdf.

Pal, Santanu, Tanmoy Chakraborty & Sivaji Bandyopadhyay. 2011. Handling multiword expressions in phrase-based Statistical Machine Translation. In *Proceedings of the 13th machine translation summit* (MT Summit 2011), 215–224. September 19-23, 2011.

Pennington, Jeffrey, Richard Socher & Christopher D. Manning. 2014. GloVe: global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014), 1532–1543. http://www.aclweb.org/anthology/D14-1162. October 25–29, 2014.

Qu, Lizhen, Gabriela Ferraro, Liyuan Zhou, Weiwei Hou, Nathan Schneider & Timothy Baldwin. 2015. Big Data Small Data, In Domain Out-of Domain, Known Word Unknown Word: the impact of word representations on sequence labelling tasks. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning* (CoNLL 2015), 83–93.

Ramisch, Carlos. 2014. *Multiword expressions acquisition: A generic and open framework* (Theory and Applications of Natural Language Processing XIV). Cham, Switzerland: Springer. 230. http://link.springer.com/book/10.1007%2F978-3-319-09207-2.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Salehi, Bahar & Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics*, vol. 1 (* SEM 2013), 266–275. June 13-14, 2013.

Salton, Giancarlo, Robert Ross & John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 194–204. Berlin, Germany: Association for Computational Linguistics.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary &

Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI:10.5281/zenodo.1471591

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 31–47. Association for Computational Linguistics. DOI:10.18653/v1/W17-1704

Schneider, Nathan, Emily Danchik, Chris Dyer & Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics* 2(1). 193–206. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/281.

Schneider, Nathan, Dirk Hovy, Anders Johannsen & Marine Carpuat. 2016. SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), 546–559. Association for Computational Linguistics. http://www.aclweb.org/anthology/S16-1084.

Scholivet, Manon & Carlos Ramisch. 2017. Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 167–175. Association for Computational Linguistics. http://www.aclweb.org/anthology/W17-1723.

Taslimipoor, Shiva, Anna Desantis, Manuela Cherchi, Ruslan Mitkov & Johanna Monti. 2016. Language resources for Italian: Towards the development of a corpus of annotated Italian multiword expressions. In *Proceedings of 3rd Italian Conference on Computational Linguistics* (CLiC-it 2016) *& 5fth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* (EVALITA 2016) (Collana dell'Associazione Italiana di Linguistica Computazionale), 5–6 December 2016. Torino: Accademia University Press. online.

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov & Afsaneh Fazly. 2017. Investigating the opacity of verb-noun multiword expression usages in context. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 133–138. Association for Computational Linguistics.

Chapter 12

Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents

Marcos Garcia
Universidade da Coruña

This chapter introduces a strategy for the automatic extraction of multilingual collocation equivalents which takes advantage of parallel corpora to train bilingual word embeddings. First, monolingual collocation candidates are retrieved using syntactic dependencies and standard association measures. Then, the distributional models are applied to search for equivalents of the elements of each collocation in the target languages. The proposed method extracts not only collocation equivalents with direct translations between languages, but also other cases where the collocations in the two languages are not literal translations of each other. Several experiments – evaluating collocations with five syntactic patterns – in English, Spanish, and Portuguese show that this approach can effectively extract large sets of bilingual equivalents with an average precision of about 85%. Moreover, preliminary results on comparable corpora suggest that the distributional models can be applied for identifying new bilingual collocations in different domains. This strategy is compared to both hand-crafted bilingual dictionaries and to probabilistic translation dictionaries learned from the same resources as the bilingual word embeddings, showing that it achieves much larger recall values while keeping high precision results.

1 Introduction

MWEs have been repeatedly classified as an important problem for developing Natural Language Processing (NLP) tools, as well as to automatically analyze

Marcos Garcia. 2018. Comparing bilingual word embeddings to translation dictionaries for extracting multilingual collocation equivalents. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 319–342. Berlin: Language Science Press. DOI:10.5281/zenodo.1469571

linguistic utterances (Sag et al. 2002). Among the different types of MWEs, processing collocations in an automatic way may pose various problems due to their intrinsic properties such as compositionality or unpredictability (Mel'čuk 1998).

From a theoretical perspective, there are at least two main views on collocations. On the one hand, there is a tendency to consider any frequent pair of words to be a collocation (Smadja 1993; Evert & Kermes 2003; Kilgarriff 2006). On the other hand, the phraseological tradition needs both a lexical restriction and a syntactic relation to consider two lexical units as a collocation.[1] From this phraseological point of view, a collocation is a restricted binary co-occurrence of lexical units between which a syntactic relation holds, and that one of the lexical units (the BASE) is chosen according to its meaning as an isolated lexical unit, while the other (the COLLOCATE) is selected depending on the base and the intended meaning of the co-occurrence as a whole, rather than on its meaning as an isolated lexical unit (Mel'čuk 1998). Thus, a noun in English such as *picture* (as a direct object) requires the verb *to take* (and not *to do*, or *to make*) in the phrase *take a picture*, while *statement* selects *to make* (*make a statement*).

In a bilingual (or multilingual) scenario, equivalent collocations are needed to produce more natural utterances in the target language(s). In this regard, the referred noun *fotografia* 'picture' would select the verb *tirar* 'to remove' in Portuguese (*tirar uma fotografia*). Similarly the Spanish *vino* 'wine' would require the adjective *tinto* (*vino tinto*), which is not the main translation of *red* (*red wine*).

The unpredictability of these structures poses problems for tasks such as machine translation, whose performance can benefit from lists of multilingual collocations or transfer rules for these units (Orliac & Dillinger 2003). In areas like second language learning, it has been shown that even advanced learners need to know which word combinations are allowed in a specific linguistic variety (Altenberg & Granger 2001; Alonso-Ramos et al. 2010). Thus, obtaining resources of multilingual equivalent collocations could be useful for a variety of applications such as those mentioned above. However, this kind of resource is scarce, and constructing them manually requires a large effort from expert lexicographers.

Since the 1990s, a number of approaches were implemented aimed at extracting bilingual collocations, both from parallel corpora (Kupiec 1993; Smadja et al. 1996; Wu & Chang 2003), and from comparable or even from non-related monolingual resources (Lü & Zhou 2004; Rivera et al. 2013), often combining statistical approaches with the use of bilingual dictionaries to find equivalents of each base.

[1] An overview of different views on collocations – both from theoretical and practical perspectives – can be found in Seretan (2011).

This chapter explores the use of distributional semantics (by means of bilingual word embeddings) for identifying bilingual equivalents of monolingual collocations: On the one hand, monolingual collocation candidates are extracted using a harmonized syntactic annotation provided by Universal Dependencies (UD),[2] as well as standard measures for lexical association. On the other hand, bilingual word embeddings are trained using lemmatized versions of noisy parallel corpora. Finally, these bilingual models are employed to search for semantic equivalents of both the base and the collocate of each collocation.

Several experiments using the OpenSubtitles2016 parallel corpora (Lison & Tiedemann 2016) in English, Portuguese, and Spanish show that the proposed method successfully identifies bilingual collocation equivalents with different patterns: *adjective-noun*, *noun-noun*, *verb-object*, *verb-subject*, and *verb-adverb*. Furthermore, preliminary results in comparable corpora suggest that the same strategy can be applied in this kind of resources to extract new pairs of bilingual collocations. In this regard, this chapter is an extended version of a previous work on bilingual collocation extraction (Garcia et al. 2017), including new collocation patterns and a larger evaluation which compares the proposed approach to probabilistic translation dictionaries (Hiemstra 1998; Simões & Almeida 2003).

Apart from this introduction, §2 includes a review of previous work on collocation extraction, especially on papers dealing with bilingual resources. Then, §3 and §4 present and evaluate the method, respectively. Finally, some conclusions and further work are discussed in §5.

2 Previous studies on collocation extraction

The extraction of monolingual collocation candidates (as well as other MWEs) from corpora is a well-known topic in corpus and computational linguistics and was the focus of a significant body of work in different languages.

In this respect, most strategies use statistical association measures on windows of n-grams with different sizes (Church & Hanks 1990; Smadja 1993). Other methods, such as the one presented by Lin (1999), started to apply dependency parsing to better identify combinations of words which occur in actual syntactic relations.

More recently, the availability of better parsers allowed researchers to combine automatically obtained syntactic information with statistical methods to extract collocations more accurately (Evert 2008; Seretan 2011).

[2] http://universaldependencies.org/

A different perspective on collocation extraction focuses not only on their retrieval, but on semantically classifying the obtained collocations, in order to make them more useful for NLP applications (Wanner et al. 2006; 2016).

Concerning the extraction of bilingual collocations, most works rely on parallel corpora to find the equivalent of a collocation in a target language. In this regard, Smadja (1992) and Smadja et al. (1996) first identify monolingual collocations in English (the source language), and then use Mutual Information (mi) and the Dice coefficient to find the French equivalents of the source collocations.

Kupiec (1993) also uses parallel corpora to find noun phrase equivalents between English and French. Their method consists of applying an expectation maximization (EM) algorithm to previously extracted monolingual collocations. Similarly, Haruno et al. (1996) obtain Japanese-English chunk equivalents by computing their mi scores and taking into account their frequency and position in the aligned corpora.

Another work which uses parallel corpora is presented by Wu & Chang (2003). The authors extract Chinese and English n-grams from aligned sentences by computing their log-likelihood ratio. Then, the competitive linking algorithm is used to decide whether each bilingual pair actually corresponds to a translation equivalent.

Seretan & Wehrli (2007) took advantage of syntactic parsing to extract bilingual collocations from parallel corpora. The strategy consists of first extracting monolingual collocations using log-likelihood, and then searching for equivalents of each base using bilingual dictionaries. The method also uses the position of the collocation in the corpus, and relies on the syntactic analysis by assuming that equivalent collocations will occur with the same syntactic relations within the collocations in both languages.

Rivera et al. (2013) present a framework for bilingual collocation retrieval that can be applied (using different modules) to both parallel and comparable corpora. As in other works, monolingual collocations based on n-grams are extracted in a first step, and then bilingual dictionaries (or WordNet, in the comparable corpora scenario) are used to find the equivalents of the base in the aligned sentence or in a small window of adjacent sentences of the source collocation.

A different approach, which uses non-related monolingual corpora for finding bilingual collocations, was presented in Lü & Zhou (2004). Here, the authors apply dependency parsing and the log-likelihood ratio for obtaining English and Chinese collocations. Then, they search for translations using word translation equivalents with the same dependency relation in the target language (using the EM algorithm and a bilingual dictionary).

Although not focused on collocations, Fung (1998) applied methods based on distributional semantics to build bilingual lexica from comparable corpora. This approach takes into account that in this type of resources the position and the frequency of the source and target words are not comparable, and also that the translations of the source words might not exist in the target document.

Similarly, the strategy presented in this chapter leverages noisy parallel corpora for building bilingual word embeddings. However, with a view to applying it to other resources such as comparable corpora, it identifies equivalents without using information about the position of the collocations or their comparative frequency in the corpora. Furthermore, it does not take advantage of external resources such as bilingual dictionaries, making it easy to extend to other languages. Garcia et al. (2018) had introduced a naive version of this approach, including experiments in Portuguese and Spanish with just one collocation pattern.

3 A new method for bilingual collocation extraction

This section presents the proposed method for automatically extracting bilingual collocations from corpora. First, the approach for identifying candidates of monolingual collocations using syntactic dependencies is briefly described. Then, the process of creating the bilingual word embeddings is shown, followed by the strategy for discovering the collocation equivalents between languages.

3.1 Monolingual dependency-based collocation extraction

Early works on n-gram based collocation extraction already pointed out the need for syntactic analysis to better identify collocations in corpora (Smadja 1993; Lin 1999). Syntactic analysis can, on the one hand, avoid the extraction of syntactically unrelated words which occur in small context windows. On the other hand, it can effectively establish a relation between lexical items occurring in long-distance dependencies (Evert 2008).

Besides, the method presented in this chapter assumes that most bilingual equivalents of collocations bear the same syntactic relation in both the source and the target languages, although it is not always the case (Lü & Zhou 2004).

In order to better capture the syntactic relations between the base and the collocate of each collocation, the strategy uses state-of-the-art dependency parsing. Apart from that, and aimed at obtaining harmonized syntactic information between languages, the method relies on Universal Dependencies annotation, which makes it possible to use the same strategy for extracting and analyzing the collocations in multiple languages.

3.1.1 Preprocessing:

Before extracting the collocation candidates from each corpus, a pipeline of NLP tools is applied in order to annotate the text with the desired information. Thus, the output of this process consists of a parsed corpus in CoNLL-U format,[3] where each word is assigned to its surface form, its lemma, its POS-tag and morphosyntactic features, its syntactic head as well as the UD relation of the word in context.

From this analyzed corpus, the word pairs belonging to the desired relations (collocation candidates) are extracted. We keep their surface forms, POS-tags, and other syntactic dependents which may be useful for the identification of potential collocations. Besides, a list of triples is retained in order to apply association measures, containing (i) the syntactic relation, (ii) the head, and (iii) the dependent (using their lemmas together with the POS-tags). Thus, from a sentence such as *John took a great responsibility*, the following triples (among others) are obtained:

nsubj(take$_{\text{Verb}}$,John$_{\text{PropN}}$)
amod(responsibility$_{\text{Noun}}$,great$_{\text{Adj}}$)
dobj(take$_{\text{Verb}}$,responsibility$_{\text{Noun}}$)

This information, along with the corpus size and the frequency of the different elements of the potential collocations, is stored in order to rank the candidates.

3.1.2 Collocation patterns:

In this chapter, candidates of five different syntactic patterns of collocations are extracted in three languages, Spanish (ES), Portuguese (PT), and English (EN):[4]

- Adjective—Noun (amod): these candidates are pairs of adjectives (as collocates) and nouns (as bases) where the former syntactically depends of the latter in a amod relation. Example: **killer**$_{\text{base}}$;**serial**$_{\text{collocate}}$.

- Noun—Noun (nmod): this pattern consists of two common nouns related by the nmod relation, where the head is the base and the dependent is

[3] http://universaldependencies.org/format.html

[4] In this chapter we address the European variety of Portuguese. However, even if we use a European Portuguese corpus (see §4), it contains some texts in the Brazilian dialect.

the collocate (optionally with a CASE marking dependent preposition: *of* in English, *de* in Portuguese and Spanish). Example: **rage**$_{\text{base}}$;**fit**$_{\text{collocate}}$.[5]

- Verb—Object (VOBJ): *verb-object* collocations consist of a verb (the collocate) and a common noun (the base) occurring in a DOBJ relation. Example: **care**$_{\text{base}}$;**take**$_{\text{collocate}}$.

- Subj—Verb (VSUBJ): the VSUBJ collocation pattern contains a common noun (the base, acting as a subject) and the verb it depends on (the collocate). Example: **ship**$_{\text{base}}$;**sink**$_{\text{collocate}}$.

- Verb—Adverb (ADVMOD): in this case, a collocate adverb modifies a verb (the base) in an ADVMOD relation. Example: **want**$_{\text{base}}$;**really**$_{\text{collocate}}$.

3.1.3 Identification of candidates:

For each of the five patterns of collocations, a list of potential candidates for the three languages is extracted. After that, the candidates are ranked using standard association measures that have been widely used in collocation extraction (Evert 2008).

In the current experiments, two statistical measures were selected, whose results complement each other: T-SCORE, which prefers frequent dependency pairs, and has been proved useful for collocation extraction (Krenn & Evert 2001), and MUTUAL INFORMATION, which is useful for a large corpus, even if it tends to assign high scores to candidates with very low-frequency (Pecina 2010).

The output of both association measures is merged in a final list for each language and collocation pattern, defining thresholds of *T-SCORE≥2* and *MI≥3* (Stubbs 1995), and extracting only collocations with a frequency of *f≥10.* This large value was defined to reduce the extraction of incorrect entries from a noisy corpus and from potential errors of the automatic analysis.

It must be noted that, since these lists of monolingual collocations have been built based on statistical measures of collocability, their members need not be *bona fide* collocations in the phraseological meaning. Thus, the lists can include idioms, e.g., *kick the bucket*, quasi-idioms, e.g., *big deal*, (Mel'čuk 1998), or free combinations, e.g., *buy a drink.*

[5] Some collocations belonging to this pattern are analyzed in UD – mainly in English – using the COMPOUND relation. These are not extracted in the experiments performed in this chapter.

3.2 Bilingual word embeddings

Word embeddings are low-dimensional vector representations of words which capture their distributional context in corpora. Even though distributional semantics methods have been largely used in previous years, approaches based on word embeddings gained popularity with the publication of *word2vec* (Mikolov et al. 2013). Based on the *Skip-gram* model of *word2vec*, Luong et al. (2015) proposed *BiSkip*, a model of word embeddings which learns bilingual representations using aligned corpora, thus being able to predict words crosslinguistically.

The method presented in this chapter uses lemmas instead of surface forms to identify the collocation candidates, so the bilingual models of word embeddings are also trained on lemmatized corpora. Therefore, the raw parallel corpus is lemmatized keeping the original sentence alignment.

The bilingual models are built using *MultiVec*, an implementation of *word2vec* and *BiSkip* (Berard et al. 2016). As the approach is evaluated in three languages, three different bilingual models are needed: Spanish-English, Portuguese-English, and Spanish-Portuguese.

As it will be shown, the obtained models can predict the similarity between words in bilingual scenarios by computing the cosine similarity between their vectors. As the models learn the distribution of single words (lemmas), they deal with different semantic phenomena such as polysemy or homonymy. Concerning collocations, this means that, ideally, the bilingual models could predict not only the equivalents of a base, but also to capture the (less close) semantic relation between the bilingual collocates, if they occur frequently enough in the data.

3.3 Bilingual collocation alignment

In order to identify the bilingual equivalent of a collocation in a target language, the method needs (i) lists of monolingual collocations (ideally obtained from similar resources), and (ii) a bilingual *source-target* model of word embeddings.

With these resources, the following strategy is applied: For each collocation in the source language (e.g., *lío*$_{\textit{base}}$ *tremendo*$_{\textit{collocate}}$ 'huge mess' in Spanish) the system selects its base and obtains – using the bilingual model – the *n* most similar lemmas in the target language (where *n=5* in the experiments performed in this chapter): *trouble*, *mess*, etc. Then, starting from the most similar lemma, we search in the target list for collocations containing the equivalents of the base (*trouble*$_{\text{base}}$ *little*$_{\text{collocate}}$, *trouble*$_{\text{base}}$ *deep*$_{\text{collocate}}$, *mess*$_{\text{base}}$ *huge*$_{\text{collocate}}$, *mess*$_{\text{base}}$ *fine*$_{\text{collocate}}$, etc.). If a collocation with a base equivalent is found, the cosine similarity between both collocates (*tremendo* versus *little*, *deep*, *huge*, and *fine*) is computed, and they are selected as potential candidates if their similarity is higher than a given threshold (empirically defined in this chapter as 0.65), and if the target candidate is among the *n* most similar words of the source collocate (again, *n=5*). Finally, if these conditions are met, the source and target collocations are aligned, assigning the average distance between the bases and the collocates as a confidence value, as in the following Spanish-English example: $\textit{lío}_{base}$ $\textit{tremendo}_{collocate}$ = $\textit{mess}_{base}$ $\textit{huge}_{collocate}$ → 0.721.

4 Evaluation

This section presents the experiments carried out in order to evaluate the proposed distributional method (henceforth DIS) in the three analyzed languages, using the five collocation patterns defined in §3.1. The approach presented in this chapter is compared to a baseline system (BAS), which uses hand-crafted bilingual dictionaries, and to probabilistic translation dictionaries (NAT).[6]

Corpora: Monolingual collocations were extracted from a subset of the OpenSubtitles2016 corpus (Lison & Tiedemann 2016), which contains parallel corpora from TV and Movie subtitles. This resource was selected because it is a large and multilingual parallel corpus likely to contain different types of collocations, also from an informal register, thus being useful for comparative studies.[7]

From the English, Spanish, and Portuguese corpora, those senteces which appear in the three languages were selected, for a total of 13,017,016 sentences. These sentences were tokenized, lemmatized and POS-tagged with a multilingual NLP pipeline (Garcia & Gamallo 2015), obtaining three corpora of about 91M (ES and PT), and about 98M (EN) tokens. The resulting data were enriched with syntactic annotation using statistical models trained with MaltParser (Nivre et al. 2007) on version 1.4 of the UD treebanks (Nivre et al. 2016).

Collocations: From each corpus, five patterns of collocation candidates were extracted: AMOD, NMOD, VOBJ, VSUBJ, and ADVMOD. For each language and pattern, a single list of collocations was obtained by merging the MI and T-SCORE outputs as explained in §3.1. Table 1 shows the number of filtered collocations in each case (*colls*).

[6] The extractions of these three methods are available at http://www.grupolys.org/~marcos/pub/pmwe-dis.tar.bz2

[7] Note, however, that OpenSubtitles2016 includes non-professional translations with some noisy elements such as typos or case inconsistencies, among others.

Table 1: Number of unique input dependencies for each syntactic pattern (*deps*), and final monolingual collocation candidates (*colls*).

Lg	AMOD		NMOD		VOBJ		VSUBJ		ADVMOD	
	deps	*colls*	*deps*	*colls*	*deps*	*colls*	*deps*	*colls*	*deps*	*colls*
ES	373K	13,870	644K	5,673	423K	17,723	287K	4,914	124K	5,526
PT	361K	12,967	709K	5,643	544K	20,984	283K	3,927	142K	6,660
EN	381K	14,175	517K	3,133	483K	15,492	264K	2,663	162K	6,711

Another version of each corpus was created only with the lemma of each token, keeping the original sentence alignments. These corpora were used for training three bilingual word embeddings with MultiVec, with 100 dimensions and a window-size of eight words: ES-EN, ES-PT, and PT-EN.[8]

Baseline (Bas): The performance of the method described in §3.3 was compared to a baseline which follows the same strategy, but uses bilingual dictionaries instead of the word embeddings models. Thus, the Bas method obtains the equivalents of both the base and the collocate of a source collocation, and verifies whether there is a target collocation with the translations. The bilingual dictionaries provided by the *apertium* project were used for these experiments (Forcada et al. 2011).[9]

The Spanish-Portuguese dictionary has 14, 364 entries, and the Spanish-English one contains 34, 994. The Portuguese-English dictionary (not provided by *apertium*) was automatically obtained by transitivity from the two other lexica, with a size of 9, 160 pairs.

Probabilistic translation dictionaries (Nat): The distributional method was also compared to probabilistic translation dictionaries. Probabilistic dictionaries are bilingual resources which contain, for each word in a source language, possible translations in the target language together with the probability of the translation being correct. To obtain these dictionaries NATools was used, which is a set of tools to work with parallel corpora that can be utilized for different tasks such as sentence and word alignment, or to extract bilingual translation dictionaries by means of statistical methods (Simões & Almeida 2003). The probabilistic dictionaries are obtained by applying the EM algorithm on sparse matrices of

[8] These models are available at http://www.grupolys.org/~marcos/pub/mwe17_models.tar.bz2

[9] SVN revision 75,477, https://svn.code.sf.net/p/apertium/svn/

bilingual word co-occurrences, previously built from parallel corpora (Hiemstra 1998).

For a better comparison to the DiS model, Nat dictionaries were extracted from the same lemmatized resources used for training the bilingual word embeddings. Thus, this method only differs from DiS in the bilingual resources used to search for equivalents of the bases and the collocates.[10]

4.1 Results

With a view to knowing the performance of Bas, Nat, and DiS in the different scenarios, 100 bilingual collocation pairs were randomly selected from each language and pattern, creating a total of 45 lists (15 from each of the three methods).[11]

Three reviewers worked during the evaluation process. Each bilingual collocation pair was labeled as (i) correct, (ii) incorrect, or (iii) dubious, which includes pairs where the translation might be correct in some contexts even if they were not considered faithful translations.[12] Correct collocation equivalents are those pairs where the monolingual extractions were considered correct, both in terms of co-occurrence frequency and of collocation pattern classification, and whose translations were judged by the reviewers as potential translations in a real scenario. Two reviewers labeled each collocation pair in the Bas and DiS outputs, achieving 92% and 83% inter-annotator agreement, respectively, with an average $\kappa = 0.39$, which indicates the difficulty of this kind of annotation. Pairs with correct/incorrect disagreement were discarded for the evaluation. Those with at least one dubious label were checked by a third annotator, deciding in each case whether they were correct, incorrect, or dubious. This third annotator evaluated the outputs of Nat using exactly the same guidelines.

From these data, the precision values for each case were obtained by dividing the number of correct collocation equivalents by the number of correct, incorrect, and dubious cases (so dubious cases were considered incorrect). Recall (r) was obtained by multiplying the precision values (p) for the number of extracted equivalents (e), and dividing the result by the lowest number of input collocations for each pair (i, see Table 1). For instance, the Spanish-Portuguese baseline

[10] After preliminary evaluations, the translation probability thresholds of both lexical units were empirically defined as 0.1.

[11] Except for baseline extractions with less than 100 elements, where all of them were selected.

[12] Some of these dubious equivalents are actual translations in the original corpus, such as the Spanish-English *copa de champaña* 'champagne cup', which was translated as *cup of wine*, even if they are semantically different.

Table 2: Number of bilingual extractions of the baseline, Nat, and DiS systems.

Pattern	model	ES-PT	ES-EN	PT-EN
amod	Bas	657	248	213
	Nat	1,329	1,113	1,005
	DiS	9,464	7,778	7,083
nmod	Bas	320	32	43
	Nat	704	138	136
	DiS	3,867	890	917
vobj	Bas	529	183	241
	Nat	1,443	1,461	1,544
	DiS	12,887	8,865	9,206
vsubj	Bas	188	27	55
	Nat	382	346	323
	DiS	2,522	1,344	1,298
advmod	Bas	58	19	22
	Nat	113	104	106
	DiS	3,721	2,301	2,412

recall for the amod pattern was estimated as follows (see Table 1, Table 2, and Table 3): $r = \frac{p*e}{i} = \frac{99*657}{12,967} = 5.01$.[13] Finally, f-score values (the harmonic mean between precision and recall) were obtained for each case, and the macro-average results were calculated for each language, pattern, and approach.

Table 2 contains the number of bilingual collocation equivalents extracted by each method in the 15 settings from the input lists of monolingual data (Table 1). These results clearly show that the baseline approach extracts a lower number of bilingual equivalents. Nat obtains much more bilingual collocations than Bas, but both methods extract less equivalents than the distributional approach. This might have happened due to the size of the dictionaries in Bas and because of

[13] Note that these recall results assume that every collocation in the shortest input list of each pair has an equivalent on the other language, which is not always the case. Thus, more realistic recall values (which would need an evaluation of every extracted pair) will be higher than the ones obtained in these experiments.

Table 3: Precision, recall and f-score of the baseline (Bas) system (*average* is macro-average).

Pattern		ES-PT	ES-EN	PT-EN	average
	P	99.0	95.8	97.9	97.6
amod	R	5.0	1.7	1.6	2.8
	F1	9.6	3.4	3.2	5.4
	P	97.8	100	91.7	96.5
nmod	R	5.5	1.0	1.3	2.6
	F1	10.5	2.0	2.5	5.1
	P	98.7	100	92.1	96.9
vobj	R	3.0	1.2	1.4	1.9
	F1	5.7	2.3	2.8	3.6
	P	93.8	96.3	92.7	94.3
vsubj	R	4.5	1.0	1.9	2.5
	F1	8.6	1.9	3.8	4.8
	P	96.7	100	95.7	97.4
advmod	R	1.0	0.3	0.3	0.6
	F1	2.0	0.7	0.6	1.1
	P	97.2	98.4	94.0	96.5
average	R	3.8	1.0	1.3	2.1
	F1	7.3	2.1	2.6	4.0

the internal properties of the collocations in both Bas and Nat, where the collocates may not be direct translations of each other. Moreover, with all three strategies, the bilingual extractions including English are smaller than the Spanish-Portuguese ones.

Concerning the performance of the three approaches, Table 3 (Bas), Table 4 (Nat), and Table 5 (DiS) contain the precision, recall and f-score for each language pair and collocation pattern. Bas obtains high-precision results for every language and collocation pattern (91.7% in the worst scenario), with a macro-average value of 96.5%. These results are somehow expected due to the quality of the hand-crafted dictionaries. However, because of the poor recall numbers, the general performance of Bas is low, achieving F-scores around 4%. Interest-

Table 4: Precision, recall and f-score of the probabilistic (Nat) system (*average* is macro-average).

Pattern		ES-PT	ES-EN	PT-EN	average
	P	92.5	92.5	83.3	89.5
amod	R	9.5	7.4	6.5	7.8
	F1	17.2	13.8	12.0	14.3
	P	91.1	98.7	91.4	93.7
nmod	R	11.4	4.4	4.0	6.6
	F1	20.2	8.3	7.6	12.1
	P	95.2	80.0	92.7	89.3
vobj	R	7.8	7.5	9.2	8.2
	F1	14.3	13.8	16.8	15.0
	P	82.4	78.6	79.2	80.0
vsubj	R	8.0	10.2	9.6	9.3
	F1	14.6	18.1	17.1	16.6
	P	59.2	78.8	83.3	73.8
advmod	R	1.2	1.5	1.3	1.3
	F1	2.4	2.9	2.6	2.6
	P	84.1	85.7	86.0	85.3
average	R	7.6	6.2	6.1	6.6
	F1	13.8	11.4	11.2	12.1

ingly, the size of the dictionary does not seem crucial to the results of the baseline. In this respect, the Spanish-Portuguese results are much better, especially in terms of recall, than Spanish-English, whose dictionary is more than twice as large. Also, the Portuguese-English results are slightly better than the Spanish-Portuguese ones, the latter being obtained using a dictionary built by transitivity.

The use of probabilistic translation dictionaries (Nat) increases the recall by a factor of more than three when compared to the baseline, but with a cost in precision, which drops, in average, from 96.5% to 85.3%. However, these differences allow the Nat approach to obtain much better F-scores than Bas. When looking at the different collocation patterns, it is worth noting that while amod, nmod, and vobj have precision values of about 90%, vsubj, and especially adv-

Table 5: Precision, recall and f-score of DiS system (*average* is macro-average).

Pattern		ES-PT	ES-EN	PT-EN	average
amod	P	92.9	92.0	90.5	91.8
	R	67.8	51.6	49.5	56.3
	F1	78.4	64.3	64.0	68.9
nmod	P	93.8	88.0	90.0	90.6
	R	64.3	25.0	26.3	38.5
	F1	76.3	38.9	40.1	51.9
vobj	P	90.1	84.0	83.9	86.2
	R	66.0	48.1	49.9	54.7
	F1	76.5	61.2	62.6	66.7
vsubj	P	80.3	81.2	74.1	78.5
	R	51.6	41.0	36.1	42.9
	F1	62.8	54.5	48.6	55.3
advmod	P	77.6	83.3	67.4	76.1
	R	52.2	34.7	24.4	37.1
	F1	62.4	49.0	35.8	49.1
average	P	86.9	85.7	81.2	84.6
	R	60.4	40.1	37.3	45.9
	F1	71.3	53.6	50.2	58.4

mod (also with very low recall values) do not surpass 80% (with one case, ES-PT, with < 60%). As it will be shown in §4.2, some preprocessing issues might be the source of the some errors of advmod extractions.

As for the DiS model, its precision is again lower than Bas and very similar to the Nat approach, with average results of 84.6%. However, the distributional strategy finds much more bilingual equivalents than the dictionaries, so recall values increase to an average of more than 45%. Again, vsubj and advmod show worse precision values than the other three patterns. Besides, the nmod extractions of the pairs including English have very low recall when compared to the other results. This might be due to not extracting nouns analyzed as compound (§3.1). As for the other two methods, the DiS Spanish-Portuguese results are better than the two other language pairs, so the linguistic distance seems to play an important role in bilingual collocation extraction.

The method proposed in this chapter assigns a confidence value (obtained from the cosine similarity between the vectors of the base and the collocate equivalents) to each bilingual pair of collocations. In this respect, Figure 1 plots the average performance and confidence curves versus the total number of extracted pairs. This figure shows that by using a high confidence value (> 90%), it is possible to extract about 40, 000 bilingual pairs with a high degree of precision. Besides, filtering the extraction with confidence values higher than 90% does not increase the precision of the system. This suggests that the errors produced in the most confident pairs arise due to factors other than semantic similarity, such as different degrees of compositionality.

However, as the confidence value decreases, the precision of the extraction also gets worse, despite the rise in the number of extractions which involves higher recall and consequently better f-score.

Finally, all the bilingual collocations extracted by DɪS were merged into a single list with the three languages, thus obtaining new bilingual equivalents (not extracted directly by the system) by transitivity.[14] This final multilingual resource has 74, 942 entries, 38, 629 of them with translations in all three languages.

4.2 Error analysis

The manually annotated lists of bilingual collocations were used to perform an error analysis of the DɪS system. These errors were classified in five types depending on their origin. Table 6 contains, for each error type, the macro-average rates of each collocation pattern as well as the final distribution of the error typology.

1. **Bilingual model (*BiModel*)**: Though useful, the bilingual word embedding approach produces some errors such as the identification of antonyms that have a similar distribution, which can align opposite collocation equivalents, such as the Portuguese-English pair $tecido_{base}$ $vivo_{collocate}$ = $tissue_{base}$ $dead_{collocate}$, instead of *living tissue*, where the extracted equivalent of the collocate *vivo* (‘living’ – in this context – or ‘alive’, in Portuguese) was *dead*. In most cases, however, the system obtained similar (but not synonymous) collocations, such as $chá_{base}$ $preto_{collocate}$ ‘black tea’ in Portuguese aligned to $coffee_{base}$ $black_{collocate}$ ‘black coffee’ in English.

[14] The merging process obtained 6, 969 new bilingual collocation equivalents not present in the original extractions, and it also includes more than one translation for some collocations.

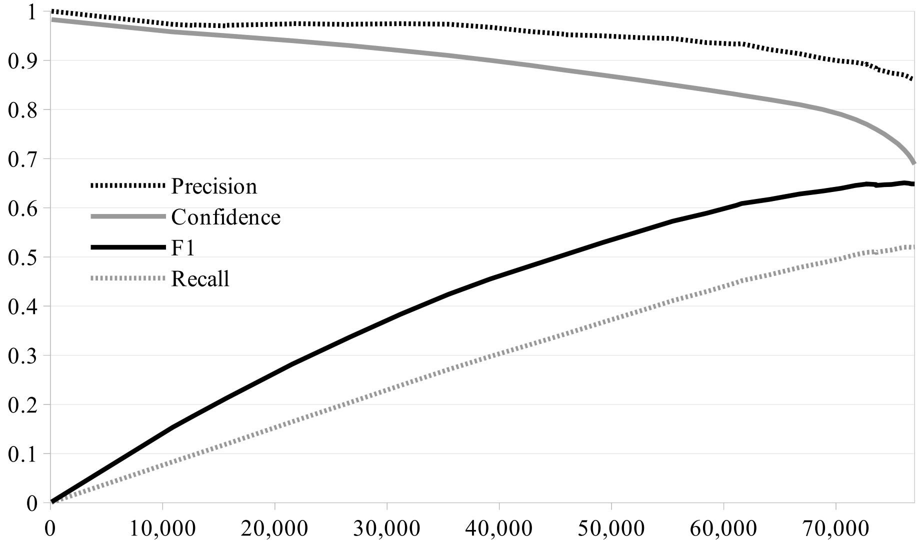


Figure 1: Average precision, recall, f-score, and confidence curves (from 0 to 1) versus total number of extractions of the DɪS model.

2. **Monolingual extraction (*MonoExtract*)**: The extraction of base and collocate pairs produced incorrect collocations such as *plan*$_{base}$ *figure*$_{collocate}$, instead of obtaining the phrasal verb *figure out* as collocate.

3. **Preprocessing (*NLP*)**: Several errors derived from issues produced by the NLP pipeline, such as POS-tagging or dependency parsing: e.g., *pain*$_{\mathrm{Noun}}$, *end*$_{\mathrm{Verb}}$ was labeled as dobj (instead of nsubj). A special case of preprocessing errors was the analysis of some Portuguese and Spanish adverbs ending in −*mente* (*-ly* adverbs in English), whose suffix was wrongly removed during the extraction process: e.g. *brutalmente* 'brutally' ⟶ *brutal*. These issues – which can be easily corrected – caused the alignment of incorrect Spanish and Portuguese collocations with English candidates, such as the Portuguese-English pair *matar*$_{base}$ *brutal*$_{collocate}$ = *kill*$_{base}$ *brutally*$_{collocate}$ instead of *matar*$_{base}$ *brutalmente*$_{collocate}$ = *kill*$_{base}$ *brutally*$_{collocate}$. This was the main source of errors of the advmod relation.

4. **Lemmatization and gender (*Gender*)**: The lemmatization of some words differs from language to language, so working with lemmas instead of tokens also might involve some errors. For instance, the Spanish word *hija* 'daughter' is lemmatized as *hijo* 'son' (also in Portuguese: *filha*, *filho*),

while in English *son* and *daughter* appear as different entries. Thus, some bilingual collocations differ in the gender of their bases, such as the Spanish-English pair *hijo*$_{base}$ *encantador*$_{collocate}$ = *daughter*$_{base}$ *lovely*$_{collocate}$ instead of *hijo*$_{base}$ *encantador*$_{collocate}$=*son*$_{base}$ *lovely*$_{collocate}$.

5. **Other errors (*Other*)**: Some other errors were caused by mixed languages in the original corpus. For example, the verb form *are*, in English, was analyzed as a form of the verb *arar* 'to plow' in Spanish. Some errors also arose from noise and misspellings in the corpora (proper nouns with lowercase letters, etc.).

It is worth mentioning that, in general, the error type distribution was similar across the different collocation patterns, showing much higher variation between different patterns of the same language pair. For instance, the distribution of Spanish-English AMOD errors is similar to the Portuguese-English AMOD one, while the typology of the Spanish-Portuguese NMOD errors is different to those of Spanish-Portuguese AMOD equivalents.

Table 6: Error rate of each of the defined error types of DIS system (*average* is macro-average).

Type	AMOD	NMOD	VOBJ	VSUBJ	ADVMOD	*average*
BiModel	70.57	93.52	59.23	45.74	32.61	60.33
MonoExtract	0	0	21.43	21.85	44.94	17.64
NLP	8.34	0	16.96	11.48	20.49	11.45
Gender	21.10	2.78	2.38	19.07	0	9.07
Other	0	3.70	0	1.85	1.96	1.50

Among the different errors produced by the presented method, an interesting case are *incongruent* collocations (Nesselhauf 2003). These expressions are those where the translation of both elements is not coherent, such as the English-Portuguese pair *requirement*$_{base}$ *meet*$_{collocate}$ = *condição* $_{base}$ *cumprir*$_{collocate}$, in which the verb *to meet* is usually translated into Portuguese as *conhecer*, not as *cumprir*. For these collocation equivalents to be correctly extracted by our method, they should appear with some frequency in the training corpus, which is not always the case. This fact may lead us to explore new compositional models, aimed at learning the distribution of the whole collocation, and not of its constituents, in further work.

4.3 Comparable corpora

A final experiment was carried out in order to find out (i) whether the bilingual word embeddings – trained on the same parallel corpora as those used for extracting the collocations – could be successfully applied to align collocations obtained from different resources, and (ii) the performance of the proposed method on comparable corpora.

Therefore, the same strategy for monolingual collocation extraction was applied in the Spanish and Portuguese *Wikipedia Comparable Corpus 2014*.[15] Then, we calculated the semantic similarity between the collocations using the same word embedding models as in the previous experiments.

From these corpora, filtered lists of 89, 285 and 140, 900 candidate collocations in Portuguese and Spanish were obtained, from 140M, and 80M of tokens respectively. From the 59, 507 bilingual collocations obtained by the DIS approach, 150 Spanish-Portuguese pairs were randomly selected and evaluated.

The precision of the extraction was 87.25%, with a recall of 58.15% (again computed using the whole set of monolingual collocations), and 69.79% f-score. These results are in line with those obtained on the OpenSubtitles Spanish-Portuguese pair (about 2% lower), so the method works well on different corpora and domains. It is worth noting that 49, 259 of the extracted collocation equivalents (83%) had not been retrieved from the OpenSubtitles corpus.

This last experiment shows that (i) the bilingual word embeddings can be used to identify collocation equivalents in different corpora than those used for training, and that (ii) they can also be applied to corpora of different domains to obtain previously unseen multilingual collocations.

5 Conclusions

This chapter presents a new strategy to automatically discover multilingual collocation equivalents from both parallel and comparable corpora. First, monolingual collocation candidates of five different patterns are extracted using syntactic analysis provided by harmonized UD annotation, together with a combination of standard association measures. Besides, bilingual word embeddings are trained on lemmatized parallel corpora. These bilingual models are then used to find distributional equivalents of both the base and the collocate of each source collocation in the target language.

The performed experiments, using noisy parallel corpora in three languages, showed that the proposed method achieves an average precision of about 85%,

[15] http://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

with reasonable recall values. A systematic comparison to translation dictionaries pointed out that the distributional approach achieves similar precision results with much higher recall values than the probabilistic dictionaries. Furthermore, the evaluation showed that setting up a confidence value as a threshold is useful for retaining only high-quality bilingual equivalents, which could benefit the work on multilingual lexicography.

Finally, preliminary tests using comparable corpora suggested that the bilingual word embeddings can be efficiently applied to different corpora than those used for training, discovering new bilingual collocations not present in the original resources.

The multilingual resources generated by the proposed method can be used in several scenarios in which MWEs play an important role, such as machine translation or second language learning. In this respect, corpora from various registers and linguistic varieties could be used in order to obtain a wider diversity of collocation equivalents that can be useful for different purposes.

The work presented in this chapter enables us to propose a number of directions for further work. First, the results of the error analysis should be taken into account in order to reduce both the issues produced by the NLP pipeline, and those which arise from the word embedding models. On the one hand, understanding collocations as directional combinations may lead us to evaluate other association measures which are not symmetrical, e.g., *Delta-P*. On the other hand, it could be interesting to evaluate other approaches for the alignment of bilingual collocations which make use of better compositionality models, and which effectively learn the semantic distribution of collocations as single units, in order to deal with cases of incongruent collocation equivalents.

Abbreviations

EM	expectation maximization	NLP	natural language processing
EN	English	PT	Portuguese
MI	mutual information	ES	Spanish
MWE	multiword expression	UD	Universal Dependencies

Acknowledgements

This work has been supported by the Spanish Ministry of Economy, Industry and Competitiveness through the project with reference FFI2016-78299-P, by two *Juan de la Cierva* grants (FJCI-2014-22853 and IJCI-2016-29598), and by a 2017 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation.

References

Alonso-Ramos, Margarita, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vázquez Veiga, Estela Mosqueira Suárez & Sabela Prieto González. 2010. Towards a motivated annotation schema of collocation errors in learner corpora. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the 7th Conference on International Language Resources and Evaluation* (LREC 2010), 3209–3214. European Language Resources Association (ELRA).

Altenberg, Bengt & Sylviane Granger. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22(2). 173–195.

Berard, Alexandre, Christophe Servan, Olivier Pietquin & Laurent Besacier. 2016. MultiVec: A multilingual and multilevel representation learning toolkit for NLP. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), 4188–4192. European Language Resources Association (ELRA).

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

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, vol. 2, 1212–1248. Berlin: Mouton de Gruyter.

Evert, Stefan & Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics*, vol. 2 (EACL 2003), 83–86.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation* 25(2). 127–144.

Fung, Pascale. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas. Machine Translation and the Information Soup* (AMTA 1998), 1–17. Springer.

Garcia, Marcos & Pablo Gamallo. 2015. Yet another suite of multilingual NLP tools. In Sierra-Rodríguez, José-Luis and Leal, José Paulo and Simões, Alberto (ed.), *Languages, applications and technologies. Communications in computer and information science* (International Symposium on Languages, Applications and Technologies (SLATE 2015)), 65–75. Springer.

Garcia, Marcos, Marcos García-Salido & Margarita Alonso-Ramos. 2017. Using bilingual word-embeddings for multilingual collocation extraction. In *Proceedings of the 13th Workshop on Multiword Expressions* (MWE '17), 21–30. Association for Computational Linguistics.

Garcia, Marcos, Marcos García-Salido & Margarita Alonso-Ramos. 2018. Discovering bilingual collocations in parallel corpora: A first attempt at using distributional semantics. In Irene Doval & María Teresa Sánchez Nieto (eds.), *Parallel corpora for contrastive and translation studies: New resources and applications.* John Benjamins Publishing Company.

Haruno, Masahiko, Satoru Ikehara & Takefumi Yamazaki. 1996. Learning bilingual collocations by word-level sorting. In *Proceedings of the 16th Conference on Computational Linguistics*, vol. 1 (COLING 1996), 525–530. Association for Computational Linguistics.

Hiemstra, Djoerd. 1998. Multilingual domain modeling in twenty-one. Automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Computational linguistics in the Netherlands 1997: Selected papers from the eighth clin meeting*, 41–57.

Kilgarriff, Adam. 2006. Collocationality (and how to measure it). In Elisa Corino, Carla Marello & Cristina Onesti (eds.), *Proceedings of the 12th EURALEX international congress*, vol. 2, 997–1004.

Krenn, Brigitte & Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *Proceedings of the ACL Workshop on Collocations*, 39–46. Association for Computational Linguistics.

Kupiec, Julian. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on association for computational linguistics* (ACL 1993), 17–22. Association for Computational Linguistics.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (ACL 1999), 317–324.

Lison, Pierre & Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Pro-*

ceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), 923–929. European Language Resources Association (ELRA).

Lü, Yajuan & Ming Zhou. 2004. Collocation translation acquisition using monolingual corpora. In *Proceedings of the 42nd annual meeting of the association for computational linguistics* (ACL 2004), 167–174. Association for Computational Linguistics.

Luong, Minh-Thang, Hieu Pham & Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the First Workshop on Vector Space Modeling for Natural Language Processing* (VSM-NLP) at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT 2015), 151–159. Association for Computational Linguistics.

Mel'čuk, Igor A. 1998. Collocations and lexical functions. In Anthony Paul Cowie (ed.), *Phraseology. Theory, analysis and applications*, 23–53. Oxford: Clarendon Press.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013.* arXiv preprint arXiv:1301.3781.

Nesselhauf, Nadja. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics* 24(2). 223–242.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), 1659–1666. European Language Resources Association (ELRA). 23-28 May, 2016.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(02). 95–135.

Orliac, Brigitte & Mike Dillinger. 2003. Collocation extraction for Machine Translation. In *Proceedings of ninth machine translation summit* (MT Summit IX), 292–298.

Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1-2). 137–158.

Rivera, Oscar Mendoza, Ruslan Mitkov & Gloria Corpas Pastor. 2013. A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, 18–25.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Seretan, Violeta. 2011. *Syntax-based collocation extraction* (Text, Speech and Language Technology). Dordrecht, Heidelberg, London, New York: Springer.

Seretan, Violeta & Eric Wehrli. 2007. Collocation translation based on sentence alignment and parsing. In *Actes de la 14e conference sur le traitement automatique des langues naturelles* (TALN 2007), 401–410. IRIT Press.

Simões, Alberto Manuel & José João Almeida. 2003. NATools – a statistical word aligner workbench. *Procesamiento del lenguaje natural* 31. 217–224.

Smadja, Frank. 1992. How to compile a bilingual collocational lexicon automatically. In *Proceedings of the AAAI Workshop on Statistically-Based NLP Techniques*, 57–63. AAAI Press.

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

Smadja, Frank, Kathleen R. McKeown & Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22(1). 1–38.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language* 2(1). 23–55.

Wanner, Leo, Bernd Bohnet & Mark Giereth. 2006. Making sense of collocations. *Computer Speech & Language* 20(4). 609–624.

Wanner, Leo, Gabriela Ferraro & Pol Moreno. 2016. Towards distributional semant-ics-based classification of collocations for collocation dictionaries. *International Journal of Lexicography* 30. 167–186.

Wu, Chien-Cheng & Jason S. Chang. 2003. Bilingual collocation extraction based on syntactic and statistical analyses. In *Proceedings of the 15th Conference on Computational Linguistics and Speech Processing*, 1–20. Association for Computational Linguistics and Chinese Language Processing.

Chapter 13

Exploiting multilingual lexical resources to predict MWE compositionality

Bahar Salehi
The University of Melbourne

Paul Cook
University of New Brunswick

Timothy Baldwin
The University of Melbourne

Semantic idiomaticity is the extent to which the meaning of a multiword expression (MWE) cannot be predicted from the meanings of its component words. Much work in natural language processing on semantic idiomaticity has focused on compositionality prediction, wherein a binary or continuous-valued compositionality score is predicted for an MWE as a whole, or its individual component words. One source of information for making compositionality predictions is the translation of an MWE into other languages. This chapter extends two previously-presented studies – Salehi & Cook (2013) and Salehi et al. (2014) – that propose methods for predicting compositionality that exploit translation information provided by multilingual lexical resources, and that are applicable to many kinds of MWEs in a wide range of languages. These methods make use of distributional similarity of an MWE and its component words under translation into many languages, as well as string similarity measures applied to definitions of translations of an MWE and its component words. We evaluate these methods over English noun compounds, English verb-particle constructions, and German noun compounds. We show that the estimation of compositionality is improved when using translations into multiple languages, as compared to simply using distributional similarity in the source language. We further find that string similarity complements distributional similarity.

Bahar Salehi, Paul Cook & Timothy Baldwin. 2018. Exploiting multilingual lexical resources to predict MWE compositionality. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 343–373. Berlin: Language Science Press. DOI:10.5281/zenodo.1469573

1 Compositionality of MWEs

Multiword expressions (hereafter MWEs) are combinations of words which are lexically, syntactically, semantically or statistically idiosyncratic (Sag et al. 2002; Baldwin & Kim 2010). Much research has been carried out on the extraction and identification of MWEs[1] in English (Schone & Jurafsky 2001; Pecina 2008; Fazly et al. 2009) and other languages (Dias 2003; Evert & Krenn 2005; Salehi et al. 2012). However, considerably less work has addressed the task of predicting the meaning of MWEs, especially in non-English languages. As a step in this direction, the focus of this study is on predicting the compositionality of MWEs.

An MWE is fully compositional if its meaning is predictable from its component words, and it is non-compositional (or idiomatic) if not. For example, *stand up* "rise to one's feet" is compositional, because its meaning is clear from the meaning of the components *stand* and *up*. However, the meaning of *strike up* "to start playing" is largely unpredictable from the component words *strike* and *up*.

In this study, following McCarthy et al. (2003) and Reddy et al. (2011), we consider compositionality to be graded, and aim to predict the *degree* of compositionality. For example, in the dataset of Reddy et al. (2011), *climate change* is judged to be 99% compositional, while *silver screen* is 48% compositional and *ivory tower* is 9% compositional. Formally, we model compositionality prediction as a regression task.

An explicit handling of MWEs has been shown to be useful in NLP applications (Ramisch 2012). As an example, Carpuat & Diab (2010) proposed two strategies for integrating MWEs into statistical machine translation. They show that even a large scale bilingual corpus cannot capture all the necessary information to translate MWEs, and that in adding the facility to model the compositionality of MWEs into their system, they could improve translation quality. Acosta et al. (2011) showed that treating non-compositional MWEs as a single unit in information retrieval improves retrieval effectiveness. For example, while searching for documents related to *ivory tower*, we are almost certainly not interested in documents relating to elephant tusks.

Our approach is to use a large-scale multi-way translation lexicon to source translations of a given MWE and each of its component words, and then model the semantic similarity between each component word and the MWE.[2] We consider similarity measures based on distributional similarity from monolingual

[1] In this chapter, we follow Baldwin & Kim (2010) in considering MWE "identification" to be a token-level disambiguation task, and MWE "extraction" to be a type-level lexicon induction task.

[2] Note that we will always assume that there are two component words, but the method is easily generalisable to MWEs with more than two components.

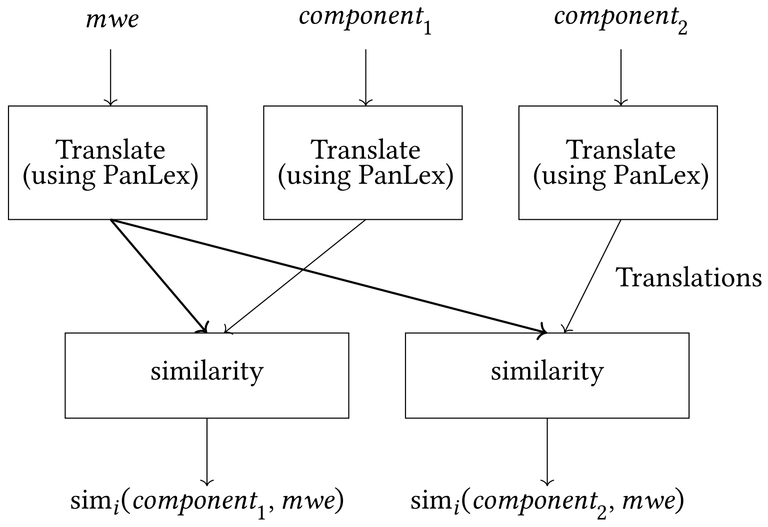


Figure 1: Outline of our approach to computing the similarity of translations of an MWE with each of its component words, for a given target language. sim$_i$ is the similarity between the first or second component of the MWE, and the MWE itself, based on either string or distributional similarity, as measured using language *i*.

corpora for the source language and each of the target languages, as well as string similarity measures applied to definitions of translations of an MWE and its component words as shown in Figure 1. We then consider a variety of approaches to combining similarity scores from the various languages to produce a final compositionality score for the source language expression, as illustrated in Figure 2. We hypothesise that by using multiple translations we will be able to better predict compositionality, and that string similarity measures will complement distributional similarity. Our results confirm our hypotheses, and we further achieve state-of-the-art results over two compositionality prediction datasets.

This chapter combines two previous works – Salehi & Cook (2013) and Salehi et al. (2014) – and extends them in the following ways:

- two new string similarity measures in §4.1.1;
- updated results in §4.2 for the method of Salehi & Cook (2013) such that they are now comparable with the results of the method of Salehi et al. (2014) in §6 – previously these results were not comparable because they used different cross-validation folds during evaluation;
- new results for a dataset of German noun compounds based on the string similarity methods in §4.2;

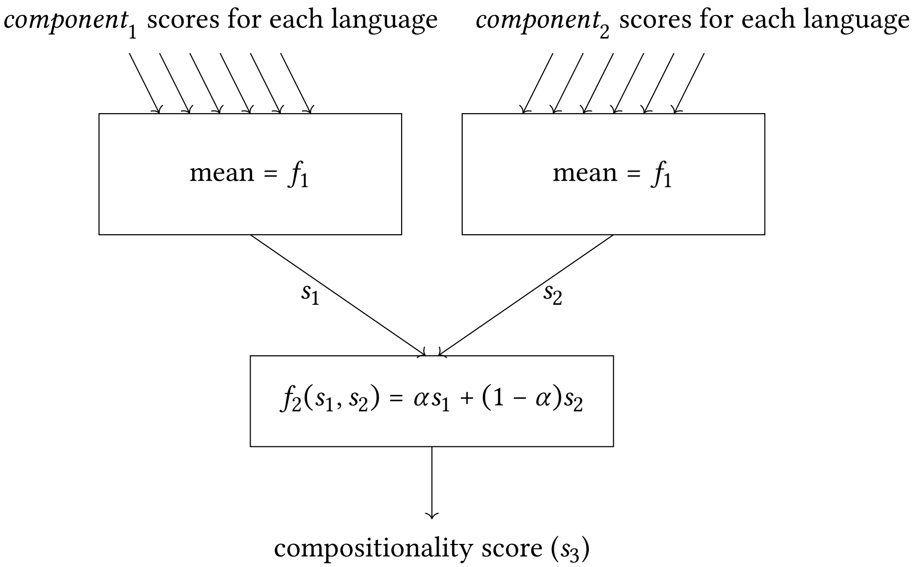

$component_1$ scores for each language
$component_2$ scores for each language
mean = f_1
mean = f_1
s_1
s_2
$f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha)s_2$
compositionality score (s_3)

Figure 2: Outline of the method for combining similarity scores from multiple languages, across the components of the MWE.

- additional error analysis in §4.2.1 for English verb-particle constructions;
- two new translation-based similarity approaches, and results for these methods, in §4.2.2;
- experiments considering an alternative translation dictionary in §5;
- analysis of the impact of window size on the distributional similarity approach in §6.1.1.

2 Related work

Most recent work on predicting the compositionality of MWEs can be divided into two categories: language/construction-specific and general-purpose. This can be at either the token-level (over token occurrences of an MWE in a corpus) or type-level (over the MWE string, independent of usage). The bulk of work on compositionality has been language/construction-specific and operated at the token-level, using dedicated methods to identify instances of a given MWE, and specific properties of the MWE in that language to predict compositionality (Lin 1999; Kim & Baldwin 2007; Fazly et al. 2009).

General-purpose token-level approaches such as distributional similarity have been commonly applied to infer the semantics of a word/MWE (Schone & Jurafsky 2001; Baldwin et al. 2003; Reddy et al. 2011). These techniques are based on the assumption that the meaning of a word is predictable from its context of use, via the neighbouring words of token-level occurrences of the MWE. In order to predict the compositionality of a given MWE using distributional similarity, the different contexts of the MWE are compared with the contexts of its components, and the MWE is considered to be compositional if the MWE and component words occur in similar contexts.

Identifying token instances of MWEs is not always easy, especially when the component words do not occur sequentially. For example, consider *put on* in ***put** your jacket **on***, and ***put** your jacket **on** the chair*. In the first example *put on* is an MWE, while in the second example, *put on* is a simple verb with prepositional phrase and not an instance of an MWE. Moreover, if we adopt a conservative identification method, the number of token occurrences will be limited and the distributional scores may not be reliable. Additionally, for morphologically-rich languages, it can be difficult to predict the different word forms a given MWE type will occur across, posing a challenge for our requirement of no language-specific preprocessing.

Pichotta & DeNero (2013) proposed a token-based method for identifying English phrasal verbs based on parallel corpora for 50 languages. They show that they can identify phrasal verbs better when they combine information from multiple languages, in addition to the information they get from a monolingual corpus. This finding lends weight to our hypothesis that using translation data and distributional similarity from each of a range of target languages, can improve compositionality prediction. Having said that, the general applicability of their method is questionable – there are many parallel corpora involving English, but for other languages, this tends not to be the case.

In the literature, compositionality has been viewed as either compositionality of the whole MWE as one unit (McCarthy et al. 2003; Venkatapathy & Joshi 2005; Katz 2006; Biemann & Giesbrecht 2011; Farahmand et al. 2015), or compositionality relative to each component (Reddy et al. 2011; Hermann et al. 2012; Schulte im Walde et al. 2013). There have also been studies which focus only on one component of the MWE. For example, Korkontzelos & Manandhar (2009) induce the most probable sense of an MWE first, and then measure the semantic similarity between the MWE and its semantic head. This approach of considering only the head component has been shown to be quite accurate for English verb-particle constructions (Bannard et al. 2003). However, this might not always be the case. For example, as shown in Reddy et al. (2011), the compositionality of the first noun (the modifier) has more impact than the second noun (the head) for English noun compounds.

Elsewhere, a lot of work has been done on specific types of MWE in specific languages. In English, studies have been done specifically on VPCs (McCarthy et al. 2003; Bannard et al. 2003), verb+noun MWEs (Venkatapathy & Joshi 2005; McCarthy et al. 2007; Fazly et al. 2009), noun compounds (Reddy et al. 2011), and adjective+noun compounds (Vecchi et al. 2011). There have also been studies focusing on a specific language other than English, such as Arabic (Saif et al. 2013) and German (Schulte im Walde et al. 2013). This chapter investigates language independent approaches applicable to any type of MWE in any language.

3 Resources

In this section, we describe the datasets used to evaluate our method and the multilingual dictionary it requires. These are the same resources as used by Salehi & Cook (2013) and Salehi et al. (2014).

3.1 Datasets

We evaluate our proposed method over three datasets (two English, one German), as described below.

3.1.1 English noun compounds (ENC)

Our first dataset is made up of 90 binary English noun compounds, from the work of Reddy et al. (2011). Each noun compound was annotated by multiple annotators using the integer scale 0 (fully non-compositional) to 5 (fully compositional). A final compositionality score was then calculated as the mean of the scores from the annotators. If we simplistically consider 2.5 as the threshold for compositionality, the dataset is relatively well balanced, containing 48% compositional and 52% non-compositional noun compounds.

Spearman correlation was used to get an estimate of inter-annotator agreement. The average correlation for compound compositionality was $\rho = 0.522$. This score was slightly higher for the compositionality of components ($\rho = 0.570$ for the first component and $\rho = 0.616$ for the second component).

3.1.2 English verb-particle constructions (EVPC)

The second dataset contains 160 English verb-particle constructions (VPCs), from the work of Bannard (2006). In this dataset, a verb-particle construction consists of a verb (the head) and a prepositional particle (e.g. *hand in*, *look up* or *battle on*). For each component word (the verb and particle, respectively), multiple annotators were asked whether the VPC entails the component word. In order to translate the dataset into a regression task, we calculate the overall compositionality as the number of annotations of entailment for the verb, divided by the total number of verb annotations for that VPC. That is, following Bannard et al. (2003), we only consider the compositionality of the verb component in our experiments. The Kappa score between the multiple annotators is 0.372 for verb and 0.352 for the particle component.

3.1.3 German noun compounds (GNC)

Our final dataset is made up of 246 German noun compounds (von der Heide & Borgwaldt 2009; Schulte im Walde et al. 2013). Multiple annotators were asked to rate the compositionality of each German noun compound on an integer scale of 1 (non-compositional) to 7 (compositional). The overall compositionality score is then calculated as the mean across the annotators. Note that the component words are provided as part of the dataset, and that there is no need to perform decompounding. This dataset is significant as it is non-English and because of the fact that German has relatively rich morphology, which we expect to impact on the identification of both the MWE and the component words.

3.2 Multilingual dictionary

To translate the MWEs and their components, we use PanLex (Baldwin et al. 2010). This online dictionary is massively multilingual, covering more than 1353 languages. The translations are sourced from handmade electronic dictionaries. It contains lemmatised words and MWEs in a large variety of languages, with lemma-based (and less frequently sense-based) links between them.

For each MWE dataset (see §3.1), we translate each MWE, and its component words, from the source language into many target languages. These translations will be used in §4 and §6. In instances where there is no direct translation in a given language for a term, we use a pivot language to find translation(s) in the target language. For example, the English noun compound *silver screen* has direct translations in only 13 languages in PanLex, including Vietnamese (*màn bac*) but not French. There is, however, a translation of *màn bac* into French (*cinéma*), allowing us to infer an indirect translation between *silver screen* and *cinéma*. In this way, if there are no direct translations into a particular target language, we search for a single-pivot translation via each of our other target languages, and combine them all together as our set of translations for the target language of interest.

4 String similarity

In this section we present our string similarity-based method for predicting compositionality, followed by experimental results using this method. This section extends Salehi & Cook (2013) as described in §1.

4.1 Compositionality prediction based on string similarity

We hypothesize that compositional MWEs are more likely to be word-for-word translations in a given language than non-compositional MWEs. Hence, if we can locate the translations of the components in the translation of the MWE, we can deduce that it is compositional. As an example of our method, consider the English-to-Persian translation of *kick the bucket* as a non-compositional MWE and *make a decision* as a semi-compositional MWE (Table 1).[3] By locating the translation of *decision* (*tasmim*) in the translation of *make a decision* (*tasmim gereftan*), we can deduce that it is semi-compositional. However, we cannot locate any of the component translations in the translation of *kick the bucket*. Therefore, we conclude that it is non-compositional. Note that in this simple example, the match is word-level, but that due to the effects of morphophonology, the more likely situation is that the components don't match exactly (as we observe in the case of *khadamaat* and *khedmat* for the *public service* example), which motivates our use of string similarity measures which can capture partial matches.

4.1.1 String similarity measures

We consider the following string similarity measures to compare the translations. In each case, we normalize the output value to the range $[0, 1]$, where 1 indicates identical strings and 0 indicates completely different strings. We will indicate the translation of the MWE in a particular language t as mwe^t, and the translation of a given component in language t as $component^t$.

[3] Note that the Persian words are transliterated into English for ease of understanding.

Table 1: English MWEs and their components with their translation in Persian. Direct matches between the translation of an MWE and its components are shown in **bold**; partial matches are shown in *italics*.

English	Persian translation
kick the bucket	mord
kick	zad
the	–
bucket	satl
make a decision	**tasmim** gereft
make	sakht
a	yek
decision	**tasmim**
public service	*khadamaat* **omumi**
public	**omumi**
service	*khedmat*

Longest common substring (LCS): The LCS measure finds the longest common substring between two strings. For example, the LCS between ABABC and BABCAB is BABC. We calculate a normalized similarity value based on the length of the LCS as follows:

$$\frac{\mathrm{LCS}(mwe^t, component^t)}{\min(\mathrm{len}(mwe^t), \mathrm{len}(component^t))} \tag{13.1}$$

Levenshtein (LEV1): The Levenshtein distance calculates the number of basic edit operations required to transform one word into the other. Edits consist of single-letter insertions, deletions or substitutions. We normalize LEV1 as follows:

$$1 - \frac{\mathrm{LEV1}(mwe^t, component^t)}{\max(\mathrm{len}(mwe^t), \mathrm{len}(component^t))} \tag{13.2}$$

Levenshtein with substitution penalty (LEV2): One well-documented feature of Levenshtein distance (Baldwin 2009) is that substitutions are in fact the combination of an addition and a deletion, and as such can be considered to be two edits. Based on this observation, we experiment with a variant of LEV1 with this penalty applied for substitutions. Similarly to LEV1, we normalize as follows:

$$1 - \frac{\mathrm{LEV2}(mwe^t, component^t)}{\mathrm{len}(mwe^t) + \mathrm{len}(component^t)} \tag{13.3}$$

Smith Waterman (SW): This method is based on the Needleman-Wunsch algorithm,[4] and was developed to locally-align two protein sequences (Smith & Waterman 1981). It finds the optimal similar regions by maximizing the number of matches and minimizing the number of gaps necessary to align the two sequences. For example, the optimal local sequence for the two sequences below is `AT--ATCC`, in which "–" indicates a gap:

> Seq1: `**ATGCATCC**CATGAC`
> Seq2: `TCT**ATATCC**GT`

As the example shows, it looks for the longest common string but has a built-in mechanism for including gaps in the alignment (with penalty). This characteristic of SW might be helpful in our task, because there may be morphophonological variations between the MWE and component translations (as seen above in the *public service* example). We normalize SW similarly to LCS:

$$\frac{\mathrm{len(alignedSequence)}}{\min(\mathrm{len}(mwe^t), \mathrm{len}(component^t))} \tag{13.4}$$

The aligned sequence is the combination of the common characters in the optimal local sequence we found using SW. In the above example, the aligned sequence is `ATATCC`.

Jaccard and Dice similarity: For further analysis, we experiment with Jaccard and Dice similarity, which are well-known for measuring the similarity between two sentences or bodies of text (Gomaa & Fahmy 2013). Both methods view the texts as sets of words, with similarity based on the size of the intersection between the sets, but differ in the way they are normalized. In our case, we expect relatively low overlap at the word level due to morphophonology, and therefore

[4] The Needleman-Wunsch (NW) algorithm was designed to align two sequences of amino-acids (Needleman & Wunsch 1970). The algorithm looks for the sequence alignment which maximizes the similarity. As with the LEV score, NW minimizes edit distance, but also takes into account character-to-character similarity based on the relative distance between characters on the keyboard. We exclude this score because it is highly similar to the LEV scores and we did not obtain encouraging results using NW in our preliminary experiments.

calculate Jaccard (J) and Dice (D) at the character- instead of word-level as follows:

$$J = \frac{|mwe^t \cap component^t|}{|mwe^t| + |component^t| - |mwe^t \cap component^t|} \tag{13.5}$$

$$D = \frac{2 * |mwe^t \cap component^t|}{|component^t| + |mwe^t|} \tag{13.6}$$

4.1.2 Calculating compositionality

Given the string similarity scores calculated between the translations for a given component word and the MWE, we need some way of combining scores across component words. First, we measure the compositionality of each component within the MWE (s_1 and s_2):

$$s_1 = f_1(\text{sim}_1(w_1, mwe), ..., \text{sim}_i(w_1, mwe)) \tag{13.7}$$

$$s_2 = f_1(\text{sim}_1(w_2, mwe), ..., \text{sim}_i(w_2, mwe)) \tag{13.8}$$

where sim is a similarity measure, sim_i indicates that the calculation is based on translations in language i, and f_1 is a score combination function.

Then, we compute the overall compositionality of the MWE (s_3) from s_1 and s_2 using f_2:

$$s_3 = f_2(s_1, s_2) \tag{13.9}$$

Since we often have multiple translations for a given component word/MWE in PanLex, we exhaustively compute the similarity between each MWE translation and component translation, and use the highest similarity as the result of sim_i. If an instance does not have a direct/indirect translation in PanLex, we assign a default value, which is the mean of the highest and lowest annotation score for the dataset under consideration. Note that word order is not an issue in our method, as we calculate the similarity independently for each MWE component.

We consider simple functions for f_1 such as mean, median, product, minimum and maximum. f_2 was selected to be the same as f_1 in all situations, except when we use mean for f_1. Here, following Reddy et al. (2011), we experimented with weighted mean:

$$f_2(s_1, s_2) = \alpha s_1 + (1 - \alpha)s_2 \tag{13.10}$$

Based on 3-fold cross-validation, we chose $\alpha = 0.7$ for ENC.[5] We found $\alpha = 0.7$ is also optimal for GNC.

Since we do not have judgements for the compositionality of the full VPC in EVPC (we instead have separate judgements for the verb and particle), we cannot use f_2 for this dataset. Bannard et al. (2003) observed that nearly all of the verb-compositional instances were also annotated as particle-compositional by the annotators. In line with this observation, we use s_1 (based on the verb) as the compositionality score for the full VPC.

4.1.3 Language selection

Our method is based on the translation of an MWE into many languages. First, we chose 54 languages for which relatively large corpora were available.[6] The coverage, or the number of instances which have direct/indirect translations in PanLex, varies from one language to another. In preliminary experiments, we noticed that there is a high correlation (between roughly $r = 0.6$ and 0.8 across the three datasets) between the usefulness of a language and its translation coverage on MWEs. Therefore, we excluded languages with MWE translation coverage of less than 50%. Based on nested 10-fold cross-validation in our experiments, we select the 10 most useful languages for each cross-validation training partition, based on the Pearson correlation between the given scores in that language and human judgements.[7] The 10 best languages are selected based only on the training set for each fold. (The languages selected for each fold will later be used to predict the compositionality of the items in the testing portion for that fold.)

4.2 Results

As mentioned above, we perform nested 10-fold cross-validation to select the 10 best languages on the training data for each fold. The selected languages for a given fold are then used to compute s_1 and s_2 (and s_3 for NCs) for each instance

[5] We considered values of α from 0 to 1, incremented by 0.1.

[6] In §6 these corpora will be used to compute distributional similarity. Note that the string similarity methods of interest here do not rely on the availability of large corpora.

[7] Note that for VPCs, we calculate the compositionality of only the verb part, because we don't have the human judgements for the whole VPC.

Table 2: Correlation (r) on each dataset, for each string similarity measure. The best correlation for each dataset is shown in boldface.

Method	ENC	EVPC	GNC
SW	**0.644**	0.349	0.379
LCS	**0.644**	**0.385**	0.372
LEV1	0.502	0.328	0.318
LEV2	0.566	0.327	**0.389**
Jaccard	0.474	0.335	0.299
Dice	0.557	0.331	0.370
Unsupervised (family)	0.556	0.257	0.164
Unsupervised (coverage)	0.642	0.323	0.343

in the test set for that fold. The scores are compared with human judgements using Pearson's correlation.

We experimented with five functions for f_1, namely mean, median, product, maximum and minimum. Among these functions, mean performed consistently better than the others, and as such we only present results using mean in Table 2.

For ENC, LCS and SW perform best, while for EVPC, LCS performs best with SW being the next best measure. Both LCS and SW look for a sequence of similar characters, unlike LEV1 and LEV2, which are not affected by match contiguity. For GNC, LEV2, SW and LCS perform better than LEV1. However, unlike the other two datasets, LEV2 is the best performing method, and SW is slightly better than LCS.

For all datasets, Jaccard and Dice perform worse than SW and LCS. This shows that, despite being useful in measuring the similarity between sentences, these two measures do not perform well in this compositionality prediction task. The relatively poor performance of these measures could be because, unlike the other measures, Jaccard and Dice are calculated independently of the order of characters. Dice performs better than Jaccard for ENC and GNC, while Jaccard performs slightly better than Dice for EVPC.

The results support our hypothesis that using multiple target languages rather than one, results in a more accurate prediction of MWE compositionality. Our best result using the 10 selected languages on ENC is $r = 0.644$, as compared to the best single-language correlation of $r = 0.543$ for Portuguese. On EVPC, the best LCS result for the verb component is $r = 0.385$, as compared to the best single-language correlation of $r = 0.342$ for Lithuanian. For GNC, the best correlation of $r = 0.389$ is well above the highest correlation of a single language of roughly $r = 0.32$.

In §6 we will combine this string similarity approach with an approach based on distributional similarity, and compare it against a baseline and state-of-the-art approaches.

4.2.1 Error analysis

We analysed items in ENC which have a high absolute difference (more than 2.5) between the human annotation and our scores (using LCS and mean). The words are *cutting edge*, *melting pot*, *gold mine* and *ivory tower*, which are non-compositional according to ENC. After investigating their translations, we came to the conclusion that the first three MWEs have word-for-word translations in most languages. Hence, they disagree with our hypothesis that word-for-word translation is a strong indicator of compositionality. The word-for-word translations might be because of the fact that they have both compositional and non-compositional senses, or because they are calques (loan translations). However, we have tried to avoid such problems with calques by using translations into several languages.

For *ivory tower* ("a state of mind that is discussed as if it were a place")[8] we noticed that we have a direct translation into 13 languages. Other languages have indirect translations. By checking the direct translations, we noticed that, in French, the MWE is translated to *tour* and *tour d'ivoire*. A noisy (wrong) translation of *tour* "tower" resulted in wrong indirect translations for *ivory tower* and an inflated estimate of compositionality.

We repeat the same error analysis for the EVPC dataset. The items with a high difference between the human annotation and our scores are: *carry out*, *drop out*, *get in*, *carry away*, *wear down* and *turn on*. All of these items are annotated as non-compositional. These VPCs also have a compositional sense beside the non-compositional meaning. Also, as with the ENC dataset, we have problems of calques. For example, *drop out* when translated to German (*ausfallen*) includes the word *fallen*, which is one of the translations of *drop*.

4.2.2 Unsupervised approach

The proposed translation-based string similarity approach has been supervised so far, in that the best target languages are selected based on training data. In this

[8] This definition is from Wordnet 3.1.

section, we propose two unsupervised approaches in which: (1) only the target languages of the same language family as the source language are considered; and (2) only the 10 target languages with the highest translation coverage are considered.

Languages of the same family: We hypothesize that translations into target languages in the same language family as the source language might be particularly useful for compositionality prediction for MWEs in the source language. To test this hypothesis, we consider an unsupervised approach in which only languages in the same family as the source language are used when computing the compositionality scores.

In this unsupervised approach, LCS scores of the languages of the same family as the source language (here Germanic, for both the English and German datasets) are considered. The Germanic languages among our 54 languages are: English, German, Danish, Dutch, Icelandic, Luxembourgish, Norwegian and Swedish.

Results for this unsupervised approach are shown in Table 2 ("Unsupervised (family)"). This approach performs substantially worse than the corresponding supervised approach based on LCS, for each dataset. This drop in performance could be because almost none of the 10 best languages selected in the supervised approach are in the same language family as the source language. The shared languages between the supervised approach and this approach are Dutch and Norwegian for ENC, English for GNC. There is no shared language between the two approaches when using EVPC.

Languages with the highest translation coverage: In the proposed supervised setup, the best target languages are those whose scores have the highest correlation with gold-standard annotations. According to our experiments, we showed that there is a strong correlation between being a good language for this compositionality prediction task and its coverage in PanLex (in the range of roughly $0.6 < r < 0.8$ across the three datasets). In other words, the target languages to which most of the source language MWEs have a translation in PanLex, result in higher correlation for compositionality prediction.

We now consider an unsupervised approach, in which only the 10 target languages with the highest translation coverage are considered. The results of this unsupervised approach, again using LCS, are shown in Table 2 ("Unsupervised (coverage)"). According to the results, despite the lower correlation scores for the proposed unsupervised method, this method is comparable to the supervised

Table 3: The 10 languages with the highest translation coverage for ENC, EVPC and GNC. Languages also selected by the supervised approach are shown in **boldface**.

ENC	EVPC	GNC
German	German	**English**
Finnish	Finnish	Japanese
French	French	French
Italian	Italian	Italian
Russian	Japanese	Russian
Spanish	Hungarian	Hungarian
Portuguese	Dutch	Dutch
Japanese	**Polish**	Turkish
Chinese	Chinese	Chinese
Czech	**Czech**	**Czech**

approach. Therefore, in the case of not having a training set for a group of MWEs (no matter in what language or what type of MWE), we suggest using the target languages to which the majority of those MWEs have a translation.

The 10 languages with highest correlation for ENC, EVPC and GNC are shown in Table 3. There is some overlap between the list of languages with the highest coverage and the 10 best languages selected in our supervised approach, as shown in boldface for each dataset.

5 An alternative multilingual dictionary

In this section we consider the same string similarity-based approach to predicting compositionality as in §4.1, but using an alternative multilingual dictionary to PanLex, specifically dict.cc.[9]

dict.cc is a translation dictionary that provides translations for both English and German into 26 languages spoken in Europe. It is a crowd-sourced dictionary, with translations being contributed, and refined, by users. Due to the relatively small number of languages it covers, relying on dict.cc goes against our goals of developing compositionality prediction methods that are applicable to any language; we could not use dict.cc to predict the compositionality of, for example, a French MWE, because translations are not available for French into many languages (only English and German). Nevertheless, by considering the

[9] https://www.dict.cc/

Table 4: Correlation (r) on each dataset, for each string similarity measure, using dict.cc and PanLex as the translation dictionary. The best correlation for each dataset is shown in boldface.

Dictionary	Method	ENC	EVPC	GNC
dict.cc				
	SW	.269	.217	.514
	LCS	.251	.262	**.523**
	LEV1	.181	.161	.482
	LEV2	.163	.189	.474
	Jaccard	.158	.127	.442
	Dice	.230	.192	.420
PanLex				
	SW	**.559**	**.294**	.270
	LCS	.551	.276	.290
	LEV1	.388	.274	.276
	LEV2	.512	.281	.262
	Jaccard	.459	.241	.267
	Dice	.541	.235	.197

use of an alternative translation dictionary (which is applicable to the English and German datasets we use for evaluation) we can learn whether our approach to predicting compositionality implicitly relies on information particular to PanLex, or whether an alternative dictionary can be substituted in its place.

We chose target languages available in dict.cc that overlap with the set of 54 target languages used in experiments with PanLex in §4.1. This resulted in 22 target languages. We introduced this restriction, as opposed to using all languages available in dict.cc, to allow us to compare PanLex and dict.cc when using the exact same set of target languages.

Results for the string similarity-based approach to predicting compositionality, using dict.cc and PanLex, each with the same 22 target languages, are shown in Table 4. The 10 best languages are selected using the same method as in §4.1.3.

For each translation dictionary and dataset, the best method is always one of either SW or LCS, and in many cases these are the top two methods (with the exceptions being EVPC and GNC using PanLex). These methods were also found to perform well in §4.2 when using PanLex and 54 target languages. This

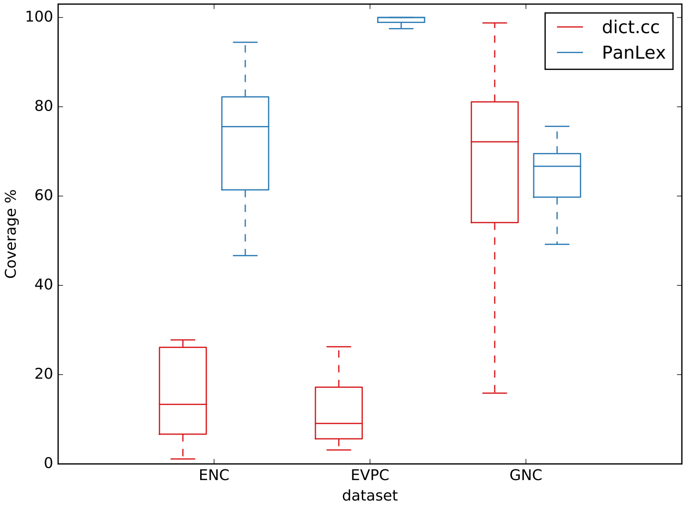


Figure 3: Boxplots showing the percentage of expressions in each dataset covered by dict.cc and PanLex, over the 22 target languages.

demonstrates that the methods are robust to the choice of specific translation dictionary, and when the number of target languages is substantially reduced.

There are, however, substantial differences between the results using different translation dictionaries. For any combination of dataset and method, the results using PanLex are always better than those using dict.cc for ENC and EVPC, while for GNC, the results using dict.cc are always better. To understand why this is the case, for each dataset and dictionary, and for each of the 22 target languages, we computed the proportion of expressions for which translations are available. Boxplots illustrating these findings are shown in Figure 3. On average across the target languages, many more expressions are covered by PanLex than dict.cc for ENC and EVPC, while for GNC the coverage is higher for dict.cc. For example, according to Figure 3, for EVPC the coverage for almost all of the 22 target languages is close to 100% in PanLex.

Because it in keeping with our goal of building methods for compositionality prediction that are applicable to any language, and because it gives the best results in two out of three cases for the datasets used for evaluation, we will only consider PanLex as the translation dictionary for the remainder of this chapter.

6 Distributional similarity

In this section we describe a method for predicting compositionality based on the same framework as in §4, but using distributional similarity instead of string similarity. This section extends Salehi et al. (2014) as described in §1.

6.1 Compositionality prediction based on distributional similarity

To predict the compositionality of a given MWE, we first measure the semantic similarity between the MWE and each of its component words using distributional similarity based on a monolingual corpus in the source language. We then repeat the process for translations of the MWE and its component words into each of a range of target languages, calculating distributional similarity using a monolingual corpus in the target language. We additionally use supervised learning to identify which target languages (or what weights for each language) optimise the prediction of compositionality. We hypothesise that by using multiple translations – rather than only information from the source language – we will be able to better predict compositionality. We further optionally combine our proposed approach with the LCS-based string similarity method from §4.

Below, we detail our method for calculating distributional similarity in a given language, the different methods for combining similarity scores into a single estimate of compositionality, and finally the method for selecting the target languages to use in calculating compositionality.

6.1.1 Calculating distributional similarity

We collected monolingual corpora for each of the 52 languages (51 target languages + 1 source language) from XML dumps of Wikipedia. These languages are based on the 54 target languages used in §4, excluding Spanish because we happened not to have a dump of Spanish Wikipedia, and also Chinese and Japanese because of the need for a language-specific word tokeniser. The raw corpora were preprocessed using the WP2TXT toolbox[10] to eliminate XML tags, HTML tags and hyperlinks, and then tokenisation based on whitespace and punctuation was performed. The corpora vary in size from roughly 750M tokens for English, to roughly 640K tokens for Marathi.

In order to be consistent across all languages and to be as language-independent as possible, we calculate distributional similarity in the following manner for a given language.

[10] http://wp2txt.rubyforge.org/

Table 5: Results of distributional similarities using 10 best languages on ENC dataset (N is window size)

Context window	Correlation (r)
Sentence	0.425
Window (N=3)	0.175
Window (N=3, with positional index)	0.031

Tokenisation is based on whitespace delimiters and punctuation; no lemmatisation or case-folding is carried out. Token instances of a given MWE or component word are identified by full-token n-gram matching over the token stream. We assume that all full stops and equivalent characters for other orthographies are sentence boundaries, and chunk the corpora into (pseudo-)sentences on the basis of them. For each language, we identify the 51st–1050th most frequent words, and consider them to be content-bearing words, in the manner of Schütze (1997). This is based on the assumption that the top-50 most frequent words are stop words, and not a good choice of word for calculating distributional similarity over. That is not to say that we can't calculate the distributional similarity for stop words, however (as we will for the EVPC dataset) they are simply not used as the dimensions in our calculation of distributional similarity.

We form a vector of content-bearing words across all token occurrences of the target word, on the basis of these 1000 content-bearing words. Our preliminary results on selecting the best context window size are shown in Table 5. According to this table, for predicting the compositionality using the best 10 languages, the sentence context window results in a higher correlation. We use sentence boundaries as the context window in the rest of our experiments. According to Weeds (2003) and Padó & Lapata (2007), using dependency relations with the neighbouring words of the target word can better predict the meaning of the target word. However, in line with our assumption of no language-specific preprocessing, we just use word co-occurrence. Finally, distributional similarity is calculated over these context vectors using cosine similarity.

6.1.2 Calculating compositionality

The procedure of calculating the compositionality is similar to what we used in §4.1.2: after translating the MWE and its components into multiple languages and measuring the distributional similarity between the translations of the MWE and its components (Figure 1), we find the best languages according to the training set. Then, we combine the scores from those best languages and finally calculate a combined compositionality score from the individual distributional similarities between each component word and the MWE. Based on our findings in §4.1.2, we combine the component scores using the weighted mean (Figure 2):

$$\text{Compositionality} = \alpha s_1 + (1 - \alpha)s_2 \tag{13.11}$$

where s_1 and s_2 are the scores for the first and the second component, respectively. We use different α settings for each dataset, based on the settings from §4.1.2.

We experiment with a range of methods for calculating compositionality, as follows:

CS_{L1}: calculate distributional similarity using only distributional similarity in the source language corpus. (This is the approach used by Reddy et al. (2011), as discussed in §2.)

CS_{L2N}: exclude the source language and compute the mean of the distributional similarity scores for the best-N target languages. The value of N is selected according to training data, as detailed in §6.1.3.[11]

CS_{L1+L2N}: calculate distributional similarity over both the source language (CS_{L1}) and the mean of the best-N languages (CS_{L2N}), and combine via the arithmetic mean.[12] This is to examine the hypothesis that using multiple target languages is better than just using the source language.

$CS_{SVR(L1+L2)}$: train a support vector regressor (SVR: Smola & Schölkopf (2004)) over the distributional similarities for all 52 languages (source and target languages).

CS_{string}: calculate string similarity using the LCS-based method of §4. LCS is chosen because, in general, it performs better than the other string similarity measures.

[11] In the case that no translation (direct or indirect) can be found for a given source language term into a particular target language, the compositionality score for that target language is set to the average across all target languages for which scores can be calculated for the given term. If no translations are available for any target language (e.g. the term is not in PanLex) the compositionality score for each target language is set to the average score for that target language across all other source language terms.

[12] We also experimented with taking the mean over all the languages – target and source – but found it best to combine the scores for the target languages first, to give more weight to the source language.

$CS_{string+L1}$: calculate the mean of the string similarity (CS_{string}) and distributional similarity in the source language.

CS_{all}: calculate the mean of the string similarity (CS_{string}) and distributional similarity scores (CS_{L1} and CS_{L2N}).

6.1.3 Selecting target languages

We experiment with two approaches for combining the compositionality scores from multiple target languages.

First, in CS_{L2N} (and CS_{L1+L2N} and CS_{all} that build off it), following the approach from §4.1.3, we use training data to rank the target languages according to Pearson's correlation between the predicted compositionality scores and the gold-standard compositionality judgements. However, in this case, based on this ranking, we take the best-N languages (instead of the best-10 languages as in §4.1.3) and again combine the individual compositionality scores by taking the arithmetic mean. We select N by determining the value that optimises the correlation over the training data. In other words, the selection of N and accordingly the best-N languages are based on nested cross-validation over training data, independently of the test data for that iteration of cross-validation.

Second in $CS_{SVR(L1+L2)}$, we take the compositionality scores from the source and all 51 target languages, combine them into a feature vector, and train an SVR over the data using LIBSVM.[13]

6.2 Results

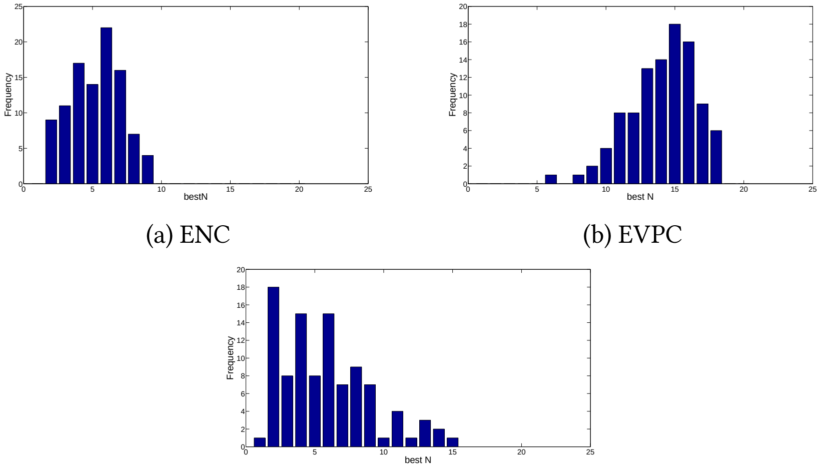
All experiments are carried out using 10 iterations of 10-fold cross validation, randomly partitioning the data independently on each of the 10 iterations, and averaging across all 100 test partitions in our presented results (Table 6). In the case of CS_{L2N} and other methods that make use of it (i.e. CS_{L1+L2N} and CS_{all}), the languages selected for a given training fold are then used to compute the compositionality scores for the instances in the test set.

Figure 4 shows histograms of the number of times each N is selected over 100 folds on ENC, EVPC and GNC datasets, respectively. From the histograms, $N = 6$, $N = 15$ and $N = 2$ are the most commonly selected settings for ENC, EVPC and GNC, respectively. That is, multiple languages are generally used, but more languages are used for English VPCs than either of the compound noun datasets.

[13] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Table 6: Pearson's correlation on the ENC, EVPC and GNC datasets

Method	Summary of the Method	ENC	EVPC	GNC
CS_{L1}	Source language	0.700	0.177	0.141
CS_{L2N}	Best-N target languages	0.434	0.398	0.113
CS_{L1+L2N}	Source + best-N target languages	0.725	0.312	0.178
$CS_{SVR(L1+L2)}$	SVR (Source + all 51 target languages)	**0.744**	0.389	0.085
CS_{string}	String Similarity	0.644	0.385	**0.372**
$CS_{string+L1}$	CS_{string} $+CS_{L1}$	0.739	0.360	0.353
CS_{all}	CS_{L1} + CS_{L2N} + CS_{string}	0.732	**0.417**	0.364



(a) ENC

(b) EVPC

(c) GNC

Figure 4: Histograms displaying how many times a given N is selected as the best number of languages over each dataset. For example, according to the GNC chart, there is a peak for $N = 2$, which shows that over 100 folds, the best-2 languages achieved the highest correlation on 18 folds.

Further analysis reveals that 32 (63%) target languages for ENC, 25 (49%) target languages for EVPC, and only 5 (10%) target languages for GNC have a correlation of $r \geq 0.1$ with gold-standard compositionality judgements. On the other hand, 8 (16%) target languages for ENC, 2 (4%) target languages for EVPC, and no target languages for GNC have a correlation of $r \leq -0.1$.

6.2.1 ENC results

English noun compounds are relatively easy to identify in a corpus,[14] because the components occur sequentially, and the only morphological variation is in noun number (singular vs. plural). In other words, the precision for our token matching method is very high, and the recall is also acceptably high. Partly as a result of the ease of identification, we get a high correlation of $r = 0.700$ for CS_{L1} (using only source language data). Using only target languages (CS_{L2N}), the results drop to $r = 0.434$, but when we combine the two (CS_{L1+L2N}), the correlation is higher than using only source or target language data, at $r = 0.725$. When we combine all languages using SVR, we achieve our best results on this dataset of $r = 0.744$, an improvement over the previous state of the art of Reddy et al. (2011) ($r = 0.714$). These last two results support our hypothesis that using translation data can improve the prediction of compositionality. The results for string similarity on its own (CS_{string}, $r = 0.644$) are slightly lower than those using only source language distributional similarity, but when combined with CS_{L1+L2N} (i.e. CS_{all}) there is a slight rise in correlation (from $r = 0.725$ to $r = 0.732$).

6.2.2 EVPC results

English VPCs are hard to identify. As discussed in §2, VPC components may not occur sequentially, and even when they do occur sequentially, they may not be a VPC. As such, our simplistic identification method has low precision and recall (hand analysis of 927 identified VPC instances would suggest a precision of around 74%). There is no question that this is a contributor to the low correlation for the source language method (CS_{L1}; $r = 0.177$). When we use target languages instead of the source language (CS_{L2N}), the correlation jumps substantially to $r = 0.398$.

When we combine English and the target languages (CS_{L1+L2N}), the results are actually lower than just using the target languages, because of the high weight on the target language, which is not desirable for VPCs, based on the source language results. Even for $CS_{SVR(L1+L2)}$, the results ($r = 0.389$) are slightly below the target language-only results. This suggests that when predicting the compositionality of MWEs which are hard to identify in the source language, it may actually be better to use target languages only. The results for string similarity (CS_{string}: $r = 0.385$) are similar to those for CS_{L2N}. However, as with the ENC

[14] Although see Lapata & Lascarides (2003) for discussion of the difficulty of reliably identifying low-frequency English noun compounds.

dataset, when we combine string similarity and distributional similarity (CS_{all}), the results improve, and we achieve the state of the art for the dataset.

In Table 7, we present classification-based evaluation over a subset of EVPC, binarising the compositionality judgements in the manner of Bannard et al. (2003). Our method achieves state-of-the-art results in terms of overall F-score and accuracy.

Table 7: Results (%) for the binary compositionality prediction task on the EVPC dataset

Method	Precision	Recall	F-score (β = 1)	Accuracy
Bannard et al. (2003)	60.8	66.6	63.6	60.0
CS_{string}	**86.2**	71.8	77.4	69.3
CS_{all}	79.5	**89.3**	**82.0**	**74.5**

6.2.3 GNC results

German is a morphologically-rich language, with marking of number and case on nouns. Given that we do not perform any lemmatisation or other language-specific preprocessing, we inevitably achieve low recall for the identification of noun compound tokens, although the precision should be nearly 100%. Partly because of the resultant sparseness in the distributional similarity method, the results for CS_{L1} are low (r = 0.141), although they are lower again when using target languages (r = 0.113). However, when we combine the source and target languages (CS_{L1+L2N}) the results improve to r = 0.178. The results for $CS_{SVR(L1+L2)}$, on the other hand, are very low (r = 0.085). Ultimately, simple string similarity achieves the best results for the dataset (r = 0.372), and this result actually drops slightly when combined with the distributional similarities.

To better understand the reason for the lacklustre results using SVR, we carried out error analysis and found that, unlike the other two datasets, about half of the target languages return scores which correlate negatively with the human judgements. When we filter these languages from the data, the score for SVR improves appreciably. For example, over the best-3 languages overall, we get a correlation score of r = 0.179, which is slightly higher than CS_{L1+L2N}.

We further investigated the reason for getting very low and sometimes negative correlations with many of our target languages. We noted that about 24% of the German noun compounds in the dataset do not have entries in PanLex.

This contrasts with ENC where only one instance does not have an entry in PanLex, and EVPC where all VPCs have translations in at least one language in PanLex. We experimented with using string similarity scores in the case of such missing translations, as opposed to the strategy described in §3.2. The results for $CS_{SVR(L1+L2)}$ rose to $r = 0.269$, although this is still below the correlation for just using string similarity.

Our results on the GNC dataset using string similarity to measure the compositionality of the whole compound are competitive with the state-of-the-art results ($r = 0.45$) using a window-based distributional similarity approach over monolingual German data by adding the modifier and head predictions (Schulte im Walde et al. 2013).[15] Note, however, that their method used part-of-speech information and lemmatisation, where ours does not, in keeping with the language-independent philosophy of this research. Furthermore, as shown in §5, our string similarity measure can be substantially improved on GNC by using a multilingual dictionary with higher coverage for the expressions in this dataset.

7 Conclusion

This chapter presented an extension of two previous studies – Salehi & Cook (2013) and Salehi et al. (2014) – that proposed supervised and unsupervised methods to predict the compositionality of MWEs based on measures of string similarity between the translations of an MWE, and translations of its component words, into many target languages, and based on distributional similarity between an MWE and its component words, both in the original source language and under translation.

In experiments using the string similarity approach, we showed that information from translations into multiple target languages can be effectively combined to give improvements over using just a single target language. We also showed that string similarity measures which capture information about character sequences perform better than measures that do not. From the experiments on unsupervised approaches, we learned that languages of the same family as the source language cannot predict the compositionality of MWEs as well as the languages for which we have good translations coverage.

For distributional similarity, our experimental results showed that incorporating information from translations into target languages improved over using

[15] Additionally, Schulte im Walde et al. (2013) showed that their method achieves the state-of-the-art results ($r = 0.65$) in predicting the compositionality of each individual component within the compound.

distributional similarity in just the source language. Furthermore, we learned that there is a strong complementarity between approaches based on string and distributional similarity.

Abbreviations

MWE	multiword expression
ENC	English Noun Compound dataset of Reddy et al. (2011)
EVPC	English Verb-Particle Construction dataset of Bannard et al. (2003)
GNC	German Noun Compound dataset of Schulte im Walde et al. (2013)
LCS	longest common substring
LEV1	Levenshtein
LEV2	Levenshtein with substitution penalty
SW	Smith Waterman algorithm

Acknowledgements

We thank the anonymous reviewers for their valuable comments, and the editors for their time and effort in compiling this volume.

References

Acosta, Otavio, Aline Villavicencio & Viviane Moreira. 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' 11), 101–109. Association for Computational Linguistics.

Baldwin, Timothy. 2009. The hare and the tortoise: Speed and accuracy in translation retrieval. *Machine Translation* 23(4). 195–240.

Baldwin, Timothy, Colin James Bannard, Takaaki Tanaka & Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE '03), 89–96. Association for Computational Linguistics. DOI:10.3115/1119282.1119294

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

Baldwin, Timothy, Jonathan Pool & Susan M. Colowick. 2010. PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations* (COLING '10), 37–40. Association for Computational Linguistics.

Bannard, Colin James. 2006. *Acquiring phrasal lexicons from corpora*. University of Edinburgh dissertation.

Bannard, Colin James, Timothy Baldwin & Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE '03 1), 65–72. Association for Computational Linguistics. DOI:10.3115/1119282.1119291

Biemann, Chris & Eugenie Giesbrecht. 2011. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the Distributional Semantics and Compositionality Workshop* (DISCo 2011) *in conjunction with ACL 2011*, 21–28.

Carpuat, Marine & Mona Diab. 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*, 242–245. Association for Computational Linguistics. http://www.aclweb.org/anthology/N10-1029.

Dias, Gaël. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (MWE '03), 41–48. Association for Computational Linguistics.

Evert, Stefan & Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language* 19(4). 450–466.

Farahmand, Meghdad, Aaron Smith & Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions* (MWE '15), 29–33. Association for Computational Linguistics.

Fazly, Afsaneh, Paul Cook & Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1). 61–103. http://aclweb.org/anthology/J09-1005.

Gomaa, Wael H & Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68(13). 13–18.

Hermann, Karl Moritz, Phil Blunsom & Stephen Pulman. 2012. An unsupervised ranking model for noun-noun compositionality. In *Proceedings of the First Joint*

*Conference on Lexical and Computational Semantics (*SEM)*, 132–141. June 7-8, 2012.

Katz, Graham. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (MWE '06), 12–19. Association for Computational Linguistics.

Kim, Su Nam & Timothy Baldwin. 2007. Detecting compositionality of English verb-particle constructions using semantic similarity. In *Proceedings of the 7th meeting of the Pacific association for computational linguistics* (PACLING 2007), 40–48.

Korkontzelos, Ioannis & Suresh Manandhar. 2009. Detecting compositionality in multi-word expressions. In *Proceedings of the ACL-IJCNLP 2009 Conference-Short papers*, 65–68. August 4, 2009.

Lapata, Mirella & Alex Lascarides. 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), 235–242.

Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the association for computational linguistics on computational linguistics* (ACL 1999), 317–324.

McCarthy, Diana, Bill Keller & John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 workshop on multiword expressions: Analysis, acquisition and treatment* (MWE '03), 73–80. Association for Computational Linguistics. DOI:10.3115/1119282.1119292

McCarthy, Diana, Sriram Venkatapathy & Aravind K. Joshi. 2007. Detecting compositionality of verb-object combinations using selectional preferences. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), 369–379. Association for Computational Linguistics. June 28–30, 2007.

Needleman, Saul B. & Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48(3). 443–453.

Padó, Sebastian & Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2). 161–199.

Pecina, Pavel. 2008. *Lexical association measures: Collocation extraction*. Prague, Czech Republic: Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic dissertation.

Pichotta, Karl & John DeNero. 2013. Identifying phrasal verbs using many bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2013). October 18-21, 2013.

Ramisch, Carlos. 2012. A generic framework for multiword expressions treatment: From acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop*, 61–66.

Reddy, Siva, Diana McCarthy & Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (IJCNLP), 210–218.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276/2010 (CICLing '02), 1–15. Springer-Verlag.

Saif, Abdulgabbar, Mohd Juzaiddin Ab Aziz & Nazlia Omar. 2013. Measuring the compositionality of Arabic multiword expressions. In *Proceedings of the second international multi-conference on artificial intelligence technology*, 245–256.

Salehi, Bahar, Narjes Askarian & Afsaneh Fazly. 2012. Automatic identification of Persian light verb constructions. In *Proceedings of the 13th International Conference on Intelligent Text Processing Computational Linguistics* (CICLing '12)), 201–210.

Salehi, Bahar & Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Second Joint Conference on Lexical and Computational Semantics*, vol. 1 (* SEM 2013), 266–275. June 13-14, 2013.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2014), 472–481. Gothenburg. http://aclweb.org/anthology/E/E14/E14-1050.pdf.

Schone, Patrick & Daniel Jurafsky. 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), 100–108. http://pascasarjana.mercubuana.ac.id/49/W01-0513.pdf.

Schulte im Walde, Sabine, Stefan Müller & Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, 255–265. Association for Computational Linguistics. June 13-14, 2013.

Schütze, Hinrich. 1997. *Ambiguity resolution in language learning*. Stanford, USA: CSLI Publications.

Smith, Temple F. & Michael S. Waterman. 1981. Identification of common molecular subsequences. *Molecular Biology* 147. 195–197.

Smola, Alex J. & Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and Computing* 14(3). 199–222.

Vecchi, Eva Maria, Marco Baroni & Roberto Zamparelli. 2011. Linear maps of the impossible: Capturing semantic anomalies in distributional space. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, 1–9.

Venkatapathy, Sriram & Aravind K. Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (HLT-EMNLP 2005), 771–778.

von der Heide, Claudia & Susanne Borgwaldt. 2009. Assoziationen zu Unter, Basis und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th norddeutsches linguistisches Kolloquium*, 51–74.

Weeds, Julie Elizabeth. 2003. *Measures and applications of lexical distributional similarity*. University of Sussex dissertation.

Name index

Language index

Subject index

Multiword expressions at length and in depth

The annual workshop on multiword expressions takes place since 2001 in conjunction with major computational linguistics conferences and attracts the attention of an ever-growing community working on a variety of languages, linguistic phenomena and related computational processing issues. MWE 2017 took place in Valencia, Spain, and represented a vibrant panorama of the current research landscape on the computational treatment of multiword expressions, featuring many high-quality submissions. Furthermore, MWE 2017 included the first shared task on multilingual identification of verbal multiword expressions. The shared task, with extended communal work, has developed important multilingual resources and mobilised several research groups in computational linguistics worldwide.

This book contains extended versions of selected papers from the workshop. Authors worked hard to include detailed explanations, broader and deeper analyses, and new exciting results, which were thoroughly reviewed by an internationally renowned committee. We hope that this distinctly joint effort will provide a meaningful and useful snapshot of the multilingual state of the art in multiword expressions modelling and processing, and will be a point of reference for future work.

ISBN 978-3-96110-123-8

9 783961 101238