Multiword expressions: Insights from a multi-lingual perspective

Manfred Sailer and Stella Markantonatou (editors)

Synopsis

Multiword expressions (MWEs) are a challenge for both the natural language applications and the linguistic theory because they often defy the application of the machinery developed for free combinations where the default is that the meaning of an utterance can be predicted from its structure. There is a rich body of primarily descriptive work on MWEs for many European languages but comparative work is little. The volume brings together MWE experts to explore the benefits of a multilingual perspective on MWEs. The ten contributions in this volume look at MWEs in Bulgarian, English, French, German, Maori, Modern Greek, Romanian, Serbian, and Spanish. They discuss prominent issues in MWE research such as classification of MWEs, their formal grammatical modeling, and the description of individual MWE types from the point of view of different theoretical frameworks, such as Dependency Grammar, Generative Grammar, Head-driven Phrase Structure Grammar, Lexical Functional Grammar, Lexicon Grammar.

Chapters

dependent constituents, definiteness, distribution of clitics, word order and various alternations. Their ultimate goal is to build a universal framework for Named Entity Recognition (NER).

- MWEs and the Emotion Lexicon: Typological and cross-lingual considerations

Aggeliki Fotopoulou, Voula Giouli Chapter 3

In this chapter, the discussion is aimed at studying predicates that pertain to the semantic field of emotions, the focus being on Modern Greek verbal multiword expressions (verbal MWEs) and their counterparts in French. A core lexicon of verbal MWEs denoting emotion was extracted from existing Modern Greek lexical resources; the initial list was further extended and revised manually in view of corpus evidence. A classification of MWEs is proposed based on syntactic, selectional and semantic properties; an attempt to map the expressions identified onto their French counterparts was also made. The cross-linguistic study reveals similarities and discrepancies in the two languages, and highlights the interaction between MWEs structure and their underlying semantics, in that the intensity of the emotion denoted and the degree of fixedness of the relevant expressions seem to be highly correlated in both languages.

- Flexibility of Multi-Word Expressions and Corpus Pattern Analysis

Patrick Hanks, Ismail El Marouf, Michael Oakes Chapter 4

This chapter is set in the context of Corpus Pattern Analysis (CPA), a technique developed by Patrick Hanks to map meaning onto word patterns found in corpora.
The main output of CPA is the Pattern Dictionary of English Verbs (PDEV), currently describing patterns for over 1,600 verbs, many of which are acknowledged to be multiword expressions (MWEs) such as phrasal verbs or idioms. PDEV entries are manually produced by lexicographers, based on the analysis of a substantial sample of concordance lines from the corpus, so the construction of the resource is very time-consuming. The motivation for the work presented in this chapter is to speed up the discovery of these word patterns, using methods which can be transferred to other languages. The chapter explores the benefits of a detailed contrastive analysis of MWEs found in English and French corpora with a view on English-French translation. The comparative analysis is conducted through a case study of the pair (*bite*, *mordre*), to illustrate both CPA and the application of statistical measures for the automatic extraction of MWEs. The approach adopted in this chapter takes its point of departure from the use of statistics developed initially by Church & Hanks (1989). Here, the authors look at statistical measures which have not yet been tested for their ability to discover new collocates, but are useful for characterizing verbal MWEs already found. In particular, they propose measures to characterize the mean span, rigidity, diversity, and idiomaticity of a given MWE.

- Multiword expressions and the Law of Exceptions

Koenraad Kuiper Chapter 5

This chapter proposes the existence of a linguistic universal, the Law of Exceptions. It hypothesizes that a relationship exists between the grammar of a language and its lexicon such that all regularities expressed in the grammar of a language are matched by exceptions which are manifested in the lexicon of that language. It is also proposed that lexical idiosyncrasies are of two types. Type 1 idiosyncrasies are in the nature of arbitrary restrictions on options provided in the grammar while Type 2 idiosyncrasies involve breaches of the rules of the grammar. To test this law requires an initial examination of the

linguistic domains where it might be tested. As a preliminary step to testing these ideas, this chapter is a scoping exercise looking chiefly at the structural properties of a subset of multiword expressions (MWE). It shows, following Barkema (1996), that many properties of MWEs cross-classify. The aim of the overview is then to examine domains of the morphosyntax of any language which might be analysed for sources of structural idiosyncrasy and thus to determine how individual languages might vary in this respect. Languages of exemplification are English, which has a relatively fixed word order and slight inflectional system, and, to a lesser extent Dutch and Māori, an Oceanic language.

- Choosing features for classifying multiword expressions

Éric Laporte Chapter 6

In this work, multiword expressions (MWEs) are considered a heterogeneous set with a glaring need for classifications. However, designing a satisfactory classification involves choosing features. Although in the case of MWEs many features are a priori available, not all of them are equal in terms of how reliably MWEs can be assigned to classes. Accordingly, resulting classifications may be more or less fruitful for computational use. The author outlines an enhanced classification. In order to increase its suitability for many languages, he uses previous works taking into account various languages.

- Revisiting the grammatical function "object" (OBJ and OBJ_θ)

Stella Markantonatou, Niki Samaridi Chapter 7

Free subject verb multiword expressions (MWEs) of Modern Greek and English provide data that challenge the theoretical status of the syntactic notion object. The authors compare the syntactic reflexes of three types of verbal complement: objects of typical monotransitive verbs, indirect objects of ditransitive verbs and fixed accusative noun phrases (NPs) that occur as direct complements of verbs in MWEs. Passivisation, clitic replacement, object optionality and distribution present themselves as syntactic reflexes that draw relatively clear cut lines across these three classes of verbal complements and suggest that the Grammatical Functions OBJ(ect) and OBJ(ect)θ of LFG should not be assigned to the fixed accusative NPs that occur in verb MWEs; rather a new Grammatical Function should be defined for this purpose.

- Derivation in the domain of multiword expressions

Verginica Barbu Mititelu, Svetlozara Leseva Chapter 8

Multiword expressions and derivation have rarely been discussed together, even though analyzing the interaction between them is of great importance for the study of each topic and, in general, for the study of the language and for Natural Language Processing. Derivation is a means of enriching the lexicon with both words and multiword expressions. Various types of derivation (suffixation, prefixation or both, as well as other derivational devices) can act upon either words or multiword expressions. The focus of the work presented in this chapter is the formation of multiword expressions from other multiword expressions via derivation. The authors analyze the morphological, syntactic and semantic aspects of this process, providing examples from Romanian and Bulgarian, languages, which belong to different families but have been in contact throughout their history. The study can be further extended with data from other languages. The perspective adopted is paradigmatic, but the syntagmatic approach, which can only be mentioned as further work, will add to the quality of the analysis of facts: corpus data will contextualize the phenomena discussed here and offer quantitative information about them.

- Modeling multiword expressions in a parallel Bulgarian-English newsmedia corpus

Petya Osenova, Kiril Simov Chapter 9

The paper focuses on the modelling of multiword expressions (MWE) in Bulgarian-English parallel news corpora (SETimes; CSLI dataset and PennTreebank dataset). Observations were made on alignments in which at least one multiword expression was used per language. Multiword expressions were classified with respect to the PARSEME lexicon-based (WG1) and treebank-based (WG4) classifications. The non-MWE counterparts of MWEs are also considered. The adopted approach is data-driven because the data of this study was retrieved from parallel corpora and not from bilingual dictionaries. The survey shows that the predominant translation relation between Bulgarian and English is *MWE-to-word*, and that this relation does not exclude other translation options. To formalize our observations, a catenae-based
modelling of the parallel pairs is proposed.

- Spanish multiword Expressions: Looking for a taxonomy

Carla Parra Escartín, Almudena Nevado Llopis, Eoghan Sánchez Martínez Chapter 10

The authors analyze Spanish multiword expressions (MWEs) and describe their linguistic properties. The ultimate goal of their analysis is to find an MWE taxonomy for Spanish which is suitable for Natural Language Processing purposes. As
a starting point of their study, they take the MWE taxonomy proposed by Ramisch (2012; 2015). This taxonomy distinguishes between morphosyntactic classes and other classes which cannot be considered morphosyntactic and which he calls “difficulty classes”. The authors have built and analyzed a special data set of Spanish MWEs. They add a new axis to Ramisch’s (2012; 2015) taxonomy, namely the flexibility axis which was introduced by Sag et al. (2002). In the light of the taxonomy analysis, the authors modified and adapted Ramisch’s taxonomy to Spanish MWEs. Here, the different types of MWEs in Spanish are analyzed and described along with flexibility tests for Spanish MWEs.

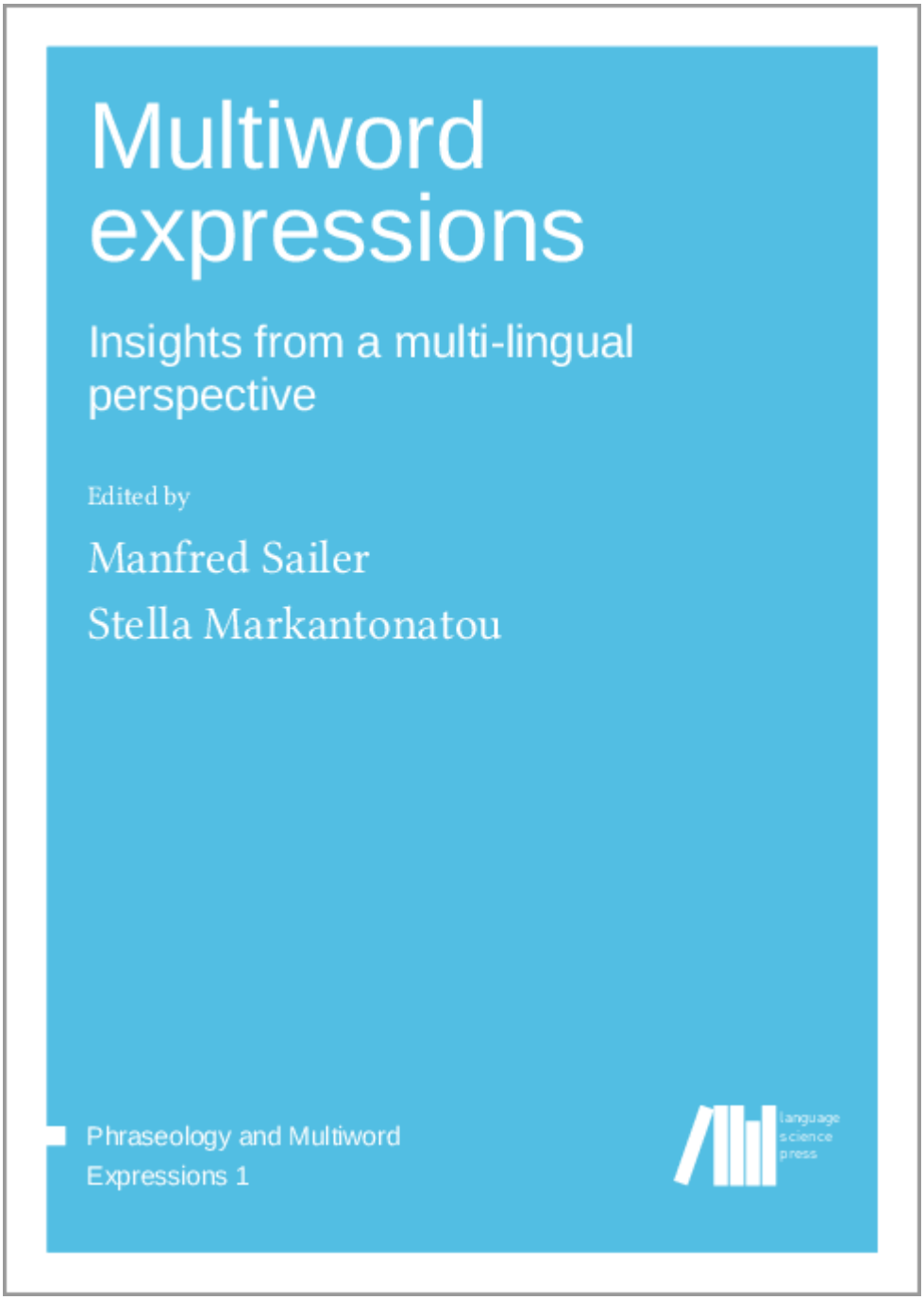
Editors

Manfred Sailer

Manfred Sailer (1969) is professor of English Linguistics at Goethe-University Frankfurt a.M. He studied general linguistics, computer science and psychology at Universität Tübingen (Master 1995, Promotion 2003) and received his postdoctoral degree (Habilitation) in English and General Linguistics at Göttingen University (2010). His mainareas of reasearch are the syntax-semantics interface, formal phraseology, negation, and the interaction of regularity and irregularity in language.

Stella Markantonatou

Stella Markantonatou (1958) has studied Chemical Engineering (National Technical University of Athens) and Linguistics (Nationaland Kapodistrian University of Athens) and holds a PhD in Linguistics (University of Essex). She is a Research Director with the Institute for Language and Speech Processing/Athena RIC, Athens, Greece. She has published on the formal study of the syntax-semantics interface, on phrase-based Machine Translation, and on the development of lexical resources.



PDF

Bibliography
Buy from Amazon.com
Buy from Amazon.co.uk
Buy from Amazon.de
Collaborative reading on Paperhive
LaTeX source on GitHub
Series
Phraseology and Multiword Expressions

Details about the available publication format: PDF

PDF
ISBN-13 (15) 978-3-96110-063-7
ISBN-13 hardcover (28) 978-3-96110-064-4
Publication date (01) 2018-05-18
doi 10.5281/zenodo.1182583