Multiword Units in Machine Translation and Translation Technology

MITKOV, Ruslan (Ed.), et al.

MITKOV, Ruslan (Ed.), et al. *Multiword Units in Machine Translation and Translation Technology*. John Benjamins, 2018

DOI: 10.1075/cilt.341
Multiword Units in Machine Translation and Translation Technology

EDITED BY
Ruslan Mitkov
Johanna Monti
Gloria Corpas Pastor
Violeta Seretan

JOHN BENJAMINS PUBLISHING COMPANY
MULTIWORD UNITS IN MACHINE TRANSLATION AND TRANSLATION TECHNOLOGY
CURRENT ISSUES IN LINGUISTIC THEORY

AMSTERDAM STUDIES IN THE THEORY AND HISTORY
OF LINGUISTIC SCIENCE – Series IV
ISSN 0304-0763

General Editor
JOSEPH C. SALMONS
University of Wisconsin–Madison
jsalmons@wisc.edu

Founder & General Editor (1975-2015)
E.F.K. KOERNER
Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin
efk.koerner@rz.hu-berlin.de

Current Issues in Linguistic Theory (CILT) is a theory-oriented series which welcomes contributions from scholars who have significant proposals that advance our understanding of language, its structure, its function and especially its historical development. CILT offers an outlet for meaningful contributions to current linguistic debate.

A complete list of titles in this series can be found on http://benjamins.com/catalog/cilt

Editorial Board
Claire Bowern (New Haven, Ct.)
Sheila Embleton (Toronto)
Elly van Gelderen (Tempe, Ariz.)
John E. Joseph (Edinburgh)
Matthew Juge (San Marcos, Tex.)
Martin Maiden (Oxford)
Martha Ratliff (Detroit, Mich.)
E. Wyn Roberts (Vancouver, B.C.)
Klaas Willems (Ghent)

Volume 341

Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor and Violeta Seretan (eds.)

Multiword Units in Machine Translation and Translation Technology
# Table of contents

About the editors vii

Multiword units in machine translation and translation technology 1

*Johanna Monti, Violeta Seretan, Gloria Corpas Pastor & Ruslan Mitkov*

## Part 1. Multiword units in machine translation

Analysing linguistic information about word combinations for a Spanish-Basque rule-based machine translation system 41

*Uxoa Iñurrieta, Itziar Aduriz, Arantza Díaz de Ilarraza, Gorka Labaka & Kepa Sarasola*

How do students cope with machine translation output of multiword units?

An exploratory study 61

*Joke Daems, Michael Carl, Sonia Vandepitte, Robert Hartsuiker & Lieve Macken*

Aligning verb + noun collocations to improve a French-Romanian FSMT system 81

*Amalia Todiraşcu & Mirabela Navlea*

## Part 2. Multiword units in multilingual NLP applications

Multiword expressions in multilingual information extraction 103

*Gregor Thurmair*

A multilingual gold standard for translation spotting of German compounds and their corresponding multiword units in English, French, Italian and Spanish 125

*Simon Clematide, Stéphanie Lehner, Johannes Graëñ & Martin Volk*

Dutch compound splitting for bilingual terminology extraction 147

*Lieve Macken & Arda Tezcan*

## Part 3. Identification and translation of multiword units

A flexible framework for collocation retrieval and translation from parallel and comparable corpora 165

*Oscar Mendoza Rivera, Ruslan Mitkow & Gloria Corpas Pastor*
On identification of bilingual lexical bundles for translation purposes: The case of an English-Polish comparable corpus of patient information leaflets

Łukasz Grabowski

The quest for Croatian idioms as multiword units

Kristina Kocijan & Sara Librenjak

Corpus analysis of Croatian constructions with the verb *doći* ‘to come’

Goranka Blagus Bartolec & Ivana Matas Ivanković

Anaphora resolution, collocations and translation

Eric Wehrli & Luka Nerima

Index
About the editors

While Prof Ruslan Mitkov is best known for his seminal contributions to anaphora resolution and automatic generation of multiple-choice tests, his original contributions and extensively cited research (250 publications) cover many other topics including but not limited to translation technology and NLP applications for language disabilities. Mitkov is author of the monograph *Anaphora resolution* (Longman) and Editor of the *Oxford Handbook of Computational Linguistics* (Oxford University Press). Current prestigious projects include his role as Executive Editor of the *Journal of Natural Language Engineering* (Cambridge University Press) and Editor-in-Chief of the *NLP book series* of John Benjamins publishers. Prof Mitkov is regular keynote speaker at international conferences and often acts as Programme Chair of scientific events; he has been principal investigator of projects with funding (exceeding €20,000,000) provided by UK funding agencies, the EC and the industry.

Dr Mitkov received his MSc from the Humboldt University in Berlin and his PhD from the Technical University in Dresden. Ruslan Mitkov is Professor of Computational Linguistics and Language Engineering at the University of Wolverhampton and Director of the Research Institute in Information and Language Processing. Mitkov is a Fellow of the Alexander von Humboldt Foundation, Germany, Distinguished Visiting Professor at the University of Franche-Comté in Besançon, France and Vice-Chair for the prestigious EC funding programme ‘Future and Emerging Technologies’. In recognition of his outstanding professional/research achievements, Prof Mitkov was awarded the title of Doctor Honoris Causa at Plovdiv University in 2011; he was also conferred Professor Honoris Causa at Veliko Tarnovo University in 2014.

Johanna Monti is currently Associate Professor of Modern Languages Teaching at the L’Orientale University of Naples where she teaches Translation Studies, Specialised Translation, Machine Translation and CAT tools. She was the Computational Linguistics Research manager of the Thamus Consortium in Salerno (FINMECCANICA Group) from 1988 till 2001 and took part in various national and international research projects (CLIA, EUROLANG, CLIPS). She was research fellow from 2008 till 2012 in Computational Linguistics at the Department of Political, Social and Communication Sciences of the University of Salerno. She received her PhD in Theories, Methodologies and Advanced applications for Communication,
Computer Science and Physics with a thesis in Computational Linguistics at the University of Salerno, Italy. She has acted as Programme Chair of various international conferences on Natural Language Processing (NLP), Machine Translation, Translation Technology, Computational Linguistics and is asked on a regular basis to review for leading Italian and international funding bodies and organisations. She was an active member of the PARSEME Cost Action and is currently acting as Principal Investigator of the PARSEME-IT project. She is member of the governing board of the I-Land inter-university research center and of the standing committee of the SIGLEX-MWE Interest Group. Her current research activities are in the field of hybrid approaches to MT, development of Linguistic Resources for NLP applications, evaluation of Translation technologies and finally new trends in translation.

Gloria Corpas Pastor has a BA in German Philology (English) from the University of Malaga and a PhD in English Philology from the Universidad Complutense de Madrid (1994). Visiting Professor in Translation Technology at the Research Institute in Information and Language Processing (RIILP) of the University of Wolverhampton, UK (since 2007), and Professor in Translation and Interpreting at the University of Malaga (since 2008). Published and cited extensively, member of several international and national editorial and scientific committees. Spanish delegate for AEN/CTN 174 and CEN/BTTF 138, actively involved in the development of the UNE-EN 15038:2006 and currently involved in several ISO Standards (ISO TC37/SC2-WG6 “Translation and Interpreting”). Regular evaluator of University programmes and curriculum design for the Spanish Agency for Quality Assessment and Accreditation (ANECA) and various research funding bodies. Director of the research group “Lexicography and Translation” since 1997 (http://www.lexytrad.en). Director of the Department of Translation and Interpreting of the University of Malaga, former President and member of the Scientific Committee of AIETI (Iberian Association of Translation and Interpreting Studies), Board member and member of the Advisory council of EUROPHRAS (European Society of Phraseology) and Vice-President of AMIT-A (Association of Women in Science and Technology of Andalusia). Her research fields include corpus linguistics, phraseology, lexicography and translation technology.

Violeta Seretan (PhD in Linguistics, University of Geneva, 2008) is a Senior Research Associate and Lecturer at the Universities of Geneva and Lausanne. She has extensive experience in the fields of Computational Linguistics and Multilingual Natural Language Processing. Since 2000, when she earned her MSc degree in Computer Science from the University of Iasi, she has conducted research on topics as varied as language analysis, lexical semantics, phraseology acquisition, information extraction, text simplification, machine translation, pre-editing, post-
editing, and machine translation evaluation (among others). She has authored a book (“Syntax-Based Collocation Extraction”, Springer, 2010) and over 50 peer-reviewed publications in international journals and conferences in the field of Human Language Technology. She has been a coordinator of the FP7 European project ACCEPT; a member of the COST Action PARSEME; and a participant in a number of nationally and locally funded research projects. She has held positions in several institutions, including the University of Edinburgh (UK), the Fuji Xerox Palo Alto Laboratory (USA), and the University of Lausanne (CH).