

Language technologies for a multilingual Europe

TC3 III

Edited by

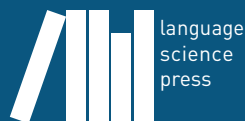
Georg Rehm

Felix Sasaki

Daniel Stein

Andreas Witt

Translation and Multilingual Natural
Language Processing 5



Translation and Multilingual Natural Language Processing

Editors: Oliver Czulo (Universität Leipzig), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Reinhard Rapp (Johannes Gutenberg-Universität Mainz)

In this series:

1. Fantinuoli, Claudio & Federico Zanettin (eds.). New directions in corpus-based translation studies.
2. Hansen-Schirra, Silvia & Sambor Grucza (eds.). Eyetracking and Applied Linguistics.
3. Neumann, Stella, Oliver Čulo & Silvia Hansen-Schirra (eds.). Annotation, exploitation and evaluation of parallel corpora: TC3 I.
4. Czulo, Oliver & Silvia Hansen-Schirra (eds.). Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3 II.
5. Rehm, Georg, Felix Sasaki, Daniel Stein & Andreas Witt (eds.). Language technologies for a multilingual Europe: TC3 III.
6. Menzel, Katrin, Ekaterina Lapshinova-Koltunski & Kerstin Anna Kunz (eds.). New perspectives on cohesion and coherence: Implications for translation.
7. Hansen-Schirra, Silvia, Oliver Czulo & Sascha Hofmann (eds.). Empirical modelling of translation and interpreting.
8. Svoboda, Tomáš, Łucja Biel & Krzysztof Łoboda (eds.). Quality aspects in institutional translation.
9. Fox, Wendy. Can integrated titles improve the viewing experience? Investigating the impact of subtitling on the reception and enjoyment of film using eye tracking and questionnaire data.
10. Moran, Steven & Michael Cysouw. The Unicode cookbook for linguists: Managing writing systems using orthography profiles.

ISSN: 2364-8899

Language technologies for a multilingual Europe

TC3 III

Edited by

Georg Rehm

Felix Sasaki

Daniel Stein

Andreas Witt

Georg Rehm, Felix Sasaki, Daniel Stein & Andreas Witt (eds.). 2018. *Language technologies for a multilingual Europe: TC3 III* (Translation and Multilingual Natural Language Processing 5). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/106>

© 2018, the authors

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/>

ISBN: 978-3-946234-73-9 (Digital)

978-3-946234-77-7 (Hardcover)

ISSN: 2364-8899

DOI:10.5281/zenodo.1291947

Source code available from www.github.com/langsci/106

Collaborative reading: paperhive.org/documents/remote?type=langsci&id=106

Cover and concept of design: Ulrike Harbort

Typesetting: Felix Kopecky, Florian Stuhlmann, Iana Stefanova, Sebastian Nordhoff, Stefanie Hegele

Proofreading: Ahmet Bilal Özdemir, Alessia Battisti, Alexis Michaud, Amr Zawawy, Anne Kilgus, Brett Reynolds, Benedikt Singpiel, David Lukeš, Eleni Koutso, Eran Asoulin, Ikmi Nur Oktavianti, Jeroen van de Weijer, Matthew Weber, Lea Schäfer, Rosetta Berger, Stathis Selimis

Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono

Typesetting software: \LaTeX

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	iii
1 Editorial	
Georg Rehm, Felix Sasaki, Daniel Stein & Andreas Witt	1
2 Machine translation: Past, present and future	
Daniel Stein	5
3 The META-NET strategic research agenda for language technology in Europe: An extended summary	
Georg Rehm	19
4 Metadata for the multilingual web	
Felix Sasaki	43
5 State of the art in Translation Memory Technology	
Uwe Reinke	55
6 Authoring support for controlled language and machine translation: A report from practice	
Melanie Siegel	85
7 Integration of machine translation in on-line multilingual applications: Domain adaptation	
Mirela-Ştefania Duma & Cristina Vertan	103
8 Disambiguate yourself: Supporting users in searching documents with query disambiguation suggestions	
Ernesto William De Luca & Christian Scheel	123

Contents

9 Multilingual knowledge in aligned Wiktionary and OmegaWiki for translation applications	
Michael Matuschek, Christian M. Meyer & Iryna Gurevych	139
10 The BerbaTek project for Basque: Promoting a less-resourced language via language technology for translation, content management and learning	
Igor Leturia, Kepa Sarasola, Xabier Arregi, Arantza Diaz de Ilarraza, Eva Navas, Iñaki Sainz, Arantza del Pozo, David Baranda & Urtza Iturraspe	181
Index	205

Preface

This book, *Language Technologies for a Multilingual Europe*, is a reissue of the Special Issue on Technologies for a Multilingual Europe, which was originally published as Vol. 3, No. 1 of the Open Access online journal *Translation: Computation, Corpora, Cognition* (TC3). After the editors of TC3 had decided to transition the journal into a different format – into the Open Access book series *Translation and Multilingual Natural Language Processing* – they invited us to prepare a reissue of our compilation, originally published in 2013. While several smaller typos in the original manuscripts have been fixed, the papers in this collection have not been substantially modified with regard to the original publication, which is still, for archival reasons, available at <http://www.blogs.uni-mainz.de/fb06-tc3/vol-3-no-1-2013/>.

Since the original publication, *Multilingual Europe* has made several important steps forward. A new set of EU projects on multilingual technologies and machine translation was funded in 2015, e.g., QT21, HimL and CRACKER. The *Cracking the Language Barrier* federation (<http://www.cracking-the-language-barrier.eu>) was established as an umbrella organisation for all projects and organisations working on technologies for a multilingual Europe. At the time of writing *METANET* is organising the next *META-FORUM* conference, which is to take place in Brussels on 13/14 November 2017. One of the key topics of this conference is the Human Language Project – a large, coordinated funding programme spanning from education to research to innovation, which aims at bringing about the much needed boost in research and a paradigm shift in processing language automatically. First steps towards the Human Language Project were discussed at a workshop in the European Parliament in early 2017. Moreover, the Common Language Resources and Technology Infrastructure (CLARIN), founded in 2012 by ten European countries, doubled its number of members since then.

The editors of this special issue would like to thank the series editors, especially Oliver Culo, for the opportunity to publish a reissue of our original compilation.

Georg Rehm, Felix Sasaki, Daniel Stein, Andreas Witt

August 28, 2017

