Annotation, exploitation and evaluation of parallel corpora

TC3 I

Edited by

Silvia Hansen-Schirra

Stella Neumann

Oliver Čulo

Translation and Multilingual Natural Language Processing 3

language science press

Translation and Multilingual Natural Language Processing

Editors: Oliver Čulo (Johannes Gutenberg-Universität Mainz), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Stella Neumann (RWTH Aachen), 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. Silvia Hansen-Schirra, Neumann, Stella & Oliver Čulo (eds.). Annotation, exploitation and evaluation of parallel corpora: TC3 I.
4. Čulo, 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.

ISSN: 2364-8899

Annotation, exploitation and evaluation of parallel corpora

TC3 I

Edited by

Silvia Hansen-Schirra

Stella Neumann

Oliver Čulo

language science press

Silvia Hansen-Schirra, Stella Neumann & Oliver Čulo (eds.). 2017. *Annotation, exploitation and evaluation of parallel corpora*: *TC3 I* (Translation and Multilingual Natural Language Processing 3). Berlin: Language Science Press.

This title can be downloaded at:
http://langsci-press.org/catalog/book/103
© 2017, the authors
Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):
http://creativecommons.org/licenses/by/4.0/
ISBN: 978-3-946234-85-2 (Digital)
978-3-946234-89-0 (Hardcover)
978-3-946234-69-2 (Softcover)
ISSN: 2364-8899
DOI:10.5281/zenodo.283376

Cover and concept of design: Ulrike Harbort
Typesetting: Sebastian Nordhoff, Iana Stefanova, Florian Stuhlmann
Proofreading: Aleksandrs Berdicevskis, Alessia Battisti, Alexis Michaud, Alexis Palmer, Anca Gâță, Andreea Calude, Annie Zaenen, Amr El-Zawawy, Ahmet Özdemir, Daniel Riaño, Daniela Kolbe, Eitan Grossman, Elizabeth Zeitoun, Eugeniu Costezki, Ezekiel Bolaji, Francesca Di Garbo, Geert Booij, Hella Olbertz, Ikmi Nur Oktavianti, Julien Heurdier, Kleanthes Grohmann, Lars Zeige, Mario Bisiada, Martin Haspelmath, Matthew Czuba, Natsuko Nakagawa, Pierre-Yves Modicom, Rafael Nonato, Mykel Brinkerhoff, Ulrike Demske, Viola Wiegand, Wafa Abu Hatab
Fonts: Linux Libertine, Arimo, DejaVu Sans Mono
Typesetting software: XeLaTeX

Language Science Press
Habelschwerdter Allee 45
14195 Berlin, Germany
langsci-press.org

Storage and cataloguing done by FU Berlin

Language Science Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

Preface to the new edition **v**

1 Introduction
Stella Neumann, Silvia Hansen-Schirra & Oliver Čulo **1**

2 Building and querying parallel treebanks
Martin Volk, Torsten Marek & Yvonne Samuelsson **9**

3 Enriching Slovene wordnet with domain-specific terms
Špela Vintar & Darja Fišer **35**

4 Empty links and crossing lines: Querying multi-layer annotation and alignment in parallel corpora
Oliver Čulo, Silvia Hansen-Schirra, Karin Maksymski & Stella Neumann **53**

5 On drafting and revision in translation: A corpus linguistics oriented analysis of translation process data
Fabio Alves & Daniel Couto Vale **89**

6 Computerlinguistik in der Dolmetschpraxis unter besonderer Berücksichtigung der Korpusanalyse
Claudio Fantinuoli **111**

Index **147**

Preface to the new edition

This is the first of three volumes which are made up of previously published volumes of the open-access journal "Translation: Computation, Corpora, Cognition" (TC3). Digitalisation has had an immense impact on the way we share our knowledge, including on the way how researchers publish their work. TC3 was one of the very first endeavours to make open-access online publication viable in Translation Studies.

OpenAccess is still being met with quite some scepticism, but we, the former editors of TC3, Silvia Hansen-Schirra, Stella Neumann and Oliver Čulo, believe that the open access to knowledge is the right way to publish scientific results: Research, both for the community and the society at large, often funded by the public and consequently made accessible to the public. The acceptability of research findings is in part determined by how the community as well as the public is informed about this research (both its aims and its achievements).

It can, however, not be taken for granted that the results of up-to-date research are easily and freely accessible to the community or a lay audience. Another problem is to keep pace with the speed of progress in the sciences and an increasing specialization, which widen the gap between the current state of research and the accessibility to published findings. As a counter model to traditional publishing, OpenAccess straightforwardly offers a solution to this problem providing free and online access to cutting-edge, innovative research.

Did we know what we were doing when we started? Well, partially so. OpenAccess has had a lot of positive effects on the availability of results and the impact of researchers' work, but in its current, often community-based form, it also poses a challenge for researchers who engage in organising an OpenAccess journal or book series: It is they who are responsible not only for the quality of the contents (which should not and we believe will not diminish in OpenAccess), but for much of the or even the whole appearance, including the design of the publication and the quality of the type setting.

After three years with a special issue every year, the journal TC3 was transformed into the book series now called "Translation and Multilingual Natural Language Processing" (TMNLP) under the roof of LangSci Press. This move re-

flects in some sense the currently fast-changing publication landscape in both sciences and humanities. Becoming a book series at LangSci has resulted in a boost of the quality of the published volumes. Also, a stringent proofreading process has helped ensure higher consistency within and across the contributions.

The idea to re-publish the TC3 volumes as TMNLP volumes came up very early, with two goals in mind:

- making the works contributed to TC3 available in the long run, beyond just by archiving them somewhere;
- honouring the work which was put into the contributions by re-publishing them under higher quality standards.

The three volumes 3, 4 and 5 are thus not mere re-prints, but the contributions were re-edited according to LangSci guidelines and quality standards. Each volume is introduced by a dedicated introduction from the original volumes. The TC3 contributions are still available in their original format for documentary purposes under http://www.t-c3.org at the time of publication of the corresponding TMNLP volumes. Nevertheless, we believe that re-publication within LangSci will ensure enhanced impact and long-time availability, and on top of that it is a further step into the new world of open-access publishing for Translation Studies.

Germersheim and Aachen, January 2017
Oliver Čulo, Silvia Hansen-Schirra, Stella Neumann