

ZSM Studien (Volume 5)

# **Non-standard Data Sources in Corpus-based Research**

**Marcos Zampieri and Sascha Diwersy (Editors)**

April, 2013  
Shaker Verlag - Aachen, Germany



ZSM-Studien  
Schriften des Zentrums Sprachenvielfalt und Mehrsprachigkeit  
der Universität zu Köln

herausgegeben von  
Christiane M. Bongartz  
Claudia M. Riehl

Vol. 5

**Marcos Zampieri**  
**Sascha Diwersy (Eds.)**

**Non-standard Data Sources in Corpus-based Research**

Shaker Verlag  
Aachen 2013

**Bibliographic information published by the Deutsche Nationalbibliothek**

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at <http://dnb.d-nb.de>.

Copyright Shaker Verlag 2013

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publishers.

Printed in Germany.

ISBN 978-3-8440-2222-3

ISSN 1867-0830

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen

Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9

Internet: [www.shaker.de](http://www.shaker.de) • e-mail: [info@shaker.de](mailto:info@shaker.de)

## **Scientific Panel**

- Jorge Baptista - University of Algarve and INESC-ID Lisbon (Portugal)
- Michael Beißwenger - TU Dortmund University (Germany)
- Stefanie Dipper - Ruhr University Bochum (Germany)
- Lothar Lemnitzer - Berlin-Brandenburg Academy of Sciences and Humanities (Germany)
- Francesco Mambrini - Harvard University (USA)
- Nuno Mamede - Instituto Superior Técnico and INESC-ID Lisbon (Portugal)
- Alexander Mehler - University of Frankfurt (Germany)
- Marco Passarotti - Università Cattolica del Sacro Cuore, Milan (Italy)
- Jürgen Rolshoven - University of Cologne (Germany)

## **Editors**

- Marcos Zampieri - University of Cologne (Germany)
- Sascha Diwersy - University of Cologne (Germany)



# Contents

<b>Preface</b>	<b>1</b>
<b>Section 1: Papers Presented at the NOSDAC Workshop</b>	<b>3</b>
<b>1 Non-standard Data in Swiss Text Messages with a Special Focus on Dialectal Forms</b>	
<i>Simone Ueberwasser</i>	<b>5</b>
<b>2 Identification of Patterns and Document Ranking of Internet Texts: A Frequency-based Approach</b>	
<i>Marcos Zampieri, Jürgen Hermes and Stephan Schwiebert</i>	<b>25</b>
<b>3 The Digital Romansh Chrestomathy - Towards an Annotated Corpus of Romansh</b>	
<i>Claes Neuefeind</i>	<b>41</b>
<b>Section 2: Non-standard Language Resources</b>	<b>59</b>
<b>4 The Taming of a Dialect: Interlinear Glossing of Swiss German Text Messages</b>	
<i>Beni Ruef and Simone Ueberwasser</i>	<b>61</b>
<b>5 NoSta-D: A Corpus of German Non-Standard Varieties</b>	
<i>Stefanie Dipper, Anke Lüdeling and Marc Reznicek</i>	<b>69</b>
<b>6 Colonia: Corpus of Historical Portuguese</b>	
<i>Marcos Zampieri and Martin Becker</i>	<b>77</b>
<b>7 Hassle-free POS-Tagging for the Alsatian Dialects</b>	
<i>Delphine Bernhard and Anne-Laure Ligozat</i>	<b>85</b>