University Centre for Computer Corpus Research on Language Technical Papers

Volume 17 Special issue. Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003) 27 March 2003 held in conjunction with the Corpus Linguistics 2003 conference

Editors: Kiril Simov and Petya Osenova.
University Centre  
for Computer Corpus  
Research on Language  
Technical Papers  

Volume 17 Special issue.  
Proceedings of the  
Workshop on Shallow  
Processing of Large  
Corpora (SProLaC 2003)  
27 March 2003  
held in conjunction with the Corpus Linguistics 2003 conference  

Editors: Kiril Simov and Petya Osenova.
Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003)
27 March 2003
held in conjunction with the Corpus Linguistics 2003 conference
Lancaster University (UK), 28 - 31 March 2003

Supported by:

OntoText - a Sirma Laboratory for Knowledge and Language Engineering

BulTreeBank Project

Bulgarian Information Society Center of Excellence for Education, Science and Technology in 21 Century (BIS-21)

Editors: Kiril Simov and Petya Osenova

Lancaster University 2003
Workshop Organisers

Kiril Simov
Petya Osenova
BulTreeBank Project
http://www.BulTreeBank.org
Linguistic Modelling Laboratory, CLPP,
Bulgarian Academy of Sciences

Workshop Programme Committee

Michael Barlow, USA
Tomaz Erjavec, Slovenia
Silvia Hansen, Germany
Atanas Kiryakov, Bulgaria
Sandra Kuebler, Germany
Ghassan Mourad, France
Joakim Nivre, Sweden
Kemal Oflazer, Turkey
Karel Oliva, Austria
Petya Osenova, Bulgaria (co-chair)
Vladimir Petkevic, Czech Republic
Adam Przepiórkowski, Poland
Geoffrey Sampson, UK
Kiril Simov, Bulgaria (co-chair)
Milena Slavcheva, Bulgaria
Marko Tadic, Croatia
Dan Tufis, Romania
Tylman Ule, Germany
Tamas Varadi, Hungary
Nikolaj Vazov, Bulgaria
Andreas Wagner, Germany
Table of Contents

Eckhard Bick
A CG & PSG Hybrid Approach to Automatic Corpus Annotation 1

Alexander Clark
Pre-processing Very Noisy Text 12

Mark Davies
Relational n-gram databases as a basis for unlimited annotation on large corpora 23

Stefan Evert, Hannah Kermes
Annotation, Storage, and Retrieval of Mildly Recursive Structures 34

Xunlei Rose Hu and Eric Atwell
A Survey of Machine Learning Approaches to Analysis of Large Corpora 45

Adam Kilgarriff
Linguistic Search Engine 53

Borja Navarro, Montserrat Civit, Mª Antonia Martí, Raquel Marcos, Belén Fernández
Syntactic, semantic and pragmatic annotation in Cast3LB 59

Roman Ondruška, Jarmila Panevová, Jan Štepánek
An Exploitation of the Prague Dependency Treebank: A Valency Case 69

Petya Osenova and Kiril Simov
Between Chunk Ideology and Full Parsing Needs 78

Veronika Reznícková
Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora 88
# Author Index

<table>
<thead>
<tr>
<th>Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eric Atwell</td>
<td>45</td>
</tr>
<tr>
<td>Eckhard Bick</td>
<td>1</td>
</tr>
<tr>
<td>Montserrat Civit</td>
<td>59</td>
</tr>
<tr>
<td>Alexander Clark</td>
<td>12</td>
</tr>
<tr>
<td>Mark Davies</td>
<td>23</td>
</tr>
<tr>
<td>Stefan Evert</td>
<td>34</td>
</tr>
<tr>
<td>Jan Štepánek</td>
<td>69</td>
</tr>
<tr>
<td>Belén Fernández</td>
<td>59</td>
</tr>
<tr>
<td>Xunlei Rose Hu</td>
<td>45</td>
</tr>
<tr>
<td>Hannah Kermes</td>
<td>34</td>
</tr>
<tr>
<td>Adam Kilgarriff</td>
<td>53</td>
</tr>
<tr>
<td>Raquel Marcos</td>
<td>59</td>
</tr>
<tr>
<td>Mª Antonia Martí</td>
<td>59</td>
</tr>
<tr>
<td>Borja Navarro</td>
<td>59</td>
</tr>
<tr>
<td>Roman Ondruška</td>
<td>69</td>
</tr>
<tr>
<td>Petya Osenova</td>
<td>78</td>
</tr>
<tr>
<td>Jarmila Paneová</td>
<td>69</td>
</tr>
<tr>
<td>Veronika Rezníková</td>
<td>88</td>
</tr>
<tr>
<td>Kiril Simov</td>
<td>78</td>
</tr>
</tbody>
</table>