

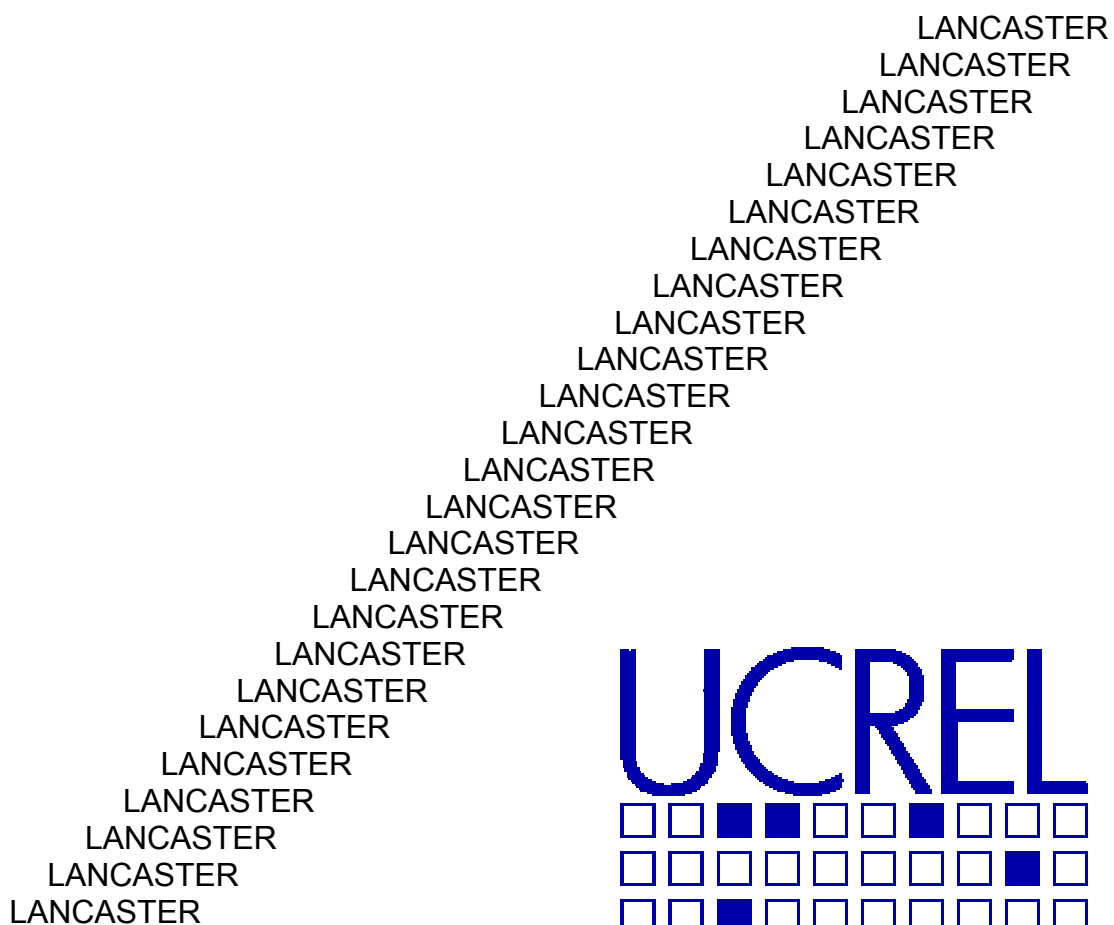
**University Centre
for Computer Corpus
Research on Language
Technical Papers**



**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.



UCREL
Computing Department
Lancaster University
Lancaster
LA1 4YR
United Kingdom
Phone: (+44) 1524 593802
Fax: (+44) 1524 593608
Email: ucrel@lancaster.ac.uk

ISBN 1-86220-134-X
Lancaster University 2003
Further copies may be obtained from
http://www.comp.lancs.ac.uk/ucrel/tech_papers.html

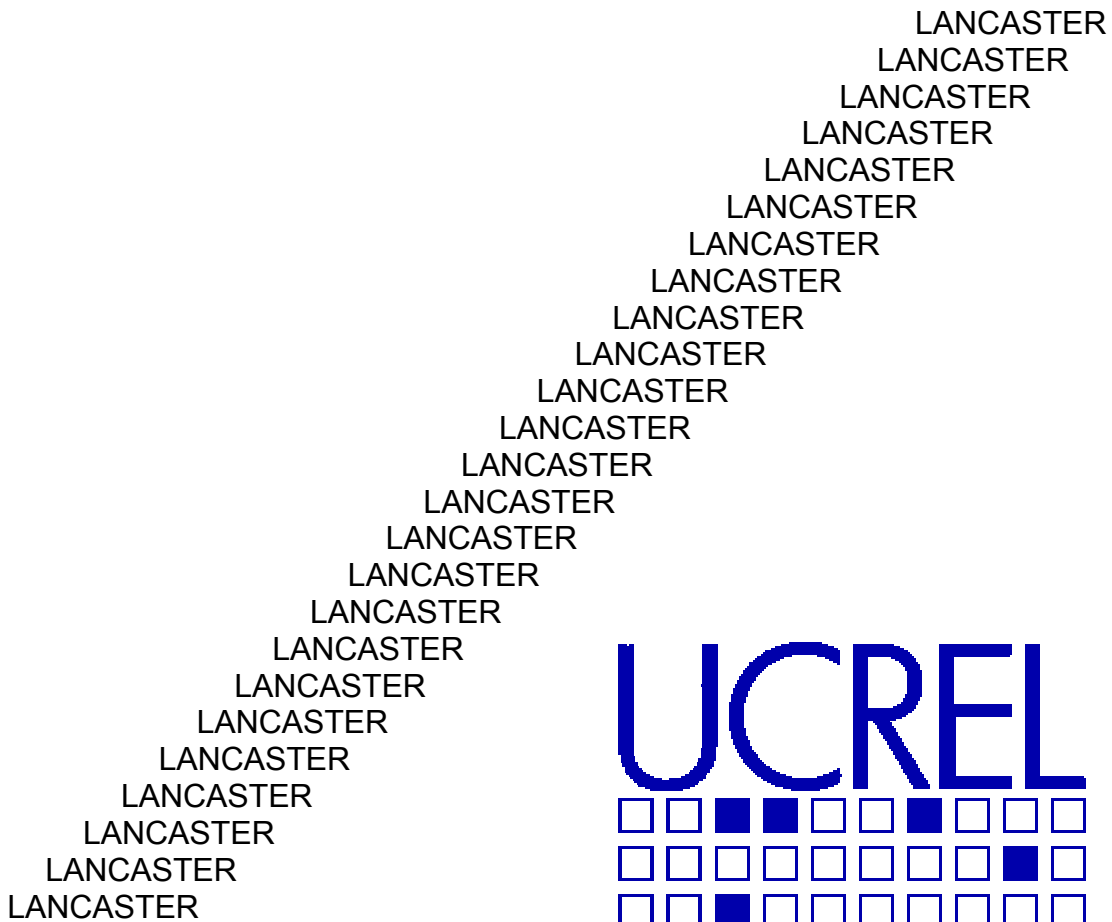
**University Centre
for Computer Corpus
Research on Language
Technical Papers**



**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.



UCREL
Computing Department
Lancaster University
Lancaster
LA1 4YR
United Kingdom
Phone: (+44) 1524 593802
Fax: (+44) 1524 593608
Email: ucrel@lancaster.ac.uk

ISBN 1-86220-134-X
Lancaster University 2003
Further copies may be obtained from
http://www.comp.lancs.ac.uk/ucrel/tech_papers.html

**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	
<i>A CG & PSG Hybrid Approach to Automatic Corpus Annotation</i>	1
Alexander Clark	
<i>Pre-processing Very Noisy Text</i>	12
Mark Davies	
<i>Relational n-gram databases as a basis for unlimited annotation on large corpora</i>	23
Stefan Evert, Hannah Kermes	
<i>Annotation, Storage, and Retrieval of Mildly Recursive Structures</i>	34
Xunlei Rose Hu and Eric Atwell	
<i>A Survey of Machine Learning Approaches to Analysis of Large Corpora</i>	45
Adam Kilgarriff	
<i>Linguistic Search Engine</i>	53
Borja Navarro, Montserrat Civit, M ^a Antonia Martí, Raquel Marcos, Belén Fernández	
<i>Syntactic, semantic and pragmatic annotation in Cast3LB</i>	59
Roman Ondruška, Jarmila Panevová, Jan Štěpánek	
<i>An Exploitation of the Prague Dependency Treebank: A Valency Case</i>	69
Petya Osenova and Kiril Simov	
<i>Between Chunk Ideology and Full Parsing Needs</i>	78
Veronika Reznícková	
<i>Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora</i>	88

Author Index

Eric Atwell	45
Eckhard Bick	1
Montserrat Civit	59
Alexander Clark	12
Mark Davies	23
Stefan Evert	34
Jan Štěpánek	69
Belén Fernández	59
Xunlei Rose Hu	45
Hannah Kermes	34
Adam Kilgarriff	53
Raquel Marcos	59
M ^a Antonia Martí	59
Borja Navarro	59
Roman Ondruška	69
Petya Osenova	78
Jarmila Panevová	69
Veronika Rezníčková	88
Kiril Simov	78