

**UNIVERSITY CENTRE FOR COMPUTER
CORPUS RESEARCH ON LANGUAGE**

Technical Papers

Volume 13 - Special issue.

**Proceedings of the Corpus Linguistics 2001
conference**

edited by

**Paul Rayson,
Andrew Wilson,
Tony McEnery,
Andrew Hardie
and
Shereen Khoja.**

ISBN 1 86220 107 2.

Lancaster University (UK), 29 March - 2 April 2001

Table of contents

Preface	vi
Bas Aarts, Evelien Keizer, Mariangela Spinillo & Sean Wallis: <i>Which or what?</i> A study of interrogative determiners in present-day English	1
Anne Abeillé, Lionel Clément, Alexandra Kinyon & François Toussenet: The TALANA annotated corpus for French : some experimental results	2
Annelie Ädel: On the search for metadiscourse units	3
Karin Aijmer: Discourse particles in contrast	13
Takanobu Akiyama: <i>John is a man of (good) vision:</i> enrichment with evaluative meanings	14
Jean-Yves Antoine & Jérôme Goulian: Word order variations and spoken man-machine dialogue in French : a corpus analysis on the ATIS domain	22
Dawn Archer and Jonathan Culpeper: Sociopragmatic annotation: New directions and possibilities in Historical Corpus Linguistics	30
Eric Atwell & John Elliott: A Corpus for Interstellar Communication	31
Manuel Barbera: From EAGLES to CT tagging: a case for re-usability of resources	40
Margareta Westergren Axelsson & Angela Hahn: The use of the progressive in Swedish and German advanced learner English: a corpus-based study	45
Anja Belz: Optimisation of corpus-derived probabilistic grammars	46
Ylva Berglund & Oliver Mason: "But this formula doesn't mean anything!"	58
Roumiana Blagoeva: Comparing cohesive devices: a corpus based analysis of conjunctions in written and spoken learner discourse	59
Hans Boas: Frame Semantics as a framework for describing polysemy and syntactic structures of English and German motion verbs in contrastive computational lexicography	64
Rhonwen Bowen: Nouns and Their Prepositional Phrase Complements in English	74
Ted Briscoe: From dictionary to corpus to self organising dictionary: Learning valency associations in the face of variation and change	79
Estelle Campione & Jean Véronis: Semi-automatic tagging of intonation in French spoken corpora	90
Pascual Cantos-Gomez: An attempt to improve current collocation analysis	100
Roldano Cattoni, Morena Danieli, Andrea Panizza, Vanessa Sandrini & Claudia Soria: Building a corpus of annotated dialogues: the ADAM experience	109
Frantisek Cermák, Jana Klimová, Karel Pala, Vladimír Petkevič: The Design of Czech Lexical Database	119
Ngoni Chipere, David Malvern, Brian Richards & Pilar Duran: Using a corpus of school children's writing to investigate the development of vocabulary diversity	126
Claudia Claridge: Approaching Irony in Corpora	134
Niladri Sekhar Dash and Bidyut Baran Chaudhuri: Corpus based Empirical Analysis of Form, Function and Frequency of Characters used in Bangla	144
Liesbeth Degand & Henk Pander Maat: Contrasting causal connectives on the Speaker Involvement Scale	158
Anne Le Draoulec & Marie-Paule Péry-Woodley: Corpus based identification of temporal organisation in discourse	159
Stefan Evert & Anke Lüdeling: Measuring morphological productivity: Is automatic preprocessing sufficient?	167

Cécile Fabre & Didier Bourigault: Linguistic clues for corpus-based acquisition of lexical dependencies	176
Richard Foley: Going out in style? <i>Shall</i> in EU legal English	185
Anna-Lena Fredriksson: Translating passives in English and Swedish: a text-linguistic perspective	196
Cécile Frérot, Géraldine Rigou & Annik Lacombe: Phraseological approach to automatic terminology extraction from a bilingual aligned scientific corpus	204
Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough, Scott Piao: The METER Corpus: A corpus for analysing journalistic text reuse	214
Hatem Ghorbel, Afzal Ballim, & Giovanni Coray: ROSETTA: Rhetorical and semantic environment for text alignment	224
Solveig Granath: Is that a fact? A corpus study of the syntax and semantics of <i>the fact that</i>	234
Benoît Habert, Natalia Grabar, Pierre Jacquemart, Pierre Zweigenbaum: Building a text corpus for representing the variety of medical language	245
Eva Hajicova & Petr Sgall: A reusable corpus needs syntactic annotations: Prague Dependency Treebank	255
David Hardcastle: Using the BNC to produce dialectic cryptic crossword clues	256
Daniel Hardt: Comma Checking in Danish	266
Raymond Hickey: Tracking lexical change in present-day English	272
Diana Hudson-Ettle, Tore Nilsson & Sabine Reich: Orality and noun phrase complexity: a corpus-based study of British and Kenyan writing in English	273
Nancy Ide & Catherine Macleod: The American National Corpus: A Standardized Resource for American English	274
Reiko Ikeo: The Positions of the Reporting Clauses of Speech Presentation with Special Reference to the Lancaster Speech, Thought and Writing Presentation Corpus	281
Wang JianDe, Chen ZhaoXiong & Huang HeYan: Multiple-Level Knowledge Discovery from Corpus	289
Yu Jiangsheng & Duan Huiming: POS Estimation of Undefined Chinese Words	290
Steven Jones: Corpus Approaches to Antonymy	297
Beom-mo Kang & Hung-gyu Kim: Variation across Korean Text Registers	311
Przemyslaw Kaszubski: Tracing idiomaticity in learner language - the case of BE.	312
Yuliya Katsnelson & Charles Nicholas: Identifying Parallel Corpora Using Latent Semantic Indexing	323
Hannah Kermes & Stefan Evert: Exploiting large corpora: A circular process of partial syntactic analysis, corpus query and extraction of lexicographic information	332
Shereen Khoja, Roger Garside & Gerry Knowles: A Tagset for the Morphosyntactic Tagging of Arabic	341
Adam Kilgarriff: Web as corpus	342
Dimitrios Kokkinakis: A Long-Standing Problem in Corpus-Based Lexicography and a Proposal for a Viable Solution	345
Julia Lavid: Using bilingual corpora for the construction of contrastive generation grammars: issues and problems	356
Maarten Lemmens: Tracing referent location in oral picture descriptions	367
Barbara Lewandowska-Tomaszczyk, Michael Oakes & Paul Rayson: Annotated Corpora for Assistance with English-Polish Translation	368
Juana Marín-Arrese, Elena Martínez-Caro & Soledad Pérez de Ayala Becerril: A corpus study of impersonalization strategies in newspaper discourse in English & Spanish	369

Mikhail Mikhailov & Miia Villikka: Is there such a thing as a translator's style?	378
Hermann Moisl & Joan Beal: Corpus analysis and results visualisation using self-organizing maps	386
Martina Möllering: Pragmatic and discursive aspects of German modal particles: a corpus-based approach	392
Rachel Muntz: Evidence of Australian cultural identity through the analysis of Australian and British corpora	393
P-O Nilsson: Investigating characteristic lexical distributions and grammatical patterning in Swedish texts translated from English	400
Julien Nioche & Benoît Habert: Using feature structures as a unifying representation format for corpora exploration	401
Matthew Brook O'Donnell, Stanley E. Porter & Jeffrey T. Reed: OpenText.org: the problems and prospects of working with ancient discourse	413
Maeve Olohan: Spelling out the Optionals in Translation: A Corpus Study	423
Constantin Orasan: Patterns in scientific abstracts	433
Gabriella Kiss and Júlia Pajzs: An attempt to develop a lemmatiser for the Historical Corpus of Hungarian	443
Byungsun Park & Beom-mo Kang: Korean grammatical collocation of predicates and arguments	452
Anselmo Peñas, Felisa Verdejo & Julio Gonzalo: Corpus-Based Terminology Extraction Applied to Information Access	458
Scott Songlin Piao & Tony McEnery: Multi-word unit alignment in English-Chinese parallel corpora	466
Katja Ploog: Syntactic change in abidjane French	476
Sylvie Porhiel: Linguistic expressions as a tool to extract information	477
Stanley E. Porter & Matthew Brook O'Donnell: Theoretical Issues for Corpus Linguistics Raised by the Study of Ancient Languages	483
Helena Raumolin-Brunberg: Temporal aspects of language change: what can we learn from the CEEC?	484
Andrea Reményi: Use logbooks and find the original meaning of representativeness	485
Antoinette Renouf: The Web as a Source of Linguistic Information	492
Jérôme Richalot: The influence of the passive on text cohesion and technical terminology	493
Rema Rossini Favretti, Fabio Tamburini & Cristiana De Santis : CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model	512
Hans-Jörg Schmid: Do women and men really live in different cultures? Evidence from the BNC	513
Josef Schmied: Exploring the Chemnitz Internet Grammar: Examples of student use	514
Mark Sebba & Susan Dray: Is it Creole, is it English, is it valid? Developing and using a corpus of unstandardised written language	522
Mirjam Sepesy Maucec & Zdravko Kacic: Language Model Adaptation For Highly-Inflected Slovenian Language In Comparison To English Language	523
Noëlle Serpollet: The mandative subjunctive in British English seems to be alive and kicking... Is this due to the influence of American English?	531
Serge Sharoff: Through the looking glass of parallel texts	543
Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, Atanas Kiryakov: CLaRK - an XML based system for corpora development	553
Kiril Simov, Gergana Popova & Petya Osenova: HPSG-based syntactic TreeBank of Bulgarian (BulTreeBank)	561
Simon Smith & Martin Russell: Determining query types for information access	562

Josef Szakos & Amy Wang: Not last, even if least: Endangered Formosan aboriginal languages and the corpus revolution	571
Elke Teich & Silvia Hansen: Methods and techniques for a multi-level analysis of multilingual corpora	572
Dan Tufis & Ana-Maria Barbu: Accurate automatic extraction of translation equivalents from parallel corpora	581
Tamás Váradi: The Linguistic Relevance of Corpus Linguistics	587
Serge Verlinde & Thierry Selva: Corpus-based vs intuition-based lexicography. Defining a word list for a French learner's dictionary	594
Jean Véronis: Sense tagging: does it make sense ?	599
Adriana Vlad, Adrian Mitrea, & Mihai Mitrea: A Corpus-Based Analysis of How Accurately Printed Romanian Obeys to Some Universal Laws	600
Martin Volk: Exploiting the WWW as a corpus to resolve PP attachment ambiguities	601
Martin Weisser: A corpus-based methodology for comparing and evaluating different accents	607
Anne Wichmann & Richard Cauldwell: Wh-Questions and attitude: the effect of context	614
Kay Wikberg: His breath a thin winter-whistle in his throat: English metaphors and their translation into Scandinavian languages	615
Karen Wu Rongquan: Public Discourse as the mirror of ideological change: A keyword study of editorials in People's Daily	616
Richard Xiao Zhonghua: A Corpus-Based Study of Interaction Between Chinese Perfective -le and Situation Types	625

Preface

All the papers in this collection are based upon talks or poster presentations given at the Corpus Linguistics (CL2001) conference, held at Lancaster University between 29th March and 2nd April 2001 organised by members of UCREL, from the Departments of Linguistics & Modern English Language and Computing.

The conference attracted over 100 participants from the language engineering and corpus linguistics communities in over 20 countries world-wide. The presentations represented a truly impressive rainbow of languages in corpus research, ranging from ancient languages, eastern and western European languages, to Semitic and Asian languages.

One of the aims of CL2001 was to celebrate the works of Geoffrey Leech who reached 65 in 2001. Geoffrey, a pioneer in the construction and exploitation of machine-readable corpora, has had considerable influence in many areas of linguistics over the years. A selection of papers from the conference will appear in an edited collection to be published in honour of Geoffrey Leech.

Paul Rayson
Andrew Wilson
Tony McEnery
Andrew Hardie
Shereen Khoja

Lancaster University, March 2001.