

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

**Technical Papers**

**Volume 16 - Special issue.**

**Proceedings of the  
Corpus Linguistics 2003 conference**

**edited by**

**Dawn Archer,  
Paul Rayson,  
Andrew Wilson,  
and  
Tony McEnery.**

**ISBN 1 86220 131 5.**

**Lancaster University (UK), 28 - 31 March 2003**

# Table of contents

<b>Preface</b>	viii
<b>Mariko Abe</b> : A Corpus-based Contrastive Analysis of Spoken and Written Learner Corpora: The Case of Japanese-speaking Learners of English	1
<b>Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., and Urizar R.</b> : Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing	10
<b>Khurshid Ahmad, Pensiri Manomaisupat, David Cheng, Tugba Taskaya, Saif Ahmad, Lee Gillam, Andrew Hippisley</b> : The mood of the (financial) markets: In a corpus of words and of pictures	12
<b>Sandra M. Aluísio, Gisele M. Pinheiro, Marcelo Finger, Maria das Graças V. Nunes, Stella E. O. Tagnin</b> : The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation	14
<b>Dawn Archer, Tony McEnery, Paul Rayson, Andrew Hardie</b> : Developing an automated semantic analysis system for Early Modern English	22
<b>Dawn Archer, Andrew Hardie, Tony McEnery, Scott Piao</b> : A corpus of seventeenth-century English news reportage: construction, encoding and applications	32
<b>Bertol Arrieta, Arantza Díaz de Ilarraza, Koldo Gojenola, Montse Maritxalar, Maite Oronoz</b> : A database system for storing second language learner corpora	33
<b>Jørg Asmussen</b> : Towards a methodology for corpus-based studies of linguistic change: Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 General Corpora of Danish	42
<b>Eric Atwell</b> : A New Machine Learning Algorithm for Neoposy: coining new Parts of Speech	43
<b>Eric Atwell, Paul Gent, Julia Medori, Clive Souter</b> : Detecting student copying in a corpus of science laboratory reports: simple and smart approaches	48
<b>Francis Henrik Aubert, Stella E. O. Tagnin</b> : A Corpus of Sworn Translations – for linguistic and historical research	54
<b>Bogdan Babych, Anthony Hartley, Eric Atwell</b> : Statistical modelling of MT output corpora for Information Extraction	62
<b>Paul Baker, Andrew Hardie, Tony McEnery, and Sri B.D. Jayaram</b> : Constructing Corpora of South Asian Languages	71
<b>Federica Barbieri</b> : The “new” quotatives in American English: A cross-register comparison	81
<b>Marco Baroni and Silvia Bernardini</b> : A preliminary analysis of collocational differences in monolingual comparable corpora	82
<b>Sabine Bartsch</b> : Investigating cross-linguistic constraints on the premodification of adjectival past participles and desubstantival adjectives. A corpus-based study of English and German	92
<b>Kate Beeching</b> : Synchronic and diachronic variation: the how and why of sociolinguistic corpora.	102
<b>Luisa Bentivogli, Christian Girardi, Emanuele Pianta</b> : The MEANING Italian Corpus	103
<b>Julie Carson-Berndsen, Ulrike Gut and Robert Kelly</b> : Discovering regularities in non-native speech	113

<b>P. Beust, S. Ferrari, V. Perlerin:</b> NLP model and tools for detecting and interpreting metaphors in domain-specific corpora	114
<b>Philippe Blache, Marie-Laure Guénot and Tristan van Rullen:</b> A corpus-based technique for grammar development	124
<b>Birte Bös:</b> Towards an integrated model of service encounters	132
<b>Roderick Bovington and Angelo Dalli:</b> Statistical analysis of the source origin of Maltese	140
<b>Lou Burnard, Tony Dodd:</b> Xara: an XML aware tool for corpus searching	142
<b>Marianna N. Christou:</b> Expressions and structures of the delexical verb ΚΑΝΩ [“MAKE” / “DO”] in Modern Greek language: A corpus-based approach to newspaper articles	145
<b>Ken Cosh and Pete Sawyer:</b> Using natural language processing tools to assist semiotic analysis of information systems	155
<b>H. Cunningham, V. Tablan, K. Bontcheva, M. Dimitrov:</b> Language engineering tools for collaborative corpus annotation	165
<b>Mark Davies:</b> Annotation without lexicons: an alternative to the standard bootstrapping approach	174
<b>Joost van de Weijer:</b> Consonant variation within words	184
<b>Debbie Elliott, Anthony Hartley and Eric Atwell:</b> Rationale for a multilingual corpus for machine translation evaluation	191
<b>John Elliott and Debbie Elliott:</b> The Human Language Chorus Corpus (HULCC)	201
<b>Jens Fauth, Hans-Jörg Schmid:</b> Detecting gender-preferential patterns of linguistic features in face-to-face communication	211
<b>Valéria D. Feltrim, Sandra M. Aluísio, Maria das Graças V. Nunes:</b> Analysis of the rhetorical structure of computer science abstracts in Portuguese	212
<b>Katerina T. Frantzi:</b> Updating LSP dictionaries with collocational information	219
<b>Robert Gaizauskas, Lou Burnard, Paul Clough and Scott Piao:</b> Using the XARA XML-Aware Corpus Query Tool to Investigate the METER Corpus	227
<b>Ana Llinares García:</b> Repetition and young learners’ initiations in the L2: a corpus-driven analysis	237
<b>Sandrine Garnier, Youhanizou Tall, Sisay Fissaha, Johann Haller:</b> Learner Corpora: Design, Development and Applications - Development of NLP tools for CALL based on learner corpora (German as a foreign language)	246
<b>Sara Gesuato:</b> The company <i>women</i> and <i>men</i> keep: what collocations can reveal about culture	253
<b>Vojko Gorjanc:</b> Tracking lexical changes in the reference corpus of Slovene texts	263
<b>Stefan Grondelaers, Dirk Speelman, Dirk Geeraerts:</b> A corpus-based approach to informality: the case of Internet chat	264
<b>Leif Grönqvist and Magnus Gunnarsson :</b> A method for finding word clusters in spoken language	265
<b>Xiaotian Guo:</b> Between Verbs and Nouns and Between the Base Form and the Other Forms of Verbs – A Contrastive Study into COLEC and LOCNESS	274
<b>Le An Ha:</b> A method for word segmentation in Vietnamese	282
<b>Silvia Hansen-Schirra:</b> Linguistic enrichment and exploitation of the Translational English Corpus	288

<b>Andrew Hardie:</b> Developing a tagset for automated part-of-speech tagging in Urdu	298
<b>Nigel Harwood:</b> Personal pronouns and academic writing: a multidisciplinary corpus-based critical pragmatic approach to EAP	308
<b>Laura Hasler, Constantin Orasan and Ruslan Mitkov:</b> Building better corpora for summarisation	309
<b>Chris Heffer:</b> Not KWIC but Quick: KeyWords in Court	319
<b>Kris Heylen and Dirk Speelman:</b> A corpus-based analysis of word order variation: The order of verb arguments in the German middle field	320
<b>Knut Hofland:</b> A web-based concordance system for spoken language corpora	330
<b>Shelley Ching-yu Hsieh:</b> The Corpus of Mandarin Chinese and German Animal Expressions	332
<b>Susan Hunston:</b> Frame, phrase or function: a comparison of frame semantics and local grammars	342
<b>Emi Izumi, Toyomi Saiga, Thepchai Supnithi, Kiyotaka Uchimoto, Hitoshi Isahara:</b> The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques	359
<b>Inés Jacob, Joseba Abaitua, Josu Gómez:</b> Automatic feeding of translation memory tools	367
<b>Steven Jones, M. Lynne Murphy:</b> Antonymy in Childhood: a corpus-based approach to acquisition	372
<b>Randall L. Jones:</b> An Analysis of Lexical Text Coverage in Contemporary German	373
<b>Stig W. Jørgensen, Carsten Hansen, Jette Drost, Dorte Haltrup, Anna Braasch, Sussi Olsen:</b> Domain specific corpus building and lemma selection in a computational lexicon	374
<b>Tomoko Kaneko:</b> How non-native speakers express anger, surprise, anxiety and grief: a corpus-based comparative study	384
<b>Sachie Karasawa:</b> Patterns of elaboration and interlanguage development: an exploratory corpus analysis of college student essays	394
<b>Hannah Kermes, Stefan Evert:</b> Text analysis meets corpus linguistics	402
<b>Adam Kilgarriff:</b> Linguistic Search Engine	412
<b>Paul Kingsbury:</b> A methodology for inducing a chronology of the Pā li Canon	413
<b>Gerry Knowles, Zuraidah Mohd Don:</b> Tagging a corpus of Malay texts, and coping with 'syntactic drift'	422
<b>Natalie Kübler and Cécile Frérot:</b> Verbs in specialised corpora: from manual corpus-based description to automatic extraction in an English-French parallel corpus	429
<b>Toshihiko Kubota:</b> A Study on Abridgement for Spoken Word Titles	439
<b>David YW Lee:</b> Spoken Academic Lexicogrammar and Discourse Patterns	440
<b>Geoffrey Leech, Martin Weisser:</b> Generic speech act annotation for task-oriented dialogues	441
<b>Agnieszka Lenko-Szymanska:</b> The curse and the blessing of mobile phones - a corpus-based study into Polish and American rhetoric strategies	447
<b>Robert Liebscher and David Groppe:</b> Rethinking context availability for concrete and abstract words: a corpus study	449
<b>Laura Löfberg, Dawn Archer, Scott Piao, Paul Rayson, Tony McEnery, Krista Varantola, Jukka-Pekka Juntunen:</b> Porting an English semantic tagger to the Finnish language	457

<b>Nadine Lucas, Bruno Crémilleux, Leny Turmel</b> : Signalling well-written academic articles in an English corpus by text mining techniques	465
<b>Anke Lüdeling and Stefan Evert</b> : Linguistic experience and productivity: corpus evidence for fine-grained distinctions	475
<b>Michaela Mahlberg</b> : High frequency nouns in English: aspects of a grammatical description	484
<b>Belinda Maia</b> : Constructing comparable and parallel corpora for terminology extraction - work in progress	485
<b>Manolis Maragoudakis, Katia Kermanidis and Nikos Fakotakis</b> : Towards a Bayesian Stochastic Part-Of-Speech and Case Tagger of Natural Language Corpora	486
<b>Kevin Mark</b> : Learner corpus building and a ‘living’ university foreign language curriculum	496
<b>Tony McEnery, Zhonghua Xiao</b> : Fuck revisited	504
<b>Dan McIntyre, Carol Bellard-Thomson, John Heywood, Tony McEnery, Elena Semino and Mick Short</b> : The Construction of a Corpus to Investigate the Presentation of Speech, Thought and Writing in Written and Spoken British English	513
<b>John McKenny</b> : Seeing the wood and the trees: Reconciling findings from discourse and lexical analysis	523
<b>Magnus Merkel, Michael Petterstedt and Lars Ahrenberg</b> : Interactive Word Alignment for Corpus Linguistics	533
<b>José María Guirao Miras Ana González Ledesma, Guillermo de la Madrid Heitzmann, Manuel Alcántara Plá, Antonio Moreno Sandoval</b> : Relating lexical items to sociolinguistic features in a spontaneous speech corpus of Spanish	543
<b>Juan M. Montero and M. Mar Duque</b> : ANESTTE: a writer’s assistant for a specific purpose language	544
<b>Olga Moudraia</b> : The Student Engineering Corpus: Analysing Word Frequency	552
<b>JoAnne Neff, Francisco Ballesteros, Emma Dafouz, Francisco Martínez, Juan-Pedro Rica</b> : Formulating Writer Stance: A Contrastive Study of EFL Learner Corpora	562
<b>Diane Nicholls</b> : The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT	572
<b>Judy Noguchi, Thomas Orr, Yukio Tono</b> : Using a dedicated corpus to identify features of professional English usage: What do “we” do in science journal articles?	582
<b>Attila Novák, Viktor Nagy, Csaba Oravecz</b> : Corpus assisted development of a Hungarian morphological analyser and guesser	583
<b>Toshifumi Oba and Eric Atwell</b> : Using the HTK speech recogniser to analyse prosody in a corpus of German spoken learners’ English	591
<b>Marija Omazic</b> : THE METACOMMUNICATIVE SETTING OF PHRASEOLOGICAL UNITS AND THEIR MODIFICATIONS – EVIDENCE FROM THE BRITISH NATIONAL CORPUS	599
<b>Nelleke Oostdijk</b> : Corpus linguistics meets language technology: deep syntactic parsing for question answering	603
<b>Maeve Paris</b> : Extending computer-assisted text analysis techniques to the detection of source code plagiarism and collusion: assisting manual inspection	611
<b>Núria Gala Pavia, Salah Aït-Mokhtar</b> : Lexicalising a robust parser grammar using the WWW	620

<b>Julien Perrez and Liesbeth Degand:</b> On the combination of corpus-based and experimental methodologies in the study of causal, contrastive and metadiscourse connectives in L1 and L2 text comprehension and production	627
<b>Scott S.L. Piao and Tony McEnery:</b> A Tool for Text Comparison	637
<b>James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro and Marcia Lazo:</b> The TIMEBANK Corpus	647
<b>Andrew Roberts and Eric Atwell:</b> The use of corpora for automatic evaluation of grammar inference systems	657
<b>Juhani Rudanko:</b> More on <i>horror aequi</i> : evidence from large corpora	662
<b>Sarah Rule, Emma Marsden, Florence Myles, Rosamond Mitchell:</b> Constructing a database of French interlanguage oral corpora	669
<b>Geoffrey Sampson:</b> Are we nearly there yet, Mum?	678
<b>Hans-Jörg Schmid, Jens Fauth:</b> Women's and men's style: fact or fiction? New grammatical evidence	679
<b>Serge Sharoff:</b> Methods and tools for development of the Russian Reference Corpus	680
<b>Bayan Abu Shawar and Eric Atwell :</b> Using dialogue corpora to train a chatbot	681
<b>Gerardo Sierra, Alfonso Medina, Rodrigo Alarcón, César A. Aguilar:</b> Towards the Extraction of Conceptual Information from Corpora	691
<b>Kiril Simov, Alexander Simov, Milen Kouylekov:</b> Constraints for corpora development and validation	698
<b>Milena Slavecheva:</b> Corpus shallow parsing: meeting point between paradigmatic knowledge encoding	706
<b>Nicholas Smith:</b> A quirky progressive? A corpus-based exploration of the <i>will + be + -ing</i> construction in recent and present day British English.	714
<b>Harold Somers:</b> Some Issues in the Mark-up of Handwriting in a Learner Corpus	724
<b>Dirk Speelman, Stefan Grondelaers, Dirk Geeraerts:</b> A profile-based calculation of region and register variation: the synchronic and diachronic status of the national variants of Dutch	733
<b>Somayajulu G. Sripada and Ehud Reiter and Jim Hunter and Jin Yu:</b> Exploiting a parallel TEXT - DATA corpus	734
<b>Asa M. Stepak:</b> A proposed mathematical theory explaining the sequence of grammatical categories	744
<b>Petra Storjohann:</b> The lexicographic use of corpora and computational tools for disambiguation	754
<b>Jozsef Szakos:</b> Cultures and Corpora: Extracting Anthropological Information from Corpora of Formosan Endangered Languages	763
<b>Jun Arata Takahashi :</b> Do we talk (or write?) differently over the Net?- A lexical enquiry into ‘a’ Net-EN -	764
<b>Kaoru Takahashi:</b> A Study of Text Types and Register Variation in the British National Corpus	773
<b>Yuri Tambovtsev:</b> The Structure of the Consonant Patterns in the Spanish Speech Sound Chain as a Clue of Typological Closeness	774

<b>Yuri Tambovtsev:</b> Phonological similarity between Basque and other world languages based on the frequency of occurrence of certain typological consonantal features	775
<b>Tess Yu-Shan Ke, Liang-Feng Chen, Chien-Chung Chen:</b> Investigation on the uses of temporal subordinators by NS and NNS in academic spoken English	780
<b>Carole Tiberius, Dunstan Brown, Greville Corbett:</b> Ambiguity in Russian Morphology	790
<b>Juhani Toivanen, Tapio Seppänen, Eero Väyrynen:</b> Creation and utilisation of the MediaTeam Emotional Speech Corpus	791
<b>Yukio Tono:</b> Learner corpora: design, development and applications	800
<b>Montserrat Civit Torruella, M<sup>a</sup> Antònia Martí Antonín, Lluís Padró Cirera :</b> Using hybrid probabilistic-linguistic knowledge to improve pos-tagging performance	810
<b>Patrick Tschorn, Anke Lüdeling:</b> Morphological knowledge and alignment of English-German parallel corpora	818
<b>Francesca Vaghi, Marco Venuti:</b> The Economist and The Financial Times. A study of movement metaphors	828
<b>Bertus van Rooy and Lande Schäfer:</b> An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus	835
<b>Tamás Váradi:</b> Shallow parsing of Hungarian business news	845
<b>Isabel Verdaguer and Anna Poch:</b> Collocational and colligational patterns in lexical sets: A corpus-based study	852
<b>Maria Verde:</b> Shedding light on SHED, CAST and THROW as nodes of extended lexical units	859
<b>Shih-Ping Wang:</b> Mutual information and corpus-based approaches to reduplicative fixed expressions	869
<b>Julie Weeds and David Weir:</b> Finding and evaluating sets of nearest neighbours	879
<b>David Wible, Ping-Yu Huang:</b> Using learner corpora to examine L2 acquisition of tense-aspect markings	889
<b>Sandra Williams and Ehud Reiter:</b> A corpus analysis of discourse relations for Natural Language Generation	899
<b>Andrew Wilson, Celia Worth:</b> Building and annotating corpora of spoken Welsh and Gaelic	909
<b>Andrew Wilson, Celia Worth:</b> Conceptual Glossaries of the Latin Vulgate Bible	918
<b>Andrew Wilson, Olga Moudraia:</b> Quantitative or Qualitative Content Analysis? Experiences from a cross-cultural comparison of female students' attitudes to shoe fashions in Germany, Poland and Russia	919
<b>Martin Wynne, Rowan Wilson, Ylva Berglund:</b> Virtual Corpora at the Oxford Text Archive	920
<b>Yang Xiaojun:</b> Survey and Prospect of China's Corpus-Based Researches	930
<b>Debra Ziegeler, Sarah Lee:</b> Analysing a Corpus-based Semantic Investigation of English Dialects	931
<b>Heike Zinsmeister, Ulrich Heid:</b> Identifying predicatively used adverbs by means of a statistical grammar model	932

## **Preface**

The papers in this collection are based upon talks or poster presentations given at the Corpus Linguistics (CL2003) conference, held at Lancaster University between 28<sup>th</sup> and 31<sup>st</sup> March 2003 organised by members of UCREL, from the Departments of Linguistics & Modern English Language and Computing. The proceedings also includes those papers presented at the pre-conference workshop “Learner Corpora: Design, Development and Applications” organised by Yukio Tono and Fanny Meunier. The three other workshops produced separate proceedings.

The conference attracted over 150 participants from the language engineering and corpus linguistics communities from many countries. The previous conference in the series, CL2001, was also held at Lancaster University in March 2001.

Dawn Archer  
Paul Rayson  
Andrew Wilson  
Tony McEnery

Lancaster University, March 2003.