English corpus linguistics on the move: Applications and implications
Organizing committee

María José López-Couso
Belén Méndez-Naya
Mario Cal-Varela
Teresa Fanego
Xabier Fernández-Polo
Paloma Núñez-Pertejo
Ignacio Palacios-Martínez

Student helpers

Cristina Blanco-García
Zeltia Blanco-Suárez
Tamara Bouso
Yolanda Joy Calvo-Benzies
Eduardo Coto-Villalibre
Tania de Dios
Fátima Faya
Susana Formoso
Vanesa Gil-Vilacoba
Lidia Gómez-García
Aleksandra Kaverina
Tamilla Mamedova
Beatriz Mato-Míguez
Alberto Monteagudo-Buceta
Noelia Ozares-Reboiras
Alba Pérez-González
María Luisa Roca-Varela
Paula Rodríguez-Abruñeiras
Iria-Gael Romay
Mario Serrano-Losada
Iván Tamaredo-Meira
Vera Vázquez-López

Scientific committee

Joan Beal (Sheffield)  Julia Lavid (Complutense de Madrid)
Douglas Biber (Northern Arizona)  Hans Lindquist (Malmö)
Hubert Cuyckens (Leuven)  Christian Mair (Freiburg)
Holger Diessel (Jena)  Gabriella Mazzon (Innsbruck)
Marina Dossena (Bergamo)  Britta Mondorf (Mainz)
Sebastian Hoffmann (Trier)  Terttu Nevalainen (Helsinki)
Marianne Hundt (Zurich)  Dirk Nöel (Hong Kong)
Merja Kytö (Uppsala)  Aquilino Sánchez Pérez (Murcia)

and the members of the ICAME Executive Board

Gisle Andersen (Bergen)  Fanny Meunier (Louvain-la-Neuve)
Kristin Davidse (Leuven)  Ilka Mindt (Paderborn)
Stefan Th. Gries (Santa Barbara)  Joybrato Mukherjee (Giessen)
Hilde Hasselgård (Oslo)  Pam Peters (Sydney)
Magnus Huber (Giessen)  Paul Rayson (Lancaster)
John M. Kirk (Belfast)  Ute Römer (Atlanta)
Maria José López-Couso (Santiago)  Irma Taavitsainen (Helsinki)
Michaela Mahlberg (Nottingham)  Sali Tagliamonte (Toronto)
ICAME 34: SPONSORS

Financial support for this event has been received from:

Grant no. FFI2011-26693-C02-01: Ministry of Economy and Competitiveness; Principal Investigator: María José López-Couso.

Institutional sponsors:

Research Unit for Variation, Linguistic Change and Grammaticalization, University of Santiago de Compostela (Ref. GI-1383)

Spoken English Research Team at the University of Santiago de Compostela (Ref. GI-1762)

Publishers:
Benvida

Dear ICAME 34 participants,

Last year in Leuven we invited you to take the pilgrim’s way to Santiago de Compostela. Now we would like to welcome you warmly to ICAME 34 (22-26 May 2013) and to our old and beautiful town in hopes that you will profit from the very promising conference sessions. ICAME 34 will confirm that English corpus linguistics is indeed on the move and will show its various and wide applications and implications. Your interest and contributions are essential to making this an inspiring and stimulating event.

We also encourage you to find some spare time here and there and do some sightseeing: to enjoy the streets and people, our food and wine, and the friendly and lively atmosphere in Santiago. We do hope that the weather will be “merciful” to us. Just in case, remember the fond old saying that “rain is art in Santiago”.

We wish you an enjoyable conference,

The ICAME 34 organizing committee

María José López-Couso, Belén Méndez-Naya, Mario Cal-Varela, Teresa Fanego, Xabier Fernández-Polo, Paloma Núñez-Pertejo, Ignacio Palacios-Martínez
TABLE OF CONTENTS

Benvida .......................................................................................................................................iii

PLENARY SPEAKERS

Multiethnolects and English corpus linguistics
Jenny Cheshire (Queen Mary, University of London) ................................................................. 15

The network metaphor of usage-based construction grammar
Holger Diessel (Friedrich Schiller University Jena) ................................................................. 15

Who is the/a/Ø professor of English at your university?
Marianne Hundt (University of Zurich) .................................................................................. 16

Community and identity: What corpora can tell us about academic discourse
Ken Hyland (University of Hong Kong) .................................................................................. 17

PRE-CONFERENCE WORKSHOPS

WORKSHOP 1: CROSS-LINGUISTIC STUDIES AT THE INTERFACE BETWEEN LEXIS
AND GRAMMAR
Convenors: Karin Aijmer (University of Gothenburg) and Hilde Hasselgård (University of Oslo)

Modal particles in a contrastive perspective – the case of the Swedish väl
Karin Aijmer (University of Gothenburg) ............................................................................... 21

I am wild about cabbage: Evaluative ‘semantic sequences’ and cross-linguistic (dis)continuities
Marina Bondi & Corrado Seidenari (University of Modena and Reggio Emilia).................. 21

Binominal size noun constructions in English and French: A contrastive corpus-based perspective
Lot Brems (University of Liège/K.U. Leuven) ......................................................................... 23

A contrastive analysis of downtoners, more or less
Signe Oksefjell Ebeling & Jarle Ebeling (University of Oslo) ................................................ 24

Motion into and out of in English, French and Norwegian
Thomas Egan & Anne-Line Graedler (Hedmark University College) .................................... 25

The postmodifying structure of noun phrases: A contrastive study of English and Norwegian
Johan Elsness (University of Oslo) ......................................................................................... 26

Cross-linguistic analysis of cohesion: Variation across production types and registers
Ekaterina Lapshinova-Koltunski & Kerstin Kunz (University of Saarland) ......................... 27
Intersubjective positioning and thematisation in English and Spanish: A contrastive analysis of letters to the editor

*Julia Lavid & Lara Moratón (Universidad Complutense de Madrid)* .............................................. 29

A comparable-corpus based approach to the expression of obligation across English and French

*Diana Lewis (University of Aix-Marseille I)* .................................................................................. 30

English non-finite participial clauses as seen through their Czech counterparts

*Markéta Malá & Pavlína Šaldová (Charles University in Prague)* ............................................... 31

Quite seen through its translation equivalents: A contrastive corpus-based study

*Michaela Martinková (Palacky University Olomouc)* ................................................................. 32

English positive polarity contexts into Spanish: A corpus-based study

*Rosa Rabadán (University of León)* .......................................................................................... 34

What Jesus wanted (us) to know

*Teresa Sánchez Roura (University of Santiago de Compostela)* .............................................. 35

Parallel corpus - a tool for diagnosing multifunctionality across languages: *Actually, naturally and in fact vs. their correspondences in Lithuanian*

*Aurelija Usoniene (Vilnius University), Jolanta Sinkuniene (Vytautas Magnus University & Vilnius University) & Audrone Soliene (Vilnius University)* .................................................. 37

**WORKSHOP 2: COMPILATION AND ANNOTATION OF SPOKEN CORPORA: TOWARDS BEST PRACTICE**

_Convenors: Gisle Andersen (NHH Norwegian School of Economics), John Kirk (Belfast) and Susan Lee Nacey (Hedmark University College)_

How do you spell *yeah/yeh/yea/yah*? Assessing orthographic transcription and comparability between spoken corpora

*Gisle Andersen (NHH Norwegian School of Economics)* .......................................................... 39

“*Ja, also, can you see me now?*” Designing and compiling a corpus of computer-mediated international academic English

*Stefan Diemer (Saarland University)* .......................................................................................... 40

From sociolinguistic interviews to a spoken corpus of London English: Creating the Linguistic Innovators Corpus (LIC)

*Costas Gabrielatos (Edge Hill University), Sebastian Hoffmann (University of Trier) & Eivind Torgersen (Sør-Trøndelag University College)* ............................................................... 41

The SPICE-Ireland pragmatic annotation scheme: A critical appraisal

*John M. Kirk (Belfast)* .................................................................................................................. 42

The Norwegian component of LINDSEI

*Susan Lee Nacey (Hedmark University College)* ........................................................................ 43

Categorising plurilingual user data? Challenges and solutions for POS-tagging VOICE

*Ruth Osimk-Teasdale (University of Vienna)* ............................................................................. 44
BeMaTaC: A flexible multilayer spoken dialogue corpus for contrastive SLA analyses
Simon Sauer & Anke Lüdeling (Humboldt University of Berlin) ................................................. 46

Best practices in the compilation, annotation and publication of the Research and Teaching Corpus of Spoken German (FOLK)
Thomas Schmidt (Institut für Deutsche Sprache, Mannheim) .................................................. 47

Curating and maintaining spoken legacy corpora for publication in the scientific community
Kai Wörner (HZSK University of Hamburg) .................................................................................. 48

**WORKSHOP 3: (LEARNER) CORPORA AND THEIR APPLICATION IN LANGUAGE TESTING AND ASSESSMENT**
Convenors: Marcus Callies (Justus Liebig University Giessen) and Sandra Götz (University of Bremen)

Operationalizing and assessing writing proficiency in the academic register: The Corpus of Academic Learner English (CALE)
Marcus Callies & Ekaterina Zaytseva (University of Bremen) .................................................. 51

An evaluation of Avalingua based on learner corpora
Pablo Gamallo, Marcos García (CITIUS, University of Santiago de Compostela) & Isaac González, Marta Muñoz, Iria del Río (Cilenis S.L) ................................................... 52

Operationalizing fluency in (learner) corpora: Implications for language testing and assessment
Sandra Götz (Justus Liebig University Giessen) ........................................................................ 54

Insights from a corpus of secondary school English examination essays in Lithuania
Rita Juknevičienė (Vilnius University) ....................................................................................... 55

MILE – The Marburg Corpus of Intermediate Learner English
Rolf Kreyer (University of Marburg) .......................................................................................... 56

The English Vocabulary Profile as a benchmark for comparing learner essays within and across corpora
Agnieszka Leńko-Szymańska (University of Warsaw) ............................................................... 58

Proficiency in writing online: First steps in assessing a learner corpus of computer-mediated communication among university students in Japan
Tim Marchand, Sumie Akutsu & Ben Rowlett (J. F. Oberlin University) ................................. 59

Data commentary in science writing: A corpus-based comparison of research articles and master’s theses in technical fields for formative self-assessment practices
Lene Nordrum & Andreas Eriksson (Chalmers University of Technology) ............................. 60

Profiling the oral proficiency interview (OPI) in English: An analysis of learner and native spoken language
Pascual Pérez-Paredes & María Sánchez Tornel (University of Murcia) .................................. 62
WORKSHOP 4: PROCESSING IN CORPORA: 'SUPPORT STRATEGIES' IN LANGUAGE VARIATION AND CHANGE

Convenors: Britta Mondorf (Johannes Gutenberg University of Mainz) and Javier Pérez-Guerra (University of Vigo)

Why speakers use the –ing form in certain communicative situations – a functional explanation
Granath Solveig & Michael Wherrity (Karlstad University) ......................................................... 65

Variation across three dimensions: Testing the complexity principle on adjectival data
Henrik Kaatari (Uppsala University) ............................................................................................... 66

Processing complexity of English comparative variants – experimental results
Gero Kunter (Heinrich Heine University Düsseldorf) ................................................................. 68

A cross-linguistic perspective on variation – order in English and Japanese NP conjuncts
Arne Lohmann (University of Vienna) & Tayo Takada (University of Hamburg) ......................... 69

Form does not follow function (but variation does). The origin of possessive have got
David Lorenz (Albert Ludwig University of Freiburg) ................................................................. 70

A data-driven, automatic parser-based approach to variation in English
Gerold Schneider (University of Zurich) ......................................................................................... 71

I was sat there talking all night: A corpus-based study on factors governing intra-dialectal variation in England
Ulrike Stange (Johannes Gutenberg University of Mainz) .............................................................. 73

Formulaicity as a determinant of processing efficiency: Investigating clause ordering in complex constructions
Daniel Wiechmann (RWTH Aachen University) ............................................................................ 74

FULL PAPERS

The ‘humour’ element in engineering lectures across cultures: An analysis of one category of pragmatic annotation
Siân Alsop (Coventry University) ................................................................................................. 79

General extenders on the move? Using spoken corpora to study change in real time
Gisle Andersen (NHH Norwegian School of Economics) ......................................................... 80

Superficial typology: The so-called ‘verb-second constraint’ in Old English and Old French
Kristin Bech & Christine Meklenborg Salvesen (University of Oslo) ........................................ 82

Clause fragments in spoken English
Jill Bowie & Bas Aarts (University College London) ................................................................. 83

A new analysis of proper noun modifiers in PDE NPs
Tine Breban (The University of Manchester) .............................................................................. 84
If-parenthetical clauses in the history of English: The rise of metalinguistic meaning
Laurel Brinton (University of British Columbia) ................................................................. 85

Existential and locative constructions in an expert and in a learner corpus of
promotional tourist texts
Erik Castello (University of Padua) ..................................................................................... 86

The verb phrase in Australian English: A diachronic corpus-based study
Peter Collins (University of New South Wales) ................................................................. 88

Predicting prototypes across native and learner English: The case of modal constructions
Sandra C. Deshors (New Mexico State University) ............................................................ 90

On neoclassical compounding and their constituents
Ana Díaz-Negrillo (University of Granada) ..................................................................... 91

The progressive in spoken learner language: Frequency and accuracy
Stefanie Dose & Sandra Götz (Justus Liebig University Giessen) .................................... 92

Support for end-weight: Do-support in Early Modern English affirmative declaratives
Matthias Eitelmann (Johannes Gutenberg University of Mainz) ........................................ 94

Methodological aspects of corpus pattern analysis
Ismail El Maarouf (RIILP, University of Wolverhampton) ................................................ 96

English in South Africa: The case of past-referring verb forms
Johan Elsness (University of Oslo) .................................................................................. 97

A discourse-functional perspective on the verbalization of the gerund
Lauren Fonteyn, Hendrik De Smet & Liesbet Heyvaert (K.U. Leuven) ............................ 99

So ladylike and particularly manly? Determinants of variation in intensifier usage
in Indian English
Robert Fuchs (University of Münster) ............................................................................... 101

Negotiating trust during a corporate crisis: A corpus-assisted discourse analysis of
CEOs’ public letters after the Gulf of Mexico oil spill
Matteo Fuoli (Lund University) ...................................................................................... 102

Regional variation in Middle English suffix usage
Anne Gardner (University of Zurich) ................................................................................ 103

Round brackets in Jane Austen
Victorina González-Díaz (University of Liverpool) .......................................................... 105

What would a native speaker do?
Stefan Th. Gries (University of California, Santa Barbara) & Sandra C. Deshors
(New Mexico State University) ....................................................................................... 107
Variation in pronoun frequencies in early English letters: Gender-based or relationship-based?
Mikko Hakala, Tanja Säily & Turo Vartiainen (University of Helsinki) ................................ 108

Metadiscourse in novice academic English
Hilde Hasselgård (University of Oslo) .................................................................................. 109

Passive use in academic writing: Comparing expert and student texts across four disciplines
Turo Hiltunen (University of Helsinki) ................................................................................... 111

Didn’t pay much attention in school, did you? – Tag questions as markers of (im)politeness in CMC
Sebastian Hoffmann (University of Trier), Anne-Katrin Blass (University of Trier) & Claire Hardaker (University of Central Lancashire) ........................................................ 113

A corpus of early American literature: Compilation, composition and preliminary results
Mikko Höglund & Kaj Syrjänen (University of Tampere) ...................................................... 114

Beyond (semi-)modals: Changes in the domain of obligation/necessity in Present-Day American English
Verena Jäger (Johannes Gutenberg University of Mainz) ..................................................... 115

Adjectival complementation: Genre variation and meaning
Henrik Kaatari (Uppsala University) .................................................................................... 117

From matrix clause to pragmatic marker: Information packaging and grammaticalization
Gunther Kaltenböck (University of Vienna) ........................................................................... 118

Discussing dissing and cussing: Impoliteness in online discourse
Andrew Kehoe & Ursula Lutzky (Birmingham City University) ............................................ 119

Presentational focus changes in English
Erwin Komen (Radboud University Nijmegen and SIL-International) ................................ 121

New and learner Englishes: Re-evaluating the norm orientation continuum
Samantha Laporte (Université Catholique de Louvain) & Alison Edwards (University of Cambridge) .................................................................................................... 123

Complexity in Early Modern English statutes: The diachrony of binomials and multinomials
Anu Lehto (University of Helsinki) ......................................................................................... 124

Discourse marker sequencing and grammaticalization
Arne Lohmann (University of Vienna) & Christian Koops (University of New Mexico) ..... 126

[T]he figure sixty forty seven forty five...: Modals of necessity in Hong Kong and Indian English
Lucía Loureiro-Porto (University of the Balearic Islands) ...................................................... 128
Anaphora resolution in a learner corpus of L2 Spanish
Cristóbal Lozano (University of Granada) ................................................................. 130

Everything is changed: The be-perfect in Irish English
Kevin McCafferty (University of Bergen) ................................................................. 133

Discipline-specific English vocabulary use in the 17th century
Lilo Moessner (RWTH Aachen University) .............................................................. 134

Corpus findings versus intuition: The value of diverging evidence
Sandra Mollin (University of Heidelberg) ................................................................. 135

Parallelism and contrast: A multifaceted investigation
M. Lynne Murphy (University of Sussex), Anu Koskela (De Montfort University)
& Steven Jones (The University of Manchester) ....................................................... 137

Perfectly regular or totally chaotic? The recent history of English maximizers
Nadja Nesselhauf (University of Heidelberg) ......................................................... 138

Multilingual practices in the history of written English: Developing a corpus
Arja Nurmi, Päivi Pahta & Jukka Tykkö (University of Tampere) ............................ 139

A corpus-based analysis of anaphoric one and its implications
John Payne, (The University of Manchester), Geoffrey K. Pullum (Brown University),
Barbara C. Scholz† (University of Edinburgh) & Eva Berlage (University of Hamburg).... 141

English-Spanish contrastive analysis: Aids for writing up minutes in an L2
Isabel Pizarro, (University of Valladolid), Rosa Rabadán (University of León)
& Marlén Izquierdo (University of León) ............................................................... 142

‘Hapax Legomena’. The silent majority in text
Antoinette Renouf (Birmingham City University) ................................................... 143

Linking learner corpus and experimental data in studying second language learners’
knowledge of verb-argument constructions
Ute Römer (Georgia State University), Audrey Roberson (Georgia State University), Nick
Ellis (University of Michigan) & Matthew Brook O’Donnell (University of Michigan) ...... 145

Investigating light verb constructions in contemporary British and Irish English
Patricia Ronan (University of Lausanne) & Gerold Schneider (University of Zurich) ...... 146

Conversational narrative as drama: From story climax to utterance climax
Christoph Ruehlemann (LMU Munich) ................................................................. 147

The elimination of homography in thirteen editions of The Kalender of Shepherdes
Hanna Rutkowska (Adam Mickiewicz University) .................................................. 149

ICE-EA2.0 - Between new methodologies and old compatibilities
Josef Schmied (Chemnitz University of Technology) .............................................. 150
Re-examining the levelling of present perfect and simple past in World Englishes
Elena Seoane (University of Vigo) & Cristina Suárez-Gómez (University of the Balearic Islands) ................................................................. 151

An enhanced vista on recent change in standard English: Findings from BLOB-1901 and the updated Brown family of corpora
Nicholas Smith (University of Salford) .................................................................................................................. 152

The Middle English development of OE ē and éo: Spelling evidence
Gjertrud F. Stenbrenden (University of Oslo) ........................................................................................................ 153

Hijo de puta/Son of a bitch. On Madrid and London teenagers’ use of rude vocatives
Anna-Brita Stenström (University of Bergen) ..................................................................................................... 155

Moving towards multimodality: Implications for corpus linguistics of a study of medical articles
Caroline Tagg & Mel Evans (University of Birmingham)...................................................................................... 156

Sir, here’s a woman Ø would speake with you: Zero subject relativizers in Early and Late Modern English
Gunnel Tottie (University of Zurich) & Christine Johansson (Uppsala University) ........................................ 157

In absolute detail: The development of English absolutes from adverbial to additional-context marker
Nikki van de Pol & Hubert Cuyckens (K.U. Leuven) .................................................................................... 158

Correlative constructions in English: The þa … þa construction
Ans van Kemenade & Meta Links (Radboud University Nijmegen) ............................................................ 160

Taboo language and swearing in 18th and 19th century English. A diachronic study based on the Old Bailey Corpus
Bianca Widlitzki & Magnus Huber (Justus Liebig University Giessen) ..................................................... 161

Developments in the use of the English present perfect: 1750-present
Xinyue Yao (University of New South Wales) ...................................................................................................... 162

SOFTWARE DEMONSTRATIONS

Extending Dependency Bank with statistical measures
Hans Martin Lehmann & Gerold Schneider (University of Zurich) ................................................................. 167

An innovative web interface for the 14-million-word Old Bailey Corpus, 1720-1913
Magnus Nissel & Magnus Huber (Justus Liebig University Giessen) ........................................................... 168

Enlarging the stock of English diachronic resources for deep corpus analysis: A diachronic lexical database
Hagen Peukert (University of Hamburg) .......................................................................................................... 169
WORK-IN-PROGRESS REPORTS

English urban vernaculars, 1400-1700
Anita Auer & Moragh Gordon (Utrecht University) ............................................................. 173

[T]he next thing Mudiyanse knew was that he was given a slap across his cheek:
Frequency, types and genre-sensitivity of light-verb constructions in Sri Lankan English
Tobias Bernaisch (Justus Liebig University Giessen) ........................................................... 173

Towards an automatic annotation of discourse presentation
Beatrix Busse (University of Heidelberg) & Michaela Mahlberg
(University of Nottingham) .................................................................................................... 175

On the use of the split infinitive in the Asian varieties of English
Javier Calle-Martín & Jesús Romero-Barranco (University of Málaga) ......................... 176

Keywords and clusters: Can they help a translator? A case study in Irving’s A Widow
for One Year and its Czech translation
Anna Čermáková (Institute of the Czech National Corpus, Charles University in Prague). 177

Cause-and-effect in history writing from the Old English to the Early Modern
English period
Claudia Claridge (University of Duisburg-Essen) ............................................................... 178

Corpus pragmatics across genres: Commitments in emails and spoken language
Rachele De Felice, Jill Bowie & Sean Wallis (University College London) ......................... 179

Nouns and verbs in advanced Dutch EFL writing: A quantitative and qualitative
longitudinal study
Pieter de Haan (Radboud University Nijmegen) .............................................................. 180

Motion to and motion through: Evidence from a multilingual corpus
Thomas Egan (Hedmark University College) ....................................................................... 181

Expletive subjects in L2 English: Can we reset the parameter with corpus data?
Esther Ferrandis & Amaya Mendikoetxea (Autonomous University of Madrid) ............. 182

“Alcohol consumption” vs. “popping in for a pint”: A study of discourse on alcohol
in parliamentary debate
Gregory Garretson (Uppsala University) ............................................................................. 184

Sociopragmatic XML tagging in a corpus of transcribed speech
Matt Gee, Ursula Lutzky & Robert Lawson (Birmingham City University) ...................... 185

Complement extraposition: What is it a diagnostic of?
Caroline Gentens (K.U. Leuven), Kristin Davidse (K.U. Leuven) & Lieselotte Brems
(University of Liège/K.U. Leuven) .................................................................................... 186

To have a glimpse at other nations… Register specificity of light verb constructions
in advanced EFL writing
Caroline Gerckens (Leibniz University Hannover & Université Catholique de Louvain) & Arian Shahrokny-Prehn (Leibniz University Hannover) .................................................. 187

Phraseological patterns of anticipatory *it* with extraposed fused relatives
*Sara Gesuato (University of Padua) ..................................................................................... 188

ICAME data and CLARIN, a progress report
*Knut Hofland (Uni Research AS, Bergen) ............................................................................. 190

*Kind-of* and *sort-of*: Pragmatic discourse markers in the SPICE-Ireland Corpus
*John Kirk (Belfast) ................................................................................................................ 191

Tracing the grammaticalization of *I think*
*Daniela Kolbe-Hanna (University of Trier)......................................................................... 191

Ongoing changes and global Englishes: Exploring new corpus resources
*Mikko Laitinen (Linnaeus University)................................................................................... 193

Information structure in the development of discourse markers and comment clauses
*Diana M. Lewis (Aix-Marseille University)........................................................................... 194

Tracking eye movements when reading Dickens’s characters: A pilot study
*Michaela Mahlberg & Kathy Conklin (University of Nottingham)....................................... 195

Order of recipe ingredients in early English medicine: Evidence of medieval practical intertextuality and literacy practices?
*Martti Mäkinen (Hanken School of Economics)................................................................. 196

*Keep it in that place versus if you could keep it in that place*: On the variation between imperatives and isolated *if*-clauses
*Beatriz Mato-Míguez (University of Santiago de Compostela) ............................................ 198

The pragmatic marker *sure* across varieties of English: When the transcription is the corpus
*Irina Pandarova (Leuphana University Lüneburg)................................................................. 200

Newspapers, politics, and CanE: The Corpus of Early 19th-Century Ontario
Newspaper English (CENCONE)
*David Reuter (University of Kiel) ......................................................................................... 201

System and content morphemes in intrasentential codeswitching in Spanish/English and Tagalog/English
*Susana Sotillo (Montclair State University) ......................................................................... 202

Styles of late modern medical writing for professional and lay audiences
*Irma Taavitsainen (University of Helsinki) ........................................................................... 204

The grammatical distinctiveness of written and spoken Canadian English
*Uwe Vosberg (University of Kiel).......................................................................................... 204
Capturing patterns of linguistic interaction in a parsed corpus: An insight into the empirical evaluation of grammar
Sean Wallis (University College London) ................................................................. 206

POSTERS

Tag questions outside the dialogue in fiction
Karin Axelsson (University of Skövde) ........................................................................... 211

Too fascinating a topic to neglect: A corpus study of the Big Mess Construction
Anne-Katrin Blass (University of Trier) ..................................................................... 212

High-frequency verbs and their collocations: Corpus findings and language-pedagogical implications
Martina Bredenbröcker (University of Paderborn) ................................................. 213

A look at participial constructions with get in East Asian Englishes:
The Hong Kong variety
Eduardo Coto-Villalibre (University of Santiago de Compostela).......................... 214

Lexical bundles in British and West African English
Sabrina Edler (University of Münster) ......................................................................... 216

Eating disorders terminology from scientific journals to mass media:
A preliminary approach
Fátima Faya-Cerqueiro (University of Castilla la Mancha) ....................................... 217

Spoken style, written style – ditransitives in speech and the written standard
Johanna Gerwin (University of Kiel) .......................................................................... 218

Eh oh I think that is eh...: I think functioning as an utterance launcher in spoken learner English
Hege Larsson Aas (Hedmark University College) ....................................................... 220

Hanken Corpus of Academic Written English for Economics
Martti Mäkinen (Hanken School of Economics) & Turo Hiltunen
(University of Helsinki) .............................................................................................. 221

Linguistic change in its social contexts in eighteenth-century English
Terttu Nevalainen, Minna Palander-Collin, Tanja Säily, Mikko Hakala, Teo Juvonen, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Anni Sairio (University of Helsinki) ................................................................................................. 222

A quantitative approach to empty referential subjects in Old English
Kristian Andersen Rusten (University of Bergen) ...................................................... 224

Was and were in the English counterfactual conditionals: Classification and probability prediction
Myoung Hyoun Song (Seoul National University) ...................................................... 225
PLENARY SPEAKERS
**Multiethnolects and English corpus linguistics**

**JENNY CHERISHIRE** (Queen Mary, University of London)

In multiethnic working class areas of inner London a new variety of English is emerging, spoken by young people of diverse ethnicities. We refer to this way of speaking as Multicultural London English (MLE), seeing it as a variable repertoire of innovative phonetic, lexical, grammatical and discourse features. These features suggest overlapping influences from many different languages as well as second and interlanguage varieties of English. Young people draw on these features as a stylistic resource, and in the absence of a consistent target model the features are a target for non-native speakers acquiring English within their peer groups. Bilingual young people lead in the use of the innovations, but they are also used by white speakers from traditional ‘Cockney’ families, especially those speakers with multiethnic friendship groups.

We have compiled a 2.5 million word corpus of London English, which is now available from the UK Economic and Social Data Service and from Sketchengine. I will briefly describe the corpus, and discuss the sociohistorical context in which MLE developed. I will show how corpus linguistic methods have helped identify many of the innovative grammatical and discourse features. These include a new first person pronoun (e.g. *I don’t care what my girl looks like, it’s her personality man’s looking at*), new discourse markers (e.g. *I’m going to get food after this still*) and a new quotative expression (*this is my mum “what are you doing?”*).

Corpus linguistic methods have also suggested some of the processes through which new grammatical and discourse features arise, giving insights, therefore, to the vexed question of how a language change may begin.

I will suggest that Multicultural London English and similar multiethnolects such as Kiezdeutsch in Berlin or Rinkebysvenska in Stockholm are typologically distinct language varieties that can be considered alongside, for example, dialects or creoles. They are a type of contact variety, characterized by innovations common to contact varieties more generally, such as the simplification of inflections, the emergence of new forms driven by information structuring processes, and the development of ‘camouflaged features’.

**The network metaphor of usage-based construction grammar**

**HOLGER DIESEL** (Friedrich Schiller University Jena)

There is a long tradition in linguistics to analyze grammar without reference to usage and development; but this tradition has been challenged by usage-based linguists who have argued that grammar is fundamentally grounded in the language user’s experience with concrete linguistic tokens. In this approach, grammar is commonly characterized as a ‘network’ (or ‘structured inventory’, Langacker 2008) of fluid constructions that are in principle always changing under the influence of general cognitive processes involved in language use. The network metaphor of usage-based grammar stresses the dynamic nature of constructions and syntactic categories; however, it does not really explain the general organization of grammar in the usage-based approach. In this paper, I consider five different types of ‘links’ between constructions and other linguistic elements that underlie the network metaphor of usage-based construction grammar, namely the links between (i) schematic constructions and constructs (i.e.
utterances), (ii) constructions and words, (iii) constructions and syntactic categories (e.g. grammatical relations), (iv) phrasal constructions and clausal constructions, and (v) between constructions of the same type (e.g. the transitive construction and the passive construction). Combining evidence from a wide range of studies, the paper seeks to develop a more structured approach to the usage-based analysis of grammatical phenomena.

Who is the/a/Ø professor of English at your university?

MARIANNE HUNDT (University of Zurich)

Variable article use in the recent history of English is a somewhat under-researched area. According to Denison (1998: 188), “[t]here are only minor changes in the use or omission of articles.” He mentions variable article use with nouns denoting illnesses, titles like Prime Minister or President in conjunction with proper names, and individual nouns like percentage or drama. Evidence from a recently digitized diary by a Victorian clerk (Nathaniel Bryceson) indicates that variability in the use of articles may, in fact, be an area of diachronic variation meriting further research:

(1) … I followed her down a street to the Liverpool Road … (Bryceson, Sunday 22nd March 1846)
(2) … Stopped at home all the afternoon looking over maps and books. … (Bryceson, Sunday 19th July 1846)
(3) The Queen Victoria has been married six years this day. (Bryceson, Tuesday 10th February 1846)

Predicate nominals designating single role holders like president, prime minister, queen or pope allow for variable article use. The following examples illustrate variation (zero, indefinite, definite) with professor in this syntactic environment:

(4) Ralph F. Young is Ø professor of history, Temple University, Philadelphia, Pa. (COHA, 2004)
(5) Noam Chomsky is a professor of linguistics at M.I.T., a New Left ideologue (American Power and the New Mandarins), a leading protester against the war in Viet Nam - and a Hedgehog. (COHA, 1970)
(6) He is the Professor of Eloquence in the University; and his lectures are regularly attended by thousands of intelligent and fashionable auditors. (COHA, 1830).

Berezowski (2009) uses data from the BNC to investigate the use or absence of the (in)definite article in such constructions. A pilot study investigating eight nouns using data from COHA indicates that there is also diachronic variation: a general tendency towards greater use of definite and indefinite articles in these constructions. However, there also seems to be considerable lexical variation. Furthermore, the structure of the NP (pre- and post-modification) may also play a role.

The present paper investigates the same set of nouns but includes a larger variety of copula verbs than the pilot study. A variable rule analysis will show which of the factors (copula verb, lexical item, modification within the NP, construal, extra-linguistic facts) best explains variable article usage in this area of English grammar. The results will be used to discuss whether Berezowski’s (2009) account, namely that incomplete grammaticalisation is the reason for variable article usage in these constructions, also holds in a study of real-time diachronic change.
Community and identity: What corpora can tell us about academic discourse

KEN HYLAND (University of Hong Kong)

To many outsiders, corpus linguistics is often seen as a dreary quantitative method for sad IT geeks, but I want to argue here that it can contribute to our understanding of two of the most controversial concepts in the social sciences: community and identity. With the emergence of community-oriented views of literacy in recent years, greater attention has been given to the specific contexts of language use, so we have learnt that texts are successful only when they employ conventions that other members of the community find familiar and convincing. Because of this, corpus studies have become invaluable in revealing how language choices help construct both arguments and disciplines. Moreover, because writers negotiate representations of themselves through the discourses of their communities, corpus studies also contribute to a new way of conceptualising identity. Essentially, the study of academic discourse shows how we choose our words to connect with others and present ideas in ways that make most sense to them. By privileging certain ways of making meanings, repeated uses of language help to perpetuate the norms and thinking of disciplinary communities and so encourage the performance of certain kinds of professional identities. Communities thus constrain identity choices but they also indicate the ways we relate independent beliefs to shared experience. In this way, the production of texts is always the production of community and of self.
PRE-CONFERENCE WORKSHOPS
Modal particles in a contrastive perspective – the case of the Swedish väl

KARIN AIJMER (University of Gothenburg)

Väl in Swedish is known as a modal particle. Modal particles are characteristic of some Northern European languages including Swedish. They are associated with epistemic rather than other types of modality. Formally the particles are short, unstressed elements which are typically used together with other modal elements (måste väl) but can also be autonomous. They are intriguing because they are not clearly associated with a single epistemic meaning but are multifunctional. On the other hand, English does not have modal particles. This makes it interesting to study how English copes with the different meanings of väl. Parallel corpora represent an opportunity both to study the correspondences of väl in other languages and to get a blueprint of its semantics, sub-senses, implicatures and connotations. In my contribution the correspondences of väl will be studied on the basis of the English-Swedish Parallel Corpus.

It is typical of väl and other modal particles that they have a large number of translations reflecting different contextual values. The particles are adverb-like but are translated in many different ways. The modal translations of väl include other modal auxiliaries (must, would, can, could), modal adverbs (probably), modal predicates (I think, I suppose, I’m sure). We can assume that all the translations express epistemic modality or represent specialised functions associated with epistemic modality. Some translations suggest for example that epistemic modality may be difficult to distinguish from interactive or discourse meanings. Väl can, for instance, be hearer-oriented and translated as a question or as a tag question. It is interesting to note that väl was often not translated at all. This reflects the fact that väl frequently co-occurred together with a modal auxiliary or another modal element (in 25% of the examples it co-occurred with a modal auxiliary). The particle cannot be translated in such examples since there is no natural correspondence in English. On the basis of the translations we can get some insight into the paradigm of epistemic modal markers in English and what the most salient realisations are. The translations suggest that epistemic modality is a richly diversified category consisting of ‘core epistemic aspects’ and extensions from them.

I am wild about cabbage: Evaluative ‘semantic sequences’ and cross-linguistic (dis)continuities

MARINA BONDI & CORRADO SEIDENARI (University of Modena and Reggio Emilia)

The main objective of our research is to investigate how evaluation is realized through phraseology in two comparable corpora (1.5 million words each) collecting Italian and English texts from the web genre of blogs.
Our assumption is that the ‘evaluatively charged’ nature of blogs’ posts – and comments especially – makes them ideal territory for an analysis of the language of evaluation (Myers 2010, Baron 2008, Herring et al. 2004). The investigation follows up from previous research (Bondi & Seidenari 2012) on 1st person pronouns and adjectives (I, me, my; mi, io, me), which score at the very top of the keyword-lists extracted from the two blog corpora. The pervasiveness of such ‘subjectivity markers’ reflects the high degree of subjectivity of blogs and can be taken as an indication of the significant role that the writer’s ‘self’ plays in the phraseological realizations of evaluative meaning.

The present paper focuses on the lexico-grammar of two comparable sets of verbal word-forms that feature prominently in our two keyword-lists. The sets include verbs that may be categorized in very general, rough terms as epistemics (think, guess, looks, sounds, etc; credo, penso, pare, sembra, etc.), attitudinals (like, hate, sucks, etc.; piace, adoro, preferisco, etc.), quotatives (say, wonder, add, etc.; dico, ripeto, chiedo, etc.) and linking verbs.

The analysis follows a two-step process. First, we perform a concordance analysis in order for the recurrent phraseological arrangements of the selected keywords to stand out, both in terms of preferred lexis (i.e. typical collocates) and syntax (i.e. grammar patterns). Second, following Hunston (2008) and Groom (2010), we provide a comprehensive semantic interpretation of all the structural information collected together in terms of abstract evaluative ‘semantic sequences’, i.e. “sequences of meaning elements, rather than […] formal sequences” (Hunston 2008: 271). Phraseology is thus identified by aligning typical recurrent surface arrangements with strings of prototypical meaning elements such as ‘entity or process evaluated’, ‘evaluation’ and different ‘sources of evaluation’.

The phraseological arrangements identified in our analysis (‘basic sequences’, ‘framed sequences’, ‘dialogic sequences’ and ‘concessive sequences’) display a considerable amount of lexical and syntactic variation across the Italian and the English samples. However, our results attest for substantial semantic continuities over and above lexis-syntactic and inter-linguistic mismatches. Semantically, the overall prosodic workings of the sequences stand firm across the corpora, irrespective of the language involved.

The kinds of evaluative meanings that characterize English and Italian blogs may be regarded as essentially the same and semantic sequences can be shown to be useful tools for cross-linguistic analysis of specific areas of meaning.

References


Binominal size noun constructions in English and French: A contrastive corpus-based perspective
LOT BREMS (University of Liège/K.U. Leuven)

Binominal syntagms are a type of complex NP attested in many European languages involving two nominal elements and possibly linked by some linking element, i.e. (determiner) (modifier) NP1 + of/van/von/de/di/Ø + (determiner) (modifier) NP2. This paper will focus on one subtype, viz. size noun (SN) constructions, and compare them in English and French.

Brems (2011) argued that English SNs synchronically display functional/formal variation, as the result of grammaticalization. In addition to a head noun function with lexical meaning (1), bunch f.i. has come to function as quantifier (2), and evaluative expression (3):

(1) […] a very large bunch of flowers with a neat little card […] (WB)
(2) She is sitting in the kitchen with a bunch of other women from the area (WB)
(3) The report is just a bunch of gobbledygook (Internet)

Corpus-analysis showed that each use displays systematic co-selection patterns between the SN and pre or postnominal collocates (Stubbs 1995): between (sets of) determiners/adjectives/quantifiers and SNs and postnominally between the SN+ of and the (premodified) N. This is why they were argued to be collocationally constrained constructions, i.e. templates in which some positions are predetermined or ‘filled in’, viz. by collocations.

The present paper takes a contrastive perspective, hitherto under-researched, the central question being whether similar-looking SN-constructions in English and French function analogously. How are SN-expressions for which structural cognates exist translated between English and French and in specific genres or registers? On the basis of samples of English bunch, load(s), heap(s), handful and French bande, tas, flopée, poignée from monolingual corpora such as WordbanksOnline (WB) and Frantext (FT)and parallel corpora such as OPUS and DPC, it will be argued that they show the same range of meanings, though to different extents and in different ways:

(4) Assis sur un tas de sable, il passait et repassait la pierre sur le vif (FT)
(5) Il a des tas d’armes à feu et fait un boucan terrible (FT)
(6) Les hommes sont des bêtes, mon cher A, des tas d’atomes et de boue (FT)

In order to assess constructional and translational equivalence the following parameters will be taken into account: collocational preferences ((in)animate/abstract; (un)count), singular vs. plural SNs, polarity sensitivity, syntactic behaviour in terms of concord patterns, and discourse contexts. In French for instance both singular and plural SNs can express evaluation (6), while in English this is typically restricted to singular forms (3). Parallel corpora of different registers allow to assess translational equivalence. Is a binominal English quantifier typically translated
by a binominal alternative or not, and if so, which SNs are most equivalent: literal translations or not (poignée de vs. handful)? Data analysis shows that different SNs may be used to translate a specific use of a SN-construction, e.g. the valuing use of bunch of is translated by bande de or flopée de, whereas more neutral quantifier uses can be translated by tas de for instance.

References


DPC: <http://www.kuleuven-kulak.be/DPC/>

OPUS: <http://opus.lingfil.uu.se/>

A contrastive analysis of downtoners, more or less

Signe Oksefiell Ebeling & Jarle Ebeling (University of Oslo)

This paper reports on a contrastive study that takes the downtoner more or less as its starting point. Our attention was drawn to this pattern after having extracted 3-gram lists from the extended fiction part of the English-Norwegian Parallel Corpus, ENPC+. With an expected frequency of six, more or less actually occurs 41 times (31 pmw) in English original (source) texts. Our interest was further piqued when it was revealed that its frequency turned out to be significantly higher in the translated material (71 occurrences (50 pmw) – LL: 6.02, p < 0.05).

From a linguistic point of view, it is interesting to note that Norwegian has a cognate congruent pattern, mer eller mindre, which is used in around 54% of the translations of more or less, e.g. example (1).

\[(1) \quad \text{One of them still looked more or less intact --- } \text{(TaFr1E)}
\]
\[\qquad \text{Et av dem så fortsatt mer eller mindre intakt ut --- } \text{(TaFr1TN)}\]

However, of the 71 instances of more or less in translated (target) texts, only 15.5% had mer eller mindre as their source. In fact, 28 different Norwegian source expressions were recorded. In the opposite direction of translation, we can note that Norwegian mer eller mindre is translated into more or less in around 53% of the cases, while more or less is used as its source in around 68% of the cases. With these cross-linguistic observations as our starting point, we will investigate this apparently skewed relationship between these cognate patterns in more detail.

Downtoners may, according to Quirk et al. (1985: 445, 601), have at least two main functions, i.e. that of premodifier or subjunct adverbial. It will therefore be of interest to survey the different uses in contrast in order to establish whether function and position play a role in the choice of cross-linguistic correspondence. Equally important will be to study other correspondences, as these may give indications as to the extent of the semantic field to which the downtoners belong. This leads us to the overarching question of why two seemingly perfect cross-linguistic patterns should reach a mutual correspondence of only around 53%?
References


**Motion into and out of in English, French and Norwegian**

**THOMAS EGAN & ANNE-LINE GRAEDLER** (Hedmark University College)

In this paper we compare and contrast English and French translations of Norwegian predications of motion events containing the boundary-crossing compound prepositions *ut av* (=out of) and *inn i* (=into). According to Slobin (2006: 70), “in translations […] manner salience follows patterns of the target, rather than source language”. Thus a translation into a path-framed language will most likely itself be path-framed even in cases where the source text codes manner in the verb phrase. (See Talmy 2000 for the path-/satellite-framed distinction.) This is likely to apply to an even greater extent to cases in which the preposition codes the crossing of a boundary, given that French is said to avoid the use of manner verbs with boundary-crossing activities. Cappelle (2011) disagrees with Slobin on the question of manner salience in translated texts, maintaining that translations will often retain traces of the typology of the source text. Neither Slobin nor Cappelle, however, compare translations into different languages of one and the same source. Such data would have the obvious advantage of allowing one to compare in detail the alterations made by two sets of translators to the same set of original texts.

Our paper is based on all occurrences in the *Oslo Multilingual Corpus* (OMC) of self-motion predications containing the two Norwegian prepositions *ut av* and *inn i*, in all of which path is coded in the prepositional phrase. The verb may also code path, as in (1), it may code manner, as in (2), or it may be a neutral verb of movement, as in (3).

(1) Han hadde *falt* ut av vinduet…. (NF1)
   He’d *fallen* from the window…

(2) Jeg kjente meg litt svimmel da jeg *stupte* inn i landhandelen… (JG3)
   I was reeling a little by the time I *dived* into the village store…

(3) En fjern og ennå utydelig skikkelse tar form og *beveger seg* inn i synsfeltet…
   (BHH1)
   A vague, faraway figure takes form and *moves* into my field of vision…

We first analyse the Norwegian originals with respect to their coding of path and manner and then turn to the two sets of translations and investigate the extent to which they retain the manner/path coding choices of the source predications and, if not, what sort of alterations they make. If Slobin is correct, the French translations should exhibit a much greater degree of path coding in the verb than either the Norwegian originals or the English translations. If Cappelle is correct, there should be fewer differences between the two sets of translations. In particular we might expect there to be some tokens in which manner is coded by the verb. Finally, we compare our results to those of both Slobin and Cappelle.

References


The postmodifying structure of noun phrases: A contrastive study of English and Norwegian
Johan Elness (University of Oslo)

English noun phrases have a varied postmodifying structure. Biber et al. (1999: 606) find that by far the most common type is prepositional phrases (65-80 % in all the registers distinguished), followed by (finite) relative clauses, -ing clauses, past participle clauses, infinitive clauses and appositions. In addition, noun phrases (the conference last week), adverb phrases (our lecture yesterday), adjective phrases (animals common in Africa) and reflexive pronouns (Obama himself) may act as postmodifiers. Faarlund et al. (1997: 261ff.) offer a detailed treatment of the postmodifying patterns occurring in Norwegian noun phrases, but without any statistical information about frequencies. It is confirmed that most of the same postmodifier types occur even in Norwegian, with the addition of postposed determiner-like possessive forms (bilen min as an alternative to min bil), which in English can only be preposed (my car). There is reason to expect large differences in the frequencies of the various modifier types between English and Norwegian, connected with the fact that non-finite clauses generally are less common in Norwegian (and an equivalent of the English -ing clause hardly exists at all). This paper reports (parts of) a contrastive study of noun-phrase postmodifiers (including appositions, often not reckoned as modifiers in the strict sense) in English vs. Norwegian based on the Oslo Multilingual Corpus (OMC), mainly its English-Norwegian section, the ENPC. In some cases the special Multiple-Translation Corpus (made up of ten different Norwegian translations of one fictional and one non-fictional English text) is consulted to trace alternative translation strategies. In a few cases findings are also reported from the Swedish and German sections of the OMC for further comparison. To the extent that tagging is available, that is used to focus the comparison on relevant occurrences. Since my investigation is concerned with cross-linguistic comparison at the phrase level while the OMC browser is based on alignment at the sentence level, extensive manual post-analysis is called for (also because the tagging does not always provide reliable information about the different functions of Norwegian som, for example). For that and other purposes random samples are used. Some of the questions pursued in my study are (I do not expect to be able to deal with all of these in detail in the workshop presentation):

• Any difference according to the direction of the translation? Any sign of interference from the source language in the case of translated texts?

• Any difference between fiction and non-fiction, and between the direct speech and the narrative passages of fictional texts? (According to Biber et al. 1999, most postmodifiers are generally rare in conversation).

• Connection premodifying/postmodifying structure, especially definite/indefinite markers.

• Connection between modifier type and the syntactic function of the NP.

• Voice, especially relevant to relative clauses, and wider thematic implications.
Any sign of postmodifying -ing clauses as markers of imperfectivity? Cf. the general imperfective meaning of finite progressive verb forms in English, made up of -ing preceded by BE.

Other semantic features, such as the animate/inanimate meaning of the head noun and the restrictive/non-restrictive status of the modifier.

References


Cross-linguistic analysis of cohesion: Variation across production types and registers

EKATERINA LAPSHINOVA-KOLTUNSKI & KERSTIN KUNZ (University of Saarland)

The present study deals with linguistic variation across text production types in English and German. We focus on variation in cohesive reference, distinguishing three reference types – personal, demonstrative and comparative (cf. Halliday and Hasan (1976). Contrastive analyses of realisations and functions of these types reveal differences between English and German, cf. Kunz (2009), Kunz and Steiner (2012). We aim at examining these differences in translations, comparing translations from German into English with original texts of both languages, as we believe that this knowledge will contribute to a better understanding of multi-functionality of reference in multilingual analysis.

Lexico-grammatical devices expressing personal, demonstrative and comparative reference are automatically extracted from English originals (EO), German originals (GO), and their translations into English (ETRANS) contained in GECCo. We compare extracted data with regard to their contexts (registers): political essays and prepared speeches, letters of shareholders, manuals, tourism leaflets, popular-scientific articles and fictional texts (available for the three sub-corpora under analysis).

Extractions of normalised (per 1 Mio) distributions of all reference types in EO vs. (ETRANS) and GO do not reveal much variation: Personal reference prevails in all sub-corpora, followed by demonstrative and comparative reference, cf. Figure 1. We additionally calculate Pearson's Chi-squared test to compare EO vs ETRANS, EO vs. GO and ETRANS vs. GO. Here, we observe that variation is lower between the latter two sub-corpora, both in terms of reference type distribution and occurrence of cohesive reference across registers.

This tendency is confirmed by further tests, e.g. by correspondence analyses, cf. Baayen (2008), which reveals less variation in GO vs. ETRANS than EO vs. ETRANS and EO vs. GO for most registers which means that ETRANS is closer to GO than EO in most registers. Figure 2 additionally demonstrates that both reference types and registers contribute differently to the variation observed.

The statistical tests have to be refined in order to allow valid interpretations. Moreover, we currently elaborate an analysis of subtypes of personal, demonstrative and comparative
reference as we expect even more variation in form and function on the fine-grained level, cf. Kunz and Steiner (2012).

In our talk, we will present our methods and resources, as well as our statistical results.

![Figure 1: cohesive reference in EO, GO and ETRANS](image1.png)

Figure 1: cohesive reference in EO, GO and ETRANS

![Figure 2: cohesive reference types across registers in EO, ETRANS and GO](image2.png)

Figure 2: cohesive reference types across registers in EO, ETRANS and GO

**References**


Intersubjective positioning and thematisation in English and Spanish: A contrastive analysis of letters to the editor

JULIA LAVID & LARA MORATÓN (Universidad Complutense de Madrid)

As part of a larger project on the contrastive annotation of a wide variety of registers and genres in English and Spanish with discourse categories (Lavid 2012), and more specifically on thematisation patterns in these two languages (Lavid et al. 2012, Moratón et al. 2012), in this paper we investigate how writers of letters to the editor exploit thematisation to project themselves into their discourse and to engage their audience in the English and the Spanish discourse communities. Using a bilingual comparable corpus of a total of eighty letters (forty English and forty Spanish) from English and Spanish newspapers, we analyse the distribution of a number of expressions indicating intersubjective positioning in thematic position and the preferred thematic types conflating with these expressions in both languages. The linguistic expressions of intersubjective positioning analysed in this study are those included under the notions of stance and engagement (Hyland 2005), and the thematic types are those proposed in the contrastive model of thematisation outlined in Lavid et al. (2010). The contrastive analysis reveals interesting features in the construction of writers and readers in these two communities through the use of thematisation: a) thematised expressions of stance and engagement are more frequent in the Spanish letters than in the English ones, which suggests that this a preferred strategy for intersubjective positioning in this language in contrast with English; b) although thematised expressions of stance predominate over those of engagement in both languages, the use of engagement strategies seems to be a characterizing feature of this genre in comparison with other journalistic genres such as news reports or commentaries; c) the preferred thematised expression of stance in the English letters is ‘self-mention’, i.e. references to first person pronouns and possessive adjectives, which indicates a strong desire on the part of the English writers to both strongly identify themselves with a particular argument and to gain credit for their perspective. Spanish shows a slight preference for ‘boosters’, followed by ‘self-mention’ expressions, which suggests a strong involvement with the topic and solidarity with their audience; d) the preferred thematised expression of engagement in the English letters is ‘shared knowledge’, a strategy through which writers move the focus of the discourse away from the writer to shape the role of the reader. In the Spanish letters, by contrast, ‘inclusive pronouns’ predominate, indicating a stronger emphasis on binding writer and reader together; e) Interpersonal Themes and PreHeads are the preferred thematic types conflating with stance and engagement in both languages, followed by Thematic Heads.

We hope that this contrastive analysis contributes to the general aim of the workshop (i.e. comparing constructions rather than single words), and also to the generic characterization of the journalistic genre of letters to the editor in English and Spanish.
References


A comparable-corpus based approach to the expression of obligation across English and French

DIANA LEWIS (University of Aix-Marseille I)

A French-English comparable corpus of political discourse is used to investigate similarities and differences in the expression of obligation. The aims of the study are threefold: (i) to look at the distribution of obligation in the comparable genres in the two languages, comparing the extent to which more grammaticalized expressions are chosen over more lexical ones, (ii) to compare the contexts of use of deontic modal and semi-modal verbs in the two languages, i.e. comparing their frequencies in contextual frames characterized by agent-type, polarity, time, and event-type, and (iii) in the light of the polysemy of the quasi-translation equivalents *must* and *devoir*, to compare frequencies of deontic and non-deontic use of these verbs in the genre.

The study is based on a comparable corpus of French and English political discourse comprising 1m words in each language. A preliminary reading of a sample of the corpus identified a range of ways speakers in this genre choose to express obligation. These include grammaticalized means such as deontic modal and semi-modal verbs (*we must*, *we need to*, *nous devons*, *il nous faudra*, etc.) as well as non-grammaticalized locutions (*our task is to*, *notre responsabilité est de*, *there is no question of*, *il nous appartendra de*, etc.). The most frequent of these were then selected for further study across the whole corpus.

French political discourse is found to use a wider range than English of non-grammaticalized expressions of (implied) obligation, both impersonal and with agents, and to rely on *devoir* and *falloir* for modal expression. The main focus of the presentation will be on the second of the above aims, the comparison of modal and semi-modal verb use, including *devoir*, *falloir*, *must*, *have [got] to*, *should*, *need to*. As is well known, the traditional translation equivalents *must* and *devoir* overlap (both express deontic and epistemic modality) but differ in formal range, in the range of meanings they encompass and in register. Although *must* is more restricted formally, and its deontic use said to be being eclipsed in PDE by its epistemic use and by the rise of *have [got] to*, it turns out to be more than twice as frequent as *have [got] to* in this genre, and more
than twice as frequent as *devoir*. The impersonal construction *s'agir de* (e.g. *il s'agit/s'agira de*) regularly occurs in deontic contexts in the French texts and appears to be further grammaticalizing.

Overall, then, grammatical deontic expression, which is overt or direct, is less in evidence in the French texts than in the English ones. These findings cannot be extrapolated to the ‘general’ languages; on the contrary, the study suggests that the impact of genre on frequency and distribution must not be underestimated.

**English non-finite participial clauses as seen through their Czech counterparts**

**MARKÉTA MALÁ & PAVLÍNA ŠALDOVÁ (Charles University in Prague)**

Non-finite clauses have long been a recurrent topic in contrastive studies of Czech and English, showing the typological differences between the two languages. Czech as a predominantly ‘verbal’ language relies on the finite predicate as a carrier of grammatical categories expressed synthetically on the verb, while English displays a tendency towards nominal ways of expression, involving also non-finite verb forms. A parallel translation corpus of English and Czech has now made it possible to re-visit the topic of condensed participial clauses. Our focus, however, has shifted from the typological comparison to the exploration of the functions of participial clauses in English as they are brought to the fore by their Czech counterparts.

The counterparts of English participial clauses are rarely syntactically congruent with the originals. The formal equivalent of the participial clause in adverbial function, the Czech transgressive construction, has become obsolete, with some of the transgressives surviving due to recategorization into secondary prepositions. These counterparts highlight the same process in English, resulting in de-participial prepositions both well-established, e.g. *concerning, owing to*, and undergoing the process of change, e.g. *judging by*.

The divergent counterparts of English participial constructions can be divided into four groups: verbless counterparts, subordinate finite clauses, coordinated finite clauses, and synthetic verbal counterparts.

The verbless counterparts, especially postmodifying PPs, accentuate the semantically suppressed character of the English participle, which serves as a linking element ascribing a particular quality to the head noun, e.g.

\[... \text{eastern wing, comprising three large rooms, ...}\]

\[... \text{východní křídlo se třemi velkými pokoji, ...}\]

[... eastern wing *with three large rooms...*]}

The finite dependent clauses represent the most explicit rendering of the participial clause into Czech, attesting that the English participles convey a full scale of semantic relations, including temporal relations, manner, purpose, reason, result, condition, and concession.

Another counterpart – a finite coordinate clause – highlights the function of participial clauses on the level of information structure: the de-dynamizing, backgrounding effect of the participle may result in a shift in the mutual position of the two clauses in translation. When the clause-
order differs from the original sentence, the less dynamic clause corresponding to the participial one assumes the initial position.

Finally, what has been termed here ‘synthetic counterparts’ comprises finite Czech verbs which represent a merger of the English participle and its superordinate verb, e.g.

*Dudley came waddling toward them ...*

*Dudley se k nim přikolébal tak rychle, ....*

These constructions manifest the tendency towards analytic expression in English. As the translation shows, the superordinate verb and the participle constitute a single semantic unit: the finite verb bears grammatical categories and the general categorial meaning of directed motion while the participle specifies the type of movement (cf. also *come running / ambling / dripping, go striding / running*. The group comprises also verbs indicating duration in time (*spend/pass time*), manner of speaking (*say*) and some other verbs (e.g., *sit, stand*).

The contrastive approach, relying on Czech as an ‘auxiliary’ language, has shown the diverse functions of the English participial construction, ranging from a semantically reduced linking element resembling a preposition, via a means of syntactic compression and a means of backgrounding a particular piece of information in the sentence, to a component of a verb-nominal construction pointing out the analytic character of English.

**Sources**


**Quite seen through its translation equivalents: A contrastive corpus-based study**

MICHAELA MARTINKOVÁ (Palacký University Olomouc)

According to dictionaries, *quite* is a polysemous word: LDOCE lists 12 different senses, OALD four. Grammar books recognize the fact that *quite* may scale upward as well as downward from an assumed norm, i.e., it belongs to two different and even opposing intensifier types: amplifiers/intensifiers and diminishers/downtoners (Biber et al. 2007: 556), amplifiers and downtoners (Quirk et al. 1985: 590, 598), in Paradis’s terms (1997: 27-28) reinforcers and attenuators.

Corpus linguists have dealt with *quite* in various syntactic positions both from a diachronic (Ghesquière 2012) and synchronic perspective (Palacios 2009: 181), recently also with the help of multivariate statistics (Desagulier 2012), and predominantly within the cognitive linguistics framework (e.g. Paradis 2008, Diehl 2003). Paradis (e.g. 1997, 2001, 2008) employs the configurational concept of boundedness, common in nominal and verbal paradigms (e.g. Langacker 1987, Talmy 1988), and applies it to degree modifiers as well as collocating adjectives. Paradis (2008: 321) differentiates between totality (bounded) degree modifiers (maximizers, e.g. *absolutely*, approximators, e.g. *almost*), which modify adjectives with bounded meanings, and scalar (unbounded) degree modifiers (boosters, e.g. *very*, moderators, e.g. *fairly*, diminishers, e.g. *a bit*), which modify adjectives with unbounded meanings. The latter can also modify an adjective with a bounded meaning, in which case they coerce its bounded reading into an unbounded reading (eg. *very male*). Given the problems one has to face
when it comes to strictly categorizing the meaning of an adjective as bounded or unbounded and transitions between the readings due to ‘contextual modification’ (Paradis 1997: 59), up-to-date corpus research on monolingual corpora (with focus on British English) has been unable to deal with the fact that a polysemous degree modifier such as quite “is capable of modifying both scales and boundaries” (Paradis 2008: 325). This presentation resorts to a parallel translation corpus (InterCorp) to see which meanings of quite (in this pilot study, only in the pre-adjectival position), are “visible through translation[s]” (Johansson 2007: 28), namely through Czech translation equivalents in two bidirectional parallel InterCorp subcorpora of fiction created for this purpose (one for British and one for American English). Based on their collocational profiles in the Czech National Corpus, equivalents and definitions in Czech-English and Czech-Czech dictionaries, the Czech translation equivalents of quite modifying an adjective were sorted into categories of degree modifiers recognized by Paradis (2008). The resulting data not only shed more light on the distribution and entrenchment of the senses of quite (as recognized by grammar books and dictionaries) in British and American English, but translations with a modal particle or adverb equivalent to really or obviously also reveal the ‘emphasizer’ (for the term see e.g. Quirk et al. 1985: 583-4) function of quite. A significant number of translations by omission, especially where quite modifies an adjective of epistemic stance, leads us to ponder the role of quite in discourse.

References


Palacios Martinez, I. M. 2009. Quite frankly, I’m not quite sure that it is quite the right colour. A corpus-based study of the syntax and semantics of quite in Present-Day English. English Studies 90/2: 180-213.


English positive polarity contexts into Spanish: A corpus-based study

Rosa Rabadán (University of León)

English constructions containing positive polarity items (PPIs) (e.g. still/already + Past Participle/Continuous tenses/Past Tense) are a frequent source of problems when translating into Spanish as well as when learning English as a foreign language. The reason for this is that this is an elusive aspect of language that, unlike negative polarity contexts (Van der Wouden 1997, Zwarts 1998), goes generally unrecognized by speakers and has been largely neglected by experts (Vanden Wyngaerd 1999), particularly in Spanish (González Rodríguez 2009).

Previous work on cross-linguistic polarity (Rabadán 2011) helped unveil the fact that a syntactic approach to positive polarity contexts does not necessarily help with this mainly semantic and pragmatic phenomenon, possibly because on many occasions they lack explicit formal marks to be associated with.

This paper sets out to explore some of these contexts in English, their semantics, and the resources Spanish uses to convey their meanings so as to find out in which ways, and, if possible, why the grammars of the two languages differ.

This has been done through corpus-based contrastive work using P-ACTRES, a custom-made English-Spanish parallel corpus (http://actres.unileon.es/demo.html). It contains nearly 2.5 million words distributed among the following subcorpora: fiction, non-fiction, newspapers, magazines and miscellanea (Izquierdo, Hofland & Reigem 2008). The tertium comparationis is translation equivalence. Whenever necessary, informants have been used to elicit semantic and pragmatic implications of the positive polarity contexts.

The contrast is modelled on Krzeszowski’s (1990) classical stages, and combines qualitative and quantitative corpus analysis.

The procedure has been, first, to qualitatively identify the Spanish equivalences of the English constructions, second, to assess their relevance quantitatively in terms of their frequency of occurrence, and, third, to find out meaning-form regularities. These have been used as a basis to formulate an empirically-grounded (although necessarily partial) explanation of this phenomenon.

References

The Gospels present Jesus as engaged in teaching, in a way which astonished those who met him, as he did so ‘with authority’ (Mark 1: 27, Luke 4: 36). His sayings and parables have gained much attention, but most of the time he was actually ‘asking questions’. Did he really want to know something? Or was the use of questions his particular method of teaching with authority? Was it the case he wanted his listeners to know something?

Speech Act Theory directs attention to what utterances do, and how successfully, when we say something (Austin 1962, Searle 1975), and it is a well-known fact that an interrogative does not have the single predictable function of asking (Downing and Locke 2006). This paper presents a cross-linguistic analysis of the interrogative construction as used by Jesus in two Present Day translations of the Gospels (English and Spanish). The analysis is made from a pragmatic standpoint, studying interrogatives as speech acts in a discourse perspective, which takes into account the interactional aspect of language (Archer 2005: 2010, Jucker and Taavitsainen 2000, 2007, Kohnen 2008).

A first step is to build a corpus of all the formal interrogative locutions, together with a grammatical taxonomy of the types (Kohnen 2007, Quirk et al 1985, Biber et al 1999). Secondly, a functional analysis of what the speaker is actually doing when using these interrogatives will follow, in terms of direct and indirect speech acts, such as admonishing, examining, doubting, wondering, challenging, and so on (Wierzbicka 1987). Related to this, a final step approaches the perlocutionary effect of these interrogatives, many of which were seeking neither information nor a verbal answer, but, rather, a response from the depths of the listener’s heart.

Although Jesus’ interrogatives have been the subject of much study, this has largely been non-linguistic, but rather, devotional, aiming at meditation and reflective prayer (Dear 2004, Pennington 2005, Guthrie 2010, Honeygosky 2010). The present contrastive linguistic study can contribute not only to a better understanding of the interrogative construction in terms of multifunctionality, but also to recent work on religious discourse in general, and Biblical exegesis in particular, from the objective insight that linguistics can provide (Kohnen 2010,

References


Parallel corpus - a tool for diagnosing multifunctionality across languages: *Actually, naturally and in fact* vs. their correspondences in Lithuanian

AURELIJA USONIENE (Vilnius University), JOLANTA SINKUNIENE (Vytautas Magnus University & Vilnius University) & AUDRONE SOLIENE (Vilnius University)

The concept of multifunctionality has become inseparable from the analysis of discourse markers, pragmatic markers, modal particles and stance adverbials. Quite frequently and successfully the evasive semantic and pragmatic nature of those linguistic items is highlighted with the help of contrastive analysis based on parallel translation corpora (Aijmer & Simon-Vandenbergen 2003, Aijmer 2007, Johansson 2007, inter alia). Speaking in favour of ‘translation method’, Aijmer et al. (2006: 113) convincingly argue that it can ultimately provide “insights into the question of multifunctionality and how it relates to semantic and pragmatic polysemy“ and emphasize the need for more empirical work in different languages (ibid.).

This work-in-progress paper attempts to contribute to the existing literature on the broad range of semantic and pragmatic properties of *actually, naturally* and *in fact* (Aijmer 2002, Lewis 2006, Aijmer & Simon-Vandenbergen 2007, Grossmann & Wirth 2007, Mortier & Degand 2009, Defour et al. 2010) as based on contrast to their key translation correspondences in Lithuanian. Consider the following examples for illustration:

1. EN-orig: *He was angry, naturally, that I was giving up.*  
   LT-trans: *Žinoma, jis siuto, kad pasiduodu.*  
   Lit. ‘*Of course, he was angry that I was giving up.*’

2. EN-orig: *She was very strict, like most of the teachers, and naturally we were afraid of making mistakes.*  
   LT-trans: *Ji, kaip visos mokytojos, buvo labai griežta ir, aišku, mes bijodavome suklysti.*  
   Lit. ‘*She, like all teachers, was very strict and, clearly, we were afraid of making mistakes.*’

One of the tasks is to establish the dominant function of the expressions under analysis as well as to draw their semantic profile within the categories of epistemic modality and evidentiality. The study employs two types of corpora: a self-compiled bidirectional parallel corpus ParaCorp. (Usoniene & Soliene 2012) which consists of fictional texts (about 5 million words), and CorALiT (Corpus Academicum Lithuanicum: <http://coralit.lt/>), which is comprised of various academic texts in five science fields (about 9 million words). Academic discourse provides perfect empirical grounds to define pragmatic functions of Lithuanian correspondences of the
markers under study and, combined with the fictional corpus, to arrive at the description of their full functional potential.

The results do not only confirm the semantic and pragmatic multifunctionality of the markers under study in both languages but also point towards cases of language specific conceptualization and functional diversity, which irrevocably demonstrate the efficiency of the parallel-corpus-based analysis.

References


How do you spell yeah/yeh/yea/yah? Assessing orthographic transcription and comparability between spoken corpora

GISLE ANDERSEN (NHH Norwegian School of Economics)

In an earlier comparison of the two corpora: The Bergen Corpus of London Teenage Language (COLT) vs. Linguistics Innovators Corpus (LIV), I observed that several differences in the inventory of orthographic forms in the two corpora were due to systematic differences in methodology applied in the transcription of the two corpora, and not necessarily due to bona fide differences between the language use in the two corpora. Such systematic differences reflect different and/or inconsistent transcription conventions that seem to especially relate to the representation of:

1. non-standard/vernacular forms (innit/init)
2. phonologically reduced forms (d’ya/do you; gonna/going to)
3. discourse markers (yeah/yeh; kind of/kinda; sort of/sorta; cos/coz/’cause)
4. filled pauses (erm/uhm/umm)
5. minimal responses (mm/m)
6. standardised forms with inconsistent spelling (grandad/granddad/grand-dad)

There is a risk that the comparison of corpora that otherwise represent very similar groups of speakers may be hampered by such differences and inconsistencies. This is especially problematic for quantitative research, where a cross-corpus comparison would require a mapping of individual forms that can be said to represent a hyper-form (in some cases a lemma) in order to make qualified statements about real usage differences between speaker groups. Moreover, it is must be kept in mind that some of these near-similar orthographies may reflect potential different phonetic realisations between dialects or individual speakers (e.g. different vowel quality in erm vs. uhm?), and choice of orthographic form might suggest functional differences between forms similar at the surface level (e.g. cos = discourse marker; ’cause = conjunction?) or differences in the degree of grammaticalisation (e.g. kinda more grammaticalised than kind of?).

Naturally, it would be helpful and nice if standards and best practise documents could provide a unified way of handling forms and phenomena such as the above, but this seems somewhat unrealistic given the fluid and dynamic nature of spoken interaction, where new forms and ways of expression continuously arise. Moreover, standardisation needs to be balanced against a more general need for flexibility and adaptation in individual corpus compilation projects. Against this background, this paper firstly aims to review what existing standards and transcription conventions say about orthographic variation and vernacular forms, whether this is identified as a potential source of error, and what should be done to ensure maximal consistency. Secondly, the paper aims to further assess the degree of consistency within and between the two corpora mentioned and assess the seriousness of inconsistencies and the degree of difficulty in overcoming them.
“Ja, also, can you see me now?” Designing and compiling a corpus of computer-mediated international academic English

STEFAN DIEMER (Saarland University)

Various studies have established that as an international academic language, English has developed distinct features in lexis, syntax and pragmatics (Conrad & Mauranen 2003, Firth 1996, Howarth 1996, Mair 2003, Meierkord 1996). Several corpora have been compiled in this field, such as the *English as Lingua Franca in Academic Settings* corpus (Mauranen 2008) and the *Vienna-Oxford International Corpus of English* (Seidlhofer 2004). Recently, international spoken varieties have been attracting attention as they move further away from traditional English academic discourse (e.g. Jenkins et al. 2001). English is now the language of choice in communication between students and researchers in most of Europe, and also plays a central part in private computer-mediated communication (CMC), frequently between non-native speakers.

This presentation addresses issues related to compilation and annotation of a corpus of spoken discourse in the academic domain in an informal setting, with conversations conducted via Skype. The *CASE Sofia-Saarbrücken* corpus consists of more than 200 hours of conversations between German and Bulgarian students of English between October 2012 and January 2013 at Saarland University, Germany and Sofia University, Bulgaria. The combination of video and audio data presented a significant challenge in terms of data collection, organization and standardization. Moreover, the finished corpus should preserve pragmatic and CMC-related discourse features such as communication lag and overlap. While the corpus follows the precepts established in Wynne (2005), in particular Thompson’s (2005) guidelines for spoken corpora, several issues were encountered that went beyond the scope of established practice. These result mainly from the computer-mediated communication medium and from the international variety used by the participants. Some of the features to be presented and discussed in the workshop are:

- Technical aspects of using Skype for corpus compilation
- Recording, combining and standardizing audio and video
- Tagging technical and CMC-related features
- Identifying and transcribing non-standard language
- Conversation planning and background data
- Storage and availability of large spoken corpora

These features are also increasingly an issue in other corpora, while their consistent treatment remains problematic in the absence of best-practice guidelines. The presentation thus proposes a range of solutions for technical and compilation issues, puts forward possible transcription elements that work on the basis of Dressler’s and Kreuz’s (2005) model system, and discusses standardization and storage. It is hoped that the ensuing workshop discussion will further contribute to harmonizing best practice in relation to compilation and transcription of spoken corpora through the CMC medium.
References


From sociolinguistic interviews to a spoken corpus of London English: Creating the Linguistic Innovators Corpus (LIC)

**Costas Gabrielatos** (Edge Hill University), **Sebastian Hoffmann** (University of Trier) & **Eivind Torgersen** (Sør-Trøndelag University College)

This paper discusses the practical and methodological issues that arose in adapting a set of data that was originally compiled with a largely manual analysis in mind to be used within a quantitative, sociolinguistic research paradigm. The data in question stem from the project *Linguistic innovators: The English of adolescents in London* (2004-2007), which was the first large-scale sociolinguistic project set in London. Interview data were collected from more than 120 speakers, who were sampled according to age, gender, ethnicity and geographical location. The transcribed dataset, which contains 1.3m words, has been successfully analysed for morpho-syntactic and discourse features within a sociolinguistic research paradigm (see Cheshire et al. 2008, Cheshire & Fox 2009). At the time of compilation, it had, however, not
been envisaged that corpus linguistic analyses would be carried out on the basis of this data. Yet, this was the aim of the project *Analysing London English using corpus tools* (2008). As a result, despite the overall high accuracy of the transcriptions and the fact that the interviews had been transcribed using a standardised protocol, some modifications to the original format became necessary. In our paper, we will in a first step briefly outline these modifications. In a second step, we will then demonstrate how the new format can now be searched with standard corpus tools such as *Wordsmith* and *AntConc* as well as with Perl scripts. In particular, we will discuss how the rich levels of sociolinguistic meta-data available within LIC for all speakers can be fruitfully employed in such an endeavour, as reported in Gabrielatos et al. (2010) and Torgersen et al. (2011). This is important, as detailed information about the corpus speakers has in the past sometimes been missing in spoken language corpora, and indeed, not always been considered as that important when the goal was to have a large corpus (Sinclair 1995). Finally, we will raise some outstanding issues of corpus design that have not yet been applied in the current format of the corpus.

Certainly, neither the conversion process nor the search methodology presented in our paper are new, nor are they particularly advanced in their level of difficulty. Nevertheless, we feel that our extended experience in making use of the LIC-data for different research purposes can contribute to a general discussion of data standards beyond core corpus-linguistic data sources. In a time when corpus linguistics has “come of age”, such a discussion is no doubt beneficial for both corpus linguists and sociolinguists alike.

**References**


**The SPICE-Ireland pragmatic annotation scheme: A critical appraisal**

*JOHN M. KIRK* (Belfast)

SPICE-Ireland is an annotated version of the spoken component of ICE-Ireland, one of the national components comprising the International Corpus of English (cf. Greenbaum 1996). SPICE stands for ‘Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland’. The SPICE-Ireland corpus comprises 626,597 words of transcribed spoken language, equally divided between Northern Ireland and the Republic of Ireland. In line with other ICE corpora, the spoken material in SPICE-Ireland contains 300 texts of approx. 2,000-words from 15 spoken discourse situations. The release version is 1.2.2 (Kirk et al. 2011), which is accompanied by a User’s Guide (Kallen and Kirk 2012).
The transcriptions are annotated for the following features: utterance speech-act function, prosody (pitch movements), utterance tags, discourse markers, and quotatives. An example transcription from a face-to-face conversation – showing SPICE annotations in bold – is as follows:

<P1A-002$A><#><rep>That’s like Malcolm Dunne </rep><#><dir>Do you remember Malcolm Dunne said do you remember when we were working in X he said+ one day </dir><quote>1why is it that 2girlsyou-know%*like* 1just get 2here% and put it round 2hup% and then 2don’t put any makeup on the rest of their 2face% </quote><dir>laughter</dir><&>

<P1A-002$D><#><dir>2Malcolm%</dir>

<P1A-002$A><#><rep>Uh 2uh%</rep><#><dir>Cos he thought you-know% you 1know the way people don't rub their makeup 2in%</dir>

Pragmatic annotation cannot be done automatically and requires careful manual analysis. A full classification depends on pitch movements and discourse markers, some of which also function as utterance tags. Pragmatic use is also bound up with co-operation among the speakers, the degree of attention each speaker is paying to the face of his addressees, and other politeness phenomena.

This presentation will present the decisions taken for each component of the pioneering SPICE-Ireland annotation scheme, which was largely devised in 2003–04. At the time, as nothing so ambitious had been undertaken, and certainly not across 15 discourse situations, the emphasis was on the design of an annotation scheme which would be feasible, implementable and useable for the purposes envisaged, and which could also be completed within a relatively short space of time. Now that the corpus is released and being put to use, this paper will offer a critical appraisal of that scheme, mindful of the workshop’s subtitle: ‘towards best practice?’ Numerous typical as well as problematical examples with regard to classification and annotation will be discussed.

References


The Norwegian component of LINDSEI

SUSAN LEE NACEY (Hedmark University College)

This paper finds its starting point in observations made during the development of the Norwegian component of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) (Gilquin et al. 2010). It addresses some of the challenges inherent in the production of spoken corpora whose compilation depends upon widespread international collaboration.

The LINDSEI corpus was created under the direction of the Centre for English Corpus Linguistics (CECL) in Belgium, which specializes in the collection and use of learner and
multilingual corpora. The first edition of LINDSEI contains 554 interviews of learners from 11 different L1 backgrounds (usually 50 interviews per subcorpus). The interviews have been conducted in English, to capture the learners’ ‘interlanguage’, an idiosyncratic dialect that shares characteristics of their L1 and L2 (and sometimes other languages as well) (see Corder 1981: 17, 85).

The interviews in the individual subcorpora were recorded by researchers at academic institutions in the 11 different countries, adhering to a tri-fold structure previously determined by the CECL. Furthermore, researchers at these individual institutions were responsible for transcribing their respective interviews, following guidelines also provided by the CECL. Finally, the CECL was responsible for the collation of the 11 subcorpora into a single accessible database, with transcriptions (but not audio files) now available to researchers on a CD-ROM.

Plans for a second edition began even before the publication of the first edition, with the ambition of including the audio files, and linking the spoken and transcribed text. Moreover, data from learners with other language backgrounds will be added, thereby increasing the size and linguistic range of the corpus. One of the additional subcorpora in the expanded LINDSEI will comprise interviews of Norwegian speakers of L2 English; the creation of this Norwegian component is well underway, as a team of researchers at Hedmark University College (Norway) under my direction has now recorded 50 interviews and nearly completed all the transcriptions.

This paper focuses on the two specific areas of corpus design and orthographic transcription. It explores issues that were raised while retracing the footsteps of the first group of LINDSEI researchers through the employment of the already-established guidelines for corpus compilation and transcription. These issues include those relating to the particular Norwegian context (e.g. means of obtaining legal authorization for collection), general issues regarding compilation and transcription (e.g. the role/identity of the interviewer) and specific ones relating to the nature of learners’ interlanguage (e.g. ‘standard’ transcriptions of filled pauses). The overarching aim of this paper is to contribute to the development of ‘best practice’ in projects requiring diverse groups of researchers to create ‘comparable’ spoken corpora, so that such corpora may truly be comparable.

References


Categorising plurilingual user data? Challenges and solutions for POS-tagging VOICE

RUTH OSIMK-TEASDALE (University of Vienna)

This paper reports on the main issues encountered in the development of part-of-speech tagging (POS tagging) guidelines for the Vienna-Oxford International Corpus of English (VOICE), a corpus of spoken English as a Lingua Franca (ELF). Moreover, the technical procedures developed for tagging VOICE will be touched on and a practical demonstration of the online interface of the tagged corpus, VOICE POS Online will be given.
The well-known challenge of assignment of parts of speech to spoken language is an even more complex issue when applied to naturally occurring, plurilingual data, as is the case for ELF corpora. Such language use is common in our globalised world, where speakers exploit available linguistic repertoires, often resulting in non-codified and creative language use (e.g. Klimpfinger 2009, Pitzl, Breiteneder & Klimpfinger 2008, Pitzl 2011, Seidlhofer 2011). The problem this poses for POS-tagging also arises in some recent spoken corpora, where the recorded, naturally occurring data also contains plurilingual or non-native language usage (e.g. Rehbein, Schalowski & Wiese 2012, UíDhonnchadha, Frenda & Vaughan 2012). However, there are, to the author’s knowledge, no fully part-of-speech tagged corpora containing primarily spoken L2 data.

In the VOICE project, it was precisely this challenge that was taken up: to develop a part-of-speech tagging methodology for a corpus of spoken ELF, consisting of transcriptions of naturally occurring, highly interactive interactions among speakers from numerous different first language backgrounds. The first part of this paper will describe the 5 main issues encountered in establishing suitable tagging guidelines:

1. The lack of appropriate tagging models for spoken ELF,
2. highly interactive, spoken nature of the data, resulting in a vast number of possible tags,
3. non-canonical language use, especially that of non-codified forms and non-codified form-function relationships,
4. questions concerning the usefulness of external points of reference,
5. questions concerning the feasibility of categorising spoken, variable language.

The second part of this paper discusses the handling of these issues, which resulted in the development of specific guidelines for tagging VOICE. In conclusion, it will be proposed that these challenges and the corresponding guidelines, as well as the issues raised concerning the categorisation of spoken language in general, are relevant for the tagging of other corpora which contain naturally occurring, plurilingual language use.

References


This paper describes the construction of a deeply annotated multimodal corpus of spoken learner German and its native speaker reference corpus. While the analysis of interlanguage (Selinker 1972) is interesting on each particular linguistic level, it is even more valuable to study interactions across several levels. How do learners use prosody in connection with information structure? How do disfluencies tie in with lexical density or syntactic complexity? In order to make studies like this possible, we constructed a corpus which (a) enables maximally flexible annotation and analysis and (b) conforms to sustainable and well-described formatting standards.

The Berlin Map Task Corpus (BeMaTaC) uses a map task design (Anderson et al. 1991) where one speaker instructs another speaker to reproduce a route on a map with landmarks. The drawing hand of the instructee is recorded on video. BeMaTaC builds on ideas developed for the Hamburg Map Task Corpus HAMATAC (Schmid et al. 2010), which, however, is not consistently tokenized and documented and is therefore unsuitable for quantitative studies or further annotation. Our transcription follows a loosely orthographic scheme and is tokenized. All data is stored in an extensible and flexible multilayer standoff architecture, which allows the addition of annotation layers at any point. The freely extendable open source SaltNPepper converter framework (Zipser & Romary 2010) makes it possible to use many different dedicated annotation tools on the same data. It also allows for conversion into interchange formats such as PAULA (Dipper 2005) and the ISO-standardized GrAF (Ide & Suderman 2007). Unicode-formatted transcriptions, normalizations and phonetic/phonological annotations are created and carefully aligned with Praat (Boersma 2010), followed by automatic lemmatization and POS-tagging (Schmid 1994). The tagset employed is the STTS (Schiller et al. 1995), the de-facto standard for German. Token-based and span annotations such as disfluency tags are added using EXMARaLDA (Schmidt & Wörner 2009), while dependency structures are annotated using the MaltParser (Nivre et al. 2007).

BeMaTaC can be accessed using ANNIS (Zeldes et al. 2009), an open-source browser-based search and visualization tool for deeply annotated corpora. Using a highly flexible node-and-edge-based query language and permitting queries over multiple corpora, ANNIS can visualize data like simple token or span annotations but also syntax trees or pointing relations. BeMaTaC’s annotations and its audio and video recordings are time-aligned, permitting the playback of the corresponding sequence with a simple click on a token. All data is freely available under a Creative Commons licence.
References


**Best practices in the compilation, annotation and publication of the Research and Teaching Corpus of Spoken German (FOLK)**

**THOMAS SCHMIDT** (Institut für Deutsche Sprache, Mannheim)

FOLK (*Forschungs- und Lehrkorpus Gesprochenes Deutsch*) is the Research and Teaching Corpus of Spoken German. There is, to date, no larger, systematically stratified collection of publicly available recordings of authentic spoken interaction, let alone a consistent set of
corresponding, computer-accessible transcriptions, for German. Recognizing this, the Institute for the German Language started in 2008 to set up FOLK. Recordings for the corpus are partly collected from other sources (i.e. the institute’s spoken language archive and data donations from external projects), partly created anew for the project. The aim is to cover a broad spectrum both in terms of regional variation and in terms of different interaction types. In its first publicly available version, FOLK comprises 70 hours of transcribed audio recordings, published via the Database of Spoken German (DGD2). Towards the end of 2013, FOLK aims to reach 100 hours of audio recordings corresponding to roughly one million word tokens. The integration of video data is planned for a later stage of the corpus.

In order to ensure a high level of consistency, an efficient transcription workflow, high community acceptance and good automatic processability of the data, a group of conversation analysis researchers and a group of software developers were actively involved in the planning stage of the corpus. By coordinating corpus development, the specification of transcription conventions and the development of annotation tools in this way, FOLK also aims to contribute to the development, documentation and establishment of a best practice in the field of corpus based spoken language analysis.

My contribution to the workshop will discuss the principal elements of what we consider to be best practice. Starting with the recording where an adequate collection of communication and speaker metadata and the legal authorisation of the prospective uses of the data are paramount, I will focus on the question which transcription conventions, transcription formats and transcription tools are best suited to achieve both an economically efficient workflow and a high level of reusability of the transcribed data. Contrary to what Thompson (2004) suggests, we think that a state-of-the-art corpus workflow can, should and need not separate the phases of ‘transcription’, ‘markup’ and ‘annotation’ too radically. Also, modern data management techniques allow us to view the fundamental conflict between precision and machine-readability on the one hand and human readability and transcript layout on the other hand, in a new light. What we advocate, therefore, is to build spoken language corpora on the basis of an adequate and well-defined data model with a corresponding XML-based data format. As we will demonstrate, this opens up possibilities (and problems!) of data sharing and analysis that were hardly envisioned in Thompson’s paper. Concerning improvements on the current state of affairs, we think that a tighter integration of existing standards for transcription (i.e. conventions), data representation (i.e. digital formats) and their handling in transcription tools is the most interesting challenge. Ideally, future efforts towards a best practice would also take a much more international, multilingual and cross-disciplinary approach to transcription.

References


Curating and maintaining spoken legacy corpora for publication in the scientific community

KAI WÖRNER (HZSK University of Hamburg)

The Hamburg Centre for Language Corpora (Hamburger Zentrum für Sprachkorpora, HZSK) acts as a central point of contact for humanities researchers working (or planning to work) with – especially spoken – language corpora. The HZSK emerged from the Collaborative Research
Centre on Multilingualism (SFB 538) in Hamburg, which documented, analysed, and described language in the context of multilingualism. The projects involved used a huge variety of corpora for their research: spoken corpora as well as written ones, in different languages, transcribed to cover phenomena from diverse linguistic fields like phonetics, pragmatics, language acquisition, and others.

Towards the end of the research centre’s funding period, it became the task of the HZSK to curate these corpora into a state in which they could be reused by other researchers.

While it seems obvious that it is impossible to achieve total comparability for corpora that heterogeneous, the HZSK aimed to harmonize as many aspects of the corpora as possible.

For the spoken language corpora, this presupposed transforming the data into future-proof formats, consistent with existing and emerging standards. This involved digitizing analogue audio and video recordings and converting transcription, annotation and metadata files into formats suitable for sustainable archiving.

All of the transcriptions were converted into the XML format of EXMARaLDA – a system developed at the Research Centre on Multilingualism especially to address the needs of researchers working with multi-level annotated multilingual spoken data. EXMARaLDA’s data model is supported by a number of tools for creating, managing and analysing spoken corpora, and it allows for easy conversion into other widely used formats, like Praat, ELAN, TEI and others, as well as commonly used visualization formats.

On the content level, the curation of the corpus metadata was also a very labour-intensive task, since most projects were not only interested in different aspects of the conversations and the participating speakers, but also used arbitrary vocabularies to describe similar aspects in their corpus metadata. Also, existing metadata standards had to be considered in the process of creating a ‘common ground’ for the metadata profiles to be established.

In this workshop contribution, I will present some examples from the curated corpora, explain the most important steps taken for their curation, discuss some of the limits when trying to harmonize heterogeneous corpora and comment on the aspects that could have benefited from a best practice for the compilation and annotation of spoken corpora.

References


OPERATIONALIZING AND ASSESSING WRITING PROFICIENCY IN THE ACADEMIC REGISTER: THE CORPUS OF ACADEMIC LEARNER ENGLISH (CALE)

MARCUS CALLIES & EKATERINA ZAYTSEVA (University of Bremen)

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001, 2009), a highly influential construct in language testing and assessment, has recently been the object of criticism by language teachers and language testing experts (e.g. Alderson et al. 2004, Hawkey and Barker 2004, Hulstijn 2007, Weir 2005). In particular, the practical implementation of the framework's descriptive ‘can-do-statements’ in educational settings is problematic due to their language- and context neutral character. Thus, there is agreement that the description of proficiency levels provided in the CEFR needs further specification with regard to a) individual languages, b) learners’ abilities to use a language in specific registers, and c) a reliable yardstick for evaluating learner language performance. This paper reports on work in progress that examines how language corpora can be used to overcome some of the limitations of the CEFR by means of identifying more explicit, language-specific linguistic descriptors or ‘criterial features’ (Hawkins & Filipovic 2012) anchored in contextualized language use. We will demonstrate how language proficiency in a specific register can be operationalized in terms of linguistic descriptors based on academic writing produced by native speakers and advanced learners of English. This approach suggests a corpus-based implementation of several well-known characteristics of academic prose combined with a corpus-driven assessment of proficiency (e.g. Wulf & Gries 2011) using the Corpus of Academic Learner English (CALE). The first step is to draw up a list of characteristic features of academic prose from which linguistic descriptors are selected in terms of their ‘keyness’ (i.e. how important and characteristic they are of the register) and their operationalizability (i.e. how well they can be retrieved from the corpora and subjected to statistical analysis). Second, a feature is retrieved semi-automatically from the corpus and subjected to statistical analyses to identify clusters formed by samples of data (i.e. learner texts) that demonstrate the highest degree of similarity as to the occurrence of one or more features. In our talk, we will present a case study on a specific linguistic descriptor using a pilot version of the CALE.

REFERENCES


An evaluation of Avalingua based on learner corpora

PABLO GAMALLO, MARCOS GARCÍA (CITIUS, University of Santiago de Compostela) & ISAAC GONZÁLEZ, MARTA MUÑOZ, IRIA DEL RÍO (Cilenis S.L)

This article describes the use of two learner corpora to evaluate precision and recall of a linguistic tool, called Avalingua, based on Natural Language Processing techniques. Avalingua is a linguistic software aimed at automatically identifying and classifying spelling, lexical, and grammatical errors in written language. The objective of the tool is to analyse the linguistic errors of texts, assess the writing proficiency, and propose solutions with the aim of improving the linguistic skills of students. It makes use of natural language processing and knowledge-rich linguistic resources. It has been developed, so far, for Galician language, and can be applied to both L1 learners of Galician (i.e., language acquisition in children) and L2 language acquisition. Avalingua for Galician consists of the following modules:

- Identification of spelling and lexical errors, which is constituted by two submodules: a state of the art spelling checker, and a classifier of usual lexical errors: e.g., Spanish and Portuguese expressions, archaisms, anglicisms, hyper-authentic terms, etc.
- Identification of integers, tokens with special characters, and proper names, which must not be considered as incorrect expressions even if they are not found in a reference (gold standard) vocabulary of Galician language.
- Identification of tokens constituted by productive affixes, and frequent Out Of Vocabulary words. Even if they are not contained in the reference vocabulary, they cannot be considered as lexical errors.
- Identification of fake friends Galician-Spanish, that is, words appearing in the reference vocabulary with very marginal senses, but which are also very common Spanish words.
- Identification of different types of grammatical errors by using a dependency based syntactic parser.
- Language identification to find citations or quotations in other languages within the text.
To evaluate the performance of *Avalingua*, we made use of two learner corpora. The first one is a collection of writing texts belonging to Galician children in the 3rd year of secondary school (3rd grade of ESO). The second one consists of texts written by adult Portuguese L2 learners of Galician language. Precision was computed by counting the number of 1 correct decision made by the system divided by the total number of decisions it made. Recall is the number of correct decisions made by the system divided by the total number of errors (spelling, lexical, and grammatical errors) found in the text. *Avalingua* achieved 93% precision and 66% recall when it was evaluated using the first corpus (Galician children), and 90% precision using the second corpus (Portuguese L2 learners). The high quality performance achieved in the two tests shows that the system can be viewed as an useful tool to help teachers to assess writing proficiency of students, not only for L1 acquisition, but also for L2 learning. The article will provide a more accurate analysis of the evaluation results, by taking into account issues such as frequent lexical and grammatical mistakes, texts with more error rate, and so on.

Given the modular structure of the system, it is possible to adapt *Avalingua* to other languages, such as English. In the last two years, there has been an increased interest in developing automatic tools for identifying and correcting both spelling and grammatical errors in texts written by English learners. This growing interest has led researchers to organize two international competitions, namely HOO-2012 and CoNLL-2013 shared task, aimed at comparing the efficiency of different systems trained on a large collection of texts written by English students. Those learner corpora are completely annotated with error tags and corrections, and all annotations have been performed by professional English instructors. In future work, our objective is to apply *Avalingua* on English texts, by using as source of errors learner corpora available from HOO-2012 (http://clt.mq.edu.au/research/projects/hoo/) and CoNLL-2013 (http://www.comp.nus.edu.sg/~nlp/conll13st.html). For this purpose, we will analyse the more frequent grammatical errors found in those corpora and define appropriate correction rules with DepPattern.

**References**


Operationalizing fluency in (learner) corpora: Implications for language testing and assessment

SANDRA GÖTZ (Justus Liebig University Giessen)

Fluency is a widely used notion when speaking about – and assessing – both native and non-native speech. Previous research on fluency has shown that describing its linguistic substance "with a degree of consensus is notoriously difficult" (Hasselgren 2002: 147), because, firstly, various definitions of the concept of fluency co-exist and, secondly, fluency is an epiphenomenon to which many individual (and interrelated) factors contribute. Most of the fluency-related research so far has focused on one of the following aspects:

1. temporal variables in speech production, such as length of runs, pause ratio, speech rate, etc. (e.g. Lennon 1990, Chambers 1997, Cucchiarini et al. 2002, Gut 2009);
2. the use of prefabricated units and formulaic sequences (e.g. Pawley & Syder 1983, Wray 2002, Erman 2007);
3. certain fluency enhancement strategies that are used to cope with the planning pressure in online speech production, such as the use of discourse markers, smallwords or repeats (e.g. Hasselgren 2002, Müller 2004, Rühlemann 2006).

The present paper combines these three approaches to fluency and suggests one way of how fluency can be operationalized in learner corpus analysis quantitatively and more holistically.

Based on this taxonomy, firstly, a comprehensive contrastive analysis of the individual fluency variables of the learner data of the 86,000-word German learner corpus LINDSEI-GE and the comparable 90,000-word native speaker corpus LOCNEC will be presented in order to identify areas in which, on the one hand, advanced German learners of English still deviate strongly from the native target norm (e.g. in the range and distribution of formulaic sequences) and in which they have already approximated to the target norm on the other (e.g. in their use of repeats). Here, both the corpus totals as well as the performances of the individual speakers will be taken into consideration in order to account for clearly visible differences in proficiency levels (e.g. Pendar & Chapelle 2008) that often go undetected or tend to be disregarded in learner corpus analysis.

Secondly, I would like to present the findings of an analysis that looked at each speaker's performance in combination. This analysis reveals that a nativelike fluency performance can be characterized by very different patterns: There are fluency enhancement strategies that show in each speaker's output (e.g. an equal proportion of repeats), whereas others have shown to be rather exchangeable (e.g. speakers either use a high proportion of discourse markers or formulaic language to establish a high level of fluency). For the learner data, however, no such clear fluency clusters become visible and not all fluency variables are represented in the learner output to the same extent, which suggests that the learners have not internalized a nativelike variety of fluency enhancement strategies.

In the light of these findings, in a last step some language-pedagogical implications for language testing and assessment of fluency will be discussed.

References


**Insights from a corpus of secondary school English examination essays in Lithuania**

**Rita Jukneviciene (Vilnius University)**

The use of learner corpora is increasingly gaining more prominence in various language assessment contexts. In this talk, I would like to share experience from Lithuania and demonstrate how a learner corpus, consisting of learners’ examination papers, might shed light not only on the learners’ achievement levels but also contribute to a better understanding of the washback effect of the examination itself. More specifically, the presentation will introduce a recently started corpus compilation project in Lithuania which is being carried out at the National Examination Centre (NEC corpus). The NEC corpus consists of examination papers, namely, personal letters and essays, written at the end of the secondary school by pupils who choose to take the national examination of English. The corpus is being developed as an in-house resource for test development, analysis of linguistic features of learner language, alignment of the examination with the CEFR, training of examination markers and materials development. At its present state, it consists of ca. 150,000 words.
One of the major reasons for the compilation of the NEC corpus was the establishment of achievement levels of the learners according to the CEFR (Council of Europe 2001). For this purpose, the Cambridge English Profile project data will be used as one of the reference sources (Hawkins and Filipovic 2012, UCLES 2011). Some preliminary findings from an ongoing research project will be presented in the workshop. The second part of the presentation will deal with corpus evidence which unexpectedly revealed the washback effect of the examination. A preliminary analysis of word frequency lists and the most frequent recurrent sequences of words, also termed clusters, lexical bundles or n-grams (Scott 2008, Biber 2006), showed that there are certain sequences that re-occur in many learner texts but are quite rare in English. It turned out that these phrases are included in the assessment instruction of the writing tasks to make it easier for the markers to mark the papers. After the marking session, the assessment information goes back to schools and, undoubtedly, significantly influences the teaching process. The corpus evidence seems to indicate that this dissemination of the list of ‘good’ phrases from the assessment instruction is leading to unwanted consequences. If the teaching of English gets reduced to the teaching of the assessment instruction, the examination, undoubtedly, has the worst possible washback effect (Alderson et. al. 1995, Hughes 2003). It was only owing to the corpus data, that this unwanted practice came to light and provided convincing evidence to the examination developers and administrators. So even at its pilot version, the corpus of examination papers has been very instrumental in the revision of the assessment procedures and documents.

References


MILE – The Marburg Corpus of Intermediate Learner English

ROLF KREYER (University of Marburg)

Over the last few decades corpus linguistic methods and research have had a considerable impact on the foreign language classroom, in particular with regard to curriculum design and reference tools. Corpora have also found their way into the classrooms themselves: teachers have begun to use corpora and corpus-based findings to supplement textbooks and textbook
materials, even though there still seems to be a general “need to convince practicing teachers to use corpora and concordances in the classroom” (Römer 2006).

More recently, researchers have become increasingly interested in the language use of foreign language learners. This has led to the compilation of a fairly large number of learner corpora with learners from different L1s, such as ICLE (Dagneaux et al. 1998), Lindsei (Granger 1997) or CALE (Callies & Zaytseva 2011). What these (and other) learner corpora have in common is that they represent learner language of a high intermediate to advanced level, usually from learners in university settings. More specifically, learner language at a truly intermediate level has remained underrepresented, although the Flensburg English Classroom Corpus (FLECC, Jäkel 2010) and the International Corpus of Crosslinguistic Interlanguage (ICCI; Tono et al. 2012) are promising first steps in that direction.

The Marburg Corpus of Intermediate Learner English (MILE), a corpus currently being compiled at the University of Marburg, aims at creating a collection of written learner English from grades 9 to 12 of the German Gymnasium. The corpus is designed as a longitudinal database and will document the progress of learners in their final years of school. In addition to rich annotation regarding biographical data, the corpus will be error-tagged by a group of native and non-native university teachers of English. This will provide researchers with the opportunity to analyze the development of intermediate learner English in unprecedented detail, e.g. how does error X develop in the final three years?, how do collocational errors develop?, what influence does a year abroad have?, how large is the range of proficiency within groups of the same institutional status?, etc. The paper, in addition to presenting basic facts about the MILE corpus (e.g. the general architecture, the underlying database or the envisaged corpus-query possibilities), will focus on questions of annotation. More specifically, the paper wants to explore whether and how annotation regarding errors and mistakes together with annotation regarding ‘good’ (e.g. idiomatic, efficient, versatile) use of language can be applied to create a learner corpus that helps us to get a clearer idea of the fuzzy variable of learner proficiency.

References


Learner corpora are compiled for a variety of purposes ranging from theory-driven studies into the processes of second language acquisition (e.g. ICLE) to more practical applications for syllabus design and materials writing (e.g. Longman Learner Corpus). Among the issues which often come up in the explorations of learner data is the identification of L1-dependant and L1-independent characteristics of interlanguage. One of the recent projects in learner corpus creation, The International Corpus of Crosslinguistic Interlanguage (ICCI), had exactly this purpose as one of its major objectives.

The ICCI is a collection of essays written in L2 English by primary and secondary school students from grade 4 to grade 12. The essays were collected in Austria, Poland, Spain, Israel, China, Hong Kong, Taiwan and Japan (Tono et al. 2012). The data in the corpus were classified according to learners’ country of origin, first language and school grade level. This followed a standard procedure applied in learner corpus compilation of using learners’ L1s and their institutional status as major classification criteria.

Yet, the attempts to use the corpus for the intended purpose ended up with some problems, as the essays were not directly comparable across countries and first languages due to the variability in the proficiency levels (Leńko-Szymańska 2012, Levinsky-Aviad 2012). Thus, there is a need for a more objective measure which would allow for more meaningful comparisons among data sets.

One such measure can be the essay length. As has been asserted by numerous researchers (e.g. Wolfe-Quintero et al. 1998), the length of an essay may serve as a very rough indicator of learners’ fluency, or at least of the discrepancies in fluency between students. Since the essays in the ICCI database were written in similar conditions and on comparable topics, it is striking that the Austrian students in grade 5 produced compositions of the same length as the Spanish students who were six years older and had six more years of experience in learning English.

Another measure that has been prosed as a discriminating factor is vocabulary size. As Milton (2010) states “it would be expected that as language level increases so would the learner’s knowledge of vocabulary and the sophistication with which that vocabulary can be used.” The research within the framework of the English Profile project is geared towards distinguishing criterial features of learner language (Hawkins & Filipovic 2012) and as part of this research the English Vocabulary Profile was developed featuring lists of words that students typically know at each level – A1 to C2 – of the CEFR.

The present study is exploratory in nature. It aims to investigate to what extent the vocabulary range of learner writing can be used to distinguish between students’ levels. The analysis will concentrate on essays from three L1 backgrounds: Polish, Spanish and German at three important turning points in the respective educational systems, grades 6, 9 and 12. Each of the essays will be examined in terms of the percentage of its vocabulary characteristic of levels A1-C2 (based on the English Vocabulary Profile lists). The percentages will be compared with essay lengths to establish whether these two types of measures converge. Although this method cannot give the definite answer whether the vocabulary range alone can serve as a sufficient indicator of a learner’s proficiency, it can shed some light on the applicability of the English Vocabulary Profile lists as a benchmark for comparing learner data within and across corpora.
References


Longman Learner Corpus. Available at: <http://www.pearsonlongman.com/dictionaries/corpus/learners.html>


Proficiency in writing online: First steps in assessing a learner corpus of computer-mediated communication among university students in Japan

TIM MARCHAND, SUMIE AKUTSU & BEN ROWLETT (J. F. Oberlin University)

In recent years, scholars have found insights gained through Learner Corpus Research (LCR) to be useful in the field of interlanguage development and L1 interference. Assigning learner proficiency within LCR remains an important issue, however, as L1 interference may manifest itself differently at different levels of proficiency (Jarvis and Pavlenko 2008). There has recently been debate over the best way to assign reliable proficiency levels to corpus texts, with the useful distinction made between learner-centred methods and text-centred methods of level assignation (Carlsen 2012).
This presentation looks to contribute to this debate by describing a case study of a learner corpus compiled from the written work of university students in Japan. The course in question uses Computer-Mediated Communication (CMC) for the dual purpose of providing lesson materials online and collecting student written output to develop the learner corpus. The paper will briefly outline the setup of the course, explain how the corpus was constructed, and describe a text-centred rating scale used to assess the written output, before discussing the merits of both the rating scale and CMC itself in measuring learner proficiency in writing.

The CMC of this study takes the form of a blog where, each week, an article about a recent news item, together with supporting class materials, is posted online for students to access. Students read the news story and, after a classroom session, write their reactions to the story on the class blog. These comments form the basis of the learner corpus and are assessed using a rating scale specifically designed to help with the analysis of the corpus.

In a previous paper, Marchand (2013) has argued that the grammatical and lexical features of CMC differ significantly from both written and spoken registers of English, although overall its features may be considered as an intermediary register between the two. The presentation will therefore explain why CMC was perceived to be a suitable medium for gauging learner proficiency and interlanguage development. It will also outline the decision-making process in constructing the rating scale and show how proficiency levels were designated to individual corpus texts. Preliminary results of the corpus analysis of sub-corpora defined by these levels will be correlated with those from two learner-centred methods of assigning proficiency: 1) learner test scores throughout the course, and 2) learner educational background.

The presentation concludes by assessing the construct validity and reliability of the rating scale employed, and determining whether CMC is a useful medium for examining the development of the L2 interlanguage for Japanese speakers of English.

References


Data commentary in science writing: A corpus-based comparison of research articles and master’s theses in technical fields for formative self-assessment practices

LENE NORDRUM & ANDREAS ERIKSSON (Chalmers University of Technology)

In the wide context of English for Academic Purposes (EAP) and English for Specific Purposes (ESP), increasing attention is devoted to an integrated approach of a linguistically oriented bottom-up corpus analysis and discourse-oriented top-down analysis of macro-structural units, or moves, such as Swales’ (1990) model *Creating A Research Space* (CARS) of introduction sections in research papers (see e.g., Partington et al 2004, Biber et al. 2007, Ådel & Reppen 2008). One prerequisite for a combination of bottom-up and top-down perspectives is a genre-
specific corpus annotated for discourse moves (Flowerdew 2004, Biber et al. 2007). Although such corpora work at the expense of “a balanced and representative view of the English language as a whole” (Hafner & Candlin, 2007:307), it is argued, they enable a thicker analysis and description of disciplinary discourse than is possible in a larger, more general corpus. Importantly, these corpora have the applied benefit of providing students with a wider repertoire of disciplinary-sanctioned linguistic resources that can form the foundation for corpora-supported learning activities (Chang & Kuo 2011, Upton & Cohen 2009). Some promising work has been carried out in this vein, notably Kanoksilapatham’s (2005) analysis of typical linguistic manifestations of moves in biochemistry articles, and Stoller & Robinson’s (2012) more pedagogically oriented study of moves in chemistry articles (see also Chang and Kuo 2011, on moves in computer science research articles).

In this study, we use a combined corpus and discourse approach to shed light on how graphs, figures and tables are presented verbally in science writing (data commentary). The multi-modal nature of science writing has been pointed out as an ‘important problem’ in linguistic approaches to disciplinary discourse (Shaw 2007: 6), but remains relatively under-investigated (see e.g. Swales & Feak 2004, Wharton 2012). Two issues are particularly important in this context: modal affordance, i.e., knowledge of the strengths and weaknesses of particular modes, and transduction, i.e., translation of the visual mode to the verbal written mode (Kress 2010). From an applied perspective, research has shown that integrating verbal and visual modes represent a complex task for students (Blåsjö 2011, Wharton 2012), and that transduction practices tend to vary even between relatively closely related fields (Stoller & Robinson 2013).

The aim of the study is to discuss how formative self-assessment practices in graduate and post-graduate ESP courses at technical universities can be developed from our combined corpus and discourse analysis of data commentary. This aim involves comparisons of expert and novice texts from a discipline-specific corpus of research articles (RA) and a complimentary advanced-learner corpus of master’s theses, annotated for discourse moves. We argue that developing self-assessment practices is of crucial value to universities from two related perspectives. First, pedagogical and curricular developments in science education emphasize students’ awareness of communicative practices and variance. Second, as international and national audits of higher education become increasingly informed by student work, students’ self-assessment skills of communicative output can provide institutions in higher education with important competitive advantages.

References


In this research, we aim at profiling the OPI from a variationist perspective, using, in the main, English native speaker as well as learner data. By profiling our informants’ OPIs, we aim at shedding light into the characteristics of this particular register as manifested across different speaking tasks. This way, we set out to research one area of the assessment of proficiency that is usually neglected: that of the linguistic nature of the communicative event which is used to assess general ‘proficiency’ in a given language.

In our research corpus, we have gathered 78 interviews of English native speakers and have analysed the sub-components of the LINDSEI corpus (Gilquin, De Cock & Granger 2010). For ease and convenience of corpus inter-comparability, each interview in our research corpus includes three different speaking tasks, namely, a Personal Narrative Component, an Interaction Component and a Picture Description. Our corpora (the native speaker component and the whole LINDSEI corpus) were POS tagged and analysed using Multidimensional Analysis (Biber 1988, Conrad & Biber 2001, Biber 2003, 2006). This research methodology seeks to interpret linguistic data in the light of language variation across registers and different dimensions of use.
Multidimensional Analysis (MA) of learner language has been underused as a tool for language research and pedagogy. One of the few studies where MA was used to explore learner language is Connor-Linton & Shohamy (2001) and one of the few pedagogic applications of MA is Aguado, Pérez-Paredes & Sánchez (2012). The results of our MA corroborate that the OPI is an adequate register to develop an assessment of learner language proficiency in communicative contexts which are characterized by involved production. However, we have found that the peculiarities of the different speaking tasks in an OPI may determine the range of linguistic features which are more likely to be generated by the communicative potential of the task itself. For example, for native speakers the normalized frequency of present tense verbs (118/1000) in the Picture Description Components is considerably much higher than in the other two components, almost doubling the frequency of this feature in the Personal Narrative Component (69.9/1000). While the analysis of isolated linguistic features in language may allow for an extremely fine-grained view on learners’ use of certain morphological areas of language use, this perspective alone fails to grasp a substantial component of language as used in real-life communication purposes in real-life contexts.

This profiling is of interest in areas such as language assessment, where the OPI is widely used to evaluate the speakers’ communicative competence, but also in the field of learner language research. Our research is an example of the new possibilities which combined methodologies of learner language analysis can offer to both the SLA research community and the more general applied linguistics research field.

References


Why speakers use the –ing form in certain communicative situations – a functional explanation

GRANATH SOLVEIG & MICHAEL WHERRITY (Karlstad University)

In the act of communicating, speakers are sometimes required to choose one of two or more linguistic forms, e.g. a gerund or an infinitive complement, whose informational content is essentially the same (see Wherrity 2001). In our paper we will show how taking a monosemous approach to grammatical morphemes in English can provide an explanation for the strategies speakers use in making such choices. Our analysis draws on principles put forth by the New York-based Columbia School of Linguistics (see e.g. Reid 1991). What we are proposing is that grammatical signals such as –ing and the infinitive marker to have one rather than several meanings. Accordingly, we will argue that instead of mentally storing a host of separate senses for such signals, language users work with basic core meanings which they “press into service” in varying contexts to convey what they wish to say. In our attempt to explain how this works, we will make use of four key terms: sign, signal, basic meaning and message. A basic meaning is the synchronically invariant semantic component of the bipolar sign which remains constant in all contexts, while the signal is its physical representation. The message is what motivates speaker choice and is not linguistically encoded in the utterance. Speakers select those basic meanings which are best suited to guiding the hearer in the construction of the intended message while requiring a minimum of processing effort.

Our material is taken from several of our previous corpus studies, where we demonstrate how this approach can explain speaker variation in the choice between I don’t want you coming around here anymore and I don’t want you to come around here anymore (see Wherrity and Granath 2008, Granath and Wherrity 2012) as well as in the choice between the simple tense and the progressive with stative verbs (Granath and Wherrity 2008, 2012). One characteristic of stative verbs is said to be their inability to occur in the progressive (see e.g. Quirk et al. 1985: 198). Nevertheless, our corpus material shows that even with such a prototypical stative verb as own, speakers may select the progressive to express a specific message, as in More and more people are owning and using cell phones today. In summary, our presentation aims to show that the grammatical variation seen in these cases is a function of speakers choosing particular grammatical signals having the basic meanings which best accord with the message they wish to communicate. Speaker choice is thus not explained in terms of senses encoded in constructions (as they are said to be for the progressive in e.g Núñez-Pertejo 2004, Smitterberg 2005, Kranich 2010) nor in terms of ad hoc notions such as ‘licensed rule-breaking’ (Mair 2006: 92), but rather, in terms of the semantic contribution made by a limited number of invariant basic meanings to which resulting messages are ultimately answerable. The approach taken here makes it possible to economically account for choices involving a variety of constructions.
References


Variation across three dimensions: Testing the complexity principle on adjectival data

HENRIK KAATARI (Uppsala University)

In this paper I will discuss the implications of the complexity (transparency) principle and the distance principle (Rohdenburg 1995, 1996) on the choice between that- and to-complements. More specifically, I will examine constructions such as those exemplified in (a)-(d), thus testing the applicability of the complexity and distance principle on adjectival synchronic data.
(a) It was inevitable (that) he should be nicknamed ‘the Ferret’, although seldom in his hearing. (Extraposed)

(b) I’m happy (that) we are married. (Post-predicate)

(c) It is difficult to test a potential cure when a disease is ill-defined. (Extraposed)

(d) Yet the authorities were unable to silence the expression of political opposition. (Post-predicate)

The adjectival data in (a)-(d) is subject to variation in a number of different ways. I will concentrate on examining variation along the following three dimensions:

1. Inter-clausal variation – that- vs. to-clauses
2. Intra-clausal variation – that-omission vs. that-retention
3. Variation across constructions – extraposed vs. post-predicate clauses

These three dimensions have a bearing on the complexity principle in that they are all represented by one more explicitly marked and one less explicitly marked grammatical option. Following this, I will test four hypotheses:

1. To-clauses are expected to be found in less complex structures than that-clauses (cf. Rohdenburg 1995, 1996).
2. The presence of inserted material between the matrix and complement clause is expected to favour that-clauses, since they are more clearly marked syntactically (cf. Rohdenburg 1996).
3. That-omission is expected be found in less complex environments than that-retention. Also, that-clauses with NP subjects are hypothesized to retain that to a greater extent than that-clauses with pronominal subjects (cf. Rohdenburg 1996).
4. Extraposed clauses are expected to be found in more complex structures than post-predicate clauses following the principle of end-weight (cf. Wasow 1997, Biber et al. 1999).

Drawing on Szmrecsányi (2004), I will use sentence length as a measure of syntactic complexity.

References


Processing complexity of English comparative variants – experimental results

GERO KUNTER (Heinrich Heine University Düsseldorf)

English offers two alternatives for the comparative form of adjectives: a synthetic variant in which the comparative is realized by the suffix -er, and an analytic variant formed by a phrasal construction using more. As more comparatives provide an early cue to the following degree phrase, Mondorf (2009) argues that they are easier to identify than synthetic comparatives; also, parsing of these constructions should be easier, because each word is mapped to exactly one function.

Based on Rohdenburg’s (1996) Complexity Principle, Mondorf (2009) proposes that comparative alternation in English can be explained by the aforementioned differences: she predicts that in environments that are cognitively complex, speakers will prefer the analytic comparative over the synthetic variant, as the processing advantage of the former construction can be used to compensate for the increased processing load of the complex environment.

The present experiment tested the assumption that analytic comparatives are indeed easier to process than synthetic comparatives. 40 native speakers of English participated in a lexical decision task in which both comparative forms of 60 different adjectives were presented in a randomized order as auditory stimuli. All adjectives were attested in COCA (Davies 2008-) to allow both variants. The processing complexity of the two variants was determined by recording the latency time required to determine whether the respective stimulus existed in English.

The results show no evidence of a processing advantage of analytic comparatives over the synthetic alternatives. Instead, more comparatives yielded generally longer reaction times, which indicates that these constructions were more difficult to process than the corresponding -er comparatives. While it remains possible that analytic comparatives are easier to process in production, they do not appear to facilitate recognition. Thus, it appears to be unlikely that speakers use analytic comparatives as a compensation strategy in cognitively complex environments.

References


A cross-linguistic perspective on variation – order in English and Japanese NP conjuncts

ARNE LOHMANN (University of Vienna) & TAYO TAKADA (University of Hamburg)

While many studies investigate the variables underlying linguistic variation in one language, few works investigate their workings across languages. The present paper does so by addressing a conspicuous difference between English and Japanese. For both languages it has been shown that the ordering of words and phrases is strongly influenced by their length. However, opposing effects of this variable have been reported: while English displays a short-before-long preference, the reverse tendency seems to hold for Japanese (Yamashita and Chang 2001, Yamashita 2002).

These divergent findings can be explained within Hawkins’ Minimize Domains account (Hawkins 2004), which is based on the tenet that speakers try to minimize the distance between dependent heads of phrases. Crucially its predictions vary for head-final (e.g. Japanese) as opposed to head-initial languages (e.g. English). The results pose a serious challenge for current language production theories, however, which hold that the shorter constituent is uttered first, as it is considered more accessible (e.g. Bock and Levelt 1994). Since the findings for Japanese contradict this logic, Yamashita and Chang (2001) propose that accessibility affects the two languages differently during sentence production. They claim that for Japanese speakers semantically salient and therefore long constituents are more accessible than short constituents during the syntactic ordering stage.

We contribute to the analysis of these cross-linguistic differences by investigating order in NP conjuncts in both languages, as exemplified below ((2) is the Japanese translation of (1)).

(1) the professor and the assistant

(2) kyōju to joshu

Since empirical evidence (cf. Temperley 2005) suggests that these constructions are not influenced by domain minimization, they provide a suitable test bed for the investigation of the hypothesis of language-dependent accessibility effects. We extracted the relevant constructions from the spoken parts of the International Corpus of English (GB-edition) and the Corpus of Spontaneous Japanese, yielding ~900 data points, respectively. The data was coded with regards to ordering constraints identified in prior research (see Chang 2009, Lohmann 2011), most importantly the lengths of both NPs in the conjuncts, and analyzed quantitatively using multifactorial logistic regression. The most important result obtained is a significant short-before-long preference for both English and Japanese (Japanese length regression coefficient estimate = 0.30, SE = 0.12, z = 2.56, p= 0.01; English length regression coefficient estimate = 0.63, SE = 0.08, z = 7.46, p < 0.001). Hence no evidence for diverging accessibility effects between the two languages and thus for the aforementioned hypothesis was found. Furthermore, in contrast to prior research, the results show that Japanese exhibits varying preferences: in some syntactic contexts a short-before long ordering prevails, while in others a long-before-short ordering is preferred. We interpret this as evidence that both universal accessibility factors and the desire to minimize dependency domains drive ordering in Japanese. Both should therefore be included in a model for constituent ordering (cf. Temperley 2007).
References


Form does not follow function (but variation does). The origin of possessive have got

DAVID LORENZ (Albert Ludwig University of Freiburg)

This paper aims to assess where ‘functionality’ comes in in a process of variation and change. Functionality is here defined as any factor that serves to improve communication or processing efficiency. A purely functionalist view is represented by MacWhinney et al. (1984: 128): “The forms of natural languages are created, governed, constrained, acquired and used in the service of communication”. However, Bybee et al. argue against functionality as the driving force of language change: “[T]he relation between grammar and function is indirect and mediated by diachronic processes. The processes that lead to grammaticization occur in language use for their own sakes” (1994: 298).

In English, stative possessive meaning can be expressed by the main verb HAVE (Jim has a very old car) or by the construction HAVE got (Jim has got a very old car). Examining the beginnings of this variation, I argue that the stative possessive use of HAVE got has indeed arisen ‘for its own sake’. It emerges gradually through invited inference, increasingly detaching it from its original perfective meaning (‘have obtained’). This is shown by analyzing the use of HAVE got in the Penn-Helsinki Corpora between 1150 and 1650 through a context model based on Heine (2002). Functional explanations, such as Crowell’s (1959) ‘pattern preservation’ (got being inserted to compensate for the frequent contraction of HAVE to ‘ve/s’), fail to explain the data.

While the emergence of possessive HAVE got appears to be a largely mechanical linguistic ‘drift’, functional ‘support strategies’ come into play once the variation in expressing
‘possession’ is established (e.g. the effect of generic subjects, Tagliamonte 2003). What this suggests is that functional strategies do not initiate language change, but they determine variation.

References


A data-driven, automatic parser-based approach to variation in English

GEROLD SCHNEIDER (University of Zurich)

Grammatical variation is more pervasive than has been assumed (see e.g. Rohdenburg and Mondorf 2003). A number of suggestions has been brought forward to explain variation in psycholinguistic terms (e.g. Hawkins 2004). We suggest a data-driven model of argument structure and alternation which reduces ambiguity to a minimum. The parsing task is modeled as competition between the competence-driven syntax principle, which contains the set of permissible grammar rules, and the performance-driven idiom principle (Sinclair 1991), which uses lexical preferences to disambiguate.

Language is rarely perceived as ambiguous by native speakers. Most utterances are understood easily; unintended readings are usually not noticed except in garden path sentences. Computational linguistics has shown that ambiguity is much more widespread than commonly assumed. But disambiguation by human readers and automatic parsers share common characteristics (Schneider et al. 2005). Human readers disambiguate more efficiently, usually even without perceiving the ambiguity. Important psycholinguistic reasons for human disambiguation are the following pragmatic factors, which we investigate by using computational linguistics methods:

- Lexical preferences and collocations: readers rarely completely analyze sentences into their individual components, instead we perceive and use semi-preconstructed phrases (Sinclair 1991, Stefanowitsch and Gries 2003). Lexical expectations (Hoey 2005) guide our interpretation.
- Argument structure and alternations: when native speakers construct sentences they employ argument structures and alternations (Levin 1993) as subtle operations (Pawley and Syder 1983) that help to minimize ambiguity. Alternations are subject to strong constraints; there
is a true choice in only a minority of cases (Lehmann and Schneider 2012, Röthlisberger and Schneider, accepted).

(2a) I gave flowers to Mary.
(2b) I gave Mary flowers.

(3a) I wrote chapters for the book.
(3b) I wrote the book chapters.

While the dative shift alternation is available for (2a,b), (3b) leads to an interpretation that is different from (3a). We use computational approaches to approximate the choice context.

- Semantic context and synonyms: the context of the document is an important source of disambiguation. We show e.g. that semantic resources improve automatic parsers (Schneider 2012), and that also the choice of synonyms helps human readers to avoid ambiguity.
- Processing constraints: we discuss that accepted principles such as end-weight can also be seen as avoidance of ambiguity: in a right-branching language like English, structures which flout end-weight offer considerably more attachment cites, thus increasing ambiguity.

Sentences are rendered in the way that they are due to many complex and interacting factors, and even subtle failures increase both the human and the automatic processing load. We use data from L2 speakers and automatically forced alternations to support our hypothesis.

References


I was sat there talking all night: A corpus-based study on factors governing intra-dialectal variation in England

ULRIKE STANGE (Johannes Gutenberg University of Mainz)

When it comes to intra-dialectal variation, the factors governing the choice between functionally equivalent variants still require an exhaustive analysis. An interesting case in point concerns the stance verbs *sit* and *stand* which alternate between standard progressive forms and pseudo-passives in a number of British English dialects:

(1a) Oh, I know but god did you see him? He *was sat* there and he was going brr brr brr brr brr brr. [KCE: 4738]
(1b) Last night <pause> I *was sitting* er what was I watching? <pause> Forget what was on now. [KCJ: 217]
(2a) Well when you *'re stood* there you can see the flames [KD2: 303]
(2b) I’m not doing anything really. I *’m standing* in the kitchen going kkeerr. [KPV: 1625]

The present paper addresses the question to what extent factors such as ‘Minimize Domains’ (Hawkins 2004), the ‘Complexity Principle’ (Rohdenburg 1996), the ‘Competition Model’ (Bates & Mac Whinney 1987), the ‘Principle of Uniformity and Transparency’ (Wurzel 1987), horror aequi (Vosberg 2003) as well as the ‘Principle of Rhythmic Alternation’ (Schlüter 2005) influence the language user’s choice of pseudo-passive vs. progressive constructions. First, corpus-based analyses of the spoken section of the BNC (Q-Web edition; subcorpus S:CONV) have revealed that variation between constructions with BE *sat/stood* and BE *sitting/standing* is common in dialects of Northern and Southwest England, and to a lesser degree in the Midlands and in London (see also Klemola 1999, Kortmann & Szmyrescanyi 2004 and Rohdenburg & Schlüter 2009 for the distribution of pseudo-passives in British English dialects). Moreover, the data show considerable differences in the distribution of pseudo-passives regarding their relative frequency and the number of dialects that make use of them, with the data containing 31 occurrences of BE *stood* and 85 of BE *sat*. Of twenty British dialects distinguished in the BNC, BE *sat* was found in thirteen and BE *stood* in ten different dialects. The search for the respective competing standard structures (cf. Wood 1962) yielded a total of 81 constructions for BE *stood/standing* and 265 for BE *sat/sitting* for analysis (incl. only those dialects with pseudo-passives).

Drawing on the conversational data of the BNC, the paper aims at evaluating which of the principles above can be considered as statistically significant determinants of intra-dialectal variation between pseudo-passive and progressive structures. To this end, it will be essential to test for other factors that might also have an impact on the choice of dialectal variants, such as age and gender. Considering that pseudo-passives are classified as “ornamentally complex” (i.e. “they complicate the system vis-à-vis the standard system, without clearly yielding an added
communicative bonus”, Kortmann & Szmrecsanyi 2009: 272), the paper will also provide tentative answers as to how the observed variation could affect the processing of these constructions in discourse.

References


**Formulacity as a determinant of processing efficiency:** Investigating clause ordering in complex constructions

**Daniel Wiechmann** (RWTH Aachen University)

Languages like English permit adverbial clauses (AdvC) to either precede or follow the main clause (MC).

(i) [AdvC] *Although he cannot afford it*, [MC] *he bought her a lovely gift*. 

74
(ii) [MC] *He bought her a lovely gift, [AdvC] although he cannot afford it.*

The linear order of clausal constituents is not random but influenced by numerous semantic, discourse-pragmatic and processing-related factors (cf. Diessel 2008, Wiechmann and Kerz 2013). Processing-related accounts predict that – given a choice – speakers produce the most cost-effective (efficient) order to minimize their processing effort (e.g. Hawkins 2004). For clause ordering, such accounts hold that sentence-initial adverbial clauses impose a greater processing demand than their sentence-final counterparts and that the preference towards sentence-final positioning should increase with growing length of the AdvC. While the predicted length-related effects have been demonstrated in studies on clause order, it is not yet clear if processing demand is sufficiently well expressed through a length-based measure: usage-based theorizing and research into formulaic language has shown that sequences that co-occur with sufficient frequency can become unitary processing blocks (chunks) or cognitive routines (Bybee 2010, Wray 2008 for overviews). The overall processing demand of a clausal constituent then is (co-)determined by its degree of formulaicity.

On the basis of a comprehensive analysis of complex concessive constructions (like the ones in (i) and (ii)) in the spoken part of the BNC, the present paper sets out to assess the role of formulaicity for the positioning of adverbial clauses. Following the methodology described in Dąbrowska (2008), degrees of formulaicity are approximated by way of measuring the structural homogeneity (how many different pattern-types are instantiated) and quantifying slot/filler associations (is there a preference to use specific lexical material in particular positions). If language users aim to maximize the efficiency of their structural choices, and if the degree of formulaicity affects the processing demand of a clausal constituent, then we would expect initial AdvC to be significantly more formulaic than their sentence-final counterparts.

**References**


FULL PAPERS
The ‘humour’ element in engineering lectures across cultures: An analysis of one category of pragmatic annotation

SIÂN ALSOP (Coventry University)

Humour is one of the most difficult pragmatic devices for lecturers and students to engage with, and for researchers to identify systematically. Cultural differences in lecture delivery and reception can cause particular problems. Wang (2012), for example, identified a startling mismatch in intention and comprehension between UK lecturers’ use of humour and its reception by Chinese students, and Zhang (2005: 114) goes so far as to argue that humour is inappropriate in lectures delivered in Chinese university settings. Significant quantitative differences in the usage and distribution of humour in lectures have also been noted in the European context, for example by Reershemius (2012). In the increasingly globalised setting of academia, understanding of such cultural differences is important to all academics and students on the move.

Partial attempts have been made to identify humour within corpora of spoken academic discourse. Maynard and Leicher (2007: 112) considered ‘humor’ as a potential category when they added pragmatic annotation to a subset of MICASE speech events, but it was ultimately excluded from their system because they found it was “too numerous for us to code all instances”. There is structural mark up for laughter in both MICASE (Simpson-Vlach and Leicher 2006) and the BASE corpus (Nesi and Thompson 2006), and laughter incidents in these two corpora have been discussed by Lee (2006) and by Nesi (2012). Laughter is not necessarily indicative of humour, however, neither does it capture all instances of humour or facilitate distinction between humour types.

This paper will introduce a new system of pragmatic annotation and demonstrate its use as a means of mapping the distribution, duration and specific function of humour in the Engineering Lecture Corpus (ELC), a growing collection which currently contains 65 English-medium University-level lectures from the UK, Malaysia, New Zealand and Italy. The transcriptions of the videoed lectures have been annotated for six functional categories, including ‘humour’. This category is further divided into eleven subtypes: ‘bawdy’, ‘black’, ‘disparagement’, ‘irony’, ‘joke’, ‘mock threat’, ‘playful’, ‘sarcasm’, ‘self-deprecation’, ‘teasing’, and ‘word play’.

Considerable differences in the lecturers’ use of humour are apparent, despite commonalities of delivery language, level and topic. In the first cycle of annotation c.100 instances of the eleven subtypes of humour were encoded, but almost two thirds of these occur in the UK subcorpus, and there is a much greater spread of humour types in subcorpora from the UK and New Zealand in comparison to the Malaysian subcorpus. ‘Sarcasm’ and ‘self-deprecation’ are notably absent in the latter. Only the ‘teasing’ subtype holds roughly equal weight (in terms of occurrence, not duration) across all the ELC subcomponents.

Early findings, then, point to the cultural specificity of Engineering lectures, with implications for those delivering and receiving lectures in different cultural contexts.

References


General extenders on the move? Using spoken corpora to study change in real time

GISLE ANDERSEN (NHH Norwegian School of Economics)

There is a growing body of literature that draws on comparable corpora to investigate language change, and in what has come to be known as “short-term diachronic comparable corpus linguistics” (Leech et al. 2009), it has become customary to investigate corpora recorded at different points in time, seeing systematic differences in word frequencies and collocation patterns as indicative of ongoing language change.

Within spoken language, two corpora offer a possibility to study change in London English, namely the Bergen Corpus of London Teenage Language (COLT, Stenström et al. 2002), recorded in 1993, and the Linguistic Innovators Corpus (LIC, Cheshire et al. 2008), recorded in 2005. Due to similarities in the age, geographical distribution and ethnicity of the speakers, it is reasonable to argue that “by comparing the young speakers in the two corpora we can … study change in real time” (Torgersen et al. 2011: 97). Some studies have successfully used the two as a basis for investigating the use of pragmatic markers (Torgersen et al. 2011), tags (Pichler & Torgersen 2012), hedges (Andersen 2012b) and lexical innovation (Andersen 2012a).

This paper offers a comparative analysis of the system of general extenders (Overstreet 1999) in the two corpora, as exemplified by the highlighted phrases below:

[7] M:  just getting treated like kids all the time and all that, you're told how to dress and shit like that. (LIC)

[9] F:  they're like grungers and all ain't they (LIC)
In a previous corpus-driven investigation (Andersen 2012b), it emerged that significant differences in the n-gram patterns of these two corpora could be ascribed to possible changes in the system of general extenders, notably the emergence of a contracted form *an’all* unique to the most recent corpus, coupled with a significantly higher relative frequency of *and all*, as well as several more complex structures, especially a subset containing the lexemes *kind of* and *stuff*. Based on these preliminary results, the present paper gives a more comprehensive account of differences in the system of general extenders in the two corpora. This involves the inspection of a wide range of forms which pragmatically can be seen to perform general extender functions; i.e. to mark the content of what is said as potentially inaccurate, suggest that more could be said about a given topic, or downgrade information which is not mentioned, etc. (Overstreet 1999).

The paper aims to chart the inventory and frequency of general extenders in the two corpora and to identify notable patterns suggesting innovation. This will be based on corpus-driven observations of n-gram frequencies, combined with corpus-based searching and qualitative analyses of distributional patterns within each corpus. I also aim to look into whether the findings corroborate sociolinguistic patterns of earlier corpus-based studies (Cheshire 2007, Palacios 2011), such as the observation that there are seemingly more adjunctive than disjunctive general extenders (i.e. more with *and* than *or*) in the language of young speakers (Cheshire 2007, Levey 2007).

References


Superficial typology: The so-called ‘verb-second constraint’ in Old English and Old French

KRISTIN BECH & CHRISTINE MKELENBORG SALVESEN (University of Oslo)

Old English (OE) and Old French (OF) are both referred to as verb-second (V2) languages, i.e. languages in which the verb typically occupies second position even after a non-subject-initial element, as exemplified in (1) and (2).

(1) *Mid twam wurðscipum ge-glængde se ælmihtiga scyppend þæs mannæ sawle*

    with two dignities adorned the almighty God the man’s soul

    ‘With two dignities Almighty God adorned man’s soul’ (*Ælfric’s Lives of Saints*)

(2) *Et por ce voudroit ele mieuz qu'ele feïst Tristan morir en aucune maniere*

    and for that would she better that she did Tristan kill in some way

    ‘And because of that she would rather have Tristan killed in some way’ (*Tristan*)

Another similarity between the two languages is that they both feature numerous non-V2 sentences in which several elements, including the subject, precede the verb. However, the syntactic and information-structural characteristics of such clauses are completely different in the two languages (Bech & Salvesen, forthcoming). Furthermore, OE is a much more heterogeneous language than OF with respect to word order in general; it has word ordering possibilities that do not exist in OF.

In this paper, we focus on the postverbal field(s) of the two languages, and our paper is divided into three parts: 1) a presentation of empirical data on the postverbal position of elements in OE and OF, using a corpus of ca. 1,500 main clauses from OE, extracted from five texts, and ca. 1,800 main clauses from OF, extracted from two texts; 2) a discussion of some syntactic implications of the empirical data, using a cartographic generative framework and 3) a discussion of possible information-structural factors that may have a bearing on element position in the two languages. The texts chosen are all fairly narrative in style, since there is a difference between narrative texts and argumentative texts with respect to information structure characteristics. Information structure is in this connection related to types of nominal and adverbial elements, e.g. pronoun vs. full DP vs. clause, and adverb vs. prepositional phrase vs. clause. We will show that categorizing the two languages simply as ‘V2 languages’ is an oversimplification of a complex matter.

References

Clause fragments in spoken English

JILL BOWIE & BAS AARTS (University College London)

Standard grammatical analyses tend to focus on the written ideal of a sentence as comprising a complete clause or a combination of clauses. However, in spontaneous spoken discourse we find many units which are non-clausal in form yet function as normal and complete contributions to discourse. While these units include many instances of ‘free-standing’ discourse forms such as interjections, response markers, and social formulae (oh, yeah, hello), we focus more closely in this research on what we call clause fragments: units with phrasal or subordinate-clausal content, which are not integrated into a larger structure and which therefore need to be interpreted through links to surrounding units. An example is given below:

A: Is Bim at the Slade now or not? Did he ever get there?

B: At Camberwell. (DCPSE: DI-B11 #72–4)

To interpret B’s fragment we need to link it to the underlined phrase at the Slade within A’s first question, and judge it to replace that phrase, giving the meaning ‘Bim is at Camberwell now’ (and thus indirectly answering A’s first question).

There has been insufficient corpus research on non-clausal discourse units. To address this gap, we use spoken corpus data to address the following research questions:

- What different kinds of non-clausal units can be distinguished on the basis of their grammatical properties? What are their frequencies in spoken data, and how do the frequencies compare across different spoken genres?
- What kinds of grammatical and semantic links can be identified between clause fragments and surrounding units in the discourse, allowing their interpretation in context?

Instead of an ellipsis approach to fragments which treats them as underlyingly clausal in form, we support an alternative approach which avoids positing extensive ‘invisible’ syntactic structure (as in Culicover and Jackendoff 2005; cf. also Progovac et al. 2006). There are many examples in our data for which a strict ellipsis account is simply unworkable. By using corpus data, we can go beyond theoretical speculation to investigate empirically the strategies needed to interpret examples of this kind.

We use the Diachronic Corpus of Present-Day Spoken English (DCPSE), an 885,000-word corpus of contemporary spoken British English which covers a wide range of genres, including informal conversations, speeches, spontaneous commentaries, interviews and radio discussions. The corpus is fully parsed and allows flexible grammatical searching. We use a combination of methods: (i) automatic searching for non-clausal units; and (ii) close examination of continuous extracts. Interpretations are checked by listening to the audio recordings.

Preliminary results indicate that around 31% of text units are non-clausal, with proportions varying markedly across different spoken genres. Of these non-clausal units, around 36% are clause fragments in the sense defined above. We have identified several recurring types in terms of their grammatical/semantic links to larger units in the discourse. These will be defined and exemplified in our talk. We will discuss the strategies needed to interpret clause fragments in context.
A new analysis of proper noun modifiers in PDE NPs

TINE BREBAN (The University of Manchester)

This paper is concerned with proper noun modifiers (PNMs), such as the Bush administration, a Birkin bag. PNMs are a special set of noun modifiers and have been studied as part of this larger group by Rosenbach (2007, 2010). She argues that PNMs denoting persons in definite noun phrases e.g. the Bush administration, the Weaver cars, constitute a case of intersective gradience (Aarts 2007), combining the syntagmatic features of classifying noun modifiers (they can be separated from the definite determiner by other premodifiers, e.g. the new Weaver cars), with identifying functions equivalent to those of determining genitives. This functional equivalence is reflected in the fact that many PNMs alternate with determining genitives, e.g. the Bush administration :: Bush's administration. However, equivalence is not complete: (1) not all PNMs allow alternating, e.g. the Guggenheim museum :: *Guggenheim's museum; (2) PNMs express a wider range of relationships than genitives. This paper reports on a large-scale corpus investigation of all types of PNMs in PDE data from Collins Wordbanks Online, BNC and COCA. I put forward an alternative explanation to Rosenbach's. The identificational mechanism of PNMs such as the Bush administration is that of identifying epithets (Halliday 1994), i.e. descriptive premodifying adjectives which convey a property that is sufficient to single out (and hence identify) the referent either in the discourse context, e.g. the red car when cars of different colours are talked about, or in a wider context of shared knowledge, e.g. the White House. Similarly, these PNMs are used as ‘descriptive labels’ standing for a relation between PNM and referent, which is unexpressed, but which can be inferred and is sufficient for identification. This analysis is borne out by all distributional features of the PNMs: (1) descriptive adjectives can be preceded by other (descriptive) premodifiers; (2) the relation between referent and PN can be one of possession (in the broad sense), in which case there is a genitive alternate, but does not have to be. Moreover, like identifying epithets, PNMs are sufficient for identification only within discourse or shared knowledge. Finally, this analysis is not restricted to PNMs denoting persons, but also applicable to locations, e.g. the Ghana problem, and dates, e.g. the 1948 earthquake, which are also attested in the corpus data. I then integrate this analysis in a comprehensive classification of PNMs, which also includes classifying PNMs, e.g. the Yorkshire terrier, a Birkin bag, and descriptive PNMs, e.g. a/the Birmingham housewife, a (typical) John Major answer. The latter constitute a small set of PNMs occurring mainly in indefinite NPs, which are used to describe a characterizing feature of the referent, e.g. being from Birmingham, having the typical style of John Major's answers. PNMs can thus be used for all three functions distinguished by amongst others Teyssier (1968) for modifiers in the NP: identification, characterization and classification. This provides evidence against a simple one-to-one association of the three functions with the categories determiner, adjectives, nouns, and in favour of a function-based analysis of NP modifiers in actual usage.

References


References


If-parenthetical clauses in the history of English: The rise of metalinguistic meaning

LAUREL BRINTON (University of British Columbia)

Clauses such as if you choose/like/prefer/want/wish/will are ‘reduced parenthetical clauses’ in the sense of Schneider (2007). In Present-day English these if- parenthetical clauses (if-PCs) may all serve metalinguistic functions. For example, the OED notes that if you will (OED, s.v. will v.1, def. 17b) can mean “if you wish it to be so called”, “if you choose or prefer to call it so”, and if you like (s.v. like v.1, def. 6b) can be paraphrased “if you wish to phrase or consider something in a particular manner” (s.v. like v.1, def. 6b), as in:

1. I am tenebrous, or tenebrious if you prefer, it's all the same. (COCA: FIC)
2. the honour of the assassins as a group - as a guild, if you like - would be tarnished (STRATHY: FIC)
3. becomes a mark of quality or territorial signpost – a trademark, if you wish (coca: acad)
4. Call me an enlightened defeatist if you wish (BNC: W [fic])

Like ‘indirect conditions’ (Quirk et al. 1985: 1095), the meaning of if-PCs is “dependent on an implicit speech act of the utterance” (though the speech act may be explicit, as in (4)). In these contexts, if-PCs belong to the set of parentheticals Kaltenbock et al. (2012) term ‘conceptual theticals’.

If-PCs are analogous to clauses such as If only he would stop drinking!, which Dancygier and Sweetser (2005: 217-219) term ‘monoclausal’ if-clauses; these have lost their conditional meaning and carry only the ‘conventional meaning’ of speaker stance. They are also similar to ‘Unsubordinated clauses’ such as If you could give me a couple of 39c stamps please (see Evans 2007, Heine 2012).

Using a variety of English diachronic corpora, this paper undertakes a study of the history of if you choose/like/prefer/want/wish, exploring both their syntactic and semantic development (on if you will, see Brinton 2008). The earliest if-PCs with these verbs date from the mid-18th to mid-19th centuries, with metalinguistic meanings arising sixty to one hundred years later. Semantically they thus follow the expected development from referential/propositional > non-referential (pragmatic/procedural) meaning, from “if you choose to do so” to “if you choose to call it so”. Syntactically, we see a general change from more tightly integrated adjunct adverbial clause to more loosely integrated parenthetical disjunct. It might be postulated that if-PCs develop from fully formed clauses/sentences, via ellipsis of their complement (non-finite) clause...
(see Quirk et al. 1985: 909-910) as well as of the main clause (the apodosis); the latter is seen as
the route of development of insubordinated clauses, where ellipsis of the main clause is
followed by reanalysis of the original subordinate clause as a main clause (Evans 2007: 368-
370).

The diachronic data do not seem to bear out this intuitively plausible syntactic development.
First, reconstruction of a full clausal structure, with obligatory complement, remains
problematic given the varied complement structures found historically and the rarity of full
structures with metalinguistic meaning. Second, while explicit apodoses are common – but not
obligatory – when the if-PC has concrete meaning, no examples with an explicit apodesis can
be found when the construction has metalinguistic meaning. The paper explores alternative
derivations, such as origin in the call it x construction (see ex. 4).

References

Brinton, L. J. 2008. The comment clause in English: Syntactic origins and pragmatic
development. Cambridge: Cambridge University Press.

Cambridge: Cambridge University Press.

Evans, N. 2007. Insubordination and its uses. Finiteness: Theoretical and empirical

Unpublished paper.

848-893.


Schneider, S. 2007. Reduced parenthetical clauses as mitigators: A corpus study of spoken

Existential and locative constructions in an expert and in a learner
corpus of promotional tourist texts

ERIK CASTELLO (University of Padua)

This paper sets out to explore the use of existential and locative constructions in two corpora of
promotional tourist texts: a corpus of texts promoting English tourist resorts and written by
expert writers (about 150,000 words) and a learner corpus made up of texts that a group of
Italian university students of English wrote as part of a course assignment with the aim of
inviting potential tourists to visit the city of Padua (about 40,000 words). Existential (e.g.
“There is a book on the table”) and locative constructions (e.g. “The book is on the table”) can
be found in all registers, yet they are remarkably frequent in fiction (Biber et al., 1999). As
tourist publications are about such referents as tourist resorts, attractions and facilities, which
need to be somehow located in space and then described, it is only natural for them, too, to
include a large number of such constructions. Various studies have explored the features of
existential constructions in general English (e.g. Clark 1978, Lakoff 1987, Givón 1993, Huddleston & Pullum 2002, Martínez Insua & Palacios Martínez 2003, Halliday 2004), as well as the syntactic, semantic and pragmatic phenomena that determine their use. In this paper these constructions are viewed as colligational patterns (e.g. Hoey 2005, 2006) and a Contrastive Interlanguage Analysis (CIA) (e.g. Granger 2002) is conducted of the expert and of the learner data, in the attempt to gain insights into their characteristics and functions in both varieties. It is indeed believed that the comparison of a corpus representing the language used in the tourist sector to a corpus of texts written by novice non-native writers is not only likely to reveal learner deficiencies, but also, to some extent, the peculiarities of the professional variety. From the analysis, there emerged six main existential and locative colligational patterns, the use of which turned out to be qualitatively and quantitatively different in the two corpora (their frequencies have been normalized to enable comparisons across the corpora): (1) “there is/’s/are + noun phrase”, which the native speakers mainly use to introduce indefinite referents in bare existential clauses; (2) “there is/’s + plural/multiple head noun/s” (e.g. “There’s beach space galore, delightful gardens and parks”) or “there are + singular head noun”, which is only exploited by the experts who, thus, seem to be addressing their readers as “childlike receivers” (Dann 1996); (3) “adjunct of space + there is/’s/are + noun phrase ’, which is over-represented in the learner corpus and hardly employed in the expert one (e.g. “Next to the chapel there are the Eremitani museums’); (4) “adjunct of space + is/are + noun phrase ‘, that is, “locative inversion” (e.g. “Near the Palace Pier is the Sea-Life Centre”) (Biber et al. 1999), which is, by contrast, only used by the expert writers; (5) “(adjunct of space) + stand/lie/stretch/hang/loom + noun phrase ”, which is nearly exclusively present in the expert data (e.g. “Beyond the green belt spreads a sizeable stretch of suburbia”); (6) “is/are + situated/placed/located/set” (e.g. “The town is situated on the Atlantic Heritage Coast”), which is utilised differently in the two varieties. This paper will illustrate and discuss these findings and their implications for teaching the specific features of existential and locative constructions in promotional tourist texts and for creating data-driven ESP teaching materials (e.g. Tribble & Jones 1990, Seidlhofer 2002, Zagrebelsky 2004).

References


Much of the research conducted on Australian English has been preoccupied with the historical development of its lexis (e.g. Eagleson 1964, Ramson 1966) and accent (e.g. Bernard 1969, Collins 1975). Apart from Fritz (2007), no attempt has been made to trace the historical development of Australian English morphosyntax. A significant obstacle has been the unavailability of historical corpora for such research. This problem is now being addressed through the compilation of a diachronic corpus of 20th-century Australian English, representing the major genres of fiction, news and scientific writing. Matched with Fritz’s two-million-word Corpus of Oz Early English, it allows us to map diachronic changes in Australian English over the past two centuries.

In this paper several categories of the verb phrase in Australian English are examined. These categories are known to have been subject to diachronic change in Late Modern English: modal and quasi-modal verbs, the progressive and the present perfect. Previous studies have explored their development in British and American English (e.g. Elsness 1997, Krug 2000, Smitterberg 2005, Leech et al. 2009, Kranich 2010), but rarely in varieties other than these (one exception being Dollinger 2008). Accordingly the paper is an attempt to build bridges between English historical linguistics and world Englishes research. Systematic comparisons are drawn with British and American English using data drawn from the latest version (3.2) of A Representative Corpus of Historical English Registers.

The results suggest that Australian English has operated relatively independently of British and American English, with no clear pattern of subservience to either variety. In the case of the quasi-modals, known to be on the rise in Modern English (see, e.g. Krug 2000), Australian
English behaves more conservatively than either of the ‘supervarieties’, with lower frequencies and a milder rate of increase (particularly in the case of be going to). However, in the case of the modals, whose decline has been substantiated in a number of studies (e.g. Leech et al. 2009), Australian English emerges as more advanced than the other two varieties (particularly ought to). In the case of the progressive, which has risen consistently in all varieties, the increase has been milder in Australian English, with a resulting smaller frequency by the end of the 20th century. The results for the perfect show it giving way to the preterite since 1800, with an Australian rate of decline intermediate between the rapid American and slower British decline. The relative impact of functional and formal features is examined (e.g. competition between root and epistemic uses in the modals and quasi-modals, the development of special ‘not-solely-aspectual’ uses of the progressive, and co-occurrence of the perfect with time adverbials). Of various socio-historical factors considered, colloquialization is noted to have been particularly influential in the developments observed.

References


Predicting prototypes across native and learner English: The case of modal constructions

SANDRA C. DESHORS (New Mexico State University)

This study compares native and four learner English varieties at the level of constructions and explores prototypical instances of *may*, *can*, *must* and *will* based on their larger contexts of use. Drawing from Selinker's (1969:71) recognition that “interlanguage [...] must be dealt with as a system” and recent work in Second Language Acquisition (SLA) revealing learners' difficulties to produce native-like form-function pairings due to “a processing bias affecting learners' L2 cue learning from a life-time of prior native language usage” (Ellis & Sagarra 2011: 591), this paper specifically discusses:

i. which specific combination of grammatical features (i.e., morpho-syntactic and/or semantic features) have high cue validity in modal constructions; and

ii. whether those combinations characterize differently prototypical constructions in native and learner English and if so to what extent.

According to the usage-based theoretical framework, learning a language involves storing forms in associative maps for syllables, lexical items, constructions, and mental models and “the selection of forms is governed by cue strength within a competitive central syntactic processor” (MacWhinney 2008: 342). Additionally, “[l]earning is driven by [...] frequency distribution of exemplars within constructions and by the match of their meaning to the construction prototype” (Ellis & Ferreira-Junior 2009: 191). So to learn (native) prototypical constructions, L2 learners need to acquire and represent statistical information about the combination of linguistic features within constructions. However, learners' processing bias during L2 production influences their learning process, causing them to develop deviant L2 prototypical constructions.

In the field of learner corpus research, recent multifactorial work by Deshors & Gries (to appear) on *may/can's* co-occurrence patterns across native English and French-English interlanguage supports constructionist approaches to the modals showing how interactions between semantic and morpho-syntactic characteristics of the modals' contexts vary systematically across language varieties. The present work furthers Deshors & Gries' study by showing to what extent such variability unveils different prototypes. Methodologically, this work assumes that the prototypicality of a construction can be assessed based on the predictability of its occurrence (Gries 2003). Using multinomial logistic regression modeling, over 100,000 instances of *may, can, must* and *will* extracted from the *Louvain Corpus of Native English Essays* and the Chinese, French, German and Swedish subsections of the *International Corpus of Learner English*, annotated for 28 grammatical features, are analyzed at two description levels: case-by-case level (providing a continuum of prototypicality of investigated constructions) and a more abstract level.

Generally, the results support the notion that constructions play a crucial part in SLA. For instance, *can* constructions illustrate that the more predictable L2 constructions are, the more qualitatively similar they are to their native counterparts. Additionally, the results emphasize the need to account for both abstract and superficial description levels: although abstractly learners and natives share similar prototypical constructions, superficially, learners' linguistic representations of those constructions systematically deviate from natives'. For instance, while types of modalized lexical verbs contribute to *may* constructions in native/learner English, *state* verbs characterize native prototypes whereas *achievement* verbs characterize learner prototypes.
Ultimately, those results have direct implications for our understanding of interlanguage systems.

References


On neoclassical compounding and their constituents

**Ana Díaz-Negrillo** (University of Granada)

It has been pointed out that neoclassical compounding should be better viewed as “a kind of prototype, from which actual formations may diverge in unpredictable ways” (Bauer 1998: 409). Indeed, despite the existence of central features to the category, the output of neoclassical compounding is far from homogeneous. This internal variation within the class is largely motivated by the nature of their constituent units. Bauer (1983: 213ff) put forward the category ‘combining form’ (CF) for the bound classical elements which occur in neoclassical compounds. On the other hand, a number of papers have recently attempted at the identification of defining features for the units contained in neoclassical compounds, by trying to identify the features that differentiate CFs from affixes (Prcic 2005, 2008; see also Ludeling et al. 2002: 257-258; Amiot & Dal 2007: 324-326 on the status of neoclassical compounds constituent units). Finally, and in response to Prcic's papers, Kastovksy (2009) argued against the category ‘combining form’ altogether and in favour of their description of the units in neoclassical compounds within other existing morphological categories, in particular the ‘stem’, ‘suffix’, and ‘clipped forms’.

In their discussion, these authors also argue that some neoclassical compounds' constituent units which have traditionally been treated as CFs, are actually better viewed as suffixes. However, although both authors agree on the need of reanalysing, for example, *-logy* as a suffix, they diverge in the analysis of some other final combining forms (FCFs). Moreover, their realisations lack empirical analysis, while most distinctive features suggested for the identification of CFs/stems from suffixes are measurable.
Corpus-based examination of neoclassical compounds may throw light on the status of neoclassical compounds and, in particular, on the status of those constituents which have been claimed to diverge from the prototype in neoclassical compounding. This paper quantitatively explores the synchronic and diachronic morphological behaviour of 513 neoclassical formations from the BNC (and, in particular, of their terminal units, based on aspects considered prototypical to suffixes and CFs in the serial publications mentioned above, namely productivity, boundness and origin of the ICF and occurrence of a middle vowel). Preliminary findings confirm, first of all, that there is marked morphological variation within neoclassical compounds' constituent units; second, they identify which terminal units show a more or less marked suffix-like status according to the criteria considered; and, finally, they suggest that neoclassical compounds coined in the 20th ct., as attested in the Oxford English Dictionary, are overall less prototypical than earlier compounds. This preliminary results can be taken as an indication of the transformation that some neoclassical compounds and, in particular, certain FCFs are undergoing, which may ultimately justify their reanalysis.

References


Bauer, L. 1998. Is there a class of neoclassical compounds, and if so is it productive? Linguistics 36: 403-422.


The progressive in spoken learner language: Frequency and accuracy

STEFANIE DOSE & SANDRA GÖTZ (Justus Liebig University Giessen)

It has frequently been noted in the applied linguistic community that the progressive aspect in English is very difficult to handle for foreign language learners (e.g. Hahn et al. 2000, Mindt 1997, Rogatcheva 2012, Swan and Smith 1987, Zydatiß 1976). The variety of functions it serves, but also its often incomplete or inadequate descriptions and representations in
mainstream teaching materials (cf. Römer 2005, 2006) seem to make it a veritable challenge for learners as well as their teachers.

The progressive has been claimed to be particularly troublesome for those whose L1 does not have an equivalent form, which is the case for e.g. Polish, Swedish, and German learners (cf. Wulff and Römer 2009: 116; Zydatiß 1976: 352). In fact, a number of corpus studies have shown that learners tend to overuse the progressive form and/or frequently use it in rather ambiguous contexts (e.g. Axelsson and Hahn 2001, Leńko-Szymańska 2007, Rogatcheva 2012). However, most previous studies have only looked at written learner language so far. Learner corpus studies dealing with the spoken language have been rare.

The present paper will report on the results of a study which investigates the extent to which the learners' issues with the progressive in writing also apply to the spoken language. This includes aspects related to the frequency of the progressives as well as the degree of accuracy with which it is used. In particular, we set out to determine

   a) whether in the spoken language, overuse is an issue, too,
   b) whether the progressive is an error-prone category, and
   c) which types of error are most common.

To this end, we make use of the German error-tagged subcorpus of the Louvain International Database of Spoken English Interlanguage (LINDSEI-GE; cf. Kämmerer 2009) and the native speaker counterpart Louvain Corpus of Native English Conversation (LOCNEC, cf. De Cock 2004). This corpus pair provides a useful resource to compare German learners' use of the progressive in their spoken output to equivalent native speaker data.

We will first report on quantitative differences within the spoken corpus pair. For instance, it turns out that the frequency patterns which have been observed for learner writing are reverse in speech, as German learners actually underuse the progressive in the spoken language. However, our qualitative error analysis revealed that the error patterns do not explain the frequency patterns at all: There is a very low proportion of errors caused by a non-use of the progressive in contexts where one would be required (e.g. *these supermarkets become popular now [GE016]), and the most frequent type of error is, in fact, a misuse of the progressive in non-required contexts (e.g. *it was very difficult for me to understand the French because they are speaking so fast [GE006]). There seems to be a complex relationship between frequency and accuracy.

Finally, we found that there is great variability in the performances of the individual learners, not only in terms of the frequency of the progressives, but also as regards the number and types of error committed. As one pedagogical implication of our study, we would like to suggest that the progressive, which is typically emphasized in beginners' EFL classes, deserves again specific attention in the language classes of advanced learners.

References


Support for end-weight: Do-support in Early Modern English affirmative declaratives

MATTHIAS EITELMANN (Johannes Gutenberg University of Mainz)

For a brief period between 1550 and 1640, DO-support displays a slight trend to being used in affirmative declaratives in addition to its simultaneous grammaticalisation in negative interrogatives, affirmative interrogatives and negative declaratives (cf. Ellegård 1953). In such affirmative declaratives, the use of DO as yet lacks the unequivocal emphatic or contrastive reading it has in Present-day English. What has been claimed by various researchers to play a
decisive role in the language user's choice of DO-supported verbs as in (1) over finite verb forms as in (2) is end-weight (cf. Nevalainen 1991: 311, Wischer 2000: 148, Bækken 2002). By this, we refer to the tendency for bulkier constituents to be positioned at the end of a sentence (cf., e.g., Leech 1983, Quirk et al. 1985, Hawkins 1992). Affirmative declarative DO “lengthens the verb phrase and increases its weight” (Bækken 2002: 324):

(1) ... accompling themselfes most happye to dwell in that Towne, where a woman of such vertuous behauiour did dwele. (EEPF, 1566)
(2) ... and commanded him to conducte the gentleman, to that part of the forest, where his father dwelled. (EEPF, 1566)

While the assumption of an end-weight effect has so far remained inconclusive due to lack of data, the present paper provides an empirical study drawing on the Early English Prose Fiction Corpus. By investigating a wide array of verbs, it shows that end position is indeed more often DO-supported than mid position. This finding of an end-weight effect is corroborated by previous studies on grammatical variation phenomena (Eitelmann 2012; Mondorf 2009: 99-107), which show the more explicit yet bulkier variant to be favoured in end position as opposed to mid position. In this respect, then, the paper contributes not only to the investigation of end-weight as a determinant of grammatical variation but also to a more differentiated understanding of the term that exceeds its common treatment as a short-before-long principle in the sense of Behaghel's (1909/1910) Law of Increasing Constituents.

The question arises to what extent the preference for periphrastic DO at the end of affirmative declaratives is in line with what has been postulated as the Complexity Principle (Rohdenburg 1996) or as the Principle of Analytic Support (Mondorf forthc.). The common claim is that more explicit variants are preferred in cognitively complex environments, thus functioning as a ‘support strategy’ compensating for increased processing effort. Addressing this issue will also involve discussing how the use of affirmative declarative DO induced by end-weight generally ties in with processing-related accounts of the Early Modern English establishment of DO-support (cf. Kroch 1989, Stein 1990, Warner 2004).

References


Methodological aspects of corpus pattern analysis

Ismail El Marouf (RIILP, University of Wolverhampton)

This paper proposes to analyse and understand the practices of lexicographers in building corpus-driven dictionaries. It focuses on Corpus Pattern Analysis (henceforth CPA; Hanks 2004, Hanks 2013), a new lexicographic technique for mapping meaning onto patterns of use. Currently, the whole enterprise is being revisited with the cooperation of lexicographers and computational linguists, within the DVC project (Disambiguation of Verbs by Collocation). One of the goals of this project is to bridge gaps between lexicography and computational linguistics in the light of CPA.

CPA is in the tradition of Corpus Linguistics which follow Sinclair's works on lexical analysis (Sinclair 1966, 1991, 2004), and is based on the Theory of Norms and Exploitations (TNE; Hanks 2013). Such works stress the importance of phraseology, advocate the use of large corpora for lexical analysis, and investigate the relations between structure and meaning. In accord with these principles, CPA provides methods and models to discover the normal uses of words in corpora. This paper investigates three issues regarding the application of CPA to verbs: semantic types, syntactic structures, and figurative language.
A normal pattern has two inseparable components: a set of context clues and a set of ‘implicatures’, or glosses. Context clues include various linguistic categories used to model a pattern. For example, a semantic type groups entities according to their similarities (e.g. humans, animals, artefacts, etc.). Each pattern set of collocates is grouped with the most appropriate type (e.g. humans ask humans to do something). Such an operation is not straightforward and problems such as the relation between lexical items and semantic types, and choosing the appropriate level of generalisation, are discussed.

A normal pattern is an ordered structure with various argument slots: subjects and objects for active voice patterns, but also particles for phrasal verbs, that-clause for speech verbs, and so on. CPA provides guidelines for grouping similar structures, such as active/passive alternations. The paper analyses the criteria which prevail in lexicographic decisions for splitting and lumping patterns. It also analyses problematic cases such as the representation of optional pattern elements and the identification of anomalous syntactic structures.

CPA proposes interesting principles for representing figurative meaning. The article analyses why figurative uses may either lead to normal and conventional patterns, or may be considered as creative exploitations of a given pattern. The status of idioms in CPA is discussed with regards to both their representation in CPA and their relation to conventional patterns. The physical/abstract pattern alternation giving rise to conventional metaphors is also illustrated.

CPA has been applied on more than 700 verbs, through a computer interface designed to assist the lexicographer in building verb entries as well as in annotating corpora. The output digital dictionary, called PDEV (Pattern Dictionary of English Verbs), is literally ‘anchored’ to the corpus. The results presented in this paper are based on the analysis of a frequent English verb, to burn, on the BNC50 corpus (a sample of the British National Corpus).

References


English in South Africa: The case of past-referring verb forms

JOHAN ELSNESS (University of Oslo)

English has a strong position in South Africa, and yet it is one of a total of 11 official languages. Among the minority White population more people speak Afrikaans than English as their first language.

This paper reports an electronic elicitation test carried out at the University of Cape Town in 2011, with 232 students acting as informants. The focus is the variation between the present perfect and the preterite verb forms. This distinction is drawn differently in English compared
with many other languages: In German and French, for example, the present perfect is used much more widely than in English, while the Scandinavian languages are more similar to English. In English the preterite is the required verb form if the reference is specified as being to past time clearly separate from the deictic zero-point. The present perfect is used of past times extending all the way up to the deictic zero-point, and of times wholly in the past which are not clearly defined. In borderline cases AmE tends to use the present perfect more sparsely than BrE.

A major research question is where South African English places itself: Is it more similar to AmE or to BrE, or is it best classified as a separate, endonormative variety? A second research question is whether any differences are detectable between White and Black speakers.

To answer the first research question, results from South Africa are compared with those from similar tests previously carried out in the United States and Britain (and also Australia and New Zealand). Most of the South African results place themselves somewhere between AmE and BrE. They suggest that South Africans use the present perfect more widely than AmE speakers but not quite so extensively as speakers of BrE. There may be internal geographical variation within South Africa, however. Most of my informants came from the Western Cape (which includes Cape Town). There were also some from Natal and other provinces, but their numbers were not sufficient to warrant any firm conclusions about any such internal South African geographical differences.

As regards the second research question, quite a few results do vary according to ethnic background. Even among informants who gave English as their first language, several results are clearly different for White and Black speakers. Results were tested for statistical significance by means of the t test applied to independent samples, which showed the differences between White and Black speakers of first-language English to be significant at least at the 5% level with as many as 17 of the 66 constructions included in the test. In these cases White informants tend to give the higher average score for the present perfect, Black informants for the preterite.

The recorded differences may be interpreted as an indication that White South Africans speak a variety of English which is closer to the British model, while Black speakers are somewhat closer to the American model, or perhaps rather to what can be seen as the prevailing international model.

References


**A discourse-functional perspective on the verbalization of the gerund**

**LAUREN FONTEYN, HENDRIK DE SMET & LIESBET HEYVAERT (K.U. Leuven)**

This paper discusses the development of nominal and verbal gerunds with a focus on their discourse-functional behaviour from Late Middle to Late Modern English. Nominal gerunds, as
in (1), have the internal syntax of a noun phrase (NP), whereas verbal gerunds, as in (2), have the internal syntax of a clause.

(1) There is, first, the dryness inseparable from the learning of a language (...). (1878, PPCMBE)

(2) He was troubled, for having been forc’d, by an irresistible passion, to rob his son of a treasure, he knew, cou’d not but be extremely dear to him. (1688, PPCEME)

Verbal gerunds are commonly believed to have developed out of nominal gerunds. While the formal aspects of this verbalization process are well documented (see, among others, Jack 1988, Fanego 2004), De Smet (2008) suggests that new insights can be gained into the English gerund if its discourse-functional features are brought into the analysis more systematically. This paper zooms in on the historical interaction between the choice for either nominal or verbal gerunds, their accessibility to referent tracking, and grammatical function.

Our analysis started from the hypothesis that over time, the -ing form in the verbal gerund has become more transparent (i.e. less available) to referent tracking (making the verbal gerund less NP-like and more clause-like). When verbal gerunds are used in discourse, it is therefore especially their participants that will be tracked as referents in the following discourse, as in (3). Nominal gerunds, on the other hand, are expected to continue to function as nominal discourse referents, supporting both reference to the gerund as a whole (4) and to their participants (5).

(3) Heat, by agitating the parts of a fit Body, solicites it as it were to send forth its Effluvia. (1675-1676, PPCEME)

(4) (...) and the talking of the sailors on the deck be plainly audible. This is not, however, to assume it will be understandable. (1905, PPCMBE)

(5) Now, upon the hearing of this story two things strike us (...). The story speaks to us as a parable. (1901, PPCMBE)

To test this hypothesis, we analysed the nominal and verbal gerunds found in the suite of the Penn Parsed Corpora of Historical English and divided the data into 3 periods covering the 1420-1914 time span.

Results show that in the earlier stages of its development, the verbal gerund, like its nominal source construction, functioned as an ordinary NP in the discourse, though it seems somewhat less likely to consist of or introduce prominent information that will be picked up in the following discourse. It is suggested that this might be related to the typical grammatical functions served by verbal gerunds, i.e. in environments where additional or background information is provided (cf. also Fanego’s (2004) observation that verbal gerunds spread according to a “grammatical relations hierarchy” from peripheral adverbials and oblique complements to core complements). Our analysis of later data, however, suggests that nominal and verbal gerunds drift apart towards the Late Modern English period in terms of referent tracking, as the verbal gerund gradually adopts more clause-like behaviour, with only its nominal participants still regularly picked up as referents, while the nominal gerund rather invites tracking of the entire unit.

References


So ladylike and particularly manly? Determinants of variation in intensifier usage in Indian English

ROBERT FUCHS (University of Münster)

Previous studies have provided conflicting evidence on how age and gender influence the frequency of intensifiers. For example, Tagliamonte and Roberts (2005) found so to be more frequent with women than with men, but Ito and Tagliamonte (2003) found no difference. Conflicting results might be due to the fact that small numbers of both types and tokens were considered in these and other studies. Other factors, such as difference in age between interlocutors, whether groups consist of men or women only or are mixed, and whether speakers are equals or not have not been considered so far in quantitative research. Based on a corpus of spoken Indian English (the spoken part of the Indian component of the International Corpus of English) with data from more than 400 speakers, this study investigates which factors determine intensifier frequency and choice. Of a list of 29 intensifiers (boosters), the most frequent 6 intensifiers, together accounting for 91% of all tokens, were chosen. The number of tokens used by each speaker, as well as the number of words uttered by each speaker were determined to compute relative frequencies for each speaker and, based on this, for groups. This has the advantage that speakers who behave unlike others from the same demographic group do not unduly influence group means through their particularly high or low usage of intensifiers. Linear regression models were then computed in the statistical software package R to determine which sociolinguistic factors significantly influence the relative frequency of intensifiers in spoken Indian English.

Adopting this robust speaker-based quantitative approach, gender, group composition (all-female, all-male, mixed), age, formality, the number of interlocutors and their relationship are shown to be significant factors. Some of the results parallel what is known of other varieties of English, such as so and really being preferred by younger and female speakers (cf. Yaguchi et al. 2010 for American English and Bulgin et al. 2008 for Tyneside English). Other results are unprecedented. We found that in IndE women use more intensifiers when speaking with other women, but in mixed groups women do not differ from men in intensifier frequency. Furthermore, women use more intensifiers than men in informal, but fewer in formal situations. More intensifiers are also used if speakers are equals in the social hierarchy, and when group size is small. While age is an important factor in determining intensifier frequency, whether interlocutors differ in age has no significant influence. Rather, age interacts with gender and group composition: In female-only and male-only groups, younger speakers use half the number of intensifiers that older speakers use, but in mixed groups this trend is much smaller. These results are explained by considering intensifiers as positive politeness devices in terms of Brown and Levinson’s (1987) theory of politeness. It is argued that women use intensifiers to establish solidarity with interlocuters in all-female groups, but not in mixed groups.

References


Negotiating trust during a corporate crisis: A corpus-assisted discourse analysis of CEOs’ public letters after the Gulf of Mexico oil spill

MATTEO FUOLI (Lund University, Sweden)

This paper examines the discourse strategies deployed by BP's CEO to restore stakeholders' trust after the 2010 Gulf of Mexico disaster. Combining data-driven quantitative and qualitative analysis, the study examines how interpersonal language resources are used in the CEO's public letters to communicate a trustworthy corporate identity. The analysis employs the model of TRUST proposed by Mayer et al. (1995) and draws on Appraisal Theory (Martin and White, 2005) and Hyland's (2005) concept of metadiscourse. The results show that affect, evaluation and modality play an important role in BP's trust-building communicative efforts.

Trust, as Luhmann (1979: 4) puts it, is a “basic fact of social life”. It is a critical component of human relationships and plays a key role in any form of cooperation and exchange (Misztal 1996: 12). However, whereas the notion of trust has attracted increasing scholarly attention from a range of disciplines, including sociology, psychology and business studies, it remains largely unexplored in linguistics. Yet, as Linnell and Keselman (2011: 156) point out, “it is in interaction between people, or between individuals or groups and their environments, that trust and distrust are created, negotiated, sustained, confirmed or disconfirmed. It is in discourse that trust comes to life”.

The analysis focuses on the CEO's ‘letter to shareholders’ and ‘letter to stakeholders’ included in the corporate annual report and social responsibility report, respectively. Considering the catastrophic consequences of the accident and the reputation damage suffered by BP due to its controversial management of the crisis, the research question I examine is the following:

What discourse strategies does BP's CEO adopt in his public letters to restore stakeholders' trust?

To address this question, I have compiled and analyzed a small-size specialized corpus of CEO letters published by BP and other 4 major oil companies from 2009 to 2012, i.e. before and after the accident. The analysis combines quantitative corpus methods and qualitative discourse analysis (Baker 2006, Baker et al. 2008). It focuses on three key attributes of trustworthiness, i.e. ability, benevolence and integrity (Mayer et al. 1995), and investigates how these are communicated in the CEO letters. It is shown, for example, that ability is communicated through positive appraisals of the company's performance (We are becoming a leading player in wind energy), benevolence by demonstrating empathy (We are deeply sorry), and integrity by underlining the reliability of the information provided (You will see a continuing, relentless focus on safety) and displaying ethical commitment (Our refreshed values). The language features connected to these three components of trustworthiness were identified and coded in
the corpus using the UAM corpus tool (O'Donnell 2008) and their distribution compared across companies and diachronically. Inter-coder agreement scores were used as a measure of reliability in the annotation process (Artstein and Poesio 2008).

Preliminary results show that BP's letters contain a comparatively higher number of expressions of affect and that these expressions are more frequent in the letters published after the disaster. Also, BP's letters feature a comparatively higher number of epistemic and modal markers and personal pronouns. These differences are interpreted in light of BP's need to re-establish stakeholders' trust after the accident.

References


**Regional variation in Middle English suffix usage**

**Anne Gardner** (University of Zurich)

Textual transmissions between c.1150 and 1500 (‘Middle English’) exhibit varying degrees of regional characteristics manifest in orthography, grammar and lexis. The present investigation
focuses on Germanic suffixes or suffixoids forming abstract nouns on the basis of adjectives and nouns as illustrated in brotherhood / brotherreden, fearlac ‘fear’, fellowship, goodness / goodleik and thraldom. While valuable work has been done concerning the general and semantic development of these suffixes at various stages in the history of the English language (e.g. Dalton-Puffer 1996, Zbierska-Sawala 1998, Anderson 2000, Ciszek 2008, Palmer 2009, Trips 2009), regional variation in suffix usage has not yet been analysed systematically on the basis of modern corpora.

Evidence drawn from the Linguistic Atlas of Early Middle English (LAEME), as well as selections from the Penn-Helsinki Corpus of Middle English (PPCME2), the Middle English Grammar Corpus (MEG-C) and the Helsinki Corpus (HC) reveals that the more dominant -hood, -ness and -ship reached peaks in productivity at different points during the Middle English period, and that changes in the deadjectival paradigm did not necessarily coincide with those observed in the denominal paradigm. Main measures of productivity employed include the appearance of new types, hybrid formations and hapax legomena, as well as lexical diversity as established through permutation testing, which offers a statistical approach to interpreting type frequencies and type-token ratios (Suomela 2007). This paper will set the suffix inventories of the five major Middle English dialect areas against general suffixal developments observed in the corpora, tracing diverging characteristics in regional suffix usage through time. For instance, word formation strategies already established in Old English, but declining or only marginally productive in Middle English (-dom, -lac, -ship, denominal -ness), seem to be preserved for longer in West Midland texts than in East Midland writings; deadjectival -hood was used more productively in the East Midlands and the South than in the North; South Eastern texts feature only those -SHIP formations which had been inherited from Old English, excluding more recent coinages – for instance, instead of fellowship, fellowreden is favoured, even though at the time fellowship was already common in the other regions and -reden was on a decline both in terms of frequency and productivity.

In addition, the role of diachronic and regional variation in the occurrence of so-called parallel derivatives will be discussed, where different suffixes can be attached to the same base without apparently causing any substantial change in meaning, as in brotherhood / brotherreden, goodness / goodleik, or fairness / fairhood / fairship / fairleik ‘beauty’. It is argued that the creation and demise of individual derivatives may reflect changes in diachronic and regional suffix usage; fairship and fairleik, to give one example, were found only in texts from the early to mid-thirteenth century associated with the West Midlands where, unlike in other areas, -LEIK and -SHIP reached a considerable level of productivity at the time.

References


Round brackets in Jane Austen

VICTORINA GONZÁLEZ-DÍAZ (University of Liverpool)

Literary scholars have provided insightful accounts of the stylistic value of punctuation in literary texts. As regards the eighteenth century, the dash appears to be one of the punctuation marks with a well-established stylistic value (see Mylne 1979: 60, Moss 1981-2: 195ff, Fanning 2002, Barchas 2003). Another one – and the one that will be studied here – is the round bracket (see Lennard 1991, 1995, Parkes 1992).

As Adamson (1998: 594, 1999: 602ff) notes, (paired) round brackets undergo a process of stylistic re-evaluation in the course of the eighteenth century. In the early eighteenth century, they are considered elements that break the perspicuity of the Enlightened sentence. By contrast, by the beginning of the 19th century round brackets are positively appraised as markers of unplanned thought, spontaneity and emotion. Such re-evaluative shift has only been studied in any depth for poetry (Lennard 1991). Whether round brackets are also stylistically ‘salient’ in contemporaneous prose and, if so, whether they underwent a process of change comparable to that occurred in poetry, remain to be explored.

The aim of the present paper is thus twofold. At the particular, author-based level, it examines the use of round brackets in Jane Austen’s work with a view to investigating their (potential) stylistic value. Recent scholarship (Morini 2009) has explored the ways in which narrative devices and characterisation techniques change in the course of her literary career. Comments on Austen’s use of round brackets are much less frequent (see Lascelles 1939, Philipps 1970). Given that the stylisation of the round bracket is contemporaneous with the development of Austen’s career, an analysis along the lines provided here helps to elucidate whether the process may have in some way influenced her use of the punctuation mark. From a more general perspective, the paper serves as a pilot case-study on the use of round brackets in genres other than poetry.

Methodologically, the paper aligns itself with Corpus Stylistics, which has so far paid relatively little attention to either punctuation matters or historical literary texts (although see Burrows 1987, Fischer-Starcke 2010, Mahlberg 2012, Maschler 2009). All paired round brackets were retrieved from selected electronic versions of Jane Austen’s works and personal letters, and a
A functional typology was created on the basis of (a) the information contained within the round brackets (the parenthetical) and (b) standard functional linguistic descriptions (see Halliday and Hasan 1976, Halliday 1994 [1985]). A quantitative and qualitative comparison of the different types of bracketed parentheticals across Austen’s writings suggests that:

(a) the function of the round bracket changes noticeably throughout Austen’s literary career (from the Juvenilia to the ‘mature’ novels, and, within these latter, from the Steventon to the Chawton novels), and that

(b) those functional changes seem to be motivated by stylistic concerns; i.e. to contribute to the parodic nature of her work (Juvenilia) and to aid the ‘unobtrusive’ characterisation of secondary characters (late ‘mature’ novels; i.e. the Chawton novels).

References


For quite some time now, corpus linguists have done detailed studies of how language use of learners differs from that of native speakers; a particularly prominent theoretical framework in which such work is undertaken is Granger's (1996) Contrastive Interlanguage Analysis (CIA) model. A frequent analytical criterion by means of which learners' non-nativeness, or foreign-soundingness, is assessed is the degree to which learners over- or underuse linguistic expressions or constructions; examples include Aijmer (2002) and Cosme (2008). However, the accuracy and relevance of many such studies can suffer from the fact that mere counts of over-/underuses severely decontextualize the expressions in question. In our work, we propose a multi-step strategy to improve this traditional method and promote a new state-of-the-art method.

The first step goes beyond frequentist accounts of over- and underuses by using multifactorial regression techniques that not only analyze the effect of many linguistic factors at the same time but also allow researchers to contrast native and different learner varieties more rigorously. That is, we annotated occurrences of the expression in question – here, 3902 uses of the modals may and can – for 17 different characteristics of their use. Then, we ran a binary logistic regression that identifies which predictors' effects differ between the native speakers and the learners. In our data, for instance, the learners' use of may with perfective/progressive aspect is significantly different from that of native speakers.

The second step is more interesting. We ran a regression only on the native data and applied the resulting regression equation to the learner data. This is the statistical equivalent of asking the question “in the linguistic situation that the learner is in right now, what would a native speaker do?” This new approach provides researchers with two completely new types of data and allows them

- to compare learners' choices to native speakers' choices in comprehensively-defined and statistically-weighted contexts;
- to distinguish cases where the learner is off a bit from cases where the learner's mistake is only minor.

Using this sequence of steps, we found that the French learners are really only right on target with can in neutral aspect, but their uses of can with perfect/progressive aspect are often wide off the mark. When the French learners use may with perfect/progressive aspect, they are on average a bit off, but their uses of may with neutral aspects again differ considerably from what native speakers would have done in those examples. We demonstrate and exemplify the method in detail and discuss its potential range of applications.
Several studies have found significant differences in the way men and women use nouns and personal pronouns (e.g. Rayson et al. 1997, Argamon et al. 2003). For example, in a recent paper Säily et al. (2011) studied the letters in the *Parsed Corpus of Early English Correspondence* (PCEEC) and found that men used more nouns than women in all the periods studied (c.1415–1681), while women consistently used more pronouns than men. This was explained by the notion of gendered styles: it was suggested that women’s style was more involved, while men’s style was more informational. However, the study by Säily et al. (2011) largely ignored the effect of register (Biber 1988) and audience design (Bell 1984): people write differently in different situations and to different people, and this could also explain the differences in the distribution of nouns and pronouns. Furthermore, Säily et al. treated all personal pronouns as a single category, whereas Biber (1988) has argued that only first- and second-person pronouns are indicative of an involved style, while third-person pronouns are a feature of narrative discourse.

In this study we will provide a detailed breakdown of pronoun frequencies in the PCEEC by number and person. Our results indicate that while women use more pronouns than men overall, this is not the case for each pronoun. For example, in our 17th century pilot data there is no difference in the use of the third person masculine *he*, and it is men who in fact use *we* more frequently in their letters. Moreover, sender and recipient roles seem to have a significant effect on pronoun frequencies to the point where gender differences are leveled out (see Palander-Collin 1999 for contrary findings). For instance, letters between husbands and wives call into question the idea that women use a more involved style of writing. In 17th-century spousal correspondence, men use the first person singular *I* more often than women, while women use *you* more frequently. We suggest that this reflects a difference in audience design: women are more concerned with mitigating potentially face-threatening suggestions to their husbands, and this is reflected in the increased use of *you* (see also Palander-Collin 1999).

Finally, we will discuss the effect of corpus balance on the results. The PCEEC is not evenly balanced for writer/recipient relationships. Overall, the proportion of women’s letters to close family members is much larger than men’s, and this may result in seeming gender differences in

References


**Variation in pronoun frequencies in early English letters: Gender-based or relationship-based?**

**Mikko Hakala, Tanja Säily & Turo Vartiainen** (University of Helsinki)
language use. Focusing on pronoun data, we will investigate whether or not this uneven balance has an effect on the results.

References


PCEEC = Parsed Corpus of Early English Correspondence (tagged version). 2006.


Metadiscourse in novice academic English

HILDE HASSELGÅRD (University of Oslo)

How do learners of English signal the organisation of their academic papers? The present study explores discourse-organising metadiscourse among Norwegian learners of English within two academic disciplines: linguistics and business. Recent corpora of discipline-specific writing offer exciting possibilities of comparing lexicogrammatical features across registers (in the vein of Biber 2006 and Hyland 2008). This investigation uses the VESPA (Varieties of English for Specific Purposes Database) corpus of discipline-specific texts in English by students whose first language is Norwegian in conjunction with similar subsections of the British Academic Written English (BAWE) Corpus (Nesi & Gardner 2012) and the Michigan Corpus of Upper-Level Student Papers (MICUSP) (Ädel & Römer 2012). This combination of corpora facilitates three main types of comparison: English L1 vs. L2, registers (linguistics/business) and British vs. American writers. Additionally, it can show whether the learners are closer to British or American practice in their use of metadiscourse. (Ädel 2006 had a similar design, but without the register dimension.)

The study focuses on textual metadiscourse (Hyland 1998) or more specifically, discourse organisation (Ädel 2010), which comprises the following discourse functions: introducing / delimiting / adding to / concluding topic, marking asides, enumerating, endophoric marking, previewing, reviewing and contextualizing. Search terms comprise explicit references to the paper or parts of it, e.g. (this) paper/essay, introduction, conclusion, summarise. Enumerative and summative conjuncts, e.g. initially, secondly, finally, consequently, are relevant along with the deictic above and below. I also consider patterns of I will, shown previously (e.g. Johansson 2008: 150) to occur with reference to discourse organisation, i.e. previewing, and I have, which is expected to play a role in reviewing. The concordances obviously need to be post-edited to remove non-metadiscourse uses (e.g. In conclusion, the favoured equivalent to the Norwegian
I decided to investigate whether Tottie’s conclusion is accurate. The list of search terms will be expanded after close-reading of particularly introductory and concluding sections of some corpus texts (cf. Bondi 2010).

So far, a pilot study of these expressions in VESPA and BAWE has been conducted. The preliminary results largely concur with those of Ädel (2006): the Norwegian learners overuse most of the metadiscourse markers studied in both disciplines. This shows that learners overuse such markers not only in general argumentative essays but also in discipline-specific academic papers. The overuse of expressions referring to the author (I will / I have), which is highly significant in both disciplines, can be linked to the well-attested higher degree of writer visibility in learner texts (Paquot et al., forthcoming). Another striking example of overuse is references to the paper/essay, often accompanied by the phrase I will. Introduction/ conclusion are, however, not overused, while conjuncts are underused, especially by the business students. The preliminary results thus indicate that learners do signpost their discourse organisation, but in ways that differ quantitatively and qualitatively from native speaker writing. The proposed investigation will survey those differences more closely.

References


BAWE: <http://wwwm.coventry.ac.uk/researchnet/BAWE/Pages/BAWE.aspx>


MICUSP: <http://micusp.elicorpora.info/>


Passive use in academic writing: Comparing expert and student texts across four disciplines

TURO HILTUNEN (University of Helsinki)

Academic prose is characterised by a high frequency of passives (Biber et al: 476). In earlier research, the preference for passives in this register has been linked with the need to obscure the author's identity (Biber 2012, Dorgeloh and Wanner 2009) and reduce his/her responsibility for the assertions being made in the text (Heiderun and Wanner 2003). Moreover, passive use is frequently motivated by discourse needs (Hiltunen 2010, Hyland 1996) particularly when the agent is mentioned (Nessi 2008). In addition, the choice between active and passive voice is connected to other paradigmatic choices, including the choice to use a particular lexical verb, tense, and aspect (Biber et al: 479). For example, as observed by Swales (2004), many commonly used verbs in academic texts vary greatly in their proportional preference for the passive voice (Roger and Smith 1997: 12).

While the general characteristics of passive use in academic English are well-understood, there is relatively little work describing how it varies between texts representing different disciplines. At the same time, it is clear that ways of using the passive are different in different disciplinary contexts. Passives have previously been shown to occur at varying rates in different disciplinary settings, at least as far as research article abstracts are concerned (Hiltunen 2010, Runblad 2007).

This paper compares how passives are used in two kinds of academic texts – published academic writing and student essays – in four different disciplines: medicine, physics, law, and literary criticism. The aim of this study is to determine how frequently the passive is used in different subcorpora and what its main lexico-grammatical patterns are in each subcorpus, focusing primarily on the choice of the verb. The data for this study comes from two corpora representing these disciplines: (1) A corpus of published research articles (described in Runblad 2008 and Biber 2012) student essays extracted from the British Academic Written English (BAWE) Corpus. Both corpora have been part-of-speech tagged using CLAWS.

The study makes use of both variationist (‘Type A’) and text-linguistic (‘Type B’) research designs, as defined by Biber (2012). It compares the subcorpora with respect to how often passives are used in them, and investigates the lexico-grammatical patterns of the instances of the passive in different subcorpora. The analysis confirms that the frequency of the passive varies between subcorpora, with published articles in the ‘hard’ disciplines exhibiting the highest rates of occurrence. By contrast, there is less variation between student texts of different disciplines. The analysis of lexico-grammatical patterns provides explanations for the observed trends, which are often linked with the situational demands of different discourse contexts.

References


**Didn’t pay much attention in school, did you? – Tag questions as markers of (im)politeness in CMC**

SEBASTIAN HOFFMANN (University of Trier), ANNE-KATRIN BLASS (University of Trier) & CLAIRE HARDAKER (University of Central Lancashire)

Canonical tag questions (TQs) such as shown in (1) to (4) are a very conspicuous – and cross-linguistically rare – feature of (spoken) English. As such, they have received ample coverage in the literature, both with regard to their form (e.g. the polarity of anchor and question tag) and their functions. While it is certainly possible to use TQs to express genuine requests for information, they have been shown to perform a whole range of additional pragmatic functions; thus, they may be used to express the speaker's attitude towards their own utterance or to facilitate the flow of interaction, to name but two of the more common ones (see e.g. Tottie & Hoffmann 2006). As a general observation, TQs tend to reduce the level of directness and therefore belong to the repertoire of face-saving strategies available to speakers (see Brown & Levinson 1987).

<table>
<thead>
<tr>
<th>Anchor</th>
<th>Question Tag</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) This is fun</td>
<td>isn't it?</td>
<td>positive-negative</td>
</tr>
<tr>
<td>(2) He's not at home</td>
<td>is he?</td>
<td>negative-positive</td>
</tr>
<tr>
<td>(3) You bought those flowers</td>
<td>did you?</td>
<td>positive-positive</td>
</tr>
<tr>
<td>(4) Yes, they don't come cheap</td>
<td>don't they?</td>
<td>negative-negative</td>
</tr>
</tbody>
</table>

In addition, some researchers have pointed to uses of TQs that cannot easily be reconciled with a face-saving view of politeness, namely TQs that are employed to challenge the conversational partner by being ‘peremptory’ or ‘aggressive’ (Algeo 1990, and in more general terms Holmes 1995) – and that are therefore best described within a theory of impoliteness. However, these uses are reported to be marginal in face-to-face conversation, amounting to no more than 1% each of all instances in Tottie & Hoffmann (2006).

In our paper, we will investigate the use of tag questions in computer-mediated communication (CMC). More specifically, we will report on an analysis of several thousand tag questions retrieved from a large collection of (asynchronous) Usenet discussions. As we will show, our data exhibits quite dramatic differences from what has been observed in face-to-face interaction: more than half of all instances express at least some degree of challenge on the part of the speaker – a preference no doubt supported by the fact that contributors to Usenet discussions can choose to remain fully anonymous. Consider (5) and (6) as typical examples:

(5) Didn't pay much attention in school, did you? (alt.philosophy.debate)

(6) Building theory on flawed premises. You simply buy into everything, don't you? (alt.philosophy.debate)

We will discuss to what extent the existing typology of pragmatic functions of TQs must be extended to cover the range of uses observed in our data. As will become apparent, only a minority of instances in fact correspond to the challenging uses previously discussed in the literature.
We believe that our Usenet-based findings also carry some relevance to the study of spontaneous spoken conversation, as it seems to us that genuinely challenging contexts are likely to be greatly underrepresented in standard spoken corpora of present-day English such as the BNC.

References


**A corpus of early American literature: Compilation, composition and preliminary results**

**MIKKO HÖGLUND & KAJ SYRJÄNEN** (University of Tampere)

This paper presents the *Corpus of Early American Literature* (CEAL, pronounced /si:l/), a prose-oriented text corpus designed to be an American English counterpart for the *Corpus of Late Modern English Texts* (CLMET) (De Smet 2005). The CLMET covers British English from 1710 to 1920 and it is divided into three 70-year periods, and the same design is also employed for the CEAL. The corpus comprises unannotated raw text files collected from various online text repositories, and is roughly comparable in size to the extended version of the CLMET (De Smet 2012). Demand for this kind of diachronic corpus is evident, as it fills an important chronological gap found in many existing historical corpora of American English – the 18th century. For instance, the *Corpus of Historical American English* (COHA) (Davies 2010-) only covers the period from 1810 onwards, while the *Corpus of Early American English* (Kytö 1994) covers the preceding century (1620-1720). The 18th century is arguably an important period for American English, as it was during this time that it evolved relatively rapidly into a variety of English that is clearly distinguishable from British English.

The founding idea for the present corpus was to create an American English counterpart for the CLMET and to facilitate diachronic comparisons between British and American English. For this reason, the criteria used in compiling the corpus are based on (but are not entirely identical to) the criteria used by De Smet in the compilation of the CLMET. The compilation criteria of the CLMET were straightforward to apply to the second subpart (1780-1850) and the third subpart (1850-1920) of CEAL. However, the first subpart (1710-1780), during which time American English was only beginning to emerge, called for various compromises and revisions to the criteria because of the scarcity of material.

In the present paper we will introduce the compilation criteria of CEAL and re-evaluate earlier observations on the challenges encountered in the compilation process (Högland & Syrjänen 2010), which largely arise from the exceptional sociopolitical and cultural circumstances in colonial America. We also provide an outline of the contents of the finished product (including
the word counts, and the numbers of authors and texts), and compare some of the features of CEAL with those of the CLMET. In addition, we present a short pilot study which investigates the evolution of the *tough* construction in American and British English with the help of the two corpora.

**References**


**Beyond (semi-)modals: Changes in the domain of obligation/necessity in Present-Day American English**

VERENA JÄGER (Johannes Gutenberg University of Mainz)

For the past decades, linguists have witnessed a “wholesale reorganization” in the system of English modal auxiliaries (Bolinger 1980: 6). Particularly the Brown family of corpora has served as a basis for many investigations showing that the core modals overall have drastically declined in frequency during the last century in standard British and American English (cf. Leech 2003, 2011). This decrease is, however, is only partly compensated by an increase in the use of the so-called emerging modals (Krug 2000) or semi-modals (cf. Leech 2003, Leech et al. 2009) such as *have to, need to* and *want to*, which can express root meanings of necessity and obligation in a “more indirect and face-sensitive” (Fairclough 1992: 203) way than core modals like *must and should*. Is there, then, a general decline in the use of modal expressions in the English language? Or can we observe a growing tendency to make use of other expressions of modal meaning such as lexical verbs, adjectives or nouns?

In an attempt to answer Leech’s (2010) question “Where have all the modals gone?”, I will extend the focus beyond modals and semi-modals as the chief exponents of modality and take a closer look at other expressions of necessity and obligation that can have root meanings. Whereas epistemic adverbs, adjectives and lexical verbs have been extensively studied (cf. Hoye 1997, Nuysts 2001, Kranich & Gast to appear), often in relation to expressions of stance and as hedges (e.g. Biber 2004, Hyland 2005), the domain of root modality has generally received “little specific attention” in comparison to epistemic modality (Nuysts et al. 2010: 17), and adjectives such as *necessary, essential, critical, crucial* (cf. Van linden and Verstraete 2011, Van linden 2012), lexical verbs such as *oblige, force, demand, urge and require* (cf. Carter and McCarthy 2006), and nouns like *demand, request, or requirement* (cf. Kanté 2010) have received only fairly recently attention as carriers of
modal meanings. I will analyze the development in the use of these expressions in the 20th century using the Corpus of Historical American English. Recent changes (1990-2009) in meaning will be retraced in the ‘spoken’ section of the Corpus of Contemporary American English (COCA). It will be shown that, with the exception of the emerging modals have to and need to and a few semi-auxiliaries, a general decline in the use of all these expressions can be observed. Hence, the decrease in the use of the core modals is not counterbalanced, but, on the contrary, paralleled by a decline in the use of lexical verbs, modal adjectives and nouns of necessity and obligation – a development which seems to be indicative of a general trend towards less explicit expressions of modality in American English.

References


Adjectival complementation: Genre variation and meaning

HENRIK KAATARI (Uppsala University)

This paper presents an analysis of adjectives complemented by *that*- and *to*-clauses, exemplified in (a)-(d).

(a) It was *inevitable* that he should be nicknamed ‘the Ferret’, although seldom in his hearing.

(b) I'm *happy* that we are married.

(c) It is *difficult* to test a potential cure when a disease is ill-defined.

(d) Yet the authorities were *unable* to silence the expression of political opposition.

The questions that will be addressed in this paper are to which extent these four complementation patterns are distributed differently across genres, and to which extent these genre differences can be linked to the differing primary meanings and functions associated with these complementation patterns.

A distinction is drawn between extraposed clauses and post-predicate clauses. Extraposed clauses, such as (a) and (c), with non-referring *it* as matrix subject are distinguished from post-predicate clauses, such as (b) and (d), with a referring pronoun(b) or a noun phrase (d) as matrix subject.

In previous studies of adjectival complementation there is a clear bias towards extraposed clauses (e.g. van Linden 2012) and the analyses are usually confined to a predetermined set of adjectives (e.g. Mindt 2011). Against this background, using corpus data from the BNC (Burnard 2007), this paper takes instead a non-lexical approach by looking at *all* the adjectives found that are complemented by *that*- and *to*-clauses, based on the assumption that it is important to include all adjectives that occur in these complementation patterns to get a full understanding of their differing meanings and primary functions.

First, I will present the distribution of these four types of complementation patterns across five different genres (academic prose, popular science, news, novels and conversation), showing that there are significant differences between them in the overall distribution across genres as well as in terms of the number of different adjectives represented in each genre. Second, I will link
these differences in genre distribution to the different meanings expressed by these complementation patterns by presenting the results of a collexeme analysis (Stefanowitsch & Gries 2003). The collexeme analysis shows that the two types of that-clauses are associated with different meanings. Extraposed that-clauses are primarily associated with epistemic modality, commenting on the likelihood of the proposition of the extraposed clause being true, typically involving adjectives such as clear and likely. Post-predicate that-clauses, on the other hand, are primarily associated with evaluation, typically involving adjectives such as glad and afraid. These differences in meaning can be linked to the differing genre distribution of that-clauses as extraposed clauses are most frequent in academic prose and popular science whereas post-predicate clauses are most frequent in novels and conversation, indicating that the two complementation patterns have different functions. Similar observations are made for to-clauses. For example, extraposed to-clauses are primarily associated with dynamic modality, typically involving adjectives expressing possibility and difficulty (e.g. possible, difficult and easy). It is thus shown that the difference in meaning cuts across both clause type (that vs. to) and construction type (extraposed vs. post-predicate).

References


From matrix clause to pragmatic marker: Information packaging and grammaticalization

Gunther Kaltenböck (University of Vienna)

It is generally assumed that the principles of information packaging have an impact on how speakers arrange information in actual language use, with initial position favouring low information value and lack of complexity. This raises the question to what extent information packaging may have an influence on the underlying grammatical system, which in a usage-based, emergent view of grammar (e.g. Hopper 1987) is seen as arising from actual language use as a conventionalization of recurrent usage patterns. More precisely the question is whether frequent use of certain low-informational items in initial matrix clause position makes them prone to grammaticalization, which, according to Boye and Harder (2007), affects elements with discourse secondary usage status.

To explore the question of a potential link between information packaging and grammaticalization the paper investigates the diachronic development of matrix clauses whose communicative purpose lies not so much in the expression of semantic content but in acting as a starting point or ‘launching pad’ for a new proposition in the form of a subsequent that-clause. Potential sources for the development of such ‘presentational matrix clauses’ are all types of
matrix clause constructions, such as object clause constructions, it-extrapolation, copular sentences, and inferential matrix clause constructions, as illustrated in (1).

(1) a. (I) think (that) a student failed the exam.
   b. (It) turns out (that/like) a student failed the exam.
   c. (The) thing is (that) a student failed the exam.
   d. It's just (that/like) a student failed the exam.

Using data from the Corpus of Historical American English (COHA, Davies 2010) and additional material from the Diachronic Corpus of Present Day Spoken English (DCPSE) and the Corpus of Contemporary American English (COCA) the paper discusses various exemplary cases of grammaticalizing matrix clauses such as in (1). Although different in form and function, they are shown to be all on the same pathway of adopting discourse-organizational functions and becoming fixed pragmatic markers. The concomitant effects of this development are bleaching of their original semantic content, formal invariability, loss of the that-complementizer, reduction and fusion to a unitary element and, in some cases, positional flexibility (e.g. I think, it turns out). The result of this process is that of a formulaic pragmatic marker which introduces a finite clause. In addition to any other pragmatic functions, which may vary for each type, they all serve as convenient low-informational starting points for the presentation of a new proposition, in accordance with the principles of end-focus and end-weight. The pragmatic principle of information packaging is thus shown to affect not only the level of actual usage but also, over time, the underlying structural level, with initial matrix clauses being a preferred locus of grammaticalization.

References


Discussing dissing and cussing: Impoliteness in online discourse

ANDREW KEOHE & URSULA LUTZKY (Birmingham City University)

Previous corpus linguistic analyses of impoliteness have tended to focus on transcribed speech. This has been particularly true in studies of swearing, several of which, from McEnery et al. (2000) to Ljung (2009), have focussed on the spoken component of the British National Corpus (BNC). There is justification for this, given that swearing is often seen as a feature of unplanned, spoken language – “an outlet for frustration and pent-up emotion and a means of releasing nervous energy after a sudden shock” (Crystal 1997: 61). Indeed, later work by McEnery & Xiao (2004) found expletives to be up to twenty times more frequent in the spoken component of the BNC than in the written component.

In this paper, we examine impoliteness (Culpeper 2011: 254-255) in a written text format but one which has been shown to exhibit some of the features traditionally associated with spoken
discourse: the blog. Our analysis is based on an 80 million word sub-corpus of the Birmingham Blog Corpus (BBC; <http://www.webcorp.org.uk/blogs>), which includes both blog posts and reader comments on them. The commenting feature on blog posts allows some degree of interaction between author and reader and, more frequently, between individual readers. This opens up new possibilities for pragmatic analysis in general and, in this study, for the analysis of impoliteness in particular.

There have been few corpus linguistic studies of impoliteness in online environments, including studies on ‘flaming’, which has received some attention in the social sciences. Flaming has been defined as “the exchange of emotionally charged, hostile, or insulting messages on computer-mediated communication networks” (Thompsen 1994: 51). Moik (2007) traces the history of flaming back to the verbal duels of the Middle Ages but notes that, unlike these traditional forms, flaming does not appear to have any fixed rules or be limited in terms of style or structure. For this reason, we adopt a lexical approach, beginning with an inventory of impoliteness terms from previous studies and expanding this through collocational analysis. We illustrate how, by examining the collocates of known expletives, we are able to extract further instances of flaming which do not contain any of the words in the initial list. This approach is particularly beneficial in the study of blogs as there are profanity filters in place on some blogging platforms which prevent use of the most common swear words and force users to be creative in their use of impolite language. Our work thus offers new perspectives on issues of growing importance as online interaction becomes more widespread.

References


Presentational focus changes in English

ERWIN KOMEN (Radboud University Nijmegen and SIL-International)

Presentational focus is a form of focus that spans the core of a clause, while placing particular highlighting on the subject (Andrews 2007, Lambrecht 1994, Sasse 1987). This form of focus is used mainly to introduce a new character in a discourse, as in (1a), but it is also used to introduce an existing character in an unexpected place, as in (1b).

(1)  a. Da fieringa com Archaestrates, ealre þare þeode cyninge, then suddenly came Archaestrates all that people’s king
    mid micelre mænio his manna. [coapollo:233]
    with great company his of men
    ‘Then suddenly came Archaestrates, king of all that people, with a great company of his men.’

    b. Ongemang þissum, com ham Pafnuntius. [coeuphr:88]
    In the midst of this came home Paphnutius.
    ‘In the midst of this, Paphnutius came home.’

While the presentational focus involving reintroduction of participants, as in (1b), requires extensive manual study of a text in order to be recognized (Paphnutius comes home at a time the main character Euphrosyne is not at all expecting him), the type where referentially new subjects are introduced, as in (1a), is ideally suitable for a corpus study on syntactically annotated texts that are enriched with referential information. Such texts are currently being prepared semi-automatically by making use of the ‘Cesax’ computer program (Komen, 2012). The information is added to existing texts taken from four syntactically parsed corpora (Kroch et al. 2004, Kroch et al. 2010, Kroch and Taylor 2000, Taylor et al. 2003). One of the reasons to add this information is that it allows studying information structure changes in English (Los, 2009). The question we want to answer is in what way presentational focus strategies have changed, and what that tells us about the interaction between syntax and focus.

This paper describes a corpus research that tracks the changes in the linguistic expression of presentational focus in English by making use of the referentially enriched texts using queries written in ‘CorpusStudio’ (Komen 2012). The results show that presentational focus is characterized by (a) demarcation of the participant NP that is being introduced and (b) positioning of this participant in such a way that it complies with the Principle of Natural Information Flow (Comrie 1989, Firbas 1964, Kaiser and Trueswell 2004), while it diverges from the canonical word order where the subject precedes the finite verb. The interaction of these principles with the changing syntax of English results in different strategies for the expression of presentational focus: Old English uses the ‘late-subject’ construction, while Late Modern English has switched to an expletive construction, witness Figure 1.
The line marked ‘NoExpletiveSbj’ looks at postverbal subjects, and gives the proportion of them that are referentially new. The line labelled ‘WithExpletiveSbj’ gives the proportion of clauses that use the expletive strategy.

The results illustrate the interaction between syntax and focus: the former requires word order for the expression of grammatical functions in an amount that changes over time, while the latter creatively makes use of the focus area demarcation possibilities available at any given time, leading to different strategies for the expression of focus.

References


New and learner Englishes: Re-evaluating the norm orientation continuum

**Samantha Laporte** (Université Catholique de Louvain) &** Alison Edwards** (University of Cambridge)

The divide long believed to exist between outer and expanding circle Englishes has recently been called into question. Gilquin and Granger (2011) point out that the exposure to and use of English varies substantially within EFL countries, and Hilbert and Krug (2012) and Edwards (forthcoming) demonstrate that, within varieties, characteristics of EFL and ESL can coexist. As Hundt and Vogel (2011:161) write, “increasing globalization might eventually blur the distinction between ENL, ESL and EFL varieties”. However, although studies comparing inner, outer and expanding circle countries are indeed emerging (e.g. Hundt & Vogel 2011, Wulff & Romer 2009), due to the nature of the corpus data available from the expanding circle (e.g. ICLE), such studies tend to focus on student writing only. Against this backdrop, we expand this scope of genres and take a first step in answering Davydova's (2012) call for indigenised and learner varieties to be investigated on the same grounds, with the potential for variation of a similar nature depending on their variable extra-linguistic backgrounds.

We seek to shed further light on the nature of the continuum across EFL, ESL and ENL. Our data come from the written components of the *International Corpus of English* (ICE) for Great Britain and the USA (ENL) and Hong Kong, India and Singapore (ESL), as well as from a comparable corpus of Dutch English (EFL). The latter, to our knowledge, is the first expanding circle corpus encompassing all ICE text categories, thus allowing for comparisons across a range of genres. Inspired by Gilquin and Granger's (2011) work with ICLE, we take the preposition *into* as a case study, conducting a quantitative and qualitative analysis of its syntactic patterns, semantic distribution, lexical variation, phraseological uses and non-standard uses. We aim to test recent claims that the cline to be found in terms of norm orientation is ENL > EFL > ESL (Hundt & Vogel 2011, Van Rooy 2006) and, within ESL, that the more advanced varieties in Schneider's (2003) dynamic model will be the most dissimilar to ENL (Mukherjee & Gries 2009). In light of these claims, we hypothesise that Dutch English will be closest to the native norm, and Singapore English most distant.

Hierarchical cluster analyses in fact reveal Singapore English to be the most norm oriented, thus supporting Hundt and Vogel's (2011) assertion that such ESL varieties can show lingering exonormative trends. Moreover, Dutch English is not markedly distinct from the New Englishes, but clusters with them in different ways depending on the focus of the analysis, e.g.
like Indian English, it shows a relatively higher proportion of intransitive patterning with *into* than the other corpora. These results support Davydova's (2012) claim that learner and New Englishes should be approached in an integrated fashion. In our view, they should be seen as existing on a continuum along which individuals and groups can move depending on their norm orientation as well as their levels of proficiency in and exposure to English.

**References**


**Complexity in Early Modern English statutes: The diachrony of binomials and multinomials**

**Anu Lehto** (University of Helsinki)

The study examines the development of binomials and multinomials in Early Modern English legal writing. Binomials have been defined as words that are connected by coordinators and that are related semantically in their meaning; consequently, multinomials form longer sequences of related concepts (Gustafsson 1984: 124, Frade 2005: 134). Binomials and multinomials are common in present-day English legal writing since they add precision to the texts (Bhatia 1993:
In relation to historical legal language, binomials have been mainly studied from the semantic and synchronic points of view but it is unknown how multinomials developed diachronically in the history of legal writing (see Kopaczyk 2009).

The paper analyses binomials and multinomials from the perspective of complexity. The focus is on the frequency and variety of multinomials. As Rescher (1998 see also Biber 1992) states, one type of complexity is the presence of numerous and various different constituent elements. Multinomials have an effect on complexity since long chains of coordinated phrases add more units to the sentences. At the same time, however, multinomials make the texts more compact, as it is unnecessary to repeat the rest of the sentence. Further, systematic writing conventions with repeated multinomials affect complexity.

The data of the study is the self-compiled corpus, the *Corpus of Early Modern English Statutes* (1491-1707). The corpus begins with the first printed legal texts that were written in English and contains about 215,000 words in four subgenres, i.e. parliamentary acts, proclamations, Privy Council's orders and king's orders. As the corpus is untagged, the multinomials and binomials are located by making lexical searches for coordinators.

Coordinated multinomials are highly common in the Early Modern statutes, e.g. in the category of proclamations their frequency is over 60 in 1,000 words. Typically coordination takes place between noun phrases but coordinated verbs are also common. Further, in the category of proclamations, for instance, the frequencies of multinomials are higher in the seventeenth-century texts than in the sixteenth century. This development is in line with previous studies on complexity that estimate that historical legal writing grew in complexity and verbosity during the Early Modern period (Mellinkoff 1969: 188, Hiltunen 1990: 58).

In addition to the frequency analysis, the study pays attention to the semantic types of binomials and multinomials. Gustafsson (1975), for instance, categorises binomials in legal language into synonymous, complementary and opposite types (Gustafsson 1975, see also Kopaczyk 2009). In the data, the complementary type prevails but synonymous and opposite types are frequent as well.

The paper further analyses the diachronic findings in relation to genre conventions and extralinguistic changes within the frameworks of historical pragmatics and genre studies (Taavitsainen and Fitzmaurice 2007, Bhatia 2004, Swales 1990). In the sixteenth and seventeenth centuries, more emphasis was placed on legal drafting and all-inclusiveness since laws became to be seen as more important in maintaining social order (Baker 2003); this development is shown in the use of multinomials.

**References**


Discourse marker sequencing and grammaticalization

Arne Lohmann (University of Vienna) & Christian Koops (University of New Mexico)

A major pathway for the development of discourse markers (DMs) is the process in which elements governed by the rules of sentential syntax come to function as utterance-initial, extra-sentential markers of discourse coherence (Auer 1996, Traugott 1997). Our talk does not concern this process itself. Instead, it focuses on its outcome: the grammatical status of the resulting DMs. We ask: Does the dissociation from their syntactic sources render DMs essentially devoid of syntax? In other words, is their placement now wholly determined by discourse-functional constraints?

To address this question, we examine a less studied but pervasive aspect of DMs: their regular occurrence in two-part sequences like and so, oh well, etc. In such sequences, the question of the DMs’ relative position becomes relevant. Is the placement of a DM in either first or second position determined by each DM’s source syntax? Or does the ordering deviate from this, due to new preferences brought about by the DMs’ discourse-level function?

To date, no investigation of the sequencing behavior of DMs in conversational discourse has been reported (but see Oates 2000, Fraser 2011). We present an analysis of Schiffrin’s (1987) set of eleven high-frequency DMs (you know, I mean, well, and, but, or, so, because,
now, then, oh) in the Fisher corpus (Cieri et al. 2004, 2005), a 24-million word corpus of telephone speech. After extracting all attested instances of the 110 theoretically possible combinations (N=156,963), collocational analysis (Stefanowitsch & Gries 2003) reveals a large number of statistically strongly associated sequences. Based on each DM's total amount of significant leftward and rightward associations, we derive the sequencing hierarchy in (1).

DMs further to the left are more likely to occur in initial position, while DMs further to the right tend to occur in second position.

(1)  oh>\{well, or\}>\{and, but\}>so>because>\{now, then\}>I mean>you know

A regression model which tests the validity of the hierarchy in (1) shows that it can explain a substantial amount of the attested orderings (adjusted R-squared: 0.29).

The strong tendency for oh and well to occur in initial position can be related to their discourse function. Both respond to prior discourse, usually the prior turn (Schiffrin 1987). It also reflects their grammaticalization history, as both originally marked independent sequential moves (Auer 1996). For the remaining nine DMs, however, the ordering follows the order expected from their 'traditional' syntax: coordinators>subordinators> sentence adverbs>matrix clauses.

It appears, then, that discourse markers are not necessarily devoid of syntax, but remain strongly constrained by the grammar of their sentence-level counterparts. We interpret the 'shallow' grammaticalization of DMs, seen in their low degree of decategorialization (Hopper 1991), as stemming from an inherent flexibility in their use. As an item is recruited into the grammatical domain of ‘discourse marking’, it retains the potential to serve in its original syntactic function. This property allows speakers to deploy DMs without having to decide, at the time of utterance, whether the intended interpretation is in terms of discourse structuring or in terms of sentence grammar.

References


The figure sixty forty seven forty five...: Modals of necessity in Hong Kong and Indian English

LUCÍA LOUREIRO-PORTO (University of the Balearic Islands)

The expression of necessity in Standard Englishes has been the subject of numerous studies which arrive to a common conclusion: the semi-modals need (to), want to and have to are gaining ground to the detriment of central modals such as must (see Krug 2000, Leech 2003, Smith 2003, Nokkonen 2006, Pfaff 2010, Aarts 2011 among many others). At the same time, recent studies on so-called emergent varieties of English suggest that the same development can be observed for South East Asia Englishes (particularly those varieties spoken in the Philippines, Singapore, Hong Kong and India, cf. Collins 2009) and this tendency appears to be extended to Fiji and Samoa, where, according to Biewer (2009), deontic must is very rare. Additionally, concerning the particular use of need and need to, Lee (2001) affirms that polar need is widely accepted in Hong Kong.

Judging from the literature, it could be easily concluded that no special behaviour can be found as for the use of modals of necessity in World Englishes, as compared to the varieties spoken in UK and US, used as reference varieties. Nevertheless, some of the findings listed above have already been proved to be wrong. For example, the use of polar need is claimed to have declined in Hong Kong by van der Auwera et al. (2012). Likewise, a preliminary corpus search (using the International Corpus of English) reveals that Indian English and Hong Kong English exhibit utterly different behaviours regarding the choice of the verbs must, need and want, as only in Hong Kong English do we observe a decay in the use of must and an increase in the use of need and want, even in epistemic contexts such as “the figure sixty forty seven forty five want to have been multiplied by point nine five”. These preliminary results go against Collins' (2009) claim that all Asian Englishes exhibit the same kind of variation as British and American English as far as modals of obligation are concerned. In addition, many previous studies are singly based on frequency rather than on the specific meaning and use of each modal verb, which renders only tentative conclusions and thus leaves plenty of room for further research.

Therefore, with the aim of delving into the understanding of the status of the modals of obligation in Hong Kong and Indian Englishes, this paper adds have to to the list of verbs which should be compared to must, need and want and conducts a corpus-based study of these two New English varieties together with a parallel study for the British English component of the ICE project (ICE-GB). The aim will be threefold: a) to map the degree of similarity among the different uses of the verbs in the three linguistic varieties studied; b) to determine the grammatical status of each of the verbs in each of the varieties, i.e. its actual degree of grammaticalization; and c) to assess the role played by factors such as language contact and linguistic democratization (see Fairclough 1992, Farrelly and Seoane 2012) in the preference of one verb over another, i.e. to evaluate to what extent the differences between the varieties can be explained in sociolinguistic terms (cf. Schneider 2010).

References

Aarts, B. 2011. Shifts in English modal verb usage in recent times. Plenary lecture delivered at the 44th Meeting of the Societas Linguistica Europaea, Logroño (Spain), 8–11 September 2011.


Anaphora resolution in a learner corpus of L2 Spanish

CRISTÓBAL LOZANO (University of Granada)

While second language research has traditionally relied on (quasi)experimental data, L2 researchers have started to use corpus data recently (Myles 2005, 2007). This presentation shows how corpus data reveals unexpected L2 behaviour that has gone unnoticed in experimental studies.

Data come from *Corpus Escrito del Español L2* (CEDEL2: Lozano 2009a). CEDEL2 contains around 800,000 words coming from nearly 2,500 learners (L1 English – L2 Spanish) at all proficiency levels, as well as an equivalent Spanish native subcorpus. A subset of the corpus (very advanced learners) was tagged with *UAM Corpus Tool* (O’Donnell 2009) to investigate anaphora resolution in L2 Spanish.

The bulk of experimental research on L1 English–L2 Spanish reveals a robust pattern (Al-Kasey & Pérez-Leroux 1999, Liceras 1988, Lozano 2002): learners acquire early the fact that overt and null referential pronominal subjects can alternate syntactically, (1). But such (apparently free) alternation is constrained discursively in native Spanish: null pronouns encode topic continuity (Ø in 2), while overt pronouns encode topic-shift when a change of referent is required (él in 3). Recent experimental L2 studies indicate that learners show persistent deficits at the syntax-discourse interface (Margaza & Bel 2006, Pérez-Leroux & Glass 1999, Rothman 2009): they often produce (i) an overt pronoun in topic-continuity contexts, which causes redundancy (él in 2), and (ii) a null pronoun in topic-shift contexts, which causes ambiguity (Ø in 3). Parallel findings have been also observed in L2 Italian (Sorace & Filiaci 2006).

(1) \[
\begin{align*}
\{\text{Él}\} & \quad \text{es millonario.} \\
\{\text{Ø}\} & 
\end{align*}
\]

‘He/Ø is a millionaire.’

(2) Pedro tiene mucho dinero y \[
\begin{align*}
\{\text{Ø}\} & \quad \text{dice que} \quad \{\text{*él}\} & \quad \text{es millonario.} \\
\{\text{*Ø}\} & 
\end{align*}
\]

‘Pedro has a lot of money and *he/Ø says that *he/Ø is a millionaire.’

(3) María y Pedro viven felices, pero \[
\begin{align*}
\{\text{*Ø}\} & \quad \text{es pobre.} \\
\{\text{él}\} & 
\end{align*}
\]

‘Maria and Pedro live happily, but he/*Ø is poor.’

But CEDEL2 data reveals several important findings that have gone unnoticed in previous experimental research, suggesting that the syntax of topic-continuity and topic-shift in L2 acquisition is richer than previously assumed:

i. Learners not only use an overt pronoun to mark topic-continuity, but they also produce full NPs (Figure 1).

ii. Learners can mark topic-shift via an overt pronoun, as would be predicted for native Spanish, (él in 3), though they drastically prefer using a full NP (Fig. 2).

iii. Additionally, learners also show a tendency to produce informationally richer phrases than pragmatically required (full NP > overt pronoun in topic-shift contexts; overt pronouns and full NP in topic-continuity contexts), which runs against economy principles (Fig. 3). These deficits have to do with the number of potential antecedents of the anaphor, coupled with the gender distinction of such antecedents.
iv. Corpus data thus reveal that learners prefer being redundant and uneconomical to ambiguous, a finding not previously reported in experimental studies.

To conclude, naturalistic (corpus) data can (and should) be used as a follow-up to experimental data to explore new patterns of L2 production (see Gilquin 2007).

![Figure 1: Syntax of topic-continuity](image1)

![Figure 2: Syntax of topic-shift](image2)

![Figure 3: Pragmatics of anaphora resolution](image3)
References


Everything is changed: The be-perfect in Irish English

KEVIN MCCAFFERTY (University of Bergen)

Retention of the be-perfect with intransitive mutative verbs of motion and process is widely noted as a feature distinguishing Irish English (IrE) from other varieties (e.g., Amador-Moreno 2010). The be-perfect has been studied empirically in present-day IrE (Harris 1984, Filppula 1999, Ronan 2005), but there has to date been no diachronic study of this construction in IrE. The present study uses data from the developing Corpus of Irish English Correspondence (CORIECOR). It shows that IrE broadly followed the development in English generally (Rydén & Brorström 1987, Kytö 1997): use of auxiliary be declined in the eighteenth and nineteenth centuries, and it became restricted to particular verbs. In IrE, this appears to have occurred at a delay of about 50 years compared to BrE and AmE. However, various verbs and verb types were affected in different ways in IrE. The be-perfect with motion verbs declined sharply, with the notable exception of go (as in other varieties), while the change happened more slowly with process verbs. Also, a couple of transitive verbs – do meaning ‘finish’ and finish itself – must be included in a study of IrE because they are frequently noted as being used in the perfect with auxiliary be. These verbs go against the trend, becoming more frequent with be over the two centuries studied. This seems to have been facilitated by the fact that many intransitive verbs take an object-like complement, but the influence of the Irish language, where the equivalent of the be-perfect is found with transitive verbs, may also have affected this development. CORIECOR permits investigation of geographical differentiation between regions that became English-speaking early (largely as a result of direct settlement from Great Britain) and regions that adopted English later (as the outcome of language shift), which may help address the issue of whether Irish substrate influence on the be-perfect in Irish English.

References


Discipline-specific English vocabulary use in the 17th century

LILO MOESSNER (RWTH Aachen University)

This paper contributes to the research field of English for Specific Purposes (ESP). ESP is not a reduced form of general language, but a variety of it, which exploits the resources of general language in a characteristic way. This holds for all structural levels of language, but is most conspicuous on the lexical level. The data of ESP vocabulary studies usually come from Present-Day English (PDE). The following properties of ESP vocabulary have been identified: monoreferentiality, precision, conciseness, transparency, and lack of emotion (Beier 1980, Gotti 2003, Brand 2008). They are realised in texts by a high frequency of attributive adjectives, of forms of *be*, of long words, of nouns in general, but of nominalizations in particular, by a low frequency of personal pronouns and of words which are easily recoverable from the context, e.g. of articles and prepositions. In the first part of my paper I will test the hypothesis that the formal properties of ESP vocabulary derived from studies of PDE data can be attested in ESP-texts of the 17th century as well. If this hypothesis can be supported – and results of a pilot study on a smaller set of data point in this direction –, this will be interpreted as a sign of the diachronic stability of ESP vocabulary features in contrast to the changing preferences of grammatical patterns in ESP texts (cf. Halliday 1988, Atkinson 1999).

The analysis and interpretation of linguistic differences between texts of different disciplines is a recent feature of ESP research (Bondi 2010: 8). The only studies on texts of earlier periods which reveal differences between disciplines use the method of multidimensional analysis (Biber and Finegan 1997, Moessner forthc.). Since multidimensional analysis mainly relies on morphosyntactic features, it is not the appropriate method for vocabulary studies. They rely on wordlists and keyword analyses (cf. Scott & Tribble 2006). These methods are applied in studies which investigate disciplinary differences in PDE ESP-texts (Hyland and Tse 2007, Malavasi and Mazzi 2010). My approach in this paper is new in that it investigates disciplinary differences in the vocabulary use of Early Modern English texts of the disciplines natural science and medicine. Keyword analysis and the analysis of the collocates of keywords will be performed in the second part of my paper. I will claim that with these techniques the content matters of natural science and medicine in the 17th century can be delimited against each other and that their discipline-specific style features can be detected.

The corpus to be analysed comprises about 200,000 words. It consists of an ESP-subcorpus and a general language reference corpus. The ESP part is subdivided into a natural science subcorpus and a medical subcorpus of equal size. Both contain 20 texts of about 2,000 words each. The general reference corpus was compiled from the Early Modern English part of the *Helsinki Corpus of English Texts* by deleting the pre-1600 and post-1700 texts of the subperiods 1570-1640 and 1640-1710 and by excluding the specialized text categories and those from the oral end of the written-oral scale. The analytical tools which I will use are part of the program suite *WordSmith* (<http://www.liv.ac.uk/~ms2928/wordsmit.htm>).

References


Corpus findings versus intuition: The value of diverging evidence

SANDRA MOLLIN (University of Heidelberg)

Recent years have seen calls for triangulation of methods in order to better understand the nature of language and language processing (e.g. De Monnik 1999, Featherston 2007, Gilquin & Gries 2009). In particular, corpus linguists are increasingly turning to alternative methods as a complementation of their own methodology. After all, as Leech pointed out, corpora represent the product of language processing, but can “make no claim about the mental phenomena which underlie the production of corpus data” (Leech 1992: 113).

Intuitive judgements are one type of alternative methodology that can be compared to and contrasted with corpus findings. While intuition as such has a bad name in empirically founded theories of language both due to some inherent methodological problems and its connection with generative theories, it is now making its comeback in the form of systematically collected judgement data from large groups of informants.

The paper will review recent studies that combine corpus linguistic methodology with introspective judgements categorising them as to whether the two methods provide converging or diverging evidence. While converging evidence (as in e.g. Gries 2003) is naturally desired by researchers, since it strengthens their results, it is not the only conceivable outcome of a
combination of methods. After all, the two types of data represent different perspectives on language: corpora represent the outcome of language processing, while intuition represents meta-knowledge of language. Neither gives direct access to language storage and production. Cases of diverging evidence therefore exist, and I will argue that these cases are more useful than those representing converging evidence. Diverging evidence points up the weaknesses of each approach and leads us to a better understanding of the psychological scope of the methods.

Among the cases of diverging evidence discussed will be McGee's (2009) comparison of corpus-derived and intuitively estimated frequent collocates, Divjak's (2008) comparison of corpus frequencies and acceptability judgements in verb constructions, and Littlemore & MacArthur's (2012) comparison of corpus-extracted and elicited word senses. In all cases, results from the two data types overlap to some extent, but not perfectly. I suggest that this is at least partly due to the role that corpus frequency is accorded: corpus frequency apparently does not necessarily instantiate salience to the language user. Since neither frequency nor salience, however, are perfect proxies for mental entrenchment, charting the differences provides interesting routes for further research.

Finally, a new study is presented on the relative frequency of pairs of equivalent binomial phrases (e.g. health and safety vs. safety and health), as found in the BNC and as judged by a sample of some 70 native speakers. While there is an overall correlation between corpus data and judgement data, we find areas of divergence, especially concerning irreversible binomials. These differences may partly be explained with methodological problems (corpus size for the corpus data and psychological response strategies for the intuitive data), but also with the fundamental differences between the data types and which kind of language competence they represent.

References


Parallelism and contrast: A multifaceted investigation

M. LYNNE MURPHY (University of Sussex), ANU KOSKELA (De Montfort University) & STEVEN JONES (The University of Manchester)

This paper reports on a multifaceted evaluation of the role of grammatical, lexical and phonetic contributors to the creation of contrastive relations in context. Linguistic means for indicating contrast are many and they overlap, as illustrated in (1):

(1) drug-related offending was relatively [high in relation to alcohol], but [low in relation to cannabis]. (COCA; Davies 2008–)

Lexical, structural, and functional properties of (1) lead us understand alcohol and cannabis as standing in contrastive focus. This understanding is aided by (a) the association of alcohol/cannabis with a canonical antonym pair high/low, (b) their occurrence in parallel lexico-grammatical structures (in brackets), and (c) the use of the contrastive coordinator but to link the parallel structures. Investigations of contrast in text have mostly focused on one or two of these properties – particularly the coordinators (e.g. Marcu and Echihabi 2002). Where they have looked at the role of lexical antonyms (Spenader and Stulp 2007), they have tended to rely on a notoriously faulty tool for antonym identification: WordNet.

In this paper, we investigate the relevance of parallelism, lexical-semantic relations and contrastive connectives in generating novel contrasts by searching for small, conjoined parallel strings in three large corpora (COCA, BYU-BNC, TIME; Davies 2008–). Having searched several unlexicalized conjoined, parallel strings (e.g. ADJ-P-N-CONJ-ADJ-P-N), we identified 28 structures containing canonical antonym pairs that support what Jones (2002) calls ‘ancillary antonymy’, i.e. the creation of a secondary opposition with the support of an established opposition. These include frames such as X by day, Y by night; long on X, short on Y; easy to X, hard to Y. Searching for variations on those partially lexicalized syntactic frames allowed us to collect nearly 2000 instances of ‘double contrast’ (Umbach 2005). We then coded these for the type of semantic relation in the secondary opposition (which can be conceived as level of semantic parallelism), the contrastiveness of the conjunction (e.g. and versus but versus comma), and the parallelism of the structure. Parallelism was calculated on the basis of (a) lexical parallelism (same words in same positions), (b) grammatical parallelism (same number of words, in the same grammatical roles), and (c) morpho-phonological parallelism (shared phones or allomorphs in same positions in words in parallel positions). By investigating semantic and grammatical parallelism in several dimensions, we are able to (a) inform further discussions and attempts towards automatic identification of contrast in text (cf. Spenader and Stulp 2007), (b) examine the conditions under which context-bound antonymy arises – particularly the interaction of contrastive conjunctions with other contrastive markers (cf. Jones 2002, Davies 2012) and (c) support the argument that parallelism itself is a contrastive construction (cf. Murphy 2006, Jones et al. 2012), testing the hypothesis that more linguistic cues for parallelism equals greater contrastive power – supporting the context-specific contrast of otherwise non-antonymous items.
Perfectly regular or totally chaotic? The recent history of English maximizers

NADJA NESSELHAUF (University of Heidelberg)

This paper is intended as a contribution to the currently emerging field of diachronic collocation studies. Whereas so-called ‘boosters’ (such as very, really, so) have received a great deal of attention, from various perspectives such as sociolinguistics and phraseology, including diachronic aspects (cf. e.g. Ito & Tagliamonte 2003, Defour 2012), maximizers have been the focus of much less research and in particular the diachronic phraseological point of view has been largely neglected to date (among the few exceptions are Méndez-Naya 2008 and 2012 on the history of utterly and downright). In this paper, I am going to look at the development of the collocations of a number of frequent maximizers in American English, with a focus on perfectly, entirely, totally, and dead.

The investigation will be based on COHA (Corpus of Historical American English) and COCA (Corpus of Contemporary American English), with additional reference to the OED and the OED online quotations database. First, common collocational patterns of the major maximizers in present-day American English will be investigated (including, in addition to the above-mentioned, completely, absolutely, fully, and wholly). Typical non-syntagmatic contexts of usage will also be considered. In a second step, the emergence and diachronic development of the usage-patterns of selected maximizers will be investigated.

Preliminary results indicate that, in certain collocations, the use of perfectly is very often used to imply some kind of contrast, and that those collocations in which this is the case (e.g. perfectly
normal, or perfectly capable as in “They'd eaten the same thing, and she felt perfectly fine, but it was soon after dinner that he'd taken ill” (COHA 2008 FIC Deadly Gift) are on the increase, whereas those collocations which do not or only rarely occur in such usage (e.g. perfectly quiet) are on the decrease. Another result pertains to the usage of entirely, which is on the decline overall, but has recently been increasing in collocation with not. In this case, it seems that the increase is particularly due to contexts where the conversational implicature of not entirely tends to be ‘not at all’ as in not entirely correct, not entirely true, not entirely sure. Both these examples demonstrate the tendency of maximizers to increase in their phraseo-pragmatic specialization over time.

What will also be demonstrated is the occurrence of what I would like to term ‘collocational’ lag, which occurs when one lexical item is in the course of time replaced by another, but its use with its major collocator stays on or even increases in frequency even after the decline of the lexical item has proceeded for some time. This occurs for example in the case of dead earnest, which stays in frequent use some 60-70 years after the decline of earnest in favour of serious started; only after a certain time span serious inherits the collocator dead.

References


Multilingual practices in the history of written English: Developing a corpus

Arja Nurmi, Päivi Pahta & Jukka Tyrkkö (University of Tampere)

Our paper explores the possibilities of developing a multi-genre corpus for studying multilingual practices in the history of written English. Our previous research has shown that existing historical corpora such as the Helsinki Corpus, the Corpus of Early English Correspondence and the various subcorpora of the Corpus of Early English Medical Writing can be used to study multilingual practices, even though they have been compiled as evidence of the history of English, not the use of other languages (see e.g. Nurmi and Pahta 2004, 2010; Pahta 2003, 2004, 2007, 2011; Pahta and Nurmi 2006, 2007, 2009). It is becoming increasingly clear, however, that for furthering the systematic study of multilingual practices it will be necessary to devise a corpus for that particular purpose.
Our project, *Multilingual Practices in the History of Written English*, focuses on providing a systematic description of multilingual practices in written historical data, and testing the validity of theoretical generalizations based on present-day written and spoken data by examining large systematically collected historical datasets. We have two basic approaches. Case studies of diachronic data testing the comparability of earlier findings on present-day multilingual practices will include (1) fictional texts, where the results of e.g. Callahan (2004) will be compared against a diachronic corpus of fiction to be compiled from existing electronic texts, and (2) personal correspondence from CEEC, which will be mirrored against results from studies concerning e-mails (e.g. Hinrichs 2006). Studies like these will allow us to identify differences and similarities between historical and modern written data, and reach generalizations concerning multilingual practices in specific types of writing. The second approach consists of a more general description of multilingualism evident in writing, and will start with a pilot study identifying one or more topic domains which are particularly rich in multilingual elements; this will be followed up by a data-driven study of the domains.

In this paper we present some initial findings on multilingual practices evident in a variety of genres in the Late Modern English period, and discuss the specifics of developing a structured corpus for studying the phenomenon. Questions of particular interest include the identification of the topic domains and texts most relevant for the corpus, sampling methods for selecting representative corpora and sub-corpora, methodological issues concerning the semi-automatic identification of foreign lexis in historical English texts, and best practices for annotating multilingual passages and for the statistical quantification of language-selection phenomena.

**References**


A corpus-based analysis of anaphoric *one* and its implications

**JOHN PAYNE**, (The University of Manchester), **GEOFFREY K. PULLUM** (Brown University), **BARBARA C. SCHOLZ†** (University of Edinburgh) & **EVA BERLAGE** (University of Hamburg)

In this paper, we report on a corpus-based study of the anaphoric count noun *one* (henceforth *one*$_{ct}$), which occurs in expressions such as *the ones in the catalogue*. This study, based primarily on the BNC, COCA and the *Wall Street Journal* corpus, reveals major difficulties with the syntactically-oriented account of the potential antecedents of *one*$_{ct}$, an account which ultimately derives from Baker (1978). We replace this with a semantic analysis in which *one*$_{ct}$ is looking not for a syntactic antecedent belonging to a particular category (in Baker’s terms an antecedent which belongs to the category ‘nominal’, subsequently N-bar), but rather for an antecedent whose meaning is of the correct type.

The implications are quite radical. One is that the complement/adjunct distinction cannot be motivated within the internal structure of noun phrases. Consider (1):

1. … they point us away from the epistemological frame of reference of this chapter towards the socio-cultural *one* of the next.  

[BNC: FA3 955]

The antecedent of *one* in (1) is *frame of reference*, but this cannot be a ‘nominal’ in the required sense since if substituted for *one* it would itself be followed by the complement-like PP of the next. The analysis we propose resolves this difficulty by treating all nouns and phrases headed by nouns as belonging to a single semantic type. A phrase such as *frame of reference* belongs to the same type as *frame*, and it is this type which allows combination with a dependent PP, not the distinction between nominal and noun.

A corollary of this analysis is that the argument-structure of nouns must be treated in a different manner to the argument-structure of verbs. In our analysis, all nouns (including *one*$_{ct}$ itself) are treated as inherently non-relational until they combine with a dependent PP. This has the desirable consequence of neatly accounting for a fact which is often ignored and which clearly distinguishes nouns as a category from verbs, namely that there are no convincing cases of nouns taking syntactically obligatory dependents. A noun like *king* is typically considered a relational noun, but nothing in the GRAMMAR of English forces us to express the domain over which a king rules. The contrast with verbs, which often take obligatory complements, is a stark one.
A final implication is the discovery that over thirty years of work related to language acquisition has been based on empirical claims that conflict strikingly with the actual facts. That work aims to enlist one\textsubscript{ct} in support of the thesis known as linguistic nativism. The relevant argument is an “argument from poverty of the stimulus”, due also to Baker (1978). It has gained acceptance from philosophers and psychologists in addition to becoming received wisdom in theoretical linguistics and developmental psycholinguistics. Rethinking is in order, because the argument collapses completely once the facts are understood. Viewed from a semantic rather than a syntactic point of view, one\textsubscript{ct} is revealed as particularly easy to learn rather than puzzlingly hard.

References


English-Spanish contrastive analysis: Aids for writing up minutes in an L2

ISABEL PIZARRO (University of Valladolid), ROSA RABADÁN (University of León) & MARLÉN IZQUIERDO (University of León)

The increasingly widespread role of English as a lingua franca calls for the development of more effective tools to assist non-natives in their on-the-job writing tasks, i.e. supportive tools that extend beyond existing empty computer templates (Rabadán and Izquierdo 2011). To address this need, we have devised a writing prototype which is fed with relevant discourse, grammatical and lexical data obtained from corpus-based contrastive analyses of a given genre.

In this paper, which sets out to present an English-Spanish contrastive description of the compulsory rhetorical moves of a specific genre, meeting minutes, moves and steps for both languages are compared at different levels following a top-down approach, and relevant features at the grammatical, phraseological and terminological levels are analysed. The work was carried out using a set of rhetorical labels (tertium comparationis) and a comparable corpus. Informants were called upon to double check the validity of the results. The analysis is corpus-based to reflect current textual behaviour as well as to ensure that the results are representative of authentic language use. The corpus is a custom-made, comparable, annotated corpus (Minutes_C-ACTRES) consisting of one hundred complete texts in each language, English and Spanish, tagged at the rhetorical level with the ACTRES text tagger. Our study combines qualitative and quantitative analysis and is based on Krzeszowski’s classic stages (1990) and on genre (Swales 1990, Upton and Connor 2001, Bhatia 1993, 2004). The following procedure was applied: first, rhetorical moves were identified and then classified as either compulsory or optional in terms of their frequency of occurrence (Biber et al. 2007); subsequently, a description of their main characteristics at different linguistic levels was formulated. The grammatical keys focused on were clause types and verbs in main and subordinate clauses (tenses, mood, and voice). The lexical and phraseological keys focused on were lexical verbs, deverbal nouns, key-words and n-grams up to 6-grams. The same process was followed for each language with two main objectives: (1) the identification of linguistic patterns to feed the Minute Generator; (2) the identification of linguistic similarities and dissimilarities between English and Spanish minutes to raise awareness of the differences between languages.

Results reveal that there are significant statistical (quantitative) and qualitative differences on all the levels analyzed: rhetorical, grammatical, phraseological and terminological. The prototypical results feed a computerized writing tool (Minute Generator) designed for non-
native users who are required to write minutes in English but who are not sufficiently proficient and/or confident in L2 writing to do so effectively.

References


‘Hapax Legomena’. The silent majority in text

ANTOINETTE RENOUF (Birmingham City University)

In recent studies, we have been working towards a finer definition of neology from a modern diachronic corpus-based perspective. We have modelled the life-cycle of a word in text (Renouf 2012, forthcoming a), as well as the pattern of cumulative neologising activity which accompanies each new word which attains a threshold of popularity in news text (Renouf, forthcoming-b). Our focus so far has been on those ‘happy few’ neologisms which, for various reasons, capture the public and media imagination.

In this paper, we turn our attention to the ‘silent majority’ of neologisms which have not made an impact in our corpus (so far). The ‘hapax legomenon’ is a word occurring once only, and over half the word types in a large corpus of text are ‘hapax legomena’. Yet they do not occur randomly, but are clearly each designed to fulfil a particular role in spoken and written communication. Our aim is to conduct a pilot study of a subset of the flow of neologisms passing through news text over time, observing their relevant linguistic and extra-linguistic properties and functions, to begin to account for their obscurity.

The research method will involve the examination of neologisms taken from a particular month’s time-chunk of news text, using the WebCorp Linguist’s Search Engine (<http://www.webcorp.org.uk/lse>) (Renouf & Kehoe, forthcoming) to access their original contexts, studying the features and circumstances of occurrence of words which remain ‘hapax’ (or ‘dis’ ‘legomena’ as distinct from those which go on to recur with higher degrees of frequency at later stages in our diachronic corpus. The selected time-chunk will be prior to the
end date of our evolving corpus, allowing a sufficient time lapse for a profile of change over time, if any, to be discoverable.

The study will select from a database of all words recorded as occurring for the first time in an unbroken flow of ‘Independent’ and ‘Guardian’ mainstream news text from 1984 to end 2011. This database was created by analytical software tools developed during the APRIL project (1997-2000) (Renouf 2010). As in the earlier AVIATOR project (1990-1993) (Renouf 1993, Baayen & Renouf 1996), new word tokens are identified by a filtering process which matches them against an accumulating record of previously-occurring word types. If not present in the master wordlist, the words are recorded as new in the evolving database, at a rate of around 7,000 per month (including errors). Unlike AVIATOR, APRIL goes on to process them through a novel chart parser for morphological analysis; also for POS tagging.

The results will establish a working typology for the sample of new words selected, with attention paid to the qualities of neologicist ‘losers’ in contrast to ‘winners’. This corpus-driven analysis will throw new light on the evolving lexicon at hapax level, and complement enquiries into the ‘success’ and ‘conventionalisation’ of English neologisms (see Boulanger 1997, Kerremans 2012).

References


Linking learner corpus and experimental data in studying second language learners’ knowledge of verb-argument constructions

UTE RÖMER (Georgia State University), AUDREY ROBERSON (Georgia State University), NICK ELLIS (University of Michigan) & MATTHEW BROOK O’DONNELL (University of Michigan)

The past few years (and ICAME conferences) have seen an increase in studies that highlight the value of combining corpus and experimental evidence in the study of linguistic phenomena (consider, for example, Gilquin and Gries 2009, Römer, O’Donnell and Ellis 2012). Many of these studies utilize corpora of native speaker output (such as the BNC or ICE-GB) to derive frequency data which is then considered in relation to speaker responses or judgments collected in experimental settings. The present paper discusses the role that second language learner output, as captured in learner corpora, may play in this context.

We investigate how learner corpora and experimental data complement each other in providing insights into second language learner knowledge of 40 different verb-argument constructions (VACs), including the ‘V against n’ and the ‘V like n’ constructions. We are interested in finding out which verbs learners of different L1s most commonly associate with a particular VAC and whether/how their verb-construction associations are different from those of native speakers. Corpus sources are the International Corpus of Learner English (ICLE), from which we consider subsets of essays written by L1 Spanish and L1 German learners, and the Louvain International Database of Spoken English Interlanguage (LINDSEI), from which we consider subsets of tasks produced by speakers of the same two native languages. From both corpora we extract instances of the selected 40 VACs and examine the type and token distribution of verbs in each VAC. In addition to the corpus data, we draw on data collected in lexical production tasks in which participants complete VAC frames (e.g. ‘She against the...’) with verbs that may fill the blank (e.g. pushed, protested, ran). Previous research asked respondents to generate the first verb that came to mind in a particular VAC frame (Römer, O’Donnell and Ellis forthcoming). The current study expands on this by using a verbal fluency test that asks native speaker, L1 German, and L1 Spanish participants to generate as many verbs as possible for a given VAC frame over a span of 60 seconds. The results of fluency tests like this provide evidence about the typicality of verbs in a VAC, as more typical verbs are likely to be produced early, and by more participants, than less prototypical ones (Gruenewald and Lockhead 1980).

Lemmatized, frequency-sorted lists of verbs from the experiments are compared against each other and against the verb lists that resulted from the ICLE and LINDSEI analyses of the same VACs. The comparisons help us assess the contributions of these different types of data to a better understanding of learner VAC knowledge. We will make reference to Gries’ (forthcoming) typology of linguistic data and consider the naturalness of the data captured. Our approach also allows us to identify those VACs for which language learners supply ‘atypical’ verbs (in comparison to native speakers) and to consider why they may struggle with these particular constructions. The paper closes with a discussion of how the combination of learner corpus and experimental data leads us to a more comprehensive understanding of the extent to which VACs are entrenched in the second language learner’s mind.

References


Investigating light verb constructions in contemporary British and Irish English

PATRICIA RONAN (University of Lausanne) & GEROLD SCHNEIDER (University of Zurich)

A large amount of theoretical studies can be found on the use of light verb constructions in different languages, amongst which in the English language (e.g. Wierzbicka 1992). Lately, a number of corpus based studies has been conducted on the use of light verbs in Old-, Middle- and Early and Late Modern English. Recently the general interest in verb complementation patterns across varieties of English has increased (e.g. Mukherjee and Hoffmann 2009, Mukherjee and Gries 2009), as well as research into methodologies in automatic extraction of verb-argument extraction from corpora (e.g. O'Donnell and Ellis 2010), nevertheless there are only few comprehensive corpus-based quantitative studies on light verb constructions on contemporary varieties of English. One such study is Algeo (1995), which investigates the five most frequent light verbs in LOB and Brown. A more extensive survey is given by Hoffmann (1972), based on 732 tokens in a 1.5 mio word drama corpus from the 1950ies and 1960ies.

The proposed paper aims to redress the shortage of corpus based research on light verb constructions in contemporary English from the British Isles by examining which light verb constructions are frequent in contemporary British English, and with what frequencies they are found. The basis of the research will be formed by the ICE GB component, the results of which will on the one hand be compared to the larger BNC and on the other hand to ICE Ireland. We will use automatically parsed versions of the corpora and extract light verb occurrences by using collocation measures (Lehmann and Schneider 2009) combined with manual filtering, and we will evaluate our results. We will also test if simple verbs corresponding to light verb constructions, e.g. to kiss, can improve the automatic detection of the related light verb construction, e.g. to give somebody a kiss.

This approach will enable us to identify light verb constructions confidently in a 1 million word corpus, and test the results against the larger BNC. Possibilities for inter-variety differences will also be tested by comparison of ICE GB and ICE Ireland. This research is valuable on the one hand for determining the use and the frequency of these collocations in contemporary English as compared to earlier varieties. On the other hand, it contributes to the study of verb-object collocations in contemporary varieties of English. As such it will clarify whether different varieties of contemporary English on the British Isles display variant use of light verb constructions.
References


Wierzbicka, A. 1982. Why can you have a drink when you can't *have an eat? Language 58: 753-99.

Conversational narrative as drama: From story climax to utterance climax

CHRISTOPH RÜHLEMANN (LMU Munich)

In conversational storytelling, presentations of discourse that (might have) occurred in the told situation are pervasive. Narrators have at their disposal a wide variety of types of discourse presentation, some of which add to the dramatic impact of storytelling. For Tannen (1986: 312), for example, ‘constructed dialogue’ (Direct and Free Direct quotation) “is a means by which experience surpasses story to become drama”. Further, the dramatic character of constructed dialogue has often been linked to a narrative’s build-up towards its ‘point’ in the sense that “[t]he climax of a conversational story is often realized in dialogue” (Norrick forthcoming). In my talk, I will stress the role of discourse presentation in dramatizing a much smaller unit: the individual utterance.

The talk will be based on data from the Narrative Corpus, a specialized corpus of conversational narrative which is heavily annotated for a broad range of discourse phenomena, including different types of discourse presentation (cf. Rühlemann & O'Donnell 2012). Adapting Leech and Short’s (1981) seminal model, I will outline a model of discourse presentation in which discourse presentation types are ordered in terms of the immediacy of the
report to the hearer. Based on this model, I will investigate patterns of sequencing report units at utterance level in the NC. The focus will be on sequences of \(3 \leq n \leq 8\) report units per utterance. The issue at stake is illustrated in (1), an utterance containing a sequence of 3 report units \((n = 3)\) (the tags MVT, MDD, and MDF label different types of discourse presentation):

(1) S1 And the police came \[^{MVT}\text{and interviewed me about it.}\] That was a white Austin Metro. \[^{MDD}\text{Don't think it's the same young woman.}\] He said \[^{MDF}\text{oh the car's been broken into [unclear]}\]

(KC9-N2)

The results of the analysis show a statistically significant trend towards immediatization: narrators sequence report units progressing from less immediate to more immediate types, thus building up what could be called utterance climax. Specifically, the increased immediacy is achieved largely by means of increased use of the two types affording the greatest immediacy, Direct and Free Direct.

The findings have far-reaching implications on two planes. First, I discuss the contribution made by the immediatization trend to the creation of what Longacre (1983) referred to as ‘tension’, which can be seen as the driving force underlying the climacto-telic structure of narrative as a discourse type. Second, the systematic patterning in sequences of report units suggests the possibility that what Sinclair (1991) referred to as the ‘idiom principle’ – the co-selection of lexical items – may, in a similar fashion, operate on the higher level of, not lexical, but discourse units.

**References**


The elimination of homography in thirteen editions of The Kalender of Shepherdes

HANNA RUTKOWSKA (Adam Mickiewicz University)

This is a paper summarising the findings of a corpus-based, qualitative and quantitative comparative study which traces the process of the elimination of homography in thirteen editions of The Kalender of Shepherdes, a comprehensive compendium of prose and verse texts on a variety of subjects, e.g. astronomy, medicine, and religion, published between 1506 and 1656. The corpus contains over 0.9 million words, and constitutes a database of transcriptions prepared by the present author, and based on the facsimiles available at Early English Books Online. The analysed editions include those printed by R. Pynson (1506, STC 22408), J. Notary (c. 1518, STC 22410), W. de Worde (1528, STC 22411), W. Powell (1556, STC 22412), J. Wally (c. 1570, STC 22415; c. 1580, STC 22416; c. 1585, STC 22416.5), T. Adams (c. 1600, STC 22419; 1604, STC 22420; 1611, STC 22421; 1618, STC 22422) J. Wright (1631, STC 22423), and R. Ibbitson (1656, Wing B713).

Orthographic distinction between homophones, together with the indication of morphological spelling, vowel length, and the establishment of etymological spelling, are considered the most important variables to be taken into consideration in the analysis of the process of orthographic regularisation and standardisation in Early Modern English (Salmon 1999). A close comparative examination of the editions which are included in the corpus indicates that most printers under consideration made conscious efforts to lower the level of homography in their orthographic systems, distinguishing between homophones by means of different spellings. The process was gradual, but with regard to native words homography seems to have reached the level comparable to Present-Day English by the 1650s.

The printers’ usage has been confronted with the recommendations of sixteenth- and seventeenth-century lexicographers, grammarians and spelling reformers, including e.g. Huloet (1552), Hart (1569), Huloet-Higgins (1572), Mulcaster (1582), Clement (1587), Coote (1596), Cawdry (1604), Butler (1633), Hodges (1644), and Wharton (1654). This confrontation has revealed that language authorities’ opinions could not have triggered the tendency to eliminate homography among the printers, because this trend had started before their writings were published. However, their publications could have supported the already advanced process of spelling regularisation.

References

Butler, C. 1633. The English grammar, or the institution of letters, syllables, and words, in the English tongue whereunto is annexed an index of words like and unlike. Oxford: Printed by William Turner, for the author.

Cawdry, R. 1604. A table alphabetcall conteyning and teaching the true writing, and vnderstanding of hard vsuall English wordes..., London: Edmund Weaver.


ICE-East Africa was among the first ICE corpora to be compiled in the early 1990s. Almost 20 years after ICE-EA1 (the first ICE corpus from ESL/EIL nations) and 50 years after Kenya’s and Tanzania’s independence, it may be time to add a diachronic dimension to ICE, esp. ICE-EA. This contribution discusses whether it is possible to combine compatibility (so that we can compare ICE-EA1 with ICE-EA2) and adaptation to culture and new technologies (so that internet communication is adequately represented in what can be called ICE-EA2.0).

The main focus of the discussion is on methodology, esp. text-type selection and new internet retrieval methods. I will demonstrate options of downloading data from newspapers from Tanzania and Kenya and from Twitter. Despite new options to gather quick-and-dirty (monitor) corpora from internet media, we should not neglect the sociolinguistic issues of net identities or personas. However, diachronic ICE corpora may also be fruitful to test New English theories, because they can be used to show empirically how ‘dynamic’ such models really are. I will also show the limitations of corpus-linguistic methodologies and complement some tests with web-questionnaires and other means. Thus ICE-EA2.0 may also serve as a model for new and ongoing ICE compilations.
Re-examining the levelling of present perfect and simple past in World Englishes

ELENA SEOANE (University of Vigo) & CRISTINA SUÁREZ-GÓMEZ (University of the Balearic Islands)

This paper is concerned with the levelling of the difference between present perfect and simple past. This is one of the vernacular universals mentioned in Chambers (2004), is included in the list by Kortmann & Schneider (2004), and is a frequently attested feature (in 59% of the 78 varieties analyzed) in the recently launched Electronic World Atlas of the Varieties of English (Kortmann & Lunkenheimer 2011). Previous research has often noted the use of the simple past for the expression of perfect meaning in general (as in So what do you think of the case which I told you just now <ICE-SIN:S1A-037>) (Elsness 2009: 229, Hundt & Smith 2009), but studies have not gone on to examine the fact that only certain types of perfect meaning are favoured by this construction, while for the expression of other perfect meanings the canonical present perfect is still the dominant form (cf. Seoane and Suárez 2013). In this paper we will show that the levelling of the present perfect and the preterite is far from a generalized development but is constrained by semantic and contextual factors. Another interesting issue in need of further clarification which we will address is the use of the present perfect for the expression of an action that is completed (You know yesterday I have seen some two guys speaking with you <ICE-IND:S1A-049>).

Drawing on data from the ICE corpora available for World Englishes (Hong Kong, India, Jamaica, Kenya, The Philippines, Singapore, and Tanzania) and using the ICE-GB and the available section of the ICE-USA as benchmark corpora, we intend to examine the distribution of present perfect and simple past forms in combination with adverbials that make their time frame explicit: ago, last (+ month, week, time) and yesterday as markers of finished actions, and already, ever, never, just, recently and yet as indicators of an action that started in the past and extends to the present. We will be looking at the following issues: (i) implications of the fact that the input language is an earlier spoken variety of English, and in many cases dates from a period in which the grammaticalization of the functional distinction between the present perfect and the simple past had not taken place (Rissanen 1999: 224-225); (ii) potential influence from the superstratum (American vs British English); (iii) the typological nature of the substratum (inflecting, isolating, agglutinating) and its influence on the present perfect vs simple past choice; (iv) the import of the phase of development of each variety, following Schneider’s (2007) dynamic approach; (v) the impact of mode (spoken vs written); (vi) and, finally, the influence of text-type (academic vs popular writing, spontaneous vs scripted spoken, etc.). Examination of these factors should yield a clear picture of the determinants of present perfect and simple past levelling in World Englishes.

References


An enhanced vista on recent change in standard English: Findings from BLOB-1901 and the updated Brown family of corpora

NICHOLAS SMITH (University of Salford)

In the last two decades, the study of recent and current change in standard English, formerly a Cinderella subject in the history of the language, has enjoyed a surge in popularity. Particularly at the level of morphosyntax, a spate of studies has shown dramatic changes in the profile of structures and functions in e.g. the noun phrase, the progressive and the modals (cf. Hundt and Mair 1999, Biber and Clark 2002, Leech et al. 2009). These accounts are helping to update descriptions of grammar in the recent history of English, and to facilitate our understanding of social and internal processes of language change. Fuelling these investigations are corpora such as ARCHER (Biber et al. 1994), members of the Brown family (Hundt and Mair 1999, Leech and Smith 2005, Baker 2008), and the massive Corpus of Historical American English, COHA (Davies 2010).

Despite these developments, it is surprisingly difficult to track linguistic progress at the start of the 20th century, particularly in British English. Important research questions that still need answering include the following:

a) What was the course of the breakdown of the prescriptive shall/will rule in standard British English? For example, can we corroborate Denison’s (1998:167-8) assertion that the rule (especially the prohibition against use of predictive will with first person subjects) noticeably wanes around the end of the 19th/start of the 20th century?

b) Can we verify the claim by Biber and Clark (2002) that the expansion of noun sequences (e.g. export officer, control freak) was much more pronounced in the latter than the earlier stages of the 20th century?

This paper seeks to address such questions by utilizing a new 500,000-word (ultimately 1m-word) corpus of British English. BLOB-1901 is a turn-of-the-century ‘prequel’ to the 1930s BLOB-1931 corpus, as well as to LOB (1961), FLOB (1991), BE06 (2006), and their American equivalents Brown (1961), Frown (1991) and AmE06 (2006). Our approach has been to match these corpora as precisely as possible in sampling frame and size, such that
the Brown family now comprises one of the largest continuous corpus records of British English (BrE) spanning the last hundred years. As with the other Brown family members, we are conducting detailed qualitative and quantitative analysis of linguistic structures in BLOB-1901, with the aim of providing a coherent, century-wide picture of different types and rates of grammatical change.

Our approach is to not to focus exclusively on the Brown family corpora, but rather to compare the results, where possible, with other corpora such as COHA and ARCHER. It will be argued that despite its size disadvantage against COHA, and its narrower period coverage than ARCHER, the expanded Brown family holds up reassuringly well for a wide range of grammatical phenomena. Moreover, used in conjunction with these other corpora, it can help shed important light on parallel, converging and diverging paths of development in BrE and AmE.

References


The Middle English development of OE ē and ēo: Spelling evidence

Gjertrud F. Stenbrenden (University of Oslo)

This paper examines the Middle English (ME) developments of Old English (OE) ē and ēo. According to the handbooks, OE ē raised to [i:] in the so-called ‘Great Vowel Shift’ (GVS) in the early 15th century. OE ēo, on the other hand, has a rather more complicated history: the
The diphthong is supposed to have monophthongised to a front rounded vowel [ø:] in all dialects in late OE, spelt variously <eo>, <oe>, <u>, <ue>, etc.; this vowel then unrounded to [e:] in the 12th century in the North, the East, and the South Midlands, but remained until the 14th century in the West Midlands, when it was unrounded there also; this [e:] fell in with etymological ê and therefore participated in the ‘GVS’ later on (Wright and Wright 1928, Luick 1914-40, Jordan 1968). Hence, the PDE reflex is /i:/.

The evidence has been abstracted from three ME corpora which collectively cover the period c. 1150-1500 (i.e. LAEME, LALME, SMED); the corpora have made use of texts in local language; the language of each text has, moreover, been precisely localised. The paper aims (1) to discuss the types of evidence exploited, and the kinds of problems which attend the interpretation of orthographical material in terms of phonetic correspondences, (2) to assess whether the ME spellings in fact support the assumed developments of OE ê and êo, (3) to date and determine the chronology of the changes to OE ê and êo, with special reference to the ‘early vowel-shift hypothesis’ (Stockwell and Minkova 1988), and (4) to identify the dialects which stand out as loci of change. Recent findings and claims by Welna (2004), Lass and Laing (2005), and Stenroos (2005) are also investigated.

References


*LAEME*, see Laing, M. 2008.


*LALME*, see McIntosh, Samuels, Benskin et al. 1986.


In a previous small-scale and unpublished study of the use of ‘taboo’ and ‘non-taboo’ vocatives by Madrid and London teenagers, I found that both groups preferred the non-taboo to the taboo vocatives, and, surprisingly enough, that the English teenagers’ use of the taboo vocatives was slightly more frequent than that of the Spanish teenagers. In this paper, I widen the scope to cover a somewhat larger number of words, which I now prefer to call ‘rude’ words, considering that available definitions, which alternate between ‘vulgar’, ‘insulting’, ‘offensive’ and ‘informal’, do not mention ‘taboo’, despite the reference of the words to taboo objects. The study includes Spanish cabrón/a (‘bastard’), coño (‘cunt’), gilipollas (‘dickhead’), hijo/a (de) puta (‘son/daughter of a whore’), maricón (‘homosexual’) and puto/a (‘whore’) and English asshole, bastard, bitch, cow, cunt, dickhead, motherfucker, son of a bitch and wanker. What looks like a difference in number, the English words being more numerous than the Spanish, is an inevitable effect of the Spanish grammatical system, which gives rise to a male and female variant of some of the words.

The study is based on the Corpus Oral de Lenguaje Adolescente de Madrid (COLAm) and The Bergen Corpus of London Teenage Language (COLT), which are largely comparable, COLAm being collected with COLT as a model in order to facilitate comparative studies. My approach is typically corpus-linguistic, involving (here mainly) qualitative research on a quantitative basis. In addition, the sociolinguistic aspect is given due consideration with respect to gender and age and also to the speakers’ sociolinguistic background. Before the study proper, one chapter is devoted to labels – from ‘taboo’ to ‘rude’ – and the relationship with swearing and slang, with reference to Andersson & Trudgill (1990), Green (2000), Jay (1992), Ljung (2006), Sanmartín Sáez (2003) and Stenström (2006).
Mateo & Yus (2000), who discuss insulting taboo words in terms of ‘communicative devices’, and distinguish between ‘offense-centred’, ‘praise-centred’ and ‘interaction-centred’ insults, argue that interaction-centred insults are used to reinforce the social bonds between speaker and hearer rather than serve as insults. This is exactly the expected effect of the rude words used as address terms in teenage talk that I discuss in this paper: what seems to be insulting expressions, such as hijo de puta and motherfucker, are expected to be used as friendly communicative devices which ease the contact between speaker and hearer, acting as ‘vocativos cariñosos’ (Rodríguez 2002: 48).

References


COLT : <http://www.hd.uib.no/colt>

COLAm: <http://www.colam.org>


Moving towards multimodality: Implications for corpus linguistics of a study of medical articles

CAROLINE TAGG & MEL EVANS (University of Birmingham)

The significance of multimodality in human communication is now recognised across linguistic disciplines as diverse as stylistics, second language acquisition, and sociolinguistics. Within English corpus linguistics, a similar recognition is evident in recent developments in spoken corpora such as the CHILDES corpus and the University of Nottingham’s multimodal corpus. However, despite the prominence of the ‘multimodal turn’, corpus studies of written language continue to rely primarily on monomodal text, due in part to the practical difficulties involved in incorporating other modes. In this talk, we raise questions as to how corpus-driven, written language research can successfully account for a range of key meaning-making resources, such as typography, layout, space, colour, texture, image and embedded video and audio. Drawing on illustrative examples from our own work with genre-specific corpora ranging from sixteenth-century English manuscript letters to text message and social network site interactions, we highlight why these features are often excluded from analysis – and what is lost because of these decisions.
Our current project seeks to fill this gap in corpus linguistics by exploring the significance of different forms of data representation in a corpus of academic medical journal articles. The project is part of an ongoing collaborative research initiative between the Centre for Corpus Research at the University of Birmingham and Elsevier, one of the world's largest academic journal publishers. Elsevier's extensive catalogue of research materials provides an opportunity to conduct research into academic medical discourse, with outcomes anticipated to, firstly, raise awareness of reader engagement and practices with multimodal discourse within the medical academic community; secondly, offer new insights for Elsevier into the effect of their new interactive online interface and its impact on reader engagement with data representation in medical research articles; and thirdly, advance theories and methods used in corpus analysis. In particular, our study addresses the research question “how is Elsevier’s new online interface impacting on the ways in which readers process and interpret data representations (figures, tables and other images) in research articles?” The research comprises a number of stages or phases, including the initial identification and description of the representation of data in medical journal articles, interviews with authors and editors, eye-tracking experiments with readers, and finally the automation of the process of initial data analysis and the development of a multimodal corpus. As well as presenting the results of our initial analysis of the data, we wish to raise and discuss some of the theoretical and practical issues involved in incorporating multimodal and contextual dimensions into a corpus-based analysis, as well as accounting for reader response and processing. By engaging with these issues, this talk flags up an important and timely new direction for a field ‘on the move’.

**Sir, here’s a woman Ø would speake with you: Zero subject relativizers in Early and Late Modern English**

**GUNNEL TOTTIE (University of Zurich) & CHRISTINE JOHANSSON (Uppsala University)**

Although zero subject relativizers (ZSRs) are frequent and even sometimes de rigueur in many Present-Day regional dialects, they are dismissed by grammarians as being “of doubtful acceptability” or “slovenly” (Quirk et al 1985: 1250) or used only in “marginally non-standard usage” in “some conversational varieties” in standard PDE (Biber et al. 1999: 619), appearing mostly in a restricted range of contexts. However, as shown in a study by Lehmann (2002), which takes account of all instances where ZSRs could occur in the British National Corpus demographic sample, they account for 13% of the total of possible cases (205/1376), especially in existential *there*-constructions and clefts. ZSRs are thus still in fairly frequent use in contemporary spoken British English.

Earlier work (e.g. Erdmann 1980, Dekeyser 1986) traces the decline of ZSRs in the Early Modern English period, but no studies appear to have followed the principle of accountability. We study the use of all subject relative markers in Early and Late Modern English based on A Corpus of English Dialogues (CED) covering the period 1560-1760 and A Corpus of Nineteenth-Century English (CONCE) for the period 1800-1900, accounting for all cases where ZSR can occur in variation with surface relativizers. We limit our study to spoken-type varieties of the language (drama, i.e. comedy, and trials) and can show that the frequency of ZSRs appears to have been similar in those periods to that in contemporary English. We focus on presentative constructions, mostly existentials and clefts, where full accountability is possible, as in (1) and (2):

(1) …*there was* another party Ø was much against it…(CED3 Trials, 1640–1679)

(2) …*there was* another party *that* was much against it…

(3) …*there was* another party *who* was much against it…
(2) …it was only Berry Ø denied that he did know… (CED3 Trials, 1640–1679)
…it was only Berry that denied that he did know…
…it was only Berry who denied that he did know…

We will, however, also discuss the very few other examples where total accountability was not possible, as e.g. I haue a Priest Ø will mumble vp a marriage, (CED2 Comedy, Lording Barrey, Ram-Alley, 1611)

We also seek to establish whether the factors governing the use of zero constructions versus explicit relativizers are similar to those prevailing in Present-Day English. Examples are coded for factors that have been shown to be of importance in the study of PDE: the number and category of the antecedent in the matrix clause, the distance between antecedent and relativizer, modality of the relative clause as well as genre (drama or trials). We also consider the importance of socio-economic factors; although age data were not available, class and gender of speakers are included.

References


In absolute detail: The development of English absolutes from adverbial to additional-context marker

NIKKI VAN DE POL & HUBERT CUYCKENS (K.U. Leuven)

The absolute construction (AC) is a non-finite construction consisting of two principal elements: a (pro)nominal head and a predicate. There are two subtypes: the unaugmented AC, as in (1), and the augmented AC, which is introduced by a preposition called 'augmentor', as in (2). In Present-day English (PDE) augmentation typically makes use of with(out).

(1) So within, your very new born sleeping all the time, its body temperature can begin to drop without you knowing. (BNC)
(2) With my form filled in, I phoned. (BNC)
(3) & ymne acwxdene eodun ut on ocelebearwes dune. (Rushworth, Matthew, 26.30)
'And when they had sung a hymn (lit. 'the hymn sung'), they went out to the Mount of Olives.'
While the AC is still commonly described as expressing the same range of semantic relationships with its matrix clause as finite adverbial subordinate clauses (e.g. cause (1), anteriority (2, 3), concession etc.) and is hence considered on a par with adverbials (Berent 1973: 147, Kisbye 1972: 72), it appears that expression of these adverbial relations by ACs has become secondary in PDE. Indeed, ACs appear to have gradually assumed the function of marker of ‘any additional contextual information’ (4). Indicative of this change are ACs that lack a single, clear adverbial interpretation and/or simply form an elaboration on the matrix clause.

(4) The animals are expensive, with the average price of a goat running about 135 dollars. (Wordbanks Online)

The purpose of this paper is to examine the shift in the AC's function from adverbial to additional-context marker. First, it will detail the AC's changing functional distribution. Second, it will try to identify, and account for, the factors favoring the AC's interpretation as ‘additional-context marker’. Possible factors are:

1. the position of the AC with respect to the matrix clause (initial, medial, or final): while initial position may be said to encode accommodate (temporal, causal) background information, final position may be preferred for encoding local, additional information (see also Kortmann 1991: 169);
2. augmentation vs. non-augmentation: any positive correlation between with-augmentation and the AC's function as 'additional-context' marker may be explained by with's bleaching (or generalization) from 'accompaniment' meaning to a mere marker of circumstantial information;
3. (partial) co-reference between the AC's subject and the matrix subject: to the extent that ACs provide additional context to the information in the main clause, co-reference relations between the AC's subject and the matrix may increase.

Third, it is argued that the AC's increasing use as additional-context marker represents a shift on the subordination-coordination cline, in particular from an integrated (subordinate) structure to a quasi-coordinate structure (paraphrasable by and or whereby). Interestingly, Killie & Swan’s (2009) study on the development of participial -ing clauses shows the opposite trend (from a quasi-coordinate -ing status to an integrated, subordinate status).

Corpora used include the Helsinki Corpus, the Penn parsed corpora, COHA, Wordbanks Online and the BNC corpus.

References


Correlative constructions in English: The \( ha \ldots ha \) construction

ANS VAN KEMENADE & META LINKS (Radboud University Nijmegen)

In this paper we will explore the history of one type of correlative construction in English. This type of construction found in earlier English (and also in present-day Dutch) is typically introduced by a verb-final adverbial sub-clause, followed by a main clause with the order resumptive adverb - finite verb - subject - ..., i.e. a Verb-Second (V2) clause. An example is given in (1), with the temporal adverb \( ha \), but the construction can also be readily introduced by \( honne, mid ha \), and conditional clauses introduced by \( gif \).

(1) Old English correlative construction with \( ha \ldots ha \)

\[
\begin{align*}
& ha \he \da to him cwom, \text{ } ha \text{ } wæs \text{ } he \text{ } forht \text{ } geworden. \\
& \text{(cobede, Bede_2:9.128.17.1222)}
\end{align*}
\]

When he then to him came, then was he fearful become "When he then came to him, he had become fearful."

Van Kemenade and Los (2006) argue that in constructions such as (1), the first clause serves to locate the context in time or discourse, while the second clause relates the following event. The discourse properties of these constructions are thus reflected in their correlative syntactic form. This paper is organized as follows: we have two general background hypotheses: 1) Old English was a discourse-configurational language, as reflected in part in its V2 character and OV character, and 2) this discourse-configurational character was ‘promoted’ by the availability of the paradigms of the demonstrative pronoun and deictic adverbs (van Kemenade 2009). Against this general background, we analyse the use of this type of correlative construction and trace its development over the Old and Middle English periods. The loss of demonstrative pronouns and deictic adverbs (including \( then \) (van Kemenade 2009)) in early Middle English, and the loss of V2 over the late Middle English and early Modern English periods form the backdrop for the loss of this construction. We use a qualitative and quantitative corpus-based approach to trace the changes in the referentiality and use of deictic/temporal adverb and pronouns, and the use of V2. Using different corpora (YCOE, PPCEME2, PPCEME) we trace the history of this construction and show how its discourse function was taken over by other constructions over time. In addition, we also look at the variation (for example the optionalness of the resumptive adverb) and function of constructions like (1) in present-day Dutch using the CGN (Corpus of Spoken Dutch) and compare these to the findings for (earlier) English. The results confirm that the present-day Dutch system (and by hypothesis, the Old English system) is strongly related to the presence of V2 and its concomitant pragmatics. This suggests that the decline of this type of correlative constructions in English goes hand-in-hand with the collapse of the paradigms of demonstrative pronouns and deictic adverbs, and with the loss of V2.

References


Taboo language and swearing in 18th and 19th century English. A diachronic study based on the Old Bailey Corpus

BIANCA WIDLITZKI & MAGNUS HUBER (Justus Liebig University Giessen)

This paper investigates the use and representation of taboo language and swearing in 18th and 19th century spoken English. Although several studies have considered taboo/swearing in the history of English (e.g. Hughes 1991, 2006, McEnery 2006), many focus on providing a broad historical sketch on the topic and/or limit the discussion to written language. By contrast, our study looks at spoken language and takes a qualitative and quantitative corpus linguistic approach to the subject.

Because of its sheer size and detail of utterance-level annotation, the Old Bailey Corpus (OBC, <http://www.uni-giessen.de/oldbaileycorpus> lends itself ideally to the study of this subject. The OBC is based on the Proceedings of the Old Bailey, <http://www.oldbaileyonline.org>, London's central criminal court, whose verbatim passages are a reasonably close representation of what was said in the courtroom. The OBC spans the years 1720-1913 and contains 14 million words of spoken English. Every single utterance has a detailed mark-up for sociolinguistic (sex, social class, age), pragmatic (role in the court proceeding) and textual variables (the shorthand scribe, printer and publisher of individual Proceedings).

A first analysis of the OBC showed that, as in other genres of the time, many taboo and swear words were considered too offensive for publication and were therefore disguised (e.g. through abbreviation or by substituting part of the word with dashes) or replaced entirely (e.g. by dashes or asterisks). An inventory of taboo and swear words in the OBC was compiled by searching for these concealment strategies and by manually looking for other swear words in their vicinity. What is considered taboo and swearing is a relative matter and varies historically. Our list will therefore be analyzed for

1. changes in the categories: which expressions became more or less offensive over time?

2. changes in the sociolinguistics of swearing: were there gender and/or social class differences and how did they change over the two centuries? Other aspects relevant here are how the use of offensive language was treated in the courtroom (verbal reprimands, fines) and what strategies speakers used to report offensive language.
3. changes in the representation of taboo words/swearing. As mentioned above, several approaches can be found in the Proceedings’ publication history:

- verbatim representation: *I'll be damn'd if I do not know that man.* (18120115)
- partial deletion: *G—D— you for a B—h, if you follow me I'll stab you this Minute.* (17270705)
- complete deletion: *I called her — and — which made her so angry [...]* (17340227)

Studying the use of these techniques offers further insight into language attitudes and changing standards of acceptability in British society at the beginning of the PDE period.

The beginning of the period considered here saw the emergence of a discourse of purity that linked swearing to particular social groups, especially men belonging to the lower social classes (McEnery 2006: 167). It will therefore be interesting to see whether the OBC reflects this development of swearing and its representation in print and which variables account for the variation observed.

References


**Developments in the use of the English present perfect: 1750-present**

**XINYUE YAO** (University of New South Wales)

Corpus-based studies on Late Modern English syntax have long been focused on such well-known developments as the rise of the progressive and the *get*-passive, the regulation of the periphrastic *do*, changes in verb complementation patterns, etc. (Aarts et al. forthcoming, Mugglestone 2006, Tieken-Boon van Ostade & van der Wurff 2009). Less attention has been paid to developments concerning the present perfect construction (with the auxiliary *have*), whose substantive change from a resultative construction of the type *I have my work finished* is known to have been completed before the Late Modern English period (Carey 1994). Using multi-genre corpora, Elsness (1997) found that the present perfect has been losing ground to the preterite in American English from 1750 onwards, and such tendency has also begun to appear in British English.

The present study further explores the nature of the purported frequency decline of the present perfect using *A Representative Corpus of Historical English Registers* (ARCHER, version 3.2). To attain a higher degree of generic homogeneity, dialogic texts were drawn from the drama and fiction sections of the corpus. Tokens of present perfects were extracted and a number of contextual features were analysed and coded. A variable rule analysis was performed to examine changes in the linguistic constraints on speakers’ choice of the present perfect over the preterite.
The analysis reveals a number of significant changes in the co-occurrence patterns of the present perfect with other linguistic features, including the overall increase of temporal specifiers, in particular temporal quantifiers such as always, never and ever, decrease of verbs with direct results such as break, kill, lose, find, arrive and leave, increase of atelic and intransitive verbs, and of negative forms. The results of the variable rule analysis confirm the relevance of these changes to the choice of the present perfect over the preterite in both British and American English. It is suggested that these findings reflect a gradual functional shift of the English present perfect from expressing resultative meaning to experiential meaning: the construction is more frequently used to direct attention to the occurrence of a past situation as opposed to its present result. It is further argued that this functional shift is in line with the grammaticalisation path of ‘resultative > anterior > perfective’ that has been attested in other Indo-European languages such as French and German (Bybee & Dahl 1989, Bybee et al. 1994). However, unlike its counterparts in these languages, the English present perfect shows signs of becoming a marked perfective rather than an unmarked perfective.

References


SOFTWARE
DEMONSTRATIONS
Extending Dependency Bank with statistical measures

HANS MARTIN LEHMAN & GEROLD SCHNEIDER (University of Zurich)

In our software presentation, we will provide an overview of the state of a dependency bank framework which offers dependency searches of syntactically parsed corpora as well as regular expression queries on the original texts (Lehmann and Schneider 2012). We present a two-pronged approach permitting the direct comparison of the two retrieval strategies and the optimization of recall and/or precision. Figure 1 shows a sample result of a dependency query that specifies the verb *depend* with an attached subject and a PP with the preposition *on*.

![Dependency Analysis of BNC:A0G:2385](image)

We will then focus on new features providing statistical tests for distributions and cross-tabulations of internal and external factors.

In addition we will also discuss advances in the area of measures of surprise, where we are currently implementing and testing measures based on the distribution of lexical items in specific syntactic structures.

We will demonstrate the new features and discuss possible linguistic applications. They encompass a wide variety of phenomena at the lexis-grammar interface, such as collocations, lexical preferences and idioms, pattern and construction grammar, lexicalization phenomena, the exploration of grammatical gradience and diachronic changes.

The investigation of gradience involving lexical items in interactions requires large amounts of data. Our web-based retrieval software scales up beyond the level of for example the British National Corpus (Aston & Burnard 1998).

References


An innovative web interface for the 14-million-word *Old Bailey Corpus, 1720-1913*

**MAGNUS NISSEL & MAGNUS HUBER** (Justus Liebig University Giessen)

This software demonstration will give an overview of the functionality of the web-based search engine that has been developed for the *Old Bailey Corpus* (OBC, <http://www.uni-giessen.de/oldbaileycorpus>). The OBC is based on the *Proceedings of the Old Bailey* (<http://www.oldbaileyonline.org>), London's central criminal court. The verbatim passages of the *Proceedings* are arguably as close as we can get to what was actually said in the courtroom. The OBC spans the years 1720-1913 and contains ca. 750,000 spoken words per decade (320,000 utterances, 14 million words in total). Every single utterance has a detailed mark-up for sociolinguistic (sex, social class, age), pragmatic (speaker role in the court proceeding) and textual variables (the shorthand scribe, printer and publisher of individual *Proceedings*).

The combination of sheer size and rich annotation make the OBC a valuable resource for multivariate analyses in historical linguistics in general, and historical sociolinguistics and pragmatics in particular. One priority in creating the OBC was free access and user-friendliness, so that the creation of a customized online tool became the logical choice. However, size and annotation put considerable demands on the search interface and underlying architecture: All passages identified as spoken language were stored in a MySQL database alongside the indexed annotation data. The user interface uses JavaScript to facilitate the dynamic retrieval and processing of search results.

The OBC web interface (*OBCdb*) allows researchers to use any combination of the extralinguistic variables mentioned above to filter the results for a given search term. More importantly, each concordance line includes all available annotation data, a major advantage over standard corpus linguistic software:

<table>
<thead>
<tr>
<th>Left</th>
<th>Key</th>
<th>Right</th>
<th>Year</th>
<th>Sex/Class</th>
<th>Sex</th>
<th>Lingo</th>
<th>Age</th>
<th>Mark-Up</th>
<th>Shorthand</th>
<th>Publisher</th>
<th>Printer</th>
</tr>
</thead>
<tbody>
<tr>
<td>In ...</td>
<td>time</td>
<td>he . ...</td>
<td>1763</td>
<td>m</td>
<td>lower clerical and sales personnel</td>
<td>m</td>
<td>worker</td>
<td>null</td>
<td>Thomas Garnes</td>
<td>J. Scott</td>
<td>J. Scott</td>
</tr>
<tr>
<td>for my master.</td>
<td>when</td>
<td>after ...</td>
<td>1763</td>
<td>m</td>
<td>null</td>
<td>m</td>
<td>witness</td>
<td>null</td>
<td>Thomas Garnes</td>
<td>J. Scott</td>
<td>J. Scott</td>
</tr>
<tr>
<td>six years before</td>
<td>time</td>
<td>...</td>
<td>1763</td>
<td>m</td>
<td>higher professionals</td>
<td>m</td>
<td>witness</td>
<td>null</td>
<td>Thomas Garnes</td>
<td>J. Scott</td>
<td>J. Scott</td>
</tr>
<tr>
<td>up at supper, when</td>
<td>time</td>
<td>after ...</td>
<td>1763</td>
<td>m</td>
<td>lower professionals, clerical and sales</td>
<td>m</td>
<td>witness</td>
<td>Brackenbury, General</td>
<td>Thomas Garnes</td>
<td>J. Scott</td>
<td>J. Scott</td>
</tr>
<tr>
<td>the deceased.</td>
<td>who</td>
<td>...</td>
<td>1763</td>
<td>m</td>
<td>medium-skilled workers</td>
<td>m</td>
<td>witness</td>
<td>Shoemaker, General</td>
<td>Thomas Garnes</td>
<td>J. Scott</td>
<td>J. Scott</td>
</tr>
<tr>
<td>d卒 on me, after</td>
<td>time</td>
<td>to meet</td>
<td>1763</td>
<td>m</td>
<td>medium-skilled workers</td>
<td>m</td>
<td>witness</td>
<td>Baker, General</td>
<td>Thomas Garnes</td>
<td>J. Scott</td>
<td>J. Scott</td>
</tr>
<tr>
<td>down one on, after</td>
<td>time</td>
<td>from me</td>
<td>1763</td>
<td>m</td>
<td>lower-skilled workers</td>
<td>m</td>
<td>domestic servant, general</td>
<td>Thomas Garnes</td>
<td>J. Scott</td>
<td>J. Scott</td>
<td></td>
</tr>
</tbody>
</table>

These search results can be easily exported into spreadsheet software for further filtering and processing.

Regarding the search term itself, OBCdb allows for several levels of complexity. Searches can be performed on the plain text or a part-of-speech tagged version of the data (*CLAWS 7* tagset) and it is possible to suppress the tags in the concordance output. There are options for simple text and wildcard searches, but regular expressions are supported as well. For the latter, OBCdb offers a unique feature in the form of an assistant designed to allow users with little knowledge of regular expressions to create elaborate search terms by selecting and combining predefined elements geared towards linguistic research.

OBCdb considerably facilitates variationist corpus linguistic studies of spoken English in the 18th and 19th centuries. It is hoped that this online tool will encourage the exploration of a period that has been the Cinderella of historical linguistics so far.
Enlarging the stock of English diachronic resources for deep corpus analysis: A diachronic lexical database

HAGEN PEUKERT (University of Hamburg)

The annotation of large corpora is usually restricted to syntactic structure or word class. Pure lexical information and information on the structure of words are stored in specialized dictionaries (e.g. CELEX, Baayen, et al., 1995). Both data structures – dictionary and text corpus – can be matched to get e.g. a distribution of certain lexical information from a text. This procedure works fine for synchronic corpora of Present-Day English. What is lacking, however, is a diachronic lexical database that allows for the matching with large corpora dependent on time, thus involving entries of Middle English or Early Modern English. The OED even contains this information, but unfortunately in an unstructured format that cannot be used for automatic extractions useful to corpus linguists.

The Morphilo Toolset aims at building a representative diachronic database of English. The software consists of three components: MorExtractor, Morphilizer and MorQuery. MorExtractor uses a quite simple algorithm that dependent on the given word class and a rule set identifies the structure of the word and assigns lexical tags to it (e.g. /root or /prefix). The identification process is based on enumerated lists comprising all prefix and suffix allomorphs listed in the OED. After inputting a tagged text from a specific time, MorExtractor produces a text file, in which the structure of all words is annotated. Since the algorithm “overgeneralizes”, the text file has to be checked for wrong annotations. This tedious task is supported by the Morphilizer component. It takes each of the text files and its time specification as an input, displays the word structure in a template and allows the user to make adjustments in a comfortable way by click and drop. Each word item, its structure and its token frequency that were checked manually are written to the Morphilo-Database. For the given time frame of the text, each word type has to be processed only once. At last, MorQuery provides a comfortable search of the database. Each combination of morphemes, allomorphs, compounds, word types, time frames or corpora can be chosen from drop-down menus. It is also possible to make selections of the most frequent queries or directly type SQL commands to the prompt.

The database will be available to the linguistic community from a website. All researchers are encouraged to query the data, but also to contribute to the project by having their own diachronic corpora read in and analyzed. Since the database will have a large stock on entries by its inception, the workload for post-processing using Morphilizer for each additional new corpus will be evanescently little.

References

WORK-IN-PROGRESS REPORTS
English urban vernaculars, 1400-1700

ANITA AUER & MORAGH GORDON (Utrecht University)

This paper presents the recently launched research project “Emerging Standards: Urbanisation and the Development of Standard English, c. 1400-1700”, which aims at illuminating the complex processes involved in the emergence and development of Standard English. While the focus was traditionally on the pre-eminent urban community – London –, this project explores an alternative history of language standardisation in England by shifting the focus to regional centres. More precisely, urban vernaculars of major regional centres with high levels of literacy and text production (manuscripts as well as printed texts from 1476 onwards) will be systematically investigated over the period 1400-1700 with respect to factors such as time, text type, and migration patterns. The study will focus on the vernaculars of York (North), Bristol (Southwest), Coventry (West Midlands), and Norwich (East Anglia), which represent different Middle English dialect areas. A comparison of the results of these longitudinal studies of urban vernaculars (as well as findings from London) is expected to clarify our understanding of the origin and spread of formal written English.

In this paper we want to introduce this project in historical urban dialectology and particularly focus on the empirical data, i.e. primarily the newly compiled corpus (incl. its urban sub-corpora) for the period 1400-1700, on which the study will largely be based. We will be concerned with the text types contained in the corpus as well as the conversion of the material into XML format.

The next thing Mudiyanse knew was that he was given a slap across his cheek: Frequency, types and genre-sensitivity of light-verb constructions in Sri Lankan English

TOBIAS BERNAISCH (Justus Liebig University Giessen)

Sri Lankan English is a South Asian second-language variety displaying characteristic structural features on the levels of phonetics and phonology (Meyler 2007), morphology (Senaratne 2009), syntax (Herat 2005, 2006) and semantics (Werner & Mukherjee 2012). However, given that, among South Asian Englishes, most scholarly attention has been given to the usage-based description of Indian English (cf. e.g. Sedlatschek 2009, Schilk 2011, Lange 2012), many publications on SLE have so far remained “largely impressionistic accounts not supported by representative samples of speakers [...] or by corpus data that reflect syntactic and grammatical language in use across a range of genres” (Mendis & Rambukwella 2010: 186). In the light of this general lack of empirical insights into Sri Lankan English, the present study approaches Sri Lankan English lexicogrammar, i.e. the structural level where variety-specific structures of postcolonial Englishes have been described to surface first (cf. Schneider 2007: 46), from a corpus-based angle.

Light-verb constructions, i.e. the combination of a semantically ‘light’ verb, a determiner and a deverbal noun isomorphic with the respective simplex verb (cf. Hoffmann et al. 2011: 267), represent one feature located at the interface between lexis and grammar. (1) exemplifies a light-verb construction in Sri Lankan English.
(1) and the next thing Mudiyanse knew was that he was given a slap across his cheek <ICE-SL:W2F-003#74:1>

Earlier studies have indicated that light-verb constructions exhibit regional cross-varietal differences in relation to their frequency of occurrence (cf. Hoffmann et al. 2011: 270), their internal structure (cf. Algeo 2006: 270) and their degree of genre-sensitivity (cf. Dixon 2005: 461). Still, light-verb constructions in Sri Lankan English have so far either been studied against a language-pedagogical background (cf. Kumara & Gunawardhana forthcoming) or on the basis of data to some extent biased towards Indian English (cf. Hoffmann et al. 2011). The study at hand contrastively examines light-verb constructions in Sri Lankan English, Indian English, which has been argued to exert epicentral influences on its neighbouring varieties including Sri Lankan English (cf. Leitner 1992: 225), and British English, the historical input variety of Sri Lankan English. The light verbs give, have, put and take serve as lexical anchors to analyse the respective light-verb constructions in corpus data. For each of the varieties, text material from the International Corpus of English and from large-scale newspaper corpora is consulted to provide insights into a) the frequencies of light-verb constructions, b) their internal structure as regards light verbs and determiners and c) their genre-sensitivity in the varieties concerned to establish whether and to what extent light-verb constructions add to the variety-specific lexicogrammatical profile of Sri Lankan English.

The present study shows that light-verb constructions occur most frequently in Indian English followed by Sri Lankan English and British English. While take clearly is the most frequent light verb in Indian English, have also figures prominently in relative terms in Sri Lankan English and British English. As regards determiner realisation, British English favours indefinite articles, zero articles are attested most frequently in Indian English and there is no clear-cut preference in Sri Lankan English between indefinite and zero articles. The association between light-verb constructions and informal contexts seems to hold in British English only. Thus, the present study delineates the variety-specific structural profile of Sri Lankan English light-verb constructions on the basis of authentic corpus data.

References


Towards an automatic annotation of discourse presentation

BEATRIX BUSSE (University of Heidelberg) & MICHAELA MAHLBERG (University of Nottingham)

It has often been claimed that highly discursive phenomena, such as discourse presentation and its interplay with narration, are too idiosyncratic or narrator-oriented to identify their different modes with the help of automatic annotation (see discussion in Semino & Short 2004). Toolan (2009) addresses the question of creating a tagger for FIT (free indirect thought) presentation. He proposes an inductive procedure and outlines a set of rules for semi-automatic searches. Toolan’s (2009) suggestions are based on corpus findings. In a similar vein, the present paper will outline textual features that can help isolate speech presentation categories as a basis for semi-automatic searches. We will draw on findings mainly from Busse (2010) and Mahlberg (2012) and focus on patterns that are identifiable in a corpus of 19th-century narrative fiction. The set of form-based diagnostics that we suggest deals, for instance, with the type of reporting verbs that are typically found with direct speech and thought. Verbs in reporting clauses further seem to appear in patterns together with body language. Eye movement, for instance, can function to indicate thought presentation. Further formal features associated with speech and thought presentation categories relate to the positions in the narrative where reporting clauses appear and whether they preceed, follow or interrupt character speech and thought. As a work-in-progress report, this paper presents the diagnostic categories that we have retrieved on the basis of corpus data and that will form the basis for semi-automatic search procedures that are in the process of being developed.
On the use of the split infinitive in the Asian varieties of English

JAVIER CALLE-MARTÍN & JESÚS ROMERO-BARRANCO (University of Málaga)

A split infinitive construction denotes a particular type of syntactic tmesis in which a word or phrase, especially an adverb, occurs between the infinitive marker to and the infinitive of the verb. The earliest instances of the split infinitive in English date back to the 13th century, in which a personal pronoun, an adverb or two or more words could appear in such environments (Visser 1984, II: 1038-1045). Its use drops drastically throughout the 16th and the 17th century, but it begins to gain ground again in the 18th century, resisting the severe criticisms of grammarians from the first half of the 19th century (Calle-Martín and Miranda-García 2009: 347-364, Perales-Escudero 2011: 316-319).

In view of the late modern English development of this construction, this paper investigates the use of the split infinitive in the Asian varieties of English with the following objectives: a) to compare the distribution of the phenomenon in British English and the emerging varieties of Asian Englishes; b) to analyse the nature of the splitting adverb in terms of its typology and lexico-grammatical features; c) to ascertain the distinct bigrams and trigrams in each variety to determine whether oral registers influence the emergence of splitting combinations in written patterns; and d) to analyse if there is any underlying rhetorical pattern(s) fostering the use of particular splitting constructions. The corpus used as source of evidence is the International Corpus of English, both the British English and the Asian English components (i.e. Hong Kong, India, Singapore, Sri Lanka and The Philippines).

References


Keywords and clusters: Can they help a translator? A case study in Irving’s *A Widow for One Year* and its Czech translation

ANNA ČERMÁKOVÁ (Institute of the Czech National Corpus, Charles University in Prague)

There has been a growing interest in corpus based contrastive multilingual research in recent years, with a number of studies clearly showing the potential benefits of the resources, methodologies and tools used in corpus linguistics for translation studies (e.g. Baker 1993, Kenny 2005, Mahlberg 2007, Teubert 2004, 2007). The research in this field points especially to the benefits of parallel and comparable corpora (e.g. Johansson 2007). The present case study of Irving’s novel and its Czech translation is a work-in-progress and applies some of the approaches used in corpus stylistics (e.g. Mahlberg 2012, Fischer-Starcke 2010) from a contrastive perspective. This case study aims to explore whether identifying textual keywords, clusters (as understood in corpus linguistics (e.g. Scott 1996)) and local textual functions (Mahlberg 2007) in a literary text, could be a helpful step in the translation process of the literary work in question.

This study is a follow-up to a previous research on keywords in a literary work and their translation equivalents in two languages, Finnish and Czech (Čermáková and Farova 2010), which opened up space for further research, as it showed that with some of the keywords the translator may benefit from being aware of their keyness and keeping the translation equivalents consistent. The present case study takes this a step further by introducing a cluster analysis, which is one of the methods used in corpus stylistics by Mahlberg (2012) to discover local textual functions. Mahlberg (2007) suggests in her article that translators can benefit from an awareness of subtle lexical relationships and patterns that are difficult or even impossible to retrieve on the basis of intuition or a single text. Mahlberg (2007: 120-121) works with the concept of ‘local textual functions’ that contribute to the characterization of a text and of which the human observer may not be always aware. This is the point where the translator may benefit from corpus linguistics tools and methods.

The text used for the present case study is John Irving’s 1998 novel *A Widow for One Year* and its translation to Czech by Milada Novakova. The novel and its translation are available through the project *InterCorp*. Keywords and clusters are generated by WordSmith Tools and as a reference corpus for generating the keyword list the BNC frequency list has been used.

The study will look in detail at some of the keywords and clusters generated by WordSmith and other repetitive patterns found in their proximity and analyze how the translator works with them. The study aims to show how the translator of a literary text would benefit from pre-processing the text with methods suggested above.

References


InterCorp (Czech National Corpus – InterCorp). Institute of the Czech National Corpus FF UK, Prague. Available at <http://www.korpus.cz>


Cause-and-effect in history writing from the Old English to the Early Modern English period

CLAUDIA CLARIDGE (University of Duisburg-Essen)

The register of history writing is one with an almost continuous representation in the history of the English language, yet it has not received much linguistic attention in either historical or modern discourse studies. One exception is Coffin (2006), who takes an educational perspective on the topic. Using some of her approaches and findings I will conduct a case study of historical writing from Old to Early Modern English.

Coffin distinguishes recording, explaining and arguing genres within history writing, which include (different kinds of) causality to various degrees. While explaining genres are based on causality and arguing genres explicitly make cause-and-effect the object of a debate, recording genres could at least potentially do without any overt causality. Annalistic or chronicle writing (e.g. Anglo-Saxon Chronicle, Capgrave’s Chronicle) can be seen as fairly prototypical instances of the recording genres and thus might contain no or very little causality. However, Coffin distinguishes simple ‘recounts’ from ‘accounts’, the latter supplying more than simple chronology. Thus, there may be interesting diversity in chronicles. Non-annalistic texts, such as Bede’s Ecclesiastical History or More’s History of Richard III, have explanatory aims and thus can be supposed to contain more overt causality. The frequency development of causal items within historical works over time may help to chart genre developments in these texts.
Cause-and-effect constructions in history writing either serve to represent how events are (seen to be) connected in the external world or to provide the chronicler’s deductions about the historical significance of events (text-internal). The former type is hypothesized to dominate, in parts or even exclusively, in early history writing. It is also supposed that external causality in chronicles will be mostly based on human actions and the linking of concrete/specific events; if more abstract explanation (factors, consequences) should be found at all, it will only be in non-annalistic and in general later works.

Causality can be realised in many different forms: conjunctions (e.g. because), conjuncts (therefore), circumstantial adverbials (through his help), verbs (lead to), and nominalizations (reason). Additionally, it can be combined with evaluation as in a major effect, primarily because of. With the help of the Thesaurus of Old English and the Historical Thesaurus of English (OED), the inventory of cause-and-effect items used in the historical works will be established and their development over time charted. The hypothesis is that conjunctions and circumstantial adverbials will be the most frequent realisations, with nouns being the least common. Combined cause-evaluation will probably be rare and come in late.

The basis for the investigation are the historical sections of the Helsinki Corpus, which will be both adapted and expanded (e.g. by removing strongly literary or biblical treatments, by using longer passages from the texts already included, and by adding new texts where possible). The other learned or academic prose sections of the Helsinki Corpus will be used for frequency comparisons.

References


Corpus pragmatics across genres: Commitments in emails and spoken language

RACHELE DE FELICE, JILL BOWIE & SEAN WALLIS (University College London)

Within the fast-growing discipline of corpus pragmatics (Romero-Trillo 2008, Jucker et al. 2009, Rühlemann 2010), a range of tools can be used to analyse the lexicon, phraseology, and syntax of speech act categories, yielding a data-driven description of phenomena which have so far tended to be based mostly on contrived, invented examples. For instance, De Felice’s (2013) study presents a detailed discussion of the structure of commitments in business English emails, analysing in particular their vocabulary and phrasing. The current paper proposes to test the wider applicability of this approach by extending the analysis to spoken business English data as well, in particular ten business transactions from the DCPSE (the Diachronic Corpus of Present-Day Spoken English, a corpus of parsed spoken English from the late 1960s to the 1990s), from which instances of commitments will be manually identified. This is the first step towards a more comprehensive evaluation of how well a classification scheme of speech act categories designed for email communication can be applied to a corpus of spoken language.

The research questions to be addressed are:

a) Are the categories of commitments used in De Felice (2013) sufficient to express the range of moves found in spoken language?
b) What linguistic resources are used in different genres to encode similar discourse moves?

c) How ‘portable’ are the corpus analysis techniques developed for one type of resource within the same domain?

In particular, the analytical methodology examines the occurrence of particular linguistic features (lexical items, phrases, noun categories such as expressions of time, verb patterns) and establishes correspondences between different linguistic patterns and the discourse and pragmatic functions they fulfil in the workplace, distinguishing in particular between transactional and relational goals (cf. e.g. Koester 2004a, b). In this paper, we aim to show whether the same set of functions is found in spoken communication, and whether these are encoded using the same range of linguistic resources. As well as contributing to our understanding of this little researched speech act, this work demonstrates how practical applications of corpus pragmatics can be re-used and extended for a variety of resources.

References


Nouns and verbs in advanced Dutch EFL writing: A quantitative and qualitative longitudinal study

PIETER DE HAAN (Radboud University Nijmegen)

Dutch university students of English are generally considered to be very proficient users of EFL. However, recent studies (de Haan & van der Haagen, forthcoming; Verheijen de Haan & Los, forthcoming) have shown that their writing products are less sophisticated than those of their native English-speaking contemporaries. Subtle matters such as the unidiomatic use of intensified adjectives and adverbs or the overuse of marked pre-subject constituents can be pointed to as markers of non-sophisticated English language use.

At the same time, it is becoming increasingly clear that groups of L2 learners, also at advanced levels, cannot be regarded as homogeneous groups, but as dynamic systems, whose members show idiosyncratic development patterns (cf. Dörnyei & Murphey 2003). While groups of students are shown to develop over time towards more native-like EFL use, individual members
of these groups are seen to develop in ways that are sometimes very different from the group pattern. These individual development patterns can even be observed in the use of such seemingly straightforward word classes as nouns and verbs.

The current paper reports on a study of a small number of individual essays written by a recent cohort of students of English (2011) at Radboud University Nijmegen. A large number of student essays were collected during the first eighteen months of their university studies. These essays were PoS tagged, in order to establish general word use patterns. In a second stage, the essays were parsed using the Stanford Parser (Klein and Manning 2003). The essays of a number of randomly selected individual students were then subjected to a detailed study of noun phrase and verb phrase use. These findings were compared to the occurrence of other signs of sophisticated language use, such as the unmarked occurrence of pre-subject sentence elements, or the advanced use of adjective intensifiers.

The preliminary results suggest that students who are more aware of native usage tend to develop towards native usage themselves faster than those who are less aware of this. Writing classes should focus more on creating student awareness of typically native English usage, so that more students will be able to reach a genuinely near-native level of written production.

References


Motion to and motion through: Evidence from a multilingual corpus

THOMAS EGAN (Hedmark University College)

Ever since Talmy (1991) formulated the distinction between path- and satellite-framed languages, there has been a lot of comparative research into the coding of motion (see, for instance, the papers in Hickmann and Robert 2006). In comparative research into the ways in which languages encode motion events, the tertium comparationis has often taken the form of events in a picture book (such as the frog story; see Berman & Slobin 1994) or in short video snippets, which are subject to oral descriptions by participants in the experiment. The advent of multilingual translation corpora allows us access to a different sort of tertium comparationis, where expressions in a source language serve as grounds for the comparison of their translations into two other languages.

In this paper I take a fresh look at the satellite- and path-framed typology in a comparative study of English and French translations of Norwegian predications of motion events containing two path prepositions, til (= to) and gjennom (= through). The former codes the end-point (goal) of a path, while the latter may denote the crossing of a boundary. The data for the study come from
the *Oslo Multilingual Corpus* (OMC). The source items translated into both languages may be viewed as prompts to which the translators respond, much in the manner of the participants in laboratory experiments.

All tokens of the two Norwegian prepositions in the OMC were classified semantically and all motion tokens were extracted manually. The Norwegian originals were then set aside and comparisons drawn between the English and French renderings of the various meanings. One point that should be made about the data is that Norwegian is a satellite-framed language like English but unlike French. Moreover, all the original Norwegian tokens contain a path preposition. Note, however, that Slobin (2006: 70) claims that “in translations […] manner salience follows patterns of the target, rather than source language”. Thus, the fact that Norwegian is satellite-framed should not be of any great consequence for the results of the comparison, at least as far as coding of manner is concerned. The results of the investigation point to substantial differences between the two languages, with English conforming by and large to the satellite-framed type. The results for French are much less clear, with the picture resembling more the kind of typological hybrid described by Kopecka (2006: 97) than a rigid path-framed type.

**References**


**Expletive subjects in L2 English: Can we reset the parameter with corpus data?**

**ESTHER FERRANDIS & AMAYA MENDIKOETXEA** (Autonomous University of Madrid)

One of the main contributions of learner corpora in Second Language Acquisition research is that they provide a much wider empirical base than has previously been available. SLA studies are often conducted on a very limited number of subjects, which, as pointed out by Granger (2002: 6), raises questions about the generalizability of results. This study aims at contributing to the debate about the nature of transfer or cross-linguistic interference by focusing on the acquisition of overt expletive subjects (*it* and *there*) in L1-Spanish – L2 English grammars. English and Spanish differ as to the setting of the Null Subject Parameter. Following previous corpus and experimental work on Verb-Subject constructions in L2 English, we analysed the
role that the L1 (Spanish), a [+ pro-drop] language, has in the acquisition of overt expletives (it, there) in L2 English, a [-pro-drop] language. Experimental work in the L2 literature has shown that full resetting of the Null Subject Parameter involves acquisition of expletive subjects in L2 English (Ruiz de Zarobe 1986). A corpus study was carried out using the WriCLE corpus, a written corpus of academic essays with 700,000 words written by L1 Spanish learners of L2 English and compiled at the Universidad Autónoma de Madrid (Rollinson & Mendikoetxea 2010). 75 texts from 75 different university students of English Studies and English Philology with different proficiency levels were randomly selected and manually annotated using the software UAM Corpus Tool (version 2.7.2) (O’Donnell 2008). Our three main research questions were: (RQ1) Are overt expletives of English (it, there) problematic for Spanish learners? If so, are they problematic at all proficiency levels? (RQ2) Do Spanish natives initially transfer their L1 parameter value (use of Ø expletive) when acquiring English as a second language? (RQ3) If they do so, are learners able to reset completely their L1 parameter [+ pro-drop] to the English one [-pro-drop]? In order to answer these research questions an annotation scheme was designed intended to account for the referentiality of the subject, the type of predicate it appeared with, the word order it appeared in and its overall grammaticality or ungrammaticality. A total of 681 expletive subjects were found in the texts selected for the study (expletives there, it and use of Ø expletive) and the results found partially confirmed RQ1 and RQ2, as only expletive it was problematic in all levels, whereas expletive there was not. However, all groups used Ø expletive, which shows that Spanish learners tend to transfer their L1 parameter value. As for RQ3, we did not expect a full acquisition of this parameter and thus not a full resetting of their L1 setting. This hypothesis was confirmed by our results, which indicate that though Spanish learners of L2 English acquire referential subjects early in the acquisition process, they never fully master the use of expletive subjects, not even in advanced stages.

References


“Alcohol consumption” vs. “popping in for a pint”: A study of discourse on alcohol in parliamentary debate

GREGORY GARRETSON (Uppsala University)

Drinking is simultaneously a central part of most Western cultures and a serious source of problems. In the UK, "having a pint with friends" is for many a cornerstone of social activity, and yet being "under the influence of alcohol" is considered one of the nation's top public-health problems. In Sweden, there is a long history of strict governmental control of alcohol, stemming in part from a strong tradition of anti-alcohol activism (Johansson 2008). Drinking is therefore much talked about in these countries, using strikingly varied language, depending on context and speaker goals. But while talk about alcohol has been studied in various ways in e.g. medicine, anthropology and sociology, these varying discourses about alcohol have hardly been studied by linguists, and never systematically.

This paper presents the results of a pilot study working toward a large-scale comparative study of public discourse on alcohol in English (in Britain) and in Swedish (in Sweden). Since these two countries have much in common but also exhibit substantial differences where attitudes toward alcohol are concerned, such a comparison proves to be highly illuminating.

The study reported is a four-way comparative corpus study: General corpora of English (the BNC) and Swedish (PAROLE) are contrasted with two specially compiled corpora of parliamentary discourse, drawn from public records of debates in the UK Parliament and the Swedish Riksdag, yielding 80 million and 50 million words, respectively, or ten years' worth of transcripts of spoken English and Swedish (a not unproblematic type of data, as will be discussed). In all four corpora, speech events where alcohol is the topic of discussion were identified and analyzed, using a multi-pronged approach involving analysis of keywords (e.g. Baker 2004), collocations (e.g. Stubbs 2001), lexical bundles (e.g. Biber 2009), and recurrent discourse patterns (e.g. Teubert 2010).

The goal of the analysis is to answer the following questions: What characterizes public discourse on alcohol in parliamentary discourse as compared to general corpora? In parliamentary discourse, how do debates on alcohol differ from other debates? And how do the answers to these questions vary between a British context and a Swedish context? The results show quite clearly that patterns of discourse regarding drinking in parliamentary debates differ markedly from those in (most) other contexts, and that there are interesting differences between the British and the Swedish contexts, suggesting different cultural conceptions of the role of alcohol in society.

This study is the beginning of a larger project examining many different genres in which drinking is discussed, in order to survey the types of public discourse of alcohol in Britain and Sweden (with the ultimate goal of adding several other countries). The comparative nature of the study is seen as crucial, as the patterns of discourse that emerge as typical in one context – and in one country – can contrast starkly with those in other contexts, ultimately revealing the different ways that the many conceptualizations of alcohol (as a cultural treasure, a pastime, a consumer good, a health risk, etc.) find expression in language.

References

When embarking on a project studying the sociolinguistic parameters of spoken language, one of the decisions to be made at the outset concerns the approach to take to transcription. While several transcription conventions are well-established (for a discussion see e.g. O'Connell & Kowal 1999), one of the main limitations of existing schemes is that not all features can be marked-up precisely in computer-readable documents (e.g. square brackets and spacing to indicate overlaps and interruptions).

Consequently, when we started a project based on the British TV show *Mock the Week*, we decided to use a different methodology in the transcription process. *Mock the Week* is a weekly 30-minute comedy panel show which was first broadcast by the BBC in 2005. It usually involves two teams of three panelists who compete against one another in different tasks, comprising quick answer rounds, stand-up comedy routines and improvisation. The data, which had to be newly transcribed in the course of this project, therefore include dialogic and interactive as well as monologic passages.

The particular focus of our study is on interruptions and overlaps in the interactive parts of *Mock the Week*, paying specific attention to the gender variable. This is mainly because we would like to show whether there is empirical evidence supporting stand-up comedian Jo Brand's view that women “won't get a word in edgeways” and “don't like competing for air time” on panel shows (Brand 2009). Thus, we focus on the study of an interactive speech feature from a sociolinguistic perspective but we would also like our data to be reusable for other types of investigations in future studies.

Therefore, we decided to create an XML-tagged corpus of series five of *Mock the Week* (54,000 words), using sociopragmatic XML tagging on a turn-by-turn basis when transcribing our data. This way, we ensured that we had an easily expandable tag set, allowing the tagging of further linguistic features in the corpus, which was also transparent, making the data machine readable.

In this paper we will introduce the turn-by-turn XML annotation system used, which allows the easy extraction of the total number of words (and turns) spoken by each panel member – crucial in the study of gender. We will show how overlaps and interruptions can be marked-up precisely through the use of XML tags and how the length of an overlap can be measured with reference to the number of words. Finally, we will demonstrate how turn-by-turn tagging allows both the participant interrupting and the participant being interrupted (overlapping and overlapped) to be taken into account so as to gain further insights into the role gender may play in these occurrences.
References


Complement extraposition: What is it a diagnostic of?


Complementation constructions divide into two basic types depending on whether the reported speaker or the actual speaker is represented as committed to the modal position (in a broad sense) of the complement clause. The first type is reported speech or thought, illustrated by (1). In (1), the reported speakers, doctors, are represented as committed to the truth of the proposition that “breast reconstruction hindered detection of cancer recurrence”. This proposition is not subscribed to by the actual speaker, as signalled by erroneously.

(1) doctors erroneously believed that breast reconstruction delays detection of local recurrence of the cancer. (WB)

The second type of complement has been characterized in the literature as having ‘fact’ status according to the actual speaker (Kiparsky & Kiparsky 1971, Halliday 1994), irrespective of the represented speaker’s attitude, which may, as in (2), be rejection.

(2) Saddam did not accept that he lost the 1991 Gulf war (WB)

Whilst the difference is not overtly coded in (1) and (2), the Kiparskys (1971) claimed that factive status does have a number of overt markers, such as the fact and it in front of the complement. It is the second alleged diagnostic of factivity, illustrated in (3), that we will subject to systematic qualitative and quantitative corpus investigation in this paper.

(3) the priest ... ’s not going to like it that we already have Jessica. (WB)

Exhaustive extractions will be made of the it + that-clause structure, called complement extraposition by Bolinger (1972), from WordbanksOnline and the British National Corpus.

We will in fact challenge the hypothesis that complement extraposition correlates systematically and exclusively with factive complements. It is also used in reported speech, when the actual speaker stresses the fact that the locution was really uttered, as in (4).

(4) the newspaper ... said it that they didn’t want us. (WB)

When used in factive contexts, there is the added complication that some examples contain claims to which the actual speaker is not necessarily committed, but a represented speaker (Delacruz 1976), who may not even be the subject of the factive predicate, as in (5). Example (5) also shows that the commitment may be to a deontic (should), rather than an epistemic, position.
Our data study will thus allow us to reconceptualize the distinction between reported speech and factive constructions as hinging on the following opposition. Reported speech depicts locations that originate in the speech event described by the matrix clause, whereas factive complements designate pre-existing propositions “being manipulated, evaluated and commented on” (Langacker 1991:35) by the matrix verb.

Having deconstructed complement extraposition into a diagnostic of finer distinctions than just factivity, we will, finally, verify the more general hypothesis that complement extraposition signals textual ‘givenness’ (De Cuba & Urögdi 2010), which we will check for factive and non-factive complements as well as complements of semi-fixed expressions such as logic/word/etc. has it, which also occur in our dataset.

References


To have a glimpse at other nations… Register specificity of light verb constructions in advanced EFL writing

CAROLINE GERCKENS (Leibniz University Hannover & Université Catholique de Louvain) & ARIAN SHAHROKHY-PREHN (Leibniz University Hannover)

The following paper deals with the question of how far advanced learners of English are aware of phraseological units and use them in their written output. Our research is focussed on so-called light verb constructions (LVCs), as e.g. make a claim, have a look, take a walk, that consist of a high frequency verb, the indefinite article and a deverbal noun. Our data for this analysis is extracted from LONGDALE-GE, i.e. the German part of the Longitudinal Database of Learner English. As this database consists of authentic texts of different types (e.g. personal narrative, academic essay) by advanced learners of English, it is also possible to check whether these students are making genre-specific choices regarding their usage of these units, as “phraseological patterns may differ across genres and disciplines” (Gilquin et al. 2007: 321).
Without a doubt, idiomatic – and even more so schematic / semi-idiomatic expressions – are an essential part of language learning since they imbue the learner with a naturalness in communication that would otherwise be missing. The ‘Kerncurriculum’ (core curriculum) for the German county of Lower Saxony (2006: 19) stresses that the lexical units taught in grades 5 to 10 should include phrasal verbs, idioms, collocations, grammatical terms, etc, making it clear that scientific research in these areas has direct ramifications for language learners and teachers. However, the implications of the ubiquitous LVCs on foreign language teaching have been left largely unexplored, the only exception being Stein who states:

The question of whether the V+N construction is a mere expansion of the simple V element in a sentence and thus generally substitutable for it or whether it has a function of its own is of immediate relevance to the teaching of English as a foreign language. If the construction has a specific meaning, the structural pattern and its meaning will have to be taught. (Stein 1991: 2-3; emphasis added)

With regards to EFL teaching it is important to note that, apart from their structural and semantic patterns, LVCs are considered to be register specific, i.e. they are supposed to be absent from academic prose (Leech 2009: 170-171, Wierzbicka 1982: 757). Our research will show that there is considerable variance in the use of LVCs by advanced learners of English and that these non-native novice writers are mostly unaware of the genre-specificity of the constructions.

References


Wierzbicka, A. 1982. Why can you *have a drink* when you can't *have an eat? Language* 58: 753-799.

**Phraseological patterns of anticipatory *it* with extraposed fused relatives**

**SARA GESUATO** (University of Padua)

Clausal end-focus in English can be created through a discontinuous construction in which the subject/object slot is occupied by the pronoun *it*, and its meaning conveyed in an extraposed clause after an adjectival/noun/verb phrase (Biber et al. 1999: 155, 1019-1020; “*IT* can be seen they were working”; “I find *IT* is a bother they behave like this”).

The construction is not usually available when the extraposed element is an NP (* “*IT* was hard my job”; * “I found *IT* hard my job”*), unless this has embedded post-modification, typically a
restrictive relative clause. In this case, the head of the NP and the relative pronoun may be
distinct elements (Huddleston & Pullum 2002: 1407-1408; "IT's amazing the number of
German theologians that sided with Hitler"; Kaltenbock 1999: 61) or fused together in a wh-
word (Biber et al. 1999: 660, 684; "IT was known what to expect"; Luelsdorff 1981: 196; my
emphasis).

On the basis of BNC concordances, this paper examines the syntactic-semantic patterns of the
anticipatory-it construction, when the predicate contains an AP and the extraposed element is
headed by what, who or the+noun.

The most frequent of the three variants is that with what (355 instances; 3.2 pmw). Mostly
attested in written sources (82%), it often occurs with the verb BE in the present tense (66%; It
is very exciting what is happening in these countries, A28). The adjectives preceding what (53
lexemes) are mostly positively connoted (94%) and express such concepts as ‘evidence/certainty/truth’ (63%; obvious), ‘emotional impact’ (28%; amazing), ‘assessment’
(5%; extraordinary) and ‘value’ (3%; important). The events represented identify broad
notions: actions (do), experiences (happen), verbal exchanges (say), situations (be) and
opinions (think), the most frequent verb being do (15%). The what tends to play the role of the
direct object (73%; it is not clear what British firms in America receive in return, A55).

The extraposed elements headed by who or the+noun are infrequent (17 and 17). The former
mostly correlates with the adjective clear (15) in the Adj_Who_subj_VP pattern (16; It wasn't
clear who was comforting whom, F9X). The latter typically shows up in the Adj_the
way_relative-clause pattern (9; It is absolutely brilliant the way they do it, KE1).

Overall, the variants of the construction examined appear to specialize in the encoding of
evidentiality, epistemic modality and evaluation, the matrix clause expressing the speaker's
(present and positive) stance and the deferred unit a reported event. The discontinuous
realization of the syntactic constituent is probably licensed by its co-occurrence with evaluative
predicative adjectives - which generally trigger the anticipatory-it use (Hunston & Sinclair
2000) – and favoured by the prominent clausal expansion of the extraposed element, which
makes it comparable to clauses.

References

and written English. London: Longman.

Spoken and written English grammar and usage. Cambridge: Cambridge University Press.

Cambridge: Cambridge University Press.


Papers (VIEWS) 8/2: 48-71.

ICAME data and CLARIN, a progress report

KNUT HOFLAND (Uni Research AS, Bergen)

This presentation will give an overview of the process for including the ICAME material within CLARIN, including models for licences, the implementation of systems for federated authentication and authorization and online handling of licences. The intellectual property rights have to be clarified and we need to get acceptance from the holders to include the material in CLARIN and to select an acceptable licence. We need to input metadata to the CLARIN metadata catalogue and to convert the old data formats to the format used by our corpus search program Corpuscle (Meurer 2012).

Since the foundation of ICAME in 1977, ICAME material has been distributed by the Norwegian Computing Centre for the Humanities (NCCH) in Bergen and by its successors. The NCCH was financed by the Norwegian Research Council and later transferred to a department of Uni Research, the research company at the University of Bergen. Through the years about 1800 orders of material have been distributed on magnetic tape, microfiche, diskettes and CD-ROM. With no basic funding today and a reduced income of the sale of material, the administration in 2010 decided that Uni Research could no longer subsidise the ICAME activities. This was reported to the ICAME Advisory Board. The Board has since then investigated how the ICAME activities can continue, especially within the CLARIN project.

CLARIN (Common Language Resources and Technology Infrastructure) is an EU project where the goal is “to make language resources and technology much more accessible to all researchers working with language material, in particular in the humanities and social sciences. CLARIN is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities”. After a preparatory phase of 3 years the construction phase started in 2011 for 5 years. After this there will be an exploitation phase of at least 5 years. The construction phase is financed by the participating countries. The partners in CLARIN come from 26 countries. In 2012 CLARIN became a European Research Infrastructure Consortium (ERIC).

In 2012 a Norwegian consortium of universities and other institutions, led by Koenraad De Smedt from University of Bergen, got resources from the Norwegian Research Council for a 5-year national CLARIN project, called CLARINO. Uni Research is a partner in this project and ICAME material is one of the resources we plan to include.

References

CLARIN: <http://www.clarin.eu/external/>

ICAME: <http://icame.uib.no/>

**Kind-of and sort-of: Pragmatic discourse markers in the SPICE-Ireland Corpus**

JOHN KIRK (Belfast)

Interpretation of pragmatic discourse markers kind-of/sort-of (the research question) depends on recognition of the two megafunctions which each performs: evidential and textual, or affective and interpersonal (the approach). The use of these items is bound up with the degree of consciousness of the speaker to the communication in question and towards his choice of expression (part of the demonstrable results). The accommodation of each other’s face is part of everyday socialization, and these needs and their fulfillment come about through the use of pragmatic discourse markers such as kind-of/sort-of in our everyday speech. The examples to be discussed from the SPICE-Ireland Corpus (the data) will show that, as evidentials, they downtone infelicitous expression, engender vagueness and mitigate uncertainty (results). As affectives, they enable speakers to negotiate their addressee’s as well as their own face, positive as well as negative, enhancing and upholding as well as minimizing or counteracting threat (more results). As the examples will further show, these items are bound up with lexical (especially collocational), syntactic, and prosodic features (part of the method). Across the corpus’s 15 discourse situations, the SPICE-Ireland annotation scheme is without precedent in its guidance towards an unraveling of pragmatic intent (another part of the research question).

References


**Tracing the grammaticalization of I think**

DANIELA KOLBE-HANNA (University of Trier)

The grammaticalization and pragmatic functions of comment clauses have been studied widely (e.g., Brinton 2008, Kaltenböck 2008, Stenström 1995). In particular, a lot of attention has been paid to *I think* and its grammaticalization (Brinton 1996: 212-254, Thompson & Mulac 1991, Fischer 2007a: 297-311, 2007b, van Bogaert 2011). This paper aims to verify the existing accounts of the grammaticalization of *I think* by tracing it back to its Old English origins.

Thompson and Mulac (1991) maintain that the use of *I think* as comment clause, or, in their terms ‘epistemic parenthetical’, results from the omission of the complementizer *that*, as exemplified in (1) and (2). This omission obscures the relationship between matrix and embedded clause and allows the reanalysis of the embedded clause as main clause and of *I think* to be syntactically more flexible, so that it can occur as parenthetical, as in (3).

(1) *I think* that I was wrong
(2) *I think* I was wrong

---

191
I was wrong, *I think*

Alternatively, Brinton (1996: 212-254; 2008: 45) argues that parenthetical *I think* derives from parenthetical *as I think* on the basis of Old and Middle English data. Lopéz-Couso and Méndez-Naya (2012) observe a similar derivation of *methinks*. The examples discussed by Brinton (1996: 212-254), however, do not include any instance of *as I think* (respectively its OE and ME variants). *As I think* does occur in a quotation in the OED, cited here as (4), as an example of “adding a comment to a statement or fact” (OED Online, ‘think, v.2’), yet right after a quotation containing an Old English parenthetical use of *think* without *as* (resp. OE *swa*), cited here as (5).

(4) Yit hadde he bot o mannes del Toward himself, so as *I thinke*, Of clothinge and of mete and drinke. (Gower, Confessio Amantis, 1393)

(5) anno 1076Rogcer […] gaderade his folic ðan cnyngce to unþearfe – he þohte – ac hit wearð heom seolfan to mycclan hearme. (Anglo-Saxon Chronicle)

“This raises the question whether (5) is a unique exception or a challenge to the hypothesis that parenthetical *as*-clauses were preliminary to *I think*. This paper aims to answer this question by examining whether *think* was used in parenthetical clauses without *as* in Old and Middle English, drawing on data from the Corpus of Early English Correspondence (CEEC), the Helsinki Corpus and the Toronto Corpus. It also reassesses Thompson and Mulac’s (1991) claim that the comment clause *I think* derives from the identical matrix clause. As parenthetical clauses with *think* have not been discussed in previous research, they are expected to be rare and thus not to lend themselves to quantitative statistical analyses. This paper therefore relies on manual inspection of the data, qualitative analysis and descriptive statistics. In conclusion it will provide insight into the relationship between linguistic theory and linguistic reality.

References


Ongoing changes and global Englishes: Exploring new corpus resources

Mikko Laitinen (Linnaeus University)

There has recently been plenty of interest in recent and ongoing grammatical changes in English (e.g. Leech et al. 2009). Using evidence of a range of structures, these studies have identified various long-term trends, such as colloquialization and ‘Americanization’ in the written standard English of the late twentieth century. Digital resources in the form of the Brown corpora have provided a sound empirical basis for this kind of work because they free an observer from impressionistic and anecdotal data and enable taking the stylistic variation into account (Mair 2008).

Most of this work on recent changes has focused only on American and British English. This presentation shifts the focus to global Englishes, viz. English as a lingua franca (ELF) and other L2-varieties of English (Seidlhofer 2011, Mauranen 2012), and asks how these current grammatical changes are reflected and adopted in use in the present-day global Englishes. Such an approach is motivated because the number of speakers of English globally far exceeds the number of native speakers. There has recently been emerging interest in the changes in the other varieties (cf. Filppula (2012) on the auxiliaries in the outer circle varieties; Laitinen (2011) on patterns of verbal complementation in lingua franca Englishes).

This presentation approaches recent grammatical changes in global Englishes through the existing corpus resources. It highlights the need for developing and collecting new materials that reflect the global spread of English. The existing (global) English corpora are, for various reasons, not suitable for investigating recent and ongoing changes. For instance, the existing learner corpora are limited to few text types, and the corpus design in the ICE corpora requires heavy modifications for the global context. The two ELF corpora only contain spoken language, and relying solely on spoken evidence may lead to incomplete conclusions on variability because it is often necessary to compare spoken and written data side-by-side for determining the direction of change (cf. Mair 2007). This presentation discusses the corpus design and compilation principles of today’s written ELF corpora. The guiding principle in this work is that the corpus design should be suitable for a range of geographic contexts in which English is emerging or has emerged as a communicative medium. I will discuss the collection process of two databases that represent English uses in Sweden and in Finland. These databases are currently in compilation, with the aim that they would provide good practices for developing a new generation of global English text corpora. I will then use these two databases to provide evidence of adverbial connectives of ANY/EACH/EVERY TIME in the sense of ‘whenever’. These connectives have undergone grammaticalization in the native varieties (Brinton 2007), and I will illustrate how this development is reflected in global Englishes.
Information structure in the development of discourse markers and comment clauses

DIANA M. LEWIS (Aix-Marseille University)

This paper investigates the relationship between discourse marking and information structure, and examines the hypothesis of a developmental tendency from discourse marking towards information structure marking.

The place of information structure in the historical development and in the present-day usage of so-called discourse markers and comment clauses has yet to be well understood. Information structure can be relevant to so-called discourse markers and comment clauses in several ways. One is that information structure pressures can lead indirectly to the emergence of discourse-marking functions in an expression, for example by favouring a particular word order that leaves the expression in a relevant position that facilitates its acquiring discourse-marking implicatures. This is often related to scope expansion of the expression. Another is that once a discourse marker has become established as such, its place in its host construction often leads to its acquiring information-structuring functions, for instance, to mark the content of the clause it attaches to as more or less salient than an adjacent one. This foregrounding or backgrounding role can become a main function.

Cases of diachronic development from medial or final VP-adverb to sentence adverb and discourse marker in initial, often parenthetical, position are well known and documented. Following work on other languages where final particles are common, there has recently been
more attention focused on final-position markers in English (e.g. Barth-Weingarten and Couper-Kuhlen 2002, Haselow 2011, 2012, Kim and Jahnke 2011, Mulder and Thompson 2008), with an emphasis on subjectivity of meaning. Positional flexibility is a feature of some markers, with some expressions developing different uses in initial and final positions (e.g. then, anyway, of course, rather, after all, like, etc.).

Frequency data are drawn from historical and present-day English corpora to trace recent diachronic developments (Modern English to Present-Day English) in two expressions that come to be used in both initial and final positions with similar meanings across positions – discourse marker in fact and comment clause mind (you) – to argue that final position has become associated with backgrounding of information and that this has largely motivated the initial-final split of the markers in question. In fact, having lexicalized as a clause-final VP adverb from a clause-final PP, moved to pre-verbal position – the ENE prototypical adverb position – and then to initial position in (a) a contrastive construction with an informationally salient host and (b) an elaborative construction with an informationally backgrounded host. In fact then moved (back) to final position, but now in its Elaborative, backgrounding function. Intersubjective mind (you) likewise split between initial and final positions along, it is argued, information-structural lines. The question of how best to account for the relationship between discourse markers and information structure is discussed: whether the information structure in question is inherent in the coherence relation, as for instance in Rhetorical Structure Theory (Mann and Thompson 1988), or is conventionally expressed by the marker in its construction.

References


Tracking eye movements when reading Dickens’s characters: A pilot study

MICHAELA MAHLBERG & KATHY CONKLIN (University of Nottingham)

This paper reports on the findings of a study piloting an innovative approach within empirical literary stylistics employing psycholinguistic methods to complement corpus stylistic findings.
We examine whether hypotheses formed on the basis of established corpus stylistic research can be demonstrated to have a cognitive reality using psycholinguistic methods. Previous corpus stylistic work has investigated Dickens’s techniques of characterisation (Mahlberg 2012, Mahlb erg & Smith 2012) and identified patterns of body language and other character cues in the form of five word clusters (e.g. *his hands in his pockets*). Corpus findings suggest that such clusters can function in the text to provide vivid and authentic descriptions of characters; these patterns have been referred to as literary ‘contextualising’ patterns (Mahlberg 2012) and contrasted with patterns that ‘highlight’ or foreground specific character information. Crucially, corpus research in general has shown that corpus findings can bring to light textual phenomena that readers may not be consciously aware of. The degree of awareness of textual patterns may have implications for how readers perceive characters in a novel, as characterisation is an emergent process where textual patterns interact with readers' real world knowledge (Culpeper 2001, Stockwell 2009). To discover to what extent readers attend to body language and other patterns of textual cues for character information, we conduct an experiment to track readers’ eye movements (similar to the experiment described in Siyanova, Conklin, & Schmitt 2011) when reading sample passages from Dickens’ novels. Our initial findings indicate that the phrases and patterns under investigation are associated with reading times that are shorter than the general reading times across the sample passages. A possible explanation of this finding is the contextualising nature of the patterns. The eyetracking data will be further related to data that the participants provide on the character information they notice. In this way, we investigate relationships between conscious impressions of characters in the reader’s mind and the patterns that potentially contribute to creating these impressions. The eye-tracking data is further complemented with the results of an online questionnaire study focusing on the same textual extracts. We discuss our results in the context of previous work in cognitive stylistics and argue that the findings open up new routes for the study of charactersation in Dickens

References


Order of recipe ingredients in early English medicine: Evidence of medieval practical intertextuality and literacy practices?

MARTTI MÄKINEN (Hanken School of Economics)

This paper investigates two closely related medieval genres, medical recipe collections and herbals (a subgenre of *materia medica* genre) and the intertextual links that may be found between the order of recipe ingredients and the order in which herbs were represented in
herbals. The basic assumption is that the medieval medical community that produced and used
texts was a discourse community: they purposefully transmitted and made use of information
across several genres and language boundaries when composing texts in one genre.

Intertextuality in Middle English medical recipe collections and herbals has been studied earlier,
and a connection sharing medical recipe information between the two genres has been
established: the use of both recipes and recipe structure in herbals, and recipe paraphrase
structure in recipe collections entails a discourse community that used both earlier recipe
collections and herbals in the compilation of new texts (Mäkinen 2006). However, there are no
studies to date that have charted recipe ingredients as an intertextually active text component in
medieval medical recipe composition. These literacy practices, the ways in which the medical
community drew on the texts at their disposal, are the objective of this study (Barton and

Recipe ingredients in the Middle Ages were classified according to the humoural theory on two
axes, hot-cold and dry-moist, in several degrees; this information was, however, not dispensed
in recipe collections in general. This paper sets out with the assumption that the order of herbs
in herbals may have affected the order medical recipe ingredients; on a more abstract level, the
entailment is that the classification of herbs according to the humoural theory was reflected in
recipe collections as well. Therefore, the order of recipe ingredients most likely signals the use
of knowledge from other sources, such as herbals, in the composition of recipe collections.

This study requires the establishment of medieval herbal nomenclature and degrees in which the
plants were differentiated from each other on the humoural theory axes. The problematic nature
of these two tasks will be further elaborated in the final paper (cf. Stannard 1980: 356, Hunt
clusters will lean on studies on phraseology (cf. Altenberg 1998, Biber and Barbieri 2007).

The herb name and the herb's position on the humoural axes will be used in this study to
investigate whether in medieval medical recipes warmer and drier ingredients were listed before
moister and colder ingredients or vice versa, and if not, whether there are other patterns for
ordering the ingredient information in recipes. The material for this study will be drawn from
the corpus of Middle English Medical Texts (MEMT).

References

Altenberg, B. 1998. On the phraseology of spoken English: The evidence of recurrent word-


Biber, D. & F. Barbieri. 2007. Lexical bundles in university spoken and written registers.
English for Specific Purposes 26: 263-286.

University Press.


*Keep it in that place versus if you could keep it in that place: On the variation between imperatives and isolated if-clauses*

**Beatriz Mato-Míguez** (University of Santiago de Compostela)

As is well known, directives in English usually take the form of an imperative clause, that is, a clause with no overt grammatical subject and a verb in the base form (cf. Quirk et al. 1985: 803), as in (1):

(1) Move your head a little.
Imperative clauses can be used for a wide range of directive illocutionary acts (commands, requests, wishes, etc). Conditional clauses are also often employed to code directives, in uses which are semantically and pragmatically only marginally conditional or not conditional at all (cf. Traugott, Ter Meulen, Reilly & Ferguson 1986: 7), in large measure because the conditional clause is a way of dealing with the negative face wants of the addressee and allows the hearer’s freedom of choice (Ford 1997: 405):

(2) **If** you could move your head a little, that would be perfect.

(3) **If** you could move your head a little, I will examine you.

Among conditional clauses, one type in particular has become specialized in directive function, namely so-called isolated *if*-clauses, illustrated in (4) below, which seem to have emerged from clauses such as those in (2) and (3) through the process known as *insubordination*, that is, “the conventionalized main clause use of what, on primae facie grounds, appear to be formally subordinate clauses” (Evans 2007: 367). Isolated *if*-clauses can be identified by the absence of a main clause, which cannot be retrieved from the context by postulating some form of ellipsis, so that their grammatical status is in fact that of an independent clause. Their most common function in English is to issue polite requests, offers and orders, although more restrictedly they may also serve other functions.

(4) **A:** **If** you put your new uhm address there no there and the old address and your name <,,> (DCPSE: DI-A20)

The directive use of English imperative clauses has been extensively studied (cf. Davies 1986, among others) but to date there are no studies offering an in-depth analysis of the use of imperatives in comparison with directive isolated *if*-clauses. In the presentation proposed for ICAME 34 I will examine both types of clause on the basis of evidence extracted from the *Diachronic Corpus of Present-Day Spoken English* (Arts and Wallis 2006, 885,000 words). Apart from offering data on the frequencies of both imperative and isolated *if*-clauses, I will pay attention to aspects such as: a) the sequential ordering of the two types of clauses in those cases (e.g. (5) below) where they occur together in the course of a conversational interaction; b) whether familiarity between speakers plays a role in the preference of one type of clause over the other (isolated *if*-clauses should represent, arguably, a more polite option); c) whether differences can be observed as regards the number of addressees intended (in other words, is there one recipient or more than one?); and d) what exactly is the specific illocutionary force of each type of clause (in other words, are there observable differences between imperatives and isolated *if*-clauses in the type of speech act – permission, request, command, etc. – they most commonly code?)

(5) **A:** and Gareth **if** you ’d like to lead <,,> Right now **take** the bike off its side stand and hold its weight on the handlebars <,> (DCPSE DI-F22 0018)

**References**


The pragmatic marker *sure* across varieties of English: When the transcription is the corpus

IRINA PANDAROVA (Leuphana University Lüneburg)

My PhD project investigates pragmatic variation in the use of the discourse marker and emphasiser *sure*, which has received very little attention so far. *Sure* is known as a backchannel and an agreement marker similar to *yes* (Tottie 1991). It has also been cited as a marker of Americanness in its emphasiser function in lexical bundles of the type ‘(NP) sure + AUX’, as in *He sure is an odd fellow* (Tottie 2002: 169, Aijmer 2009). A different type of *sure*, usually in utterance-initial and -final position, has been attested for Irish English (Walshe 2009, Moreno 2006). The aim of my project is to build on the existing research on *sure* and carry out the first systematic synchronic, comparative analysis of its formal and functional characteristics in Irish, British and American English in order to fully understand its versatility and multifunctionality. My approach is both corpus-based and corpus-driven in that I attempt to identify and quantify both attested and unattested features in the use of *sure* across the three varieties.

The focus of the proposed work-in-progress report falls on the formal characteristics of *sure* in the spoken components of *ICE Ireland* and *ICE Great Britain*, as well as the *Santa Barbara Corpus of Spoken American English*. The analysis seeks to characterise *sure* in terms of syntactic position, scope and level of utterance integration. This is done by considering features at the morpho-syntactic level, contextual cues and editorial comments provided in the corpora (e.g. pauses and restarts). Crucially, the audio recordings on which the corpus transcriptions are based will be studied as well in order to determine the phonetic realisations of *sure*, different intonation patterns associated with it and its relationship to the preceding or following utterance, which cannot always be reconstructed only by relying on the transcriptions.

Unsurprisingly, the results obtained in this study show that *sure* is overwhelmingly a feature of informal spoken interaction. However, there are highly significant differences in the distribution of *sure* across the three varieties, with *sure* most frequent in Irish English and least frequent in British English. There are also considerable formal differences across the three varieties. In Irish English, *sure* is predominantly initial, with scope over the following utterance and integrated prosodically into it. Interestingly, it can introduce interrogatives and a particular type of tag clause (e.g. *He’s not at home, sure he’s not*). It can also be tagged to the end of an utterance. In British English, *sure* is usually a backchannel or marker of agreement.
This is typical for American English as well but here *sure* is also very common in medial, usually pre-finite position as an emphasizer.

References


Newspapers, politics, and CanE: The Corpus of Early 19th-Century Ontario Newspaper English (CENCONE)

**David Reuter** (University of Kiel)

CENCONE was compiled to provide evidence for linguistic differences in newspapers from opposing ends of the political spectrum (Tories vs. Whigs) in early nineteenth-century Ontario. It aims to help analyze the influence of political allegiance on the development of Canadian English (CanE) in the time of a massive exertion of influence from both Great Britain and the newly-independent USA – the two countries which shaped Canada's demographic, cultural and, especially, linguistic situation. The political stratification of the corpus sets it apart from other diachronic approaches to the variety's development (of which there are generally very few), e.g. from Dollinger's (2008) *Corpus of Early Ontario English* (CONTE). In line with the majority of previous studies, Dollinger uses the settlement history of the country as a theoretical backdrop.

The possibility of a direct connection between political allegiances and language choices in early Ontario has been suggested in various studies (e.g. Avis 1973, Chambers 2004, Dollinger 2006, 2008) but has never been thoroughly explored. CENCONE tries to fill this gap by employing the available data from newspapers, a medium which best reflects and even fuels the political animosities at the time, because nearly all newspaper editors are opinion leaders of one or the other political party. Since differences in political opinion are generally expected to be most pronounced in the periods preceding far-reaching political events, CENCONE is designed around the three events that had the greatest impact on the political development of early Ontario: the War of 1812, the Upper Canada Rebellion of 1837, and the Canadian Confederation of 1867. Therefore, CENCONE covers the following three years: 1810, 1835, and 1860. The periodization (25 years between each period) allows for a comparison with corpora such as CONTE. Each year is represented by four newspapers (two Tory papers and two Whig papers) so that CENCONE contains a total of 12 newspapers. Its size is currently
approaching 300,000 words, evenly distributed on the 12 newspapers. Showing a different behavior in British and American English (BrE and AmE), the following three variables are analyzed in the corpus: -OR vs. -OUR spelling, SHALL vs. WILL as a marker of futurity, and the variable use of the PROGRESSIVE aspect. The results do not only distinguish political inclination (Tory vs. Whig) and publication dates (1810 vs. 1835 vs. 1860), but they are also compared to the genres of letters and diaries (from CONTE) and to BrE and AmE usage at the time. Thus, a multitude of conclusions can be drawn from the analysis: It is shown

a. how political affiliation can influence language,

b. in what way early nineteenth-century Ontario English has developed,

c. to what extent CanE differed from BrE and AmE at that time,

d. to what extent newspaper articles linguistically differed from less formal written texts at that time, and

e. whether CENCONE can be considered to be an expedient tool for the analysis of the above factors.

References


Dollinger, S. 2006. The modal auxiliaries *have to* and *must* in the *Corpus of Early Ontario English*: Gradient change and colonial lag. *Canadian Journal of Linguistics* 51/2-3: 287-308.

**System and content morphemes in intrasentential codeswitching in Spanish/English and Tagalog/English**

*SUSANA SOTILLO* (Montclair State University)

Code switching (CS) or the alternation between two codes or languages within a stretch of discourse is a global phenomenon. CS is an ordinary form of communication among speakers living and interacting in bilingual/multilingual communities (Auer 1998, Myers-Scotton 1993, Myers-Scotton & Jake 2009, Labov 1972, Poplack 1980, 2001, Wei 2009, Woolard 2004, Zentella 1997). Research on code-switching has been informed by a variety of perspectives, such as Variationist Sociolinguistics, Social Network Analysis, the Matrix Language Framework (MLF), and the Conversation Analytic approach. Recent studies of digital discourse have also examined CS as a form of stylistic usage and process of identity construction (Dorleijn & Nortier 2009). Others, Androutsopoulos (2011) and Thurlow & Mrozek (2011),
have argued for a more flexible and comprehensive analysis of language in the new media. Androutsopoulos (2011) maintains that contemporary digital language research should move beyond indexing linguistic differences and sociolinguistic variations. Following Myers-Scotton & Jake’s MLF (2009) and Wei’s (2009) recent studies of system and content morphemes, the present study examines system and content morphemes in samples from a 6,000 word SMS corpus collected over a period of 24 months. The code-switched text messages are from the personal SMS networks of texters who code-switch in Spanish-English and Tagalog-English. Instances of code-switching are identified, coded, and manually tagged for further processing using *MonoConc Pro*. The SMS texting samples of code-switched system and content morphemes will be compared with code-switched data in the *ICE Philippines* corpus and *Talkbank’s Spanish-English Miami* corpus. The following research questions are addressed: (1) Are intrasentential code-switches more frequent in SMS texting than in face-to-face (F2F) conversations/interviews? (2) Is there a difference in the frequency of code-switched system and content morphemes between SMS texting and face-to-face interaction? And, (3) what is the purpose or function of code-switching in texting when compared to F2F code-switching?

**References**


Styles of late modern medical writing for professional and lay audiences

IRMA TAAVITSAINEN (University of Helsinki)

In Present-Day English styles of writing science and medicine differ greatly in texts targeted at different audiences. This paper focuses on the early phases of ‘popularization’, i.e. how medical knowledge was presented in writings for a heterogeneous lay audience as compared to writings for professional readerships in the eighteenth century. The data comes from the periodicals section of the corpus of the Late Modern English Medical Writing 1700-1800 (LMEMT, forthcoming), comprising half a million words in three different kinds of periodicals. The Edinburgh Medical Journal (EMJ, 1733-) is a pioneer specialized professional journal. Its readership consisted of highly educated medical experts. The Philosophical Transactions (PT, 1666-) can be described as a more general scientific journal, published by the Royal Society for its membership of educated men in various disciplines. The Gentleman’s Magazine (GM, 1731-) has been characterized as the first ‘modern’ magazine for a lay readership, and thus it opens up new possibilities of studying the dissemination of knowledge for broader audiences. It contains medical information of various kinds, providing an overview of what the literate eighteenth-century public was told about medicine and how medical issues were treated. It also deals with wider initiatives and impulses in society advocating public interests and health concerns in the Age of Enlightenment.

My hypothesis, based on a qualitative reading of these materials, is that the style in GM is more affective, more in line with the period style of polite society than the discourse for professional audiences. My pilot study aims at verifying this statement by studying interactive features and emotive language use in these journals. Adjectives, interjections and exclamations are particularly revealing, and a new turn of sentiment with personal affect features can be noticed in GM, especially in its verse items. PT and EMJ do not show similar features, but different strategies may be employed in appealing to the audiences; metatextual comments should be particularly revealing in this respect. I shall begin my study by assessing the wordlists and KWIC concordances, keywords and collocations, comparing the different periodicals with one another. The materials of LMEMT have not been assessed before, and the research question of diversification of styles in this period is also novel. The study will reveal how the styles of medical writing developed in the eighteenth century in these three channels for different audiences, in particular in the Gentleman’s Magazine.

The grammatical distinctiveness of written and spoken Canadian English

UWE VOSBERG (University of Kiel)

The paper compares written and spoken Canadian English (CanE) with British and US-American English (BrE and AmE) in a number of variable areas of morphology and syntax. With the exception of Chambers’ Dialect Topography online project mainly focussing on phonological and lexical (including just a few lexico-grammatical) alternatives, the Survey of Canadian English is to date probably the only empirical investigation of the present-day language to cover all Canadian provinces and thus identify something like a national standard even for a number of grammatical variation phenomena (cf. Scargill 1974). While the two studies are based on questionnaires and fail to contrast their results with the situation in both
BrE and AmE, this contribution is now concerned with a large-scale electronic corpus analysis of the following selection of four variables in all three national varieties of English:

1. Adverb marking (AdvP morphology): the intensifier *really*

2. Omission of a preposition (PP syntax): *flee (from) the country*

3. Insertion of a preposition (NP syntax): *all (of) these books*

4. Modals and quasi-modals (VP syntax): *needn’t / don’t have to / don’t need to / haven’t got to worry*

These types of construction will be analysed in the following three text collections containing spoken and written (including fictional and non-fictional) material: the British National Corpus (comprising 100m words), the American National Corpus, 2nd release (23m words), as well as a 2011 release of the Canadian Strathy corpus (48m words).

The results do not only shed light on questions referring to regularity, explicitness and the concomitant question of formality, but mainly help to identify different scenarios of national contrasts (cf. Hundt 1998 for NZE). It is well-known that BrE and AmE tend to follow the same trends in the development of grammatical alternatives. However, very often the two varieties are not affected by these tendencies to the same extent (cf. Rohdenburg & Schlüter 2009). Many analyses demonstrate that CanE often occupies a middle position in the distribution of the variants, and it is usually more inclined towards the situation in AmE (cf. *need not* or *flee from* in the written language or *all these* in the spoken language), sometimes, however, to BrE (cf. *really, need not* in the spoken language). In only very rare and therefore interesting cases, CanE falls outside the spectrum between BrE and AmE (cf. *really* or *all these* in the written language).

While it has been clear for quite some time (cf. Mair 1998: 153f) that written AmE shows a great tolerance of constructions typical of colloquial (e.g. spoken BrE) usage, the paper now suggests that CanE grammar often does not distinguish between the written and the spoken language and that the spectrum of developmental stages with respect to the behaviour of grammatical variants in written or spoken BrE, CanE and AmE may be represented in the following way:

written BrE ------ spoken BrE ≈ written CanE ≈ spoken CanE ≈ written AmE ----- spoken AmE

References


Capturing patterns of linguistic interaction in a parsed corpus: An insight into the empirical evaluation of grammar

SEAN WALLIS (University College London)

This paper concerns a thorny methodological problem at the heart of corpus linguistics: the ontological status of corpus parsing.

Numerous competing grammatical frameworks exist on paper, as algorithms and embodied in parsed corpora (Abeillé 2003). However, not only is there little agreement among linguists concerning the optimal grammatical framework to be used in a corpus, but there is no agreed methodology for determining the benefits of one grammar over another.

The result is fragmentation of methodology and annotation schemes. Some researchers (notably Sinclair 1987) take an expressly anti-annotation ‘data-driven’ approach as a cautious response; most ‘theory-driven’ researchers adopt a given framework critically. Can we turn this problem around? Can we use a parsed corpus to evaluate the parsing scheme in the corpus?

The conventional approach for empirically comparing frameworks is based on distinguishability, or the reliable retrieval of individual linguistic events from an annotated corpus. However this criterion is essentially circular, open to ‘improvement’ by including redundant terms, and fails to reflect broader structural perspectives embodied in the grammar.

In this paper we describe a design of a corpus linguistic research experiment which generates thought-provoking results in its own right and which we tentatively offer as a new way of empirically evaluating grammatical frameworks.

This experimental design attempts to capture patterns of linguistic interaction along grammatical axes. These distributional patterns can only be obtained from large ‘forests’ of trees, and results reveal whether parsing is biased by extra-grammatical constraints on natural language. As an evaluative criterion, it does not rely on atomic retrieval alone, does not risk redundancy and is no more circular than a conventional scientific reliance on auxiliary assumptions.

We argue that permutations in this type of experiment allow us to explore whether the grammatical framework is exposing or hiding extra-grammatical constraints on language production – in much the same way as different scientific equipment may bring research results into or out of focus.

We apply our design to a series of natural experiments on ICE-GB (Nelson et al. 2002) by examining the additive probability applying to repeated decisions. We find an interaction between decisions to add attributive adjective phrases under a noun phrase with a noun head, such that the probability of adding successive adjective phrases falls. On the other hand, preverbal adverb phrases do not exhibit this interaction.

We then turn to grammatical axes which depend on the clause analysis in the parsing scheme. We compare successive clauses modifying the same NP head, and embedding clauses where the NP head is the most recent one. Embedding shows a significant serial decline over depth, whereas sequential postmodification of the same head reveals a fall and then a rise in additive probability.

We argue that this result can only be explained as a natural phenomenon acting on language production expressed by the distribution of cases on these distinct clause axes. Further, this is
evidence that the clause structure reflects the language production process, and therefore can be said to represent cognitively relevant units.

References


POSTERS
Tag questions outside the dialogue in fiction

KARIN AXELSSON (University of Skövde)

Tag questions are primarily used in spoken conversation, so it is natural that most research has focused on such use (e.g. Holmes 1996, Tottie & Hoffmann 2006). In writing, they are mainly found in the representation of speech, for example in fiction dialogue (Axelsson 2011). Tag questions have also been shown to appear in non-fictional writing; Tottie and Hoffmann (2009: 311) state that such tag questions predominantly “express the writer’s attitude, or stance.”

Axelsson (2011) finds that the use of tag questions in fiction dialogue has to resemble real-life use in order to be credible, but that there are differences in the distribution of functions due to the fact that fiction needs to provide interesting reading; therefore, various challenging uses are overrepresented in fiction dialogue, whereas tag questions in certain other functions are underrepresented as trivial matters are avoided in fiction dialogue.

In fiction texts, tag questions are also sometimes found outside the dialogue. There do not seem to be any previous studies of such tag questions. The aim of this study is therefore to investigate the use of tag questions in the narrative parts of fiction texts. The data is retrieved from the British National Corpus. The analysis of 90 tag questions with declarative anchors shows that they are used in thought, writing and speech presentation as well as in pure narration.

Most instances of tag questions outside the dialogue in fiction are representations of thought, primarily free indirect thought. The character involves in self-deliberation, “‘translating’ into words a phenomenon that might have consisted of non-verbal cognitive activities” (Semino & Short 2004: 118). These tag questions are in the past in order to adapt to the past tense of the narrative in general, and the tag subject is in the third person. There are also tag questions in direct thought; as the wordings are the same as the characters would have used if they had verbalized their thoughts, there is more variation in tense and tag subjects.

As might be expected, there are also a few instances of tag questions in writing presentation, usually in letters which a character is writing or reading. More surprising perhaps is the use of tag questions in direct speech outside the dialogue. These are either instances of hypothetical speech, where tag questions are clearly formulated to be spoken but never uttered for various reasons, or parts of schizophrenic conversations between a character and one or several ‘headvoices’.

Some tag questions found in the narrative parts of novels with an I-narrator resemble the stance-marking tag questions of non-fictional writing (cf. Tottie & Hoffmann 2009 above). Moreover, in certain novels with an I-narrator, the implied reader is addressed in tag questions and sometimes even seems to have responded:

(1) It’s a funny thing how so many contactees are called George, isn’t it? No? Well please yourself then.

Tag questions are used far less frequently outside than within the dialogue in fiction; nevertheless, tag questions are exploited for many different kinds of uses outside the dialogue.
Too fascinating a topic to neglect: A corpus study of the Big Mess Construction

ANNE-KATRIN BLASS (University of Trier)

The present study is concerned with a construction that defies generalisations about the prototypical English noun phrase (NP) to such an extent that it has been termed the “Big Mess Construction” (Berman 1974). The construction is illustrated by typical examples in (1) to (3) below.

(1) Perhaps conspiracy is too strong a word. (BNC: A4M 11)
(2) This is a bizarre lapse for so successful a man. (BNC: ED7 488)
(3) Was she that lousy a lover? (BNC: A0L 3232)

The italicised NPs in these examples deviate from the canonical order of constituents in the English NP insofar as, in each case, an adjective phrase (AP) functioning as premodifier of the head of the NP (too strong, so successful, that lousy) precedes the element realising the determiner function (the indefinite article a) rather than following it. As examples (1) to (3) suggest, the possibility for the APs to occur in this fronted position is dependent on the fact that the head adjectives are in their turn premodified by one of a limited set of degree modifiers (here too, so, that). Consider the ungrammatical versions of these NPs without the degree expression (*strong a word, *successful a man, *lousy a lover).

There is as yet no comprehensive corpus-based investigation of the construction, and accounts in reference grammars tend to be cursory or fragmentary, if not completely absent. In my study, I will take a first step towards closing this gap by presenting a detailed investigation of those variants of the construction that are introduced by too on the basis of the British National Corpus (BNC).

The first part of my study is devoted to a brief descriptive analysis of 1,209 instances of the construction in the BNC. The detailed formal analysis reveals that the construction displays much more variability than previously assumed and exhibits no text-type specific preferences.

The larger part of my study is concerned with an investigation of the motivations behind the use of the construction, focusing on the domains of prosody (cf. Schlüter’s (2005) account of the
“Principle of Rhythmic Alternation”) and discourse-pragmatics. I will conclude by briefly discussing the question of whether the evaluative functions that can be discerned solely hinge on the degree modifiers that introduce the Big Mess word order or whether the Big Mess Construction can be regarded as a construction in the Construction Grammar sense (Goldberg 2006), i.e. as a form-meaning or form-function pairing.

References


High-frequency verbs and their collocations: Corpus findings and language-pedagogical implications

**Martina Bredenbröcker** (University of Paderborn)

Various studies (cf. Meunier/Granger 2008, Aijmer 2009, Römer 2009, Reppen 2010, etc.) stress the importance of transferring research results from corpus linguistics to the wide field of language pedagogy. The focus of these studies is almost always on secondary EFL classrooms from age group 11 onwards. However, to provide a solid foundation for life-long language learning as proposed in the *Common European Framework of Reference for Languages* (CEFR), it is necessary to start from an earlier age (e.g. from age group 6 onwards).

Consequently, this work-in-progress report focusses on young learners in primary classrooms. Various corpora that reflect the lexis used by eight to ten-year-old children with English as L1 are the basis for an investigation of lexical structures relevant for young learners of English. These self-compiled corpora contain children’s books, magazines and spoken classroom conversation.

The aim of the present study is to compile a basic set of collocations from adult and children’s corpora. The collocations from these two groups will be compared with respect to their language-pedagogical implications. It will be discussed whether the resulting set is applicable to the primary EFL classroom, irrespective of the learner’s mother tongue.

Frequency lists derived from these corpora will be contrasted with frequency lists coming from standard reference corpora containing mostly adult language, such as the BNC and COCA. Next, the collocations of the children’s high-frequency nouns and verbs will be analysed using different statistical measures, such as the MI-score, T-score, and others.

To give an example: Among the top five collocating nouns for the verb *go* in the BNC and COCA are *way, people, time, school* and *things*. There is only one overlap with the self-compiled corpus of children’s books, namely *school*. The other collocating nouns found in the children’s corpus (*room, bed, home and party*) obviously describe objects and concepts taken
from the childhood world of experience, i.e. basic aspects and areas of life that are relevant for children at primary school level.

By applying tried and trusted vocabulary selection criteria, the resulting set of collocations will be evaluated with regard to the needs of young learners. Among these criteria are – apart from frequency – range/coverage, learnability, availability, familiarity, coverage and regularity (cf. Nation 2001).

Finally, further language-pedagogical implications and applications (e.g. syllabus design, creation of teaching materials, pre- and in-service teacher education) will be discussed.

References


**A look at participial constructions with get in East Asian Englishes: The Hong Kong variety**

**EDUARDO COTO-VILLALIBRE** (University of Santiago de Compostela)

The English passive voice has been the subject of numerous studies and prolific debates. As we know, passive periphrases in Present-Day English can select two different auxiliaries, *be* and *get*. But what is a *get*-passive? The change from the active voice to the passive voice involves a structural rearrangement of the clause elements, where the active subject becomes the passive agent, the active object becomes the passive subject, and the agent, if mentioned at all, is introduced by the preposition *by*. Though these features apply generally to both *be*- and *get*-passives, research on the latter suggests that *get*-passives, as opposed to *be*-passives, tend to be avoided in formal English and are recurrent in conversation, occur only with dynamic verbs (e.g. *cut*, *throw*, *send*, *type*, *play*) and do not normally have an overt agent *by*-phrase. Moreover, the animate subject of *get*-passives has a higher degree of responsibility for the action described than the subject of *be*-passives, and is also commonly attributed adversative consequences, hence its frequent occurrence with predicates such as *arrest*, *blame*, *hit*, *shoot*, *kill* and the like (cf. Hatcher 1949: 436-437, Quirk et al. 1985: 161, Collins 1996: 52, Carter & McCarthy 1999: 49-52, Huddleston & Pullum et al. 2002: 1442).
Alongside central get-passives, prior research (cf. Granger 1983, Quirk et al. 1985, Svartvik 1985, Collins 1996) has identified other structures also consisting of get + past participle, and thus formally related to get-passives, but differing from them in various ways. Along a gradient based on their degree of ‘passiveness’, we can distinguish between get-passives (He got arrested by the police), on the one hand, and so-called semi- (She got excited about the new book), pseudo- (They got married), adjectival- (He got very frightened), idiomatic- (She got used to it), and reflexive get-constructions (I got myself hurt).

In two earlier papers (respectively Coto Villalibre 2012a, Coto Villalibre 2012b) I discussed the formal and semantic characteristics of those various structures in the light of the data retrieved from two corpora of Present-Day spoken British English and Indian English. In the presentation proposed for ICAME 34, on the one hand, and so-called semi- (She got excited about the new book), pseudo- (They got married), adjectival- (He got very frightened), idiomatic- (She got used to it), and reflexive get-constructions (I got myself hurt).

References


Hatcher, A. G. 1949. To get/be invited. Modern Language Notes 64/7: 433-446.


Lexical bundles in British and West African English

SABRINA EDLER (University of Münster)

This PhD project is concerned with a contrastive phraseological investigation of one inner circle variety, i.e. British English (BrE), and two outer circle varieties, i.e. Ghanaian English (GhE) and Nigerian English (NigE). While registers have been extensively investigated from a phraseological perspective for inner circle varieties such as BrE, similar “studies of lexical or lexico-grammatical co-selection phenomena in New Englishes [...] are rare” (cf. Nesselhauf 2009: 2). Essential questions this research project seeks to answer cover the following issues:

(a) Which lexical bundles (LB) structures are preferred in the registers of British, Ghanaian and Nigerian press reportage and editorials?

(b) Are these preferences solely register-specific or do they point towards a development of GhE and NigE as distinctive national standards through endonormative stabilisation? Is it thus justifiable to postulate the existence of distinctive labels such as “Ghanaian English” and “Nigerian English”?

(c) Which conclusions may be drawn from the results regarding both the classification of GhE and NigE within Schneider’s Dynamic Model and its general applicability?

(d) Which correlations can be observed between the functional and structural characteristics of the bundles investigated?

The analysis is conducted using data from two registers, press reportage and editorials, because the language of newspapers has been shown in several studies to reliably mirror phraseological pattern usage in varieties of English (cf. Tingley 1981, Jowitt 1994). For the West African varieties, the data is drawn from the online editions of five national Ghanaian and Nigerian newspapers, respectively, while the newspaper section of the British National Corpus (BNC) is used for the investigation of the British news texts. In order to pinpoint both similarities and differences between the phraseological structure of BrE, GhE and NigE with regard to the two registers, lexical bundles were chosen as the relevant linguistic variable. LBs were identified first by Biber et al. (1999: 992) and defined as “[…] combinations of words that […] recur most commonly in a given register”. Moreover, numerous studies have shown that LBs, and 4-word bundles in particular, are “good discriminators between registers” (Scott and Tribble 2006: 142) and thus a very effective means of revealing phraseological peculiarities of different registers. For the retrieval of the bundles, the free concordancer AntConc 3.2.4w will be used.

A pilot study investigating the distribution of 4-word bundles in the newspaper section of the BNC and two 400,000-word news text collections from five national online newspapers for each Ghana and Nigeria has shown that both press reportage and editorials from all three varieties alike favour LB structures dominated by a noun phrase (NP) pattern (e.g. NP + of-phrase fragments such as the end of the). In addition, while bundles built around a verbal core occur considerably less frequently in the reportage and editorial sections of all newspapers under scrutiny, the findings suggest that there are differences with regard to the frequencies of individual verbally dominated LBs (e.g. passive constructions).

References


Eating disorders terminology from scientific journals to mass media: A preliminary approach

FÁTIMA FAYA-CERQUEIRO (University of Castilla la Mancha)

The relevance of the media in the distribution of health information is noteworthy, not only through public campaigns on different aspects, but also echoing scientific discoveries. In fact, there is a good amount of literature devoted to the role played by the media in the transmission of health knowledge and their direct responsibility in health issues. Among more specific topics, we find the treatment of some health problems in newspapers, as in the case of eating disorders. In fact, newspaper coverage on childhood obesity has been studied by Westall (2011) in three Spanish newspapers in order to provide a qualitative analysis. Her study shows some examples of the inclusion of recent scientific discoveries in news reports. The observation of knowledge transfer from highly specialised contexts to more general ones may provide revealing insights into how information about a specific health topic, such as the field of eating disorders, reaches the general public.

This study aims to establish a correlation between lexicon from the semantic field of eating disorders in scientific journals and its use in mass media, and more specifically in British newspapers.

For this purpose, a corpus of scientific articles on different eating disorders, such as well-known anorexia and bulimia, will be selected from a group of prestigious scientific journals. They will be sorted by relevance in well-known databases. Wordlists of specific vocabulary will be produced together with their frequencies. Then, those lists will be checked against a selection of newspaper articles including information on the same topic, in order to analyse how specialized lexicon in this field is incorporated onto everyday language through the media.

The second corpus of articles will be drawn from renowned daily British newspapers, such as The Guardian and The Times, taking into account texts from the last decade, which will facilitate a chronological approach. Thus, we will be able (i) to check if there is a transfer of knowledge from specialised media to more generalised media in this specific discipline and topic; and (ii) whether this correlation evolves throughout the years or each medium maintains a distinctive catalogue in the selected period.

This preliminary study aims to identify some patterns in the incorporation of specific lexicon on health issues to more widespread media. The results obtained could be compared to those provided by an analysis of scientific journals and newspapers written in other languages, paying attention to different cultural backgrounds. Similarly, the time period covered could also be increased, as well as the range of illnesses observed.
Even if the catalogue of specific vocabulary extracted from newspapers may not be representative of the lexical availability in average speakers, we would be able to acknowledge the role played by the media in scientific information distribution and its actual application to everyday life.

References


Spoken style, written style – ditransitives in speech and the written standard

Johanna Gerwin (University of Kiel)

The dative alternation, i.e. the alternation between a double object construction (e.g. *She gave him a book*) and a prepositional construction (e.g. *She gave a book to him*), is often explained with recourse to language-internal determinants. These include the (morpho-)phonological characteristics of the particular ditransitive verbs (cf. e.g. Pinker 1989, Grimshaw 2005, Coppock 2009) or the semantic and pragmatic advantages of one construction over the other in certain contexts (cf. e.g. Green 1974, Bock & Warren 1985, Gropen et al. 1989, Goldberg 1992, Levin 1993, Thompson 1995, Davidse 1996). Many of these studies are not based on corpus data but reflect the intuition of individual researchers.

The present study takes a different approach to the problem of what determines the choice between these meaning-equivalent ditransitive syntagms. By means of corpus data from two spoken corpora (*Freiburg English Dialect Corpus* (FRED) and *BNCweb*) and two written corpora (LOB and F-LOB) the influence of the type of register on the choice of ditransitive is to be investigated. The data from FRED and LOB are from the 1960/70s, whereas the data from the BNCweb and F-LOB are from the 1990s. This enables a synchronic analysis of register variation in these two stages of the language. Furthermore, the four corpora allow for diachronic comparison of the development of ditransitives from the 1960s to the 1990s in both spoken and written language.

The four corpora were searched for a number of ditransitive verbs known to participate in the dative alternation (e.g. *give, tell, pay* and *deny*). The analysis shows that individual ditransitive verbs vary significantly as to whether they tend to appear in the double object construction (DOC) or the prepositional construction (PREP) and thus confirms previous research (cf. Wasow 2002, Wasow & Arnold 2003, Wolk & al. forthcoming). However, the type of register can be observed to exert a clear influence on the frequency of either construction appearing with certain ditransitive verbs, as can changes in these patterns over time with some individual verbs.

Diachronic comparison of both the spoken and the written data sets reveals that the double object construction is on the rise in both registers but is more advanced in spoken language. Reasons for the overall increase of DOC at the cost of PREP may be found in the differential use of ditransitive idiomatic expressions in spoken and written language. Whereas in spoken language DOC-idioms such as *give sb. advice* prevail, in written texts PREP-idioms such as *pay*...
attention to sb., which tend to involve a higher degree of formality, are more frequent. Paradoxically, both types of idiom often involve the high-frequency ditransitive verb give.

The study is one of a series investigating the influence of regional and stylistic factors on the choice of ditransitive construction in Present-Day English. It provides a novel approach to ditransitives in being fully corpus-based and dealing with extra-linguistic determinants on the dative alternation.

References


Wolk, C., J. Bresnan, A. Rosenbach & B. Szmrecsanyi (forthcoming). Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. Diachronica 30/2.
Eh oh I think that is eh...: I think functioning as an utterance launcher in spoken learner English

HEGE LARSSON AAS (Hedmark University College)

This poster presents preliminary results from a corpus-based investigation into the occurrences of the frequent recurrent word-combination I think used as an utterance launcher, an expression with “a special function of beginning a turn or an utterance” (Biber et al. 1999: 1073), in a corpus of advanced non-native English speech (NNS).

The initial hypothesis of the present study is that speech management in terms of “providing the speaker with a planning respite” (ibid.) in the initial position of an utterance is an important function of several frequent recurrent word-combinations in NNS, which may serve as a partial explanation for their high frequencies overall. The term ‘recurrent word-combinations’ is defined here as combinations of two or more words appearing repeatedly in identical form in the corpus, following Altenberg (1998). Particular attention is paid to the recurrent subject-verb combination I think, as well as the recurrent word-combinations in which it is embedded (e.g. I think it's), which previous studies of comparable corpora have shown to be used with very high frequencies by learners, and often overused when compared to native English speech (e.g. Aijmer 2004, De Cock 2004). Although I think is found to perform a ‘starting-point function’ also in native speaker texts (e.g. Aijmer 1997, Kärkkäinen 2003), these quantitative findings of overuse may reflect the greater tendency for learners to “cling on to certain fixed phrases and expressions which they feel confident in using” (Granger 1998: 156), a tendency which in turn makes preferred combinations likely to appear at major planning points in speech, perhaps most notably at the beginning of utterances.

The study identifies the extent to which recurrent word-combinations with I think appear in initial utterance-launching position in the corpus, and whether the occurrences in initial position are more frequently seen to take part in a ‘hesitation preface’ (Aijmer 2011: 239) or a ‘wobbly thematic springboard’ of utterances (De Cock 2004: 233), rather than being an integral part of the intended message, “presenting a personal stance relative to the information in the following complement clause” (Biber et al. 1999: 1003). The analysis classifies the material according to position and function, and detects co-occurrence with possible markers of hesitation such as incomplete clauses, filled and unfilled pauses and truncated words. The results presented here are the first in a wider investigation looking at recurrent word-combinations in this context.

The primary data for the investigation has been retrieved from the Norwegian component of the Louvain International Database of Spoken English Interlanguage (LINDSEI) containing oral interviews of advanced learners of English with a Norwegian native language background (Gilquin et al. 2010). The results presented here arise from some of the first research based into the language collected in this Norwegian subcorpus, which is still under compilation and not yet accessible to a wider public. This investigation thus adds to the corpus-based, descriptive account of this interlanguage variety, and further provides a basis for comparison with other native or non-native language corpora.

References


**Hanken Corpus of Academic Written English for Economics**

**MARTTI MÄKINEN** (Hanken School of Economics) & **TURO HILTUNEN** (University of Helsinki)

This poster will report on the compilation of a new learner corpus, *Hanken Corpus of Academic Written English for Economics* (working title), work in progress at Hanken School of Economics and University of Helsinki. Our aim is to produce a corpus of student writing in business studies and economics, written by L2 speakers of English as part of their studies. The corpus is intended to function as a resource in corpus-aided teaching of academic writing at Hanken, and as a source of data for studies on learner academic writing. The corpus is a joint enterprise shared by researchers at the Centre for Languages and Business Communication at Hanken School Economics, Helsinki, and at the Department of Modern Languages, University of Helsinki. Currently the compilers are using the corpus for two phraseological studies (Hiltunen and Mäkinen, in preparation a and b).

According to Granger and Goossens (2012), there are no academic learner corpora on Economics only. At the moment, *Hanken Corpus* contains 114 texts, with c. 450,000 words. The texts included are both student papers and Master’s theses. The corpus is still work in progress, and the number of student papers will grow by 60-70 every year. The corpus is POS-tagged with CLAWS (C7 tagset, UCREL a and b); in addition to that the texts will be annotated for metadata, containing information about the age, L1, gender, duration of English studies, etc. The corpus will also be provided in plain text ASCII and XML formats, for different purposes of use. Eventually, all the flavours of the corpus will be made available online.
The corpus is designed with two research aims in mind. On the one hand, data from Hanken corpus can be compared to published research articles, either extracted from general reference corpora or collected separately. Such a comparison will highlight the differences between novice and expert writing in the discipline of economics. On the other hand, the Hanken Corpus can be compared to the student texts by native English speakers, available in the British Academic Written English Corpus (BAWE) and the Michigan Corpus of Upper-Level Student Papers (MICUSP). In this way, it is possible to obtain information about differences between texts by students of different linguistic backgrounds. While both BAWE and MICUSP also contain texts written by L2 speakers of English, the Hanken Corpus will be able to provide more information about two L1s, Swedish and Finnish, which are the two main L1s of students studying at Hanken School of Economics.

References


Hiltunen, T. & M. Mäkinen. (in preparation a.). Formulaic language in L2 academic writing for business studies and economics.

Hiltunen, T. & M. Mäkinen. (in preparation b.). A list of academic phrases for economics: A corpus-driven study [working title].


Linguistic change in its social contexts in eighteenth-century English

TERTTU NEVALAINEN, MINNA PALANDER-COLLIN, TANJA SÄILY, MIKKO HAKALA, TEO JUVONEN, SAMULI KAISLANIEMI, MIKKO LAITINEN, MINNA NEVALA, ARJA NURMI, ANNI SAIRIO (University of Helsinki)

These posters present new results of ongoing sociolinguistic research by the C18 Research Group. The studies deal with the sociolinguistics of a dozen changes in eighteenth-century English based on data extracted from the Late Modern English part (1680-1800) of the socially stratified Corpus of Early English Correspondence (see CoRD for details). The changes explored include linguistic features whose sociolinguistic trajectories are well known up to 1700 (e.g. Nevalainen and Raumolin-Brunberg 2003), but the later developments of which have not
been fully explored. Their spread in the language community coincides with different stages of the S-curve, including changes progressing in mid-range and those nearing completion, but also stable variation and features difficult to conceptualize as sociolinguistic variables.

On the basis of earlier research, the impact of different social variables on the choice of linguistic variants may depend on the stage of the change: the most variation is observed when the process is in mid-range whereas the range of social variation narrows down when the change is nearing completion (e.g. Kurki 2004, Tagliamonte and D’Arcy 2009: 99). The main aim of the posters is thus to explore how social variables such as gender, age and social rank may co-vary with the linguistic features at different stages. In addition, the analysis accounts for individual variation by looking for outliers, who are either conservative or progressive with respect to a particular process of change (cf. Nevalainen, Raumolin-Brunberg and Mannila 2011). When interpreting the results, changes in eighteenth-century society need to be taken into account, including changes in societal structure and social climate, ongoing language standardization processes and the emerging normative influence as well as a newly developed consciousness of letter writing as a social accomplishment. These factors affect the nature of the available data and pose further methodological challenges that are tackled in new ways.

The four posters map the sociolinguistic questions and challenges in the eighteenth century context and present the results thematically as follows: 1) presentation of the 2-million-word corpus, 2) new quantitative methods for historical sociolinguistics, 3) the sociolinguistics of ongoing changes in mid-range, and 4) the sociolinguistics of changes nearing completion. The new methods to be discussed provide statistical solutions for cases that are difficult to model in terms of the S-curve including the productivity of the suffixes -ness and -ity, the development of the auxiliary do in affirmative statements, and the rise of the progressive. The ongoing changes focus on the development of the noun subject and object of the gerund, the variation of its vs. of it, and the indefinite pronouns ending in -body vs. -one. The changes nearing completion are represented by the variation of has/does/says vs. hath/doth/saith, the pronoun thou, and the indefinite pronoun -man. These multiple changes highlight the variety and complexity of sociolinguistic factors in the interpretation of language change.

References


A quantitative approach to empty referential subjects in Old English

KRISTIAN ANDERSEN RUSTEN (University of Bergen)

In the philological and historical linguistic traditions, it has been noted that Old English (OE) features numerous occurrences of finite clauses which lack an expressed referential subject, as illustrated in (1).

(1) *Wearð þa fordifen on an igerlond ut on ðære Wendelsæ.*

became then away-driven to an island out in the Mediterranean-sea

‘[He] was then driven to an island in the Mediterranean.’

(Bo 115.22)

Such ‘subjectless’ finite clauses can be analysed as containing an underlying empty pronoun functioning as subject. The phenomenon exemplified in (1) thus corresponds to the much-discussed pro-drop of generative syntax (cf. e.g. Taraldsen 1978; Jaeggli & Safir 1989). Since Present-Day English is a non-pro-drop language, the ‘presence’ of empty subjects in OE has rightly elicited scholarly interest, yet little quantitative research investigating their occurrence has been carried out. Consequently, there is scant consensus as concerns the frequency of such subjects and the mechanisms which sanction their occurrence. Mitchell (1985: 633) states that they occur ‘spasmodically’, yet van Gelderen (2000: 121) claims that “pro-drop is quite common” in OE, relating such commonness to the strength of the language’s inflectional system – an explanation Mitchell (1985) rejects.

The availability of syntactically annotated corpora such as the YCOE (Taylor & al 2003) and the YCOEP (Pintzuk & Plug 2001) facilitates large-scale empirical investigation of this phenomenon in OE. This paper presents a corpus-based, quantitative study which utilizes these resources, exploring the following research questions:

1. To what extent do empty referential subjects occur in OE, and to what extent is the occurrence sanctioned by structural–pragmatic criteria such as the distance between the empty subject and its antecedent, their respective clause types, the antecedent’s syntactic function and the ‘richness’ of the verbal inflections?

2. Why and when do overt referential subjects become obligatory in English finite clauses?

In order to answer these questions, all occurrences of empty referential subjects in 131 texts of OE prose and poetry have been collected by extracting all clauses containing the empty category *NP-NOM-PRO* from the YCOE and YCOEP, respectively. The corresponding overt subjects have been collected from the same texts by extracting all clauses containing constituents tagged *NP-NOM-PRO^N*. This contrastive data enables an investigation of the relative frequency of empty subjects according to period, genre and dialect, as well as a substantial quantitative appraisal of structural variables hypothesized to influence subject omission in OE. Using the PPCME2 (Kroch & Taylor 2000) and PPEME (Kroch et al. 2004), the structural characteristics of the decline of the phenomenon will also be traced via Middle English to Early Modern English.

The information from these sources is expected to show that empty subjects represent an extremely rarely occurring feature of early English syntax. Further, the data is expected to show...
that verbal inflections are insufficient in ‘licensing’ the occurrence of empty subjects, suggesting that morphology is outranked as an explanation by the grammatical function of the antecedent and the clause type in which the empty subject occurs.

References


Was and were in the English counterfactual conditionals: Classification and probability prediction

MYOUNG HYOUNG SONG (Seoul National University)

As Kierkegaard wished for some indicative power, observing that his life was far too subjunctive, the indicative form was has kept encroaching upon the territory previously occupied by the subjunctive form were in the English counterfactual conditional adjuncts, to the effect that were is now being regarded as “something of a fossil” (Quirk & al. 1985: 158) or “an unstable remnant of an earlier system” (Huddleston & Pullum 2002: 88). Despite the fossilization of were in counterfactuals (e.g. If I were rich, I would go for it), it was reported to be still the dominant option over was both in British and American English (Johansson 1988). This result has led me to ask such research questions as:

1. Has there been a change in the use of was and were in counterfactual conditionals, recently?
2. What predictors contribute to the currently more frequent use of were relative to the use of was?

3. How much is there a chance of having were in counterfactual conditional adjuncts, based on the important predictors?

To answer these questions, the present paper is mainly concerned with the overall (and relative) frequency of was and were in a dataset taken from the COCA corpus representing the most recent American English (1990–2000), and the BYU-BNC corpus representing the British English counterpart (1980–1993). The dataset is analyzed to investigate a recent variation in the use of was and were according to regions (British English and American English), text types (69 subtypes under 7 main genres) and other morpho-syntactic predictors: grammatical persons (1sg., 3sg.), complementation patterns (to infinitive, present participle, past participle), negation, and inversion.

The frequency data of was and were are entered into \( \chi^2 \) test for frequency difference in British English and American English, random forests for variable importance and metric scaling, and a logistic regression analysis for probability prediction in turn. The results indicate that the use of were is relatively balanced across the genres, while the use of was is biased towards the SPOKEN and FICTION genres. To our interest, there is a division of labor in spoken texts between was and were in the way that the former usually go with participles (e.g. you might cry if someone was tampering with your car), while the latter most likely occur with to infinitive (e.g. if someone were to fall, he would unquestionably die). Lastly, the overall odds ratio of was and were occurrences is 0.5 to 1 with the odds for the use of were in American English much greater than that in British English.

These findings have important implications in terms of (1) the status of were as either a fossil or a stronghold of subjunctive mood in English grammar, and (2) the morpho-syntactic and semantic similarities and differences influencing the use of the two BE-forms in counterfactual conditional adjuncts.

References


Patterns and functions of functional sentence stems in academic text
NAIXING WEI (Beihang University) & JUNYING ZHOU (Beijing University of Technology)

Formulaic language in academic English texts has been a central focus of research in corpus linguistics in recent years (see Adel 2006, Hyland 2004, 2005, Biber et al 2004, 2007, Biber 2009, etc.). Broadly speaking, previous studies of formulaic language in academic English can be regarded as falling into two major strands: the lexical bundle studies and the phraseology studies. The two strands of studies differ significantly in the methods for identifying formulaic sequences, which inevitably arrive at items of different nature. The lexical bundle studies (e.g. Biber et al 2004, 2007, Biber 2009, Cortes 2004, Hyland 2007, etc.) adopted a frequency-driven method for identifying formulaic sequences, which often arrived at structurally incomplete sequences, such as *there is a*, *the nature of the*, *has to do with*, etc., specific meanings and functions of which are difficult to pin down. By contrast, the phraseology studies (e.g. Simpson-Vlach & Ellis 2010, Schmitt & Carter 2004, Martinez & Schmitt 2012, etc.) often involved heavy human-identification of formulaic items with reference to qualitative criteria. Although studies of both strands have offered important insights, from different perspectives, for understanding academic text organization and for pedagogy in English for Academic Purposes (EAP), there is room for improvement in this area of research.

The present study sets out to identify a type of more structurally complete formulaic sequences, namely, Functional Sentence Stems (FSSs), by using an automatic identification method. These are recurrent clause-level lexico-grammatical sequences in academic research articles (e.g. *this paper describes*, *it is important to note*), which perform a wide spectrum of specific functions in organizing textual information and airing authorial opinions and attitudes. The paper aims to: (1) discover the types and specific forms of FSSs, with a special focus on their formal patternings and local functions (Mahlberg 2007: 193); (2) discuss their co-selections at different linguistic levels (Sinclair 1991, 1996, 2004, Tognini-Bonelli 2001); and (3) provide more insights and implications for pedagogy.

Methodologically, we use a new computing method for extracting and identifying phraseological sequences from corpora we have recently developed (Wei and Li, in press) to treat data from JDEST, a 6.5-million-word academic English text corpus. The method utilizes a new normalizing algorithm of ‘probability-weighted average’ to calculate the internal association of an n-gram (*n ≥ 2*) by taking a weighted average of the ‘glue’ values for all the dispersion points in the sequence. The formula for ‘probability-weighted average’ is as follows:

\[
G = P(G_i)G_i + P(G_2)G_2 + P(G_3)G_3 + \cdots + P(G_{n-1})G_{n-1}
= \sum_{i=1}^{n-1} P(G_i)G_i
\]

The algorithm, along with several other automatic procedures, has enabled us to identify frequently-occurring FSSs which are relatively structurally complete. Finally, 6274 different types of FSSs, with a total of 164,089 tokens, have been identified, with their formal patternings described and their local functions examined in detail. The study shows that FSSs are frequently occurring discourse-constructing devices in research articles and that they demonstrate a variety of co-selections which function to express specific local functions in their environment. The paper also discusses implications for academic discourse analysis and EAP pedagogy.

References


English corpus linguistics on the move: Applications and implications