

# **AACL 2008**

**American Association for Corpus Linguistics**

March 13–15, 2008  
Brigham Young University

**Organized by**

Mark Davies  
and volunteer faculty and students from Brigham Young University



## Schedule

<b>Wed</b>		<b>Pre-conference workshop: Using R for Corpus Linguistics (4188 JFSB)</b>				
	12:00	Retrieving corpus-linguistic data from corpora with R (Stefan Th. Gries)				
	4:30	Statistically analyzing data with R (Harald Baayen)				
<b>Thurs</b>	8:30	Welcome				
	8:45	Tony McEnery: <i>Corpus Linguistics and the Humanities</i> (Lee Library, auditorium)				
	10:00	Buses to Aspen Grove				
		<b>ACQUISITION 1</b> (Timpanogos A)	<b>CORPUS CREATION 1</b> (Timpanogos B)	<b>SYNTAX 1</b> (Timpanogos C)	<b>STYLISTICS 1</b> (Aspen/Fir)	
	11:00	<i>Annotation, Indexing and Querying a Multilingual, Multimodal Classroom Discourse Corpus</i> Huaqing Hong and Paul Doyle	<i>Complementing the BNC with a Corpus from the Web</i> William H. Fletcher	<i>Finnish case alternating adpositions: A corpus study</i> Sander Lestrade	<i>Representations of Islam in the British and American press 1999–2005</i> Paul Baker and Costas Gabrielatos	
	11:30	<i>A New Corpus of Student Academic Writing</i> Susan Conrad and Sarah Albers	<i>Introducing Word-Cruncher 7.1</i> Monte Shelley and James Rosenvall	<i>The diachronic development of some verbs with copulative function in Spanish.</i> Alfonso Gallegos Shibya	<i>Emotive Language and Disease Outbreak Reports</i> Mike Conway	
	12:00	<i>The Nora Corpus: a study of Arab EFL written discourse</i> Eman Al Nafjan	<i>Tubing the Web: a corpus-based study on video-communication</i> Elisabetta Adami	<i>Scrambling in Spoken Dutch</i> Geertje van Bergen and Peter de Swart	<i>Discursive construction of terrorism in Peoples Daily and The Sun before and after 9.11</i> Yufang Qian	
	12:30	<i>Towards a Multi-layered &amp; Multimodal Annotation Model of Learner Corpora</i> Yukio Tono	<i>The 360 million word BYU Corpus of American English (1990–2007)</i> Mark Davies	<i>Subject omission in Russian: A study of the Russian National corpus</i> Tatiana Zdorenko	<i>Parents, victims, suspects, victims...: The 'framing' of the McCanns in the British tabloid press</i> Del Barrett	

	1:00	LUNCH				
		<b>APPLICATIONS</b> (Timpanogos A)	<b>COMPUTATIONAL</b> <b>1</b> (Timpanogos B)	<b>CROSS-LINGUISTIC 1</b> (Timpanogos C)	<b>FREQUENCY 1</b> (Aspen/Fir)	
	2:00	<i>A Corpus-Based Model for the Work of Editors</i> Joseph Richardson	<i>Learning Appraisal Extraction Patterns</i> Ken Bloom	<i>Stock Market Jargon Metaphors in General Discourses: A Corpus-based Study of Mandarin Chinese</i> Carrie Hsin-wen Tsen and Shelley Ching-yu Hsieh	<i>Semantic frequency and the creation of pedagogical word lists: What can we learn from SemCor?</i> Dee Gardner	
	2:30	<i>Authorship Attribution : What Mixture-of-Experts Says We Don't Yet Know</i> Patrick Juola	<i>The CEPRIL Metaphor Candidate Identifier: A program for identifying metaphor in corpora</i> Tony Berber Sardinha	<i>Towards Predicting New Words from Newer Words: Lexical Borrowings in French</i> Paula Chesley	<i>Comparative and Superlative Adjectives and Textual Frequency</i> Laura Teddman	
	3:00	<i>Using Corpora in an English Usage Class</i> Don Chapman	<i>Semi-automatic Classification and Extraction of Predicates from German Text Corpora</i> Ekaterina Lapshinova-Koltunski	<i>Semantic Anglicisms in Contemporary Metropolitan French</i> Betsy Kerr	<i>Exploring Text-initial Concgrams in a Newspaper Corpus</i> Matthew Brook O'Donnell, Mike Scott and Michaela Mahlberg	
		<b>ACQUISITION 2</b> (Timpanogos A)	<b>COMPUTATIONAL</b> <b>2</b> (Timpanogos B)	<b>SYNTAX 2</b> (Timpanogos C)	<b>ENGLISH VARIATION</b> (Aspen/Fir)	
	4:00	<i>The Effect of Computer-Mediated Communication on the Acquisition of Registers of Written Academic Discourse by Creole-Speaking Children</i> Arlene Clachar	<i>Estimating the saliency of constructions using document frequencies from the web</i> Gard Jense and Christer Johansson	<i>On the development of Nouns as Internal Dependents in the Contemporary English Noun Phrase</i> Iria Pastor-Gómez	<i>Selected words, phrases, and meanings of African (American) provenience in General American: A corpus-based study</i> Radoslaw Dylewski	
	4:30	<i>Word-Frequency and Vocabulary Acquisition: An Analysis of Elementary Spanish Textbooks</i> Concepción Godev	<i>Creating Subcorpora to Explore the Topic Structure of Domain-Specific Text</i> Paul Deane	<i>A multifactorial approach to that-deletion in English complement constructions</i> Stefanie Wulff	<i>"I've never seen anything like it": An inquiry into the grammatical clustering involving never in spoken American English</i> Hongyin Tao	

	5:00	<i>Spoken Spanish in Corpora and in Text-books: Implications for Acquisition</i> Grant Goodall	<i>Intermodal cohesion and coherence in multisemiotic text</i> Sabine Bartsch	<i>A Corpus Analysis of Dative Clitic Doubling in Spanish</i> Karen Vogel and Gerald Delahunty	<i>“Ah lovely stuff, eh?” On invariant tag meanings and usage across three varieties of English</i> Georgie Columbus	
	5:45	Laurel Brinton: <i>Historical Pragmatics and Corpus Linguistics: Problems and Strategies</i> (Timpanogos Room)				
	7:00	DINNER				
	8:30	Buses return to BYU				
<b>Fri</b>	8:30	Buses leave BYU for Aspen Grove				
	9:15	Susan Hunston: <i>“You can’t deny the fact that...”: An Application of Corpus Linguistics</i> (Timpanogos Room)				
		<b>CROSS-LINGUISTIC 2</b> (Timpanogos A)	<b>COMPUTATIONAL 3</b> (Timpanogos B)	<b>SYNTAX 3</b> (Timpanogos C)	<b>STYLISTICS 2</b> (Aspen/Fir)	<b>REGISTERS 1</b> (Pine/Spruce)
	10:30	<i>A corpus-based investigation of cognate prepositions in English and Swedish</i> Kerstin Lindmark	<i>An Efficient Framework for Large Scale Cross Document Coreference (CRC)</i> Jian Huang and C. Lee Giles	<i>Corpus-based constituency tests and the structural position of auxiliary verbs</i> Jack Grieve	<i>The use of hedging devices in American English. Identifying some trends in the FROWN corpus</i> Adriana Teresa Damascelli	<i>Suggestions and Recommendations in Academic Speech</i> Luciana Diniz
	11:00	<i>Colouring COMPARA: contrastive and monolingual colour studies in English and Portuguese</i> Rosário Silva	<i>Measures of dispersion in corpus data: a critical review and a suggestion</i> Stefan Th. Gries	<i>Using corpora to quantify contexts: The case of the Portuguese Perfect</i> Patricia Amaral and Chad Howe	<i>Grammatical Expression of Stance in Outsourced Call Center Discourse</i> Eric Friginal	<i>Testing Language-Based Indicators of Deception on a Corpus of Legal Narratives</i> Eileen Fitzpatrick and Joan Bachenko
	11:30	<i>A Statistical Analysis of CEEJUS, Corpus of English Essays by Japanese University Students</i> Shin Ishikawa	<i>Using an XML database for large corpora: Introducing Cheshire3</i> Matthew Brook O’Donnell, Catherine Smith, Robert Sanderson, Clare Llewellyn and John Harrison	<i>A corpus-based study of mandative subjunctive triggers in published research articles</i> Pamela Pearson	<i>Comparing stance in qualitative and quantitative research reports</i> Bethany Ekle Gray	<i>Social Taboo, Evaluation, and Identity Construction Online</i> Mohammed Albakry

	12:00	<i>Studying phraseology and translation through the corpus and the database</i> Maria Freddi	<i>Compiling a new French frequency dictionary</i> Deryle Lonsdale and Yvon Le Bras	<i>Reflexive Pronoun as Discourse Focus Marker: The Case of Spanish morir vs. morirse</i> Jeff Turley	<i>Linguistic realizations of rhetorical structure: A corpus-based study of research articles in applied linguistics and educational technology</i> Phuong Dzung Pho	<i>The representation of time and space across historical discourse: a comparative analysis of evaluative effects</i> Marc Silver and Sara Radighieri
	12:30	LUNCH				
		<b>ACQUISITION 3</b> (Timpanogos A)	<b>CORPUS CREATION 2</b> (Timpanogos B)	<b>SEMANTICS</b> (Timpanogos C)	<b>HISTORICAL 1</b> (Aspen/Fir)	<b>REGISTERS 2</b> (Pine/Spruce)
	1:30	<i>Developing writer stance in intermediate level Japanese university writing in academic and disciplinary courses</i> Jan Minagawa	<i>Developing a Corpus for a Morphologically Rich, Endangered Language</i> Deryle Lonsdale	<i>Did the boys leave or not? Negation and quantifier scope: a corpus study</i> Gunnel Tottie	<i>A corpus-based research on the history of clitic climbing in European Portuguese</i> Aroldo Andrade	<i>Colloquialization: An Alteration in Written English</i> Ilka Mindt
	2:00	<i>The Use of Linking Adverbials in EFL Learners' Time-Constrained Spoken Monologue</i> Kornwipa Poonpon	<i>MayanWiki: Facilitating Consensus Through an Openly Editable Corpus</i> Robbie Haertel	<i>The role of constructions in the assignment of modal meanings</i> Ferdinand de Haan and Sheila Dooley	<i>"De Ahí, Por Consiguiente, Por Ende, Por Lo Tanto and Por Tanto: A Distributional Diachronic and Synchronic Analysis"</i> Arthur H. Wendorf	<i>Corpora of Spanish versus an educational text of astronomy</i> Cristina Hansen
	2:30	<i>"Students must": A corpus based look at directives in university language</i> Randi Reppen	<i>Biblia Medieval: a parallel corpus of medieval Spanish</i> Andrés Enrique-Arias and Laura Carmago	<i>Beyond the lemma: Inflection-specific constructions in English</i> Sally Rice and John Newman	<i>A diachronic process that gives birth to a Spanish discourse particle: The case of "claro"</i> Francisco Ocampo	<i>Journalistic Corpus Similarity over Time</i> Cristina Mota
	3:00	<i>Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive</i> Ute Römer and Stefanie Wulff	<i>Into the woods: First steps in a collaborative corpus of medical conversations</i> Boyd Davis and Charlene Pope	<i>Good Nouns, Bad Nouns: What the corpus says and what native speakers think</i> Philip Dilts	<i>El coche ese vs el coche de marras: The postnominal demonstrative and de marras in Diachronic and Synchronic Corpora</i> David Alexander	<i>Characterizing genre: The case of scientific texts</i> Elke Teich
	3:30	Return to BYU, OR Spend time at Sundance, and then return at 5:30				
	6:00	DINNER (At Skyroom, BYU)				
<b>Sat</b>	8:30	Harald Baayen: <i>Co-occurrence below and above the word level: exploring language at the intersection of corpus linguistics, psycholinguistics and statistics</i> (3714 Lee Library)				

		<b>HISTORICAL 2</b> (3710 Lee Library)	<b>COMPUTATIONAL 4</b> (3712 Lee Library)	<b>PATTERNS</b> (3714 Lee Library)	<b>STYLISTICS 3</b> (3716 Lee Library)		
	10:00	<i>American slang in mainstream magazine writing</i> Anna Belladelli	<i>A Corpus-Driven Pattern Dictionary for Mapping Meaning onto Use</i> Patrick Hanks	<i>'Positioning lexical bundles in university class sessions'</i> Eniko Csomay and Viviana Cortes	<i>Fat and Health Literacy: Two Revealing Terms in CADOH (Corpus of American Discourses on Health)</i> Laurel Stvan		
	10:30	<i>Recent Change in Core Grammar: a Case Study Based on Corpus Evidence</i> Juhani Rudanko	<i>A Corpus Study of Levin's Verb Classification</i> Jianguo Li, Kirk Baker and Chris Brew	<i>A Cross-sectional Analysis of Lexical Bundles in Written Medical Discourse</i> Shozo Yokoyama	<i>The Israeli-Palestinian Conflict in American, Arab, and British Media: Corpus-Based Critical Discourse Analysis</i> Magdi Kandil		
	11:00	<i>The OED as a Corpus: Looking for Dual-Form Adverbs</i> Milagros Chao Castro	<i>Collocates for Word Sense Disambiguation by means of a Discriminant Function Analysis Model: A Corpus-Based Approach</i> Moises Almela Sanchez	<i>A neo-Firthian approach to academic writing: Uncovering local patterns and local meanings in the discourse of linguistics</i> Ute Römer	<i>Using collocational profiling to investigate the construction of refugees, asylum seekers and immigrants in the UK press</i> Paul Baker and Costa Gabrielatos		
	11:30	<i>Controlling for Fads in Historical Corpora</i> Angus B. Grieve-Smith	<i>The Arrau corpus of anaphoric relations</i> Ron Artstein and Massimo Poesio	<i>Analysis of Canonical Chinese Antonym Co-occurrence</i> Xingfu Wang, Eric Ringger, Guohui Liu, and Shiping Liu	<i>Corpus-based Study of New Motherhood in Charlotte Perkins Gilman's Herland</i> Ya-Jie Chen, Hui-Chuan Lu, Kailing Liu		
	12:00	LUNCH					
		<b>ACQUISITION 4</b> (3710 Lee Library)	<b>COMPUTATIONAL 5</b> (3712 Lee Library)	<b>HISTORICAL 3</b> (3714 Lee Library)	<b>DIALECTS</b> (3716 Lee Library)		
	1:30	<i>Evaluating corpus use in language learning: State of play and future directions</i> Alex Boulton	<i>Accelerating Corpus Annotation through Active Learning</i> Peter McClanahan, Eric Ringger, Robbie Haertel, Kevin Seppi, George Busby, Deryle Lonsdale	<i>Corpus attestations and linguistic explanations: the case of the history of Spanish prepositional finite clauses</i> Manuel Delicado-Cantero	<i>Articles in Registers of Indian English: A Corpus-based Study</i> Chandrika Rogers		

2:00	<i>The Reading Class Builder: A tool for creating corpus-based teaching materials</i> Tony Berber Sardinha and Jose Lopes Moreira Filho	<i>An Efficient Query Package for Richly Annotated Discourse Corpus</i> Pengcheng Wu and Huaqing Hong	<i>Online Databases and Language Change: the Case of Spanish “dizque”</i> Viola Miglio	<i>Creole African American Vernacular English: Origins of a Dialect</i> Katherine Horwinski Healy	
2:30	<i>Corpus consultation in drafting and revising: A case study of a biomedical ESL graduate student</i> Hongmei Wu	<i>Semantic annotation of a dialog corpus</i> Silvie Cinkova	<i>Towards a Quantitative Characterization of Corpora at the Morphological Level: the use of Morphological Profiles to Measure Diachronic Change</i> Alfonso Medina	<i>The Maori presence in New Zealand English: a lexical approach</i> Marta Degani	
	<b>CORPUS CREATION 3</b> (3710 Lee Library)	<b>COMPUTATIONAL 6</b> (3712 Lee Library)	<b>VARIATION</b> (3714 Lee Library)	<b>FREQUENCY 2</b> (3716 Lee Library)	
3:30	<i>Compiling and Annotating a Corpus for Syriac</i> Eric Ringger	<i>Milk, bread and toothpaste: Adapting Data Mining techniques for the analysis of collocation at varying levels of discourse</i> Robert Sanderson, Matthew Brook O'Donnell and Clare Llewellyn	<i>The Project for the Sociolinguistic Study of the Spanish Language of Spain and the Americas (PRE-SEEA)</i> Laura Camargo	<i>A corpus-driven study of phraseological variation</i> Martin Warren	
4:00	<i>'The design of a web-based interface for the BAWE corpus'</i> Hilary Nesi	<i>Probabilistic tagging of a corpus of Mennonite Low German: A case study using QTag</i> Christopher Cox	<i>Tracking Sociohistorical Trends in the Use of Roman Letters in Chinese Newswires</i> Helena Riha and Kirk Baker	<i>Semantic Frequency: a new look at word frequency counts</i> Athelia Graham	
4:30	<i>Corpus Creation: CATE, CPEC &amp; CAHC</i> Hui-Chuan Lu, Yun-Hui Chen, Chia-Chi Tien	<i>Using a Query Language as an Annotation Tool</i> Eric J.M. Smith	<i>Verbal Imperative Variations in Qumran Legal Texts and Other Registers</i> Donald Parry	<i>Statistical Modelling of Empirical Data in Corpus Stylistics</i> Ji Meng	
5:15	Doug Biber: <i>Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles</i> (3714 Lee Library)				
6:30	DINNER				

# Abstracts

**Elisabetta Adami**

University of Verona

## **Tubing the Web: a corpus-based study on video-communication**

*YouTube*, the leading Web-site for video sharing, is daily accessed by millions of users who load their own videos and share a very peculiar way of communication; by means of a video file, language joins images and sounds and reaches virtually every (connected) corner of the world through the electronic medium, creating an intricate network of communication threads made up of both video responses and written comments. Video-interaction on the Web seems to exploit maximally the potentialities of a never-ending multi-modal, multi-directional and multi-authored hypertext, where apparently all media and modes merge together in the meaning-making process.

In this view, it becomes apparent that a purely linguistic approach cannot account for the phenomenon; indeed, any attempt at investigating this new form of communication must necessarily consider its very nature of multimodal form of interaction. Therefore, investigating it implies combining the methods of corpus analysis with the categories of multimodal analysis, together with the tools of conversation analysis and with the studies on computer-mediated communication.

Bearing this in mind, my paper will examine a corpus of videoclips, starting from a video randomly selected on *YouTube* and following the communication thread resulting from its related written comments and video responses.

Firstly, the analysis will focus on the peculiar features of video-communication, so as to hypothesize possible similarities and differences with real-world interactional exchanges and other well-established computer-mediated communication forms.

Secondly, the paper will highlight the conversational patterns that emerge in the thread, focussing particularly on the opening and closing formulas represented in the videos.

Thirdly, the corpus will be analysed in its multimodal deployment in order to investigate the ways in which the semiotic resources employed in video-interaction are combined in new and/or recognisable patterns of representation and meaning-making.

Finally, setting aside any pretension of representativeness and significance of the re-

sults, the paper will hint at some of the features that are apparently distinctive of this new form of interaction, so as to postulate some starting points for further research on the subject.

#### Selected references

- Aston, G. & Barnard, L. (eds.) (2001). *Corpora in the Description and Teaching of English*. Bologna: CLUEB.
- Baldry, A. & P. J. Thibault (2006). *Multimodal Transcription and Text Analysis. A Multimedia Toolkit and Coursebook*. London/Oakville: Equinox.
- Barnes, S. B. (2003). *Computer-Mediated Communication: Human to Human Communication across the Internet*. Boston: Allyn and Bacon.
- Finnegan, R. (2002). *Communication*. London/New York: Routledge.
- Goodman, S & Graddol, D. (eds.) (1996). *Redesigning English: New texts, new identities*. London/New York: Routledge.
- Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Arnold.
- Halliday, M. A. K., & Hasan, R. (1985). *Language Context and Text: Aspects of Language in a Social-semiotic Perspective*. Oxford, UK: Oxford University Press.
- Herring, S. C. (ed.) (1996). *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*. Amsterdam: Benjamins.
- Hundt, M., N. Nesselhauf & C. Biewer (eds.) (2007). *Corpus Linguistics and the Web*. Amsterdam/New York: Rodopi.
- Knapp, M. L. & J. A. Daly (eds.) (2002, 3<sup>rd</sup> ed.). *Handbook of Interpersonal Communication*. London: Sage.
- Kress, G. & T. van Leeuwen (2006, 2<sup>nd</sup> ed.). *Reading images. The grammar of visual design*. London/New York: Routledge.
- O'Halloran, K. L. (ed.) (2004). *Multimodal Discourse Analysis. Systemic-Functional Perspectives*. London/New York: Continuum.
- Snyder, I. (ed.) (1998). *Page to Screen. Taking Literacy into the Eelectronic Eera*. London/New York: Routledge.
- Thompson, G. & S. Hunston (eds.) (2005). *System and Corpus: Exploring connections*. London/New York: Equinox.
- Thurlow, C., L. Lengel & A. Tomic (2004). *Computer-Mediated Communication. Social Interaction and the Internet*. London: Sage.

### **Eman Al Nafjan**

King Saud bin Abdulaziz University for Health Sciences

#### **The Nora Corpus: a study of Arab EFL written discourse**

This paper presents an initial semantic and stylistic analysis of how Arabic language and culture influences the written discourse of Saudi EFL students. It is an analysis of a small-size corpus of daily journal entries written by Saudi student nurses for an English class. The analysis centres on the features and characteristics that make this particular genre of Arab EFL discourse discernible from its counterpart written by other L1 speakers. In what ways has the students' first language stamped their EFL production, whether it is in

the syntactic sentence structure or in their semantic choices. And should this influence be considered a pedagogically targeted area?

The software program Wordsmith Tools has been employed to investigate the frequency of word items and frequent and infrequent keywords that were then compared to the BNC. Also a small manual investigation of direct translations from Arabic was conducted with reference to the Bank of English. The results show that the participants have numerous areas of Arabic influence mainly in the stylistic and religious dimensions of their writing. It is argued that although the influence of Arabic language and culture is apparent in the learner corpus, not all of this influence should be viewed as pedagogically adverse.

## **Mohammed Albakry**

Middle Tennessee State University

### **Social Taboo, Evaluation, and Identity Construction Online**

The present paper examines the evaluative language of online message board groups engaged in communicating about a social taboo (having an extramarital affair, inducing vomiting to lose weight, etc), i.e. practices considered objectionable or undesirable by society and thus associated with a sense of shame (Frazer, 1990). Talking about issues of social taboo constitutes a face threatening act to participants (Goffman, 1963; Brown & Levinson, 1987) and thus it is important to study how language is used and managed to mitigate negativity and face threats in such sensitive discourse situations. Based on a personally collected corpus composed of more than 100,000 words from different board message archives and postings, the methodology of this study has two major levels: pragmatic and linguistic. First the pragmatic/social macro level explores the interactional context that produced the discourse. This level of analysis draws on Scollon's theory (1998) of Mediated Discourse Analysis (MDA), and Politeness Theory (Brown & Levinson, 1987). Second, the linguistic micro-level explores specific linguistic devices or strategies of evaluation employed in the texts. This level of analysis, informed by Tannen's work (1997) on discourse frames and Martin and White (2005) on linguistic appraisal, investigates the use of such language resources as stance adverbials, hedges, speech acts, reported speech, negations, pronouns used as references to self and others, and use of the internet-specific punctuation of emoticons. Two main research questions are addressed: (1) how do the narratives of the group members in fraught discourse situations reflect their self-evaluation and construct their textual personas? (2) what are the different communicative strategies by which posts and threads of reply create "support" for the self and the group members?

**David Alexander**  
Ohio State University

**El coche ese vs el coche de marras:  
The postnominal demonstrative and de marras in Diachronic and Synchronic Corpora**

Modern Spanish has two referring expressions to refer to given information:

1) *Dame la carta<sub>i</sub>. La carta esa<sub>i</sub> que te mencioné*  
V-give-CL-1<sup>st</sup>-s ART-the letter ART-the letter DEM-that CP-that CL-2<sup>nd</sup>-s V-1<sup>st</sup>-s-mentioned  
(Gutiérrez-Rexach 2002)

2) *[el Quijote]<sub>i</sub> [el libro de marras]<sub>i</sub>.*  
ART-the book PREP-of before.  
(CREA)

In contrast to 1a) the pre-nominal demonstrative is the default form across Romance languages (Bernstein 2001). The literature identifies this placement as having a special function where the Spanish postnominal demonstrative has a pejorative reading tied to the medial *ese* (Gutiérrez-Rexach 2002). An aim the present study is two fold: 1) to clarify the role of the postposed demonstrative and its “pejorative” use and 2) determine the origin and function of the PP *de marras*.

I investigate the postnominal demonstrative 1a) and the PP *de marras* 1b) in diachronic and synchronic corpora (CREA, CORDE, www.corpusdelespanol). This analysis is carried out via discourse frameworks: Prince 1981’s Givenness, Gundel et. al 1994’s Givenness Hierarchy in addition to the Neo-Gricean Framework Blackwell 2003. My analysis reveals the postnominal demonstrative (4) and *de marras* (5) encoders of hearer-old information.

I argue that the postnominal demonstrative is not pejorative but is used to encode hearer-old information: Anaphoric, Shared Knowledge, and Situationally Accessible In contrast to the adjoining PP *de marras* which is inherently pejorative (5).

The postnominal demonstrative first appears in the 14th-15th centuries in an inverted string:

6) a. *La tenor de la carta esta*  
ART-the tone PREP-of ART-the letter DEM-this  
(Biblia Latina, Anónimo. 1300-1400) (CORDE)

In contrast, the construction *de marras* comes from the Arabic *marras*-once, one time which first appears in the 13th century (Corominas Pascual 1980) which is then incorporated into a PP adjunct from the 15th century onwards.

The postnominal demonstrative and the PP *de marras* serve as referential expressions in the modern language. Both expressions model the addressees model of discourse, *de marras* does so in a more overt fashion via a lexical term. The lexical PP has a more affective function than syntactic inversion given that lexical terms are fixed whereas the interpretation of syntax depends on shifting discourse frames (Fillmore 1983).

While both expressions are “colloquial” *de marras* appears more regularly in the diachronic record than the postnominal demonstrative, which has a dearth of examples in the 16th and 17th centuries, given that the discourse context vital to the use of the inverted demonstrative is largely absent from such written corpora.

## Patricia Amaral and Chad Howe

University of Coimbra

### Using corpora to quantify contexts: The case of the Portuguese Perfect

Among the Present Perfects (PP) in Romance, the Portuguese PP is unique in requiring iteration of the predicate, as in (1), which can only have the reading in (1)a. Crucially, the Portuguese PP cannot have the resultative reading shown in (1)b and, unlike other Romance PPs, does not display the full range of prototypical meanings (Comrie 1976). In this paper, we present both synchronic and diachronic evidence from *O Corpus do Português* (Davies & Ferreira 2006) to study patterns of collocations relating to the Portuguese PP and their influence in the process of semantic change that the PP underwent.

- (1) Que autores brasileiros o sr. tem lido? (BR-Oral, 1997)
- |    |   |              |
|----|---|--------------|
| a. | ‘Which Brazilian authors have you been reading lately?’ | Iterative    |
| b. | *‘Which Brazilian authors have you read (once)?’        | *Resultative |

We use empirical data from historical corpora of Portuguese to disprove the proposal made by Harris (1982) with respect to his classification of perfects across Romance languages and their respective degrees of grammaticalization. Contra Harris, it is shown that the iterative interpretation is an innovative rather than a conservative feature of the Portuguese PP. Second, we show that an analysis of the contexts of use of the PP, as provided by the corpus data, proves crucial to understanding the processes of semantic change in this case. More specifically, the iterative reading arises as an invited inference licensed by contextual information, as demonstrated via different collocational tendencies (e.g. co-occurrence with adverbials), and eventually becomes conventionalized over time.

This analysis further suggests a number of innovations in the use of corpora for

the analysis of pragmatic phenomena in diachrony. Utilizing large-scale, searchable corpora greatly facilitates the ability to quantify contextual features concomitant with processes of semantic change.

## **Aroldo Andrade**

State University of Campinas (Unicamp)

### **A corpus-based research on the history of clitic climbing in European Portuguese**

The collocation of clitic object pronouns is a main research topic in Portuguese syntax, not only because it is a relevant criterion distinguishing the grammars of Brazilian and European Portuguese, but also in that the latter has changed in the beginning of the eighteenth century from a scenario of predominant proclisis to another in which enclisis is the non-marked option (Galves, Britto & Paixão de Sousa 2005). While previous studies have mainly dealt with independent clauses, this research focuses on embedded infinitive clauses forming a complex predicate with a finite verb, also referred to as a restructuring predicate. In this context, it is possible to find the phenomenon of clitic climbing, namely, the movement of an embedded clitic towards the domain of its matrix clause. In order to do so, we have collected data from the Tycho Brahe Historical Corpus of Portuguese, comprising texts from the sixteenth until the nineteenth century. The data were obtained automatically from 23 tagged texts by means of Perl scripts and from 2 parsed texts using the CorpusSearch software (Randall 2000). The work hypothesis consists in verifying whether the collocation of clitics in restructuring contexts follows the same principles valid in independent clauses. Therefore, the orders in (i) are expected in most obligatory proclisis contexts, while the options in (ii) should be found in variation contexts.

- (i) Clitic collocation in obligatory proclisis contexts
  - a. (attractor) clitic – finite verb – infinitive verb (João não as quis enviar)
  - b. (attractor) finite verb – infinitive verb + clitic (João não quis enviá-las)  
‘John did not want to send them’
  
- (ii) Clitic collocation in variation contexts
  - a. clitic – finite verb – infinitive verb (João as mandou apagar)
  - b. finite verb + clitic – infinitive verb (João mandou-as apagar)
  - c. finite verb – infinitive verb + clitic (João mandou apagá-las)  
‘John wanted to send them’

## **Ron Artstein and Massimo Poesio**

University of Southern California, University of Essex, and Università di Trento

### **The Arrau corpus of anaphoric relations**

The Arrau corpus of anaphoric relations was created at the University of Essex between 2004 and 2007. It introduces an annotation scheme specifically targeted at marking two phenomena which had been difficult to annotate: ambiguous expressions which may refer to more than one object from previous discourse, and expressions which refer to abstract entities such as events, actions and plans. The corpus consists of a mixture of genres: task-oriented dialogues from the Trains-91 and Trains-93 corpus, narratives from the Gnome corpus and English Pear Stories corpus, and newswire from the Wall Street Journal portion of the Penn Treebank.

The corpus was created using the MMAX2 tool (Mueller and Strube 2003) which allows marking text units at different levels. Each noun phrase is marked as either anaphoric, discourse-new, or non-referential. Antecedents of anaphoric NPs are marked by pointers, and anaphoric ambiguity is indicated by multiple pointers from a single anaphoric expression (Poesio and Artstein 2005). Reference to an event, action or plan is marked by a pointer from the referring NP to the clause that introduces the abstract entity (Artstein and Poesio 2006).

The Arrau corpus differs from existing corpora like MUC and ACE since it marks all NPs, not only those that refer to entities of interest like people and organizations. The annotation is richer than a division of NPs into equivalence classes which refer to the same object, but it can be converted into equivalence classes by removing ambiguous links. The corpus has been used in the development of the anaphora resolution system at the 2007 Johns Hopkins summer workshop on natural language engineering; we plan to release it to the public in the coming months.

### **References**

- Ron Artstein and Massimo Poesio. Identifying reference to abstract objects in dialogue. Brandial 2006 proceedings, pages 56-63, Potsdam, Germany, September 2006.
- Christoph Mueller and Michael Strube. Multi-level annotation in MMAX. Proceedings of the 4th SIGDIAL, pages 198-207, 2003.
- Massimo Poesio and Ron Artstein. Annotating (anaphoric) ambiguity. Corpus Linguistics proceedings, Birmingham, England, July 2005.

**Harald Baayen**  
University of Alberta

**Co-occurrence below and above the word level:  
exploring language at the intersection of corpus linguistics,  
psycholinguistics and statistics**

This paper addresses co-occurrence patterns below and above the word level. Below the word level, paradigms are the traditional structure for organizing inflectionally related words (e.g., scissor/scissors, walk/walks/walked). Recent psycholinguistic experiments (see, e.g., Moscoso del Prado et al. (2005) and Milin et al., 2007) have revealed that the processing of a given inflectional form (e.g., walks) is co-determined by the other forms in its inflectional paradigm, and that information-theoretical measures based on an extending Shannon's entropy provide an excellent tool for gauging the relevance of paradigmatic structure for lexical processing in both comprehension and production. In short, below the word level, a collocate of morphemes (such as walk-s) is not processed in isolation, but against the backdrop of its other morphemic collocates.

The hypothesis that I will explore in this presentation is that the same holds for co-occurrence above the word level. I take as point of departure the collocation approach described in Stefanowitch and Gries (2005). In their distinctive collexeme analysis of the dative alternation in English, for instance, they pit the frequencies of use of the 'paradigmatic' alternatives for a given verb (the double object construction versus the prepositional object construction) against the joint frequencies of use of the same alternatives realized with any other verb. The resulting 2 by 2 contingency table is then evaluated by means of the Fisher exact test of independence, and the resulting p-value is used to rank verbs with respect to which paradigmatic alternative they prefer.

Entropy measures may also help to come to grips with variation in co-occurrence patterns. For instance, entropy measures may capture variation in collocation preferences over and above grammatical constraints pertaining to syntactic weight, recency, animateness and definiteness in a generalized linear mixed model for the likelihood that the double object (or the prepositional object) construction is selected to express the dative in English. To illustrate this, I will make use of the data of Bresnan et al. (2007), which are available in the languageR package on the R CRAN archives (cf. Baayen, 2008).

Considered jointly, it seems that morphemes in words, and words in sentences, are processed on the basis of very similar, perhaps even the same probabilistic mechanisms, mechanisms which are exquisitely sensitive to the range of contexts in which a linguistic unit appears, and that combine fine-grained local contextual information with more global generalizations across exemplars in the mental lexicon and the mental construction (see also Levy, 2008). The challenge for future research is to develop both more refined measures and improved experimental techniques to trace the role of what is not overtly ex-

pressed in the signal on the production and comprehension of that signal, both below and above the word level.

### References

- Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R*. Cambridge University Press.
- Bresnan, J., Cueni, A., Nikitina, T. and Baayen, R. H. (2007) Predicting the dative alternation, in Bouma, G. and Kraemer, I. and Zwarts, J. (eds.), *Cognitive Foundations of Interpretation*, Royal Netherlands Academy of Sciences, 69-94.
- Levy, R. (2008) Expectation-based syntactic comprehension. Manuscript, University of Edinburgh.
- Milin, P., Filipovic Durdevic, D. and Moscoso del Prado Martin, F. (2008). The Psychological Reality of Inflectional Paradigms. To appear in *Journal of Memory and Language*.
- Moscoso del Prado Martin, F., Kostic, A. and Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing, *Cognition* 94, 1-18.
- Stefanowitsch, A. and Gries, St. Th. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1, 1-44.

## Paul Baker and Costa Gabrielatos

Lancaster University

### Representations of Islam in the British and American press 1999-2005

This paper describes the analysis of an 87 million word corpus of British newspaper articles and a 40 million word corpus of American newspaper articles which refer to the subject of Islam. Comparisons were made of different newspapers according to following major categorizations: 1) tabloids vs. broadsheets, 2) British vs. American news 3) before and after 9/11.

In order to examine representations of Islam and Muslims, the corpus was subjected to a comparative analysis, by analysing the lexis that was used most significantly in the tabloid articles, when compared to the broadsheets, and vice versa (Scott 2002, Baker 2006). Keywords, collocates and concordances were used in order to make sense of the data. Tabloids tended to focus on emotive stories about a small number of Muslim terrorists, whereas broadsheets reported stories about Muslims in a wider range of contexts, the differences between the newspapers becoming more apparent after 9/11.

When comparing British and American stories, moral panic theory (Cohen 1987, McEnery 2005) was used in order to categorise keywords between the two cultures – revealing some interesting differences in the ways that moral panics around Islam developed.

This paper also raises questions about the effects of linguistic priming on readers, and whether a corpus analysis is able to quantify bias (and indeed what constitutes bias).

## References

- Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.
- Cohen, S. (1987) *Folk Devils and Moral Panics*. London: Routledge.
- McEnery, A. (2005) *Swearing in English*. London: Routledge.
- Scott, M. (2002) 'Picturing the key words of a very large corpus and their lexical upshots – or getting at the Guardian's view of the world' in B. Kettemann & G. Marko (eds.) *Teaching and Learning by Doing Corpus Analysis*, Amsterdam: Rodopi, pp. 43-50

## Paul Baker and Costa Gabrielatos

Lancaster University

### Using collocational profiling to investigate the construction of refugees, asylum seekers and immigrants in the UK press

As the representation of minority groups in the press can construct their identity (e.g. Duffy and Rowden, 2005: 6, in Greenslade, 2005: 7), the discourses surrounding these groups have been the focus of linguistic studies (e.g. ter Wal, 2002). This paper reports on the ESRC funded project, 'Representation of refugees and asylum seekers in UK newspapers 1996-2005', which used a corpus of 140 million words (175,000 articles from 15 UK newspapers), spanning 1996-2005. The project combined critical discourse analysis and corpus linguistics approaches (Baker et al., forthcoming; Gabrielatos & Baker, forthcoming). However, this paper focuses on the contribution of corpus research to (critical) discourse analysis (e.g. Koller & Mautner, 2004; Orpin, 2005; Sotillo & Wang-Gempp, 2004), and, more specifically, on the collocational analysis of the words *refugees*, *asylum seekers*, *immigrants* and *migrants* (RASIM).

The analysis adopted the methodology in Baker & McEnery (2005) and McEnery (2006), adding the notion of *consistent collocates* (akin to *key keywords* (Scott, 2004: 115)), i.e. collocates present in at least seven out of the ten annual sub-corpora. Collocates are a suitable vehicle for the discursive presentation of a group (Baker, 2006), because they can contribute to "a semantic analysis of a word" (Sinclair, 1991: 115-116), and because "they can convey messages implicitly and even be at odds with an overt statement" (Hunston, 2002: 109). The analysis also employs the related notions of *semantic prosody* (Louw, 1993: 157), *semantic preference* (Stubbs, 2001: 65), and *discourse prosody* (ibid.: 65-66).

The clustering of consistent collocations provided evidence of systematic semantic associations, which, moreover, map onto the CDA notions of *topos* (Reisigl & Wodak, 2001: 74-76) and *topic* (Sedlak, 2001: 129-130), as well as metaphors commonly employed in racist discourse (van der Valk (2000: 234). Arguably, these patterns reveal elements of the underlying discourses relating to RASIM.

## References

- Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.
- Baker, J.P. & McEnery, A.M. (2005). A corpus-based approach to discourses of refugees and asylum seekers. *Journal of Language and Politics* 4(2), 197-226.
- Baker, P., Gabrielatos C., Khosravini, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (forthcoming). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Submitted to *Discourse & Society* (under review).
- Gabrielatos, C. & Baker, P. (forthcoming). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996-2005. *Journal of English Linguistics*.
- Greenslade, R. (2005). Seeking scapegoats: The coverage of asylum in the UK press. Asylum and Migration Working Paper 5. Institute for Public Policy Research.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Koller, V. & Mautner, G. (2004). Computer applications in critical discourse analysis. In Coffin, C., Hewings, A. & O'Halloran, K. (eds). *Applying English Grammar: Functional and corpus approaches*. London: Hodder and Stoughton, 216-28.
- Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker, M., Francis, G. & Tognini-Bonelli, E. (eds.) *Text and technology: In honour of John Sinclair*, Philadelphia and Amsterdam: John Benjamins, 157-176.
- McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present*. London: Routledge.
- Orpin, D. (2005). Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze. *International Journal of Corpus Linguistics* 10(1), 37-61.
- Reisigl, M. & Wodak, R. (2001). *Discourse and Discrimination: Rhetorics of racism and anti-semitism*. London and New York: Routledge.
- Sedlak, M. (2001). You really do make an unrespectable foreign policy. In Wodak, R. & van Dijk, T.A. (eds.) *Racism at the Top: Parliamentary discourses on ethnic issues in six European states*, 107-168.
- Sinclair, J.McH. (1991) *Corpus Concordance Collocation*. Oxford: Oxford University Press.
- Sotillo, S.M. & Wang-Gempp, J. (2004). Using corpus linguistics to investigate class, ideology, and discursive practices in online political discussions: Pedagogical applications of corpora. In Connor, U. & Upton, T.A. (eds.) *Applied Corpus Linguistics*, 91-122.
- Scott, Mike. 2004. *Oxford WordSmith Tools Version 4*. Oxford: Oxford University Press. Available online: <http://www.lexically.net/downloads/version4/wordsmith.pdf>
- Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.
- ter Wal, J. (2002). Racism and cultural diversity in the mass media: An overview of research and examples of good practice in the EU Member States, 1995-2000. European Research Centre on Migration and Ethnic Relations (ERCOMER).
- van der Valk, I. (2000). Parliamentary discourse on immigration and nationality in France. In Wodak, R. & van Dijk, T.A. (eds.) *Racism at the Top: Parliamentary discourses on ethnic issues in six European states*, 221-260.

## **Del Barrett**

King's College London

### **Parents, victims, suspects, victims ... : The 'framing' of the McCanns in the British tabloid press**

On 3 May 2007, a three-year old girl, Madeleine McCann, disappeared whilst on holiday with her parents in Portugal. The police initially treated the incident as a missing person case. The McCanns took the unusual step of instigating a high profile publicity campaign in an inter-continental series of press conferences, deliberately courting the media in order to keep Madeleine's name and picture on the front page. Four months after her disappearance, her parents were questioned, as *arguido*, following the alleged discovery of forensic evidence that they had murdered their daughter. A few weeks later a Portuguese judge indicated that the McCanns were no longer wanted for questioning, and once again the search for Madeleine began.

Through interrogation of a corpus of press articles, this study examines how the British tabloids portray the McCanns throughout this time and investigates whether the press really shows unfaltering solidarity with the parents, as suggested by headlines such as 'Kate's tears of anguish'.

The methodology involves a technique termed 'frame-tracing', which is based on combining the principles of 'lexical priming' (Hoey, 2005) and 'semantic prosody' (Louw, 1993) with the concept of 'six degrees of separation'. Frame-tracing uses chains of collocates to examine intertextuality in media discourse, in order to find other impressions, allusions and connotations, which, when taken together, construct an alternative reading to that which is expressed in the headlines. Such frames are virtually undetectable in individual texts. In a corpus of news reports, however, the emergence of such frames, and the way that they are constructed over time, can be observed. In some cases, there are only traces of a particular frame, but in others, the alternative associations are ubiquitous.

The results show a marked difference in the way that the tabloids represent the McCanns. Some publications do indeed show complete solidarity. Others, however, keep their options open by overtly showing support and yet covertly having already found the McCanns guilty, thus foreshadowing that well-known tabloid phrase: 'We told you so'.

## Sabine Bartsch

Technische Universität Darmstadt

### Verb distribution and co-occurrence in scientific text

Studies of scientific text tend to focus on those items that are deemed to hold information regarding scientific concepts, i.e. the focus tends to be on the nouns. In conjunction with the often made observation that nominalisation is highly indicative of scientific writing and that nominal forms are found with notable frequency in scientific text these nouns have received a lot of due attention especially in the context of knowledge representation and information extraction. Yet, it is indisputable that verbs are likewise central to the linguistic construal of scientific meaning, especially in view of the fact that verbs typically establish relations between concepts denoted by nouns.

This paper seeks to show that the distribution and co-occurrence of verbs is highly domain-specific and that linguistic knowledge can be exploited to extract and model (taxonomies, ontologies) domain-knowledge based on appropriately annotated corpora of authentic text.

To this end, the paper presents a corpus study of verbs in scientific and academic texts and shows their contribution to the construal of domain-specific meaning based on (a) their distribution and clustering in texts and (b) their characteristic patterns of co-occurrence within the predicate-argument structure. The study shows how patterns and alternations in different syntactic-semantic classes of verbs are characteristic of texts from particular domains and indicate shifts of information within the text (e.g. the clustering of verbs such as ‚be‘, ‚refer to‘ and other so-called relational verbs (Halliday 2004; Halliday, Matthiessen 1999) in sections presenting definitions). The second perspective (b) of the paper is concerned with the characteristic collocational patterning of those verbs within the predicate argument structure. It can be shown how these patterns can aid the identification of taxonomic as well as non-taxonomic relations in texts from special domains. The paper shows how collocational relations may be employed in building linguistically informed models of domain-knowledge.

### References

- Bateman, John. 1990. "Upper Modelling: A general organization of knowledge for natural language processing." Paper prepared for the Workshop on Standards for Knowledge Representation Systems, Santa Barbara, March, 1990. In: Proceedings of the International Language Generation Workshop (Pittsburgh, June 1990).
- Halliday, MAK, Christian M.I.M. Matthiessen. 1999. *Construing Experience through Meaning*. London, New York: Continuum.
- Halliday, M.A.K. 2004. *An Introduction to Functional Grammar*. Third edition revised by C. Matthiessen. London: Arnold.

**Anna Belladelli**

University of Verona

### **American slang in mainstream magazine writing**

Semantic proliferation and word class shifting are some of the main features of language change. If the language under scrutiny is American slang, the signs of such evolution can be traced even from a short-term diachronic perspective. Most slang words and expressions are short-lived, or rapidly lose their grip on speakers, thus becoming unfashionable and obsolete, or even dying out. Conversely, other words, also defined as *basic slang lexemes* (Moore 2004), happen to outlive the generation or sociocultural context that created them, and they begin to accumulate an increasing number of meanings, referents and uses through time. Indeed, since slang as a whole serves a variety of social and psychological functions, both for single individuals and for groups – as first outlined by Mencken (1921) and Partridge (1933), and later on by Drake (1980), Lighter (1994) and Eble (1996) – different speech communities which take on exploiting the same word will contribute their own ‘layer’ to its meaning, use and function.

Drawing on ‘mainstream’ US magazines (e.g. *Time*), a number of slang lexemes will be retrieved and analyzed. Their diachronic development throughout the 20<sup>th</sup> century will be outlined, as well as their current status as slang vocabulary, and their use in magazine writing. The present paper aims at allowing for a better understanding of how, if at all, mainstream media are able to catch up with this specific aspect of language change. In particular, it analyzes which ‘layers’ of slang permeate mainstream magazine writing, in which contexts such vocabulary is allowed, and what might be the social and cultural implications of this selective lexical appropriation.

#### Selected references

- Drake G. F. (1980), “The Social Role of Slang”, in H. Giles, W. P. Robinson, and P. M. Smith (eds.), *Language: Social Psychological Perspectives*. New York: Pergamon Press. 63-70.
- Eble C. (1996), *Slang and Sociability*. Chapel Hill and London: The University of North Carolina Press.
- Lighter J. E. (1994), *Random House Historical Dictionary of American Slang*. New York: Random House.
- Mencken H. L. (1919), “American Slang”, in Id., *American Language*. New York: Knopf. 555-589.
- Moore R. L. (2004), “We’re Cool, Mom and Dad Are Swell: Basic Slang and Generational Shifts in Values”. *American Speech*, 79(1): 59-86.
- Partridge E. (1933), *Slang Today and Yesterday*. London: Routledge.

**Tony Berber Sardinha**

Pontifical Catholic University of Sao Paulo

**The CEPRIL Metaphor Candidate Identifier:  
A program for identifying metaphor in corpora**

Traditionally, retrieving metaphor from corpora has been carried out with general-purpose corpus linguistics tools, such as concordancers, wordlisters, and frequency markedness identifiers. With a word frequency list or frequency markedness list, analysts choose words with metaphoric potential and then run concordances for these words. Analysts may also include other words of interest based on previous literature, a partial reading of the corpus data, their experience or intuition. The problem with these procedures is that they are biased, since they restrict the range of retrievable metaphors to those signaled by familiar terms, words that have received attention in previous research or have been noticed by reading portions of the corpus, and so on. To reduce such bias, we need a computer tool that can process a whole corpus, evaluate the metaphoric potential of each word in the corpus, and then inform researchers about which words seem to have a greater probability of metaphor use, so that they can choose which words to look at in detail in the corpus. In this paper, I present an online metaphor identification program that retrieves potential metaphors from both English and Portuguese corpora. This is currently the only publicly available such tool. The program works by analyzing the patterns (bundles and collocational framework) and part of speech of each word and then matching these patterns to the information in databases, which contain several kinds of information about lexis and its relationship to metaphor, all automatically extracted from extensively hand-annotated corpora. The result is an output in which each word is listed with its metaphor probability. Researchers can then inspect this output and arguably make more informed decisions about which words to focus on in their analysis.

**Douglas Biber**

Northern Arizona University

**Merging corpus linguistic and discourse analytic research goals:  
Discourse units in biology research articles**

The present study addresses the following general question: Can the goals and methods of discourse analysis be reconciled with the goals and methods of large-scale corpus-based analysis? That is, is it possible to uncover generalizable patterns of discourse organization,

based on detailed analysis of individual texts but at the same time based on analysis of all texts in a corpus?

The talk first briefly introduces two general approaches to this research problem: top-down and bottom-up analysis. Then, a particular framework for bottom-up corpus/discourse analysis is described, illustrated through an investigation of the patterns of discourse organization in a corpus of biology research articles:

- First, each text in the corpus is segmented into vocabulary-based Discourse Units using computational techniques.
- Second, analysis of all Discourse Units in the corpus is undertaken to identify the underlying Discourse Unit Types in biology research articles, based on their primary linguistic characteristics (using Multi-Dimensional analysis).
- Third, the Discourse Unit Types are interpreted in functional terms.
- Fourth, the analysis returns to the discourse organization of individual texts, analyzing their discourse structure as a sequence of Discourse Units, shifting among various Discourse Unit Types.
- Finally, the preferred general patterns of discourse organization in this genre are identified and interpreted.

## **Ken Bloom**

Illinois Institute of Technology

### **Learning Appraisal Extraction Patterns**

Research on evaluative language has worked to understand the linguistic resources by which evaluations can be expressed, and how such evaluations operate in a sociolinguistic context. Evaluative language is key to many sorts of discourse in politics, ideology, and marketing. Work on local grammars of evaluation (e.g. Hunston and Sinclair, 2000) has elucidated many recognizable patterns by which evaluative attitudes are associated with their targets in text. However, discovering such patterns by hand is a time-consuming process, which we seek to augment by automatically methods to help find these patterns.

The new area of *sentiment analysis*, which seeks to analyze opinion in natural language text using computers, includes a variety of interesting problems, including classifying reviews as positive or negative based on the opinions expressed within the reviews, performing data mining on opinions to understand public sentiment about parts of a product, analyzing political discourse, and using sentiment to predict the success of movies.

In past work, we have developed a computerized system to extract *appraisal expressions* consisting of an *attitude* and a *target*, which are connected using a syntactic linkage pattern. Our original system used manually-constructed lexicons and linkage rules tuned

to specific genres of texts (e.g., movie reviews). We are now extending this work to extract candidate attitudes and targets first, and uses these to bootstrap discovery of new linkage patterns in a corpus in an unsupervised fashion. These patterns constitute candidates for linguists to develop local grammars of evaluation, as well as a resource for automated text analysis. The extracted rules can be evaluated both by manual examination for syntactic/semantic coherence, and by applying them to sentiment classification of movie and product reviews. We will present our automated discovery technique, evaluation results, and examples of the discovered patterns.

## Alex Boulton

CRAPEL-ATILF/CNRS, Nancy Université

### Evaluating corpus use in language learning: State of play and future directions

Much discussion in language learning and teaching revolves around two main questions: *what* to learn and *how* to learn it. Corpus linguistics has had a significant impact on the former, as publishers use the findings to inform the language content of their syllabuses. On the other hand, and despite significant output in terms of research articles and publications, few teachers and still fewer learners have had any direct contact with language corpora per se. For such awareness to filter down to the end users, the major players upstream (publishers, software developers, politicians, schools, teacher trainers, etc.) have to be convinced that corpora have something to offer. While well-developed theoretical arguments show the potential is considerable, empirical evidence is surprisingly lacking (Chambers 2007).

This paper looks at over 50 studies which provide some form of evaluation of the use of corpora in language learning. Many are essentially qualitative, bringing useful insights but remaining by definition rather subjective; quantitative evidence, however, remains scarce. Furthermore, the focus is often on learners' and teachers' attitudes to the use of corpora, or on their ability to use corpus techniques, rather than on the effectiveness of corpus use. Finally, the remaining studies tend to divide into two main categories: those which are concerned with the use of corpora as a reference tool, especially in written production (including error-correction and translation); and those few which actually look at whether corpora contribute to language learning itself – “data-driven learning” proper, to use Johns' term (e.g. 1991).

This paper synthesises the research findings to date on all these questions, and argues that we cannot expect such techniques to become part of mainstream language teaching and learning practices without substantial empirical support.

## References

- Chambers, A. 2007. Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda & J. Santana (eds) *Corpora in the Foreign Language Classroom*. Amsterdam: Rodopi, p. 3-16.
- Johns, T. 1991. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns and P. King (eds) *Classroom Concordancing*. *English Language Research Journal* 4, p. 27-45.

**Laurel J. Brinton**

University of British Columbia

**Historical Pragmatics and Corpus Linguistics: Problems and Strategies**

Corpus linguistics is “the *sine qua non* of historical linguistics” (McEnery and Wilson 2001:123). Contemporary corpus linguistics has led to significant advances in historical linguistics, most notably in the speed and ease with which data can be retrieved. The English historical linguist has available for use a wide variety of corpora. However, none is entirely ideal. Only two corpora, the *Oxford English Dictionary* and the *Helsinki Corpus*, provide the full diachronic span from Old English to the present day. The OED quotation bank, though not a corpus strictly speaking, can—with caution—be fruitfully used by the historical linguist (Hoffmann 2004). At only 1.5 million words for 1000 years of language history, the *Helsinki Corpus*, a balanced general-purpose corpus, may prove too small for some types of searches. Apart from these sources, the historical English linguist must cobble together a variety of corpora from the individual periods of English, ranging from the *Dictionary of Old English Corpus* containing almost all extant Old English texts, to the *Middle English Dictionary* (sharing many of the weaknesses of the OED), to the rich Chadwyck-Healey corpora designed primarily for the literary scholar (and quite user-unfriendly for the linguist).

After a review of the historical corpora available to the English linguist, this paper explores some of the problems encountered by a scholar wishing to apply corpus linguistic methodology in the field of historical pragmatics. I articulate the strategies that I have adopted in my work on pragmatic markers and, more recently, on comment clauses in the history of English (Brinton forthcoming). As a case study, I explore the development of the comment clause (*as*) *you say* in the history of English. The use of a mixed qualitative/quantitative corpus-based approach allows for a detailed, empirically based description of the rise of (*as*) *you say*; at the same time, it permits testing of the “matrix clause hypothesis”, the prevailing theory concerning the origin of comment clauses that has been extrapolated from Thompson and Mulac’s synchronic work on *I think/guess*. Frequency counts of the presumed source construction (i.e., *you say that* S) in the earlier periods cast

doubt on the validity of the matrix clause hypothesis. Corpus data suggest a more nuanced view of the rise of this comment clause, namely, that a variety of structures, including relative/adverbial *as you say*, main clause *you say*, and *you say* following a fronted element all contributed to its genesis.

### References

- Brinton, Laurel J. Forthcoming (2008). *The Comment Clause in English: Syntactic Origins and Pragmatic Development*. (Studies in the English Language.) Cambridge: Cambridge University Press.
- Hoffmann, Sebastian. 2004. Using the OED quotations database as a corpus—a linguistic appraisal. *ICAME Journal* 28:17-30.
- McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. 2nd ed. (Edinburgh Textbooks in Linguistics.). Edinburgh: Edinburgh University Press.
- Thompson, Sarah and Anthony Mulac. 1991. A quantitative perspective on the grammaticalization of epistemic parentheticals in English. In Elizabeth Closs Traugott and Bernd Heine, eds. *Approaches to Grammaticalization*, Vol. 2, 313-329. Amsterdam and Philadelphia: John Benjamins.

## Laura Camargo

Universitat de les Illes Balears

### Project for the Sociolinguistic Study of the Spanish Language of Spain and the Americas (PRESEEA)

The aim of this paper is to present the *Project for the Sociolinguistic Study of the Spanish Language of Spain and the Americas* (PRESEEA)<sup>1</sup>. PRESEEA represents the creation of a macrocorpus of different urban varieties of Spanish spoken today in the world. PRESEEA is a global and open project –which in a short period of time is expected to become the most significant of its kind– with participants of universities and linguistic institutes all throughout the Hispanic world, including cities in the USA like Miami. Its main goal is to set up a sociolinguistic documentation service and, more particularly, to create synchronic sociolinguistic *corpora* which illustrate the forms of the urban Spanish varieties spoken nowadays in the world. So far, 33 working groups are actively working in the project. Materials provided by the groups, following the general methodological guidelines of the project, would constitute PRESEEA corpus. In order to create a corpus of spoken language which includes relevant linguistic information –phonetics, grammar and discourse (Moreno Fernández 2005a)– it would be necessary to bear in mind the following principles:

- (1) A basic sociolinguistic methodology. Involved teams are committed to collecting sociolinguistic materials considering the methodology of the project. This procedure guarantees a collection of homogeneous and comparable samples.

---

<sup>1</sup> The acronym stands for the name of the Project in Spanish: *Proyecto para el Estudio Sociolingüístico del Español de España y de América*.

- (2) Materials' edition and publication. All linguistic materials collected by the PRESEEA groups must comply with the established transcription guidelines (Moreno Fernández 2005b).

In most cases the groups are still collecting the data using the technique of semiguide interviews, but some of these oral materials have already served as the basis for various sociolinguistic studies. The transcriptions from the PRESEEA interviews will be freely accessible online in both searchable text and audio recording formats (*see* <<http://www.linguas.net/preseea>>).

After presenting the general project, I will briefly introduce the PRESEEA-Palma de Mallorca corpus. Palma, the capital city of the island of Majorca, is a bilingual community. Both Catalan and Spanish equally enjoy an official status and are widely spoken by its population. Following the methodology of the project, the sampling in this community is being carried out using a uniform quota method (3 per section). For analysis of informing candidates, sex (M/F), age (3 generations: 20-34; 35-54; 55+), and level of education (high, medium or low) were taken into consideration.

#### References

- Blas-Arroyo, J. L. (2007). "Spanish and Catalan in the Balearic Islands". *International Journal of the Sociology of Language*, 184, 79-93.
- Moreno-Fernández, F. (2005a). "Project for the Sociolinguistic Study of Spanish from Spain and America (PRESEEA). A corpus with a grammar and discourse bias". In Takagaki, T. *et al.* (eds.), *Corpus-Based Approaches to Sentence Structures*. Amsterdam: John Benjamins, 265-288.
- Moreno-Fernández, F. (2005b). "Corpus para el estudio del español en su variación geográfica y social. El corpus PRESEEA". *Oralia*, 8, 123-139.
- Romera, M. (2006). "Aspectos metodológicos de la recogida de datos en situaciones de contacto de lenguas: la selección de hablantes en el caso de Palma". Paper read at the *I Jornadas de Lingüística Hispánica de la Universitat de les Illes Balears*.

### **Milagros Chao Castro**

University of Santiago de Compostela

#### **The OED as a Corpus: Looking for Dual-form adverbs**

Ungerer (1988) offers a list of 62 dual-form adverbs understanding this concept as an item which presents two adverbial variants, namely a basic form without any suffix and a derived form in -ly (e.g. *strong/strongly*). Other authors have also dealt with these adverbial forms, such as Quirk et al. (1985), Nevalainen (1997), or Huddleston & Pullum (2002), and they have analyzed their formation and use in different periods of the English language. However, this paper does not try to explain the morphological and syntactic characteristics of these items, but to enlarge the list provided by Ungerer by looking for

all the dual-form adverbs depicted in the *OED* in the different periods of the history of the English language, including both those of native origin and those taken from other languages. In doing so, I will also support the possibility of using the *OED* as a historical corpus (Fischer 1997, Plag 1999, 2003, or Mair 2004).

### References

- FISCHER, A. 1997. The *Oxford English Dictionary* on CD-ROM as a Historical corpus: *To wed* and *to marry* Revisited. In U. Fries, V. Müller and P. Schneider (eds.). *From Ælfric to The New York Times: Studies in English Corpus Linguistics*. Amsterdam: Rodopi, 161-71
- HUDDLESTON, R. & G. K. PULLUM (2002). *The Cambridge Grammar of the English Language*. Cambridge: C.U.P.
- MAIR, C. (2004). "Corpus linguistics and grammaticalization theory: Statistics, frequencies and beyond." In Lindquist, H. and C. Mair (eds.). *Corpus Approaches to Grammaticalization in English*. Amsterdam & Philadelphia: John Benjamins, 121-150.
- NEVALAINEN, T. (1997). "The processes of adverb derivation in Late Middle and Early Modern English." In Rissanen, M. et al. (eds.). *Grammaticalization at Work. Studies of long-term developments in English*. Berlin: Mouton de Gruyter, 145-190.
- OED = The Oxford English Dictionary on CD-ROM* (1989). [2<sup>nd</sup> ed.]. Ed. by John A. Simpson & Edmund S.C. Weiner. Oxford: O.U.P.
- PLAG, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation*. Berlin & New York: Mouton de Gruyter.
- PLAG, I (2003). *Word-Formation in English*. Cambridge: C.U.P.
- QUIRK, R., S. GREENBAUM, G. LEECH, AND J. SVARTVIK (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- UNGERER, F. (1988). *Syntax der englischen Adverbialen* (Linguistische Arbeiten 215). Tübingen: Niemeyer.

## Don Chapman

Brigham Young University

### Using Corpora in an English Usage Class

Corpora have become a natural place to test prescriptive rules. Writers since at least Horace have noted the importance of actual language use in establishing correctness, and for over a century, linguists have buttressed that concept with deeper theoretical knowledge about language and attempts to measure actual usage. For a scholarly tradition that has tried to use such expedients as letters written to the government (Fries 1940), *OED* entries (Quinn 1980), and dictionary citation files (Merriam-Webster 1994), the availability of large, searchable corpora is a godsend. Indeed, Finnegan (1999) predicted that corpora would become an increasingly important tool in debates about usage.

Finnegan's prediction is being realized by frequent articles in journals like *English Today* and *American Speech* that test the validity of prescriptive rules against actual language

use in corpora. Similarly, *The Cambridge Guide to English Usage* makes extensive use of corpora. But as important as these scholarly studies are, there remains much work to do in educating the public on the value of corpora. Readers who have been trained in consulting usage books that provide very little evidence at all, will need some instruction on why corpus-based evidence is important and how that evidence can be assessed. A college usage course is one of the venues for such instruction.

The most important activity in teaching the value of corpora is having students use corpora to measure the validity of prescriptions. This task is easy enough for students to understand, but it introduces several complications that students must overcome as they carry it out. One difficulty is in designing suitable searches. Usage prescriptions come in several varieties, ranging from those that proscribe a form in all circumstances, like *ain't*, to those that prefer one form over another depending on the circumstances, like *flagrant vs. blatant*. Students need help in figuring out which search will best fit which prescription being investigated. More important, students need help interpreting the results of their search. What does it mean for a proscribed form to occur 40% of the time in a corpus when compared to the frequency of the prescribed form? What about 20% frequency? 5%? 2.5%? Again, students, probably expecting clear winners and losers in their searches need help in telling what their numbers mean. Finally, students need help in analyzing the corpora they use for their search. What does it mean if the frequencies vary depending on the corpus searched? What differences should they expect between corpora of published vs. unpublished writings? What kinds of corpora will work best for their investigations? These questions and their place in a usage class will be explained and illustrated in this paper.

### References

- Finegan, Edward. 1999. "English Grammar and Usage." *The Cambridge History of the English Language*. Vol. 4.
- Fries, Charles. 1940. *American English grammar : the grammatical structure of present-day American English with especial reference to social differences or class dialects*. New York: Appleton.
- Merriam-Webster's Dictionary of English Usage. 1994. Springfield, MA: Merriam-Webster.
- Quinn, Jim. 1980. *American Tongue and Cheek: A Populist Guide to our language*. New York: Pantheon.

### **Ya-Jie Chen, Hui-Chuan Lu, and Kailing Liu**

National Cheng Kung University

#### **Corpus-based Study of New Motherhood in Charlotte Perkins Gilman's *Herland***

By taking advantage of rapidly developed computational technologies and applying the corpus-based approach, this paper aims to examine literary theories and feminist works

from a linguistic perspective. The present study not only reinforces the qualitative results of previous studies (Golden 1996, Gilbert and Gubar 1988, Gough 1998, Bergman 1999) on literary texts from linguistic analysis, but also provides quantitative evidence to support feminist literary criticism.

Charlotte Perkins Gilman is a prominent figure in first wave feminism, which focused on absolute rights such as suffrage, while the topic of equality appears as the main issue in the second wave ([http://en.wikipedia.org/wiki/Second\\_wave\\_feminism](http://en.wikipedia.org/wiki/Second_wave_feminism)). In order to study the concept of “motherhood” widely developed by second wave feminism (early 1960s to late 1980s) and farsightedly addressed by Gilman in terms of Val Gough’s (1998) “maternal separatism” and “professionalization of child care”, we will compare the distributions and tendencies of expression usage in her major works, *Herland* and *Yellow Wall Paper*, with the British National Corpus (100,000,000 words). Utilizing three programs (WordList, KeyWords and Concord) provided by Mike Scott’s Oxford WordSmith Tools and the coding system of the workbench, ATLAS.ti, we will contrast the morphologically, syntactically and semantically annotated data according to the principles Gilman tried to express. With the result of analysis, we can derive the patterns and clusters of similarities and differences such as expression and style variations among her works, the balanced and representative corpus with special focus on the studied theme. The final goal of this paper will be to show how the corpus-based approach can be effective for literary studies both quantitatively and qualitatively, and also to demonstrate how two different academic areas can be combined to produce wider implications.

## Paula Chesley

University of Minnesota

### **Towards Predicting New Words from Newer Words: Lexical Borrowings in French**

This study examines new lexical borrowings in a French newspaper corpus in order to predict lexical integration of borrowings. I search for all attestations of new lexical borrowings – borrowings not yet in a French dictionary – in a corpus of 21,560 sentences (T1 corpus; Abeillé et al., 2003). To quantify the degree of lexical integration, the T1 frequency rates are cross-checked in the online archives of another French newspaper from a later date (T2 corpus). Predictor variables tested for correlations with the frequency rates in the T2 corpus include:

- (1) Source language of the borrowing;
- (2) Sense pattern of the borrowings (monosemous vs. polysemous). A polysemous sense pattern is hypothesized to correlate with higher T2 frequencies;

- (3) Cultural context of the borrowings. The cultural context details the relationship between the culture of the denotatum and language. Lexical borrowings can either remain *restricted* to a culture that typically corresponds to their language, or they are *unrestricted* in that they need not correspond to a particular culture. Unrestricted borrowings are postulated to correspond to higher T2 frequencies;
- (4) Number of syllables of the borrowing, with fewer syllables hypothesized to correlate with higher T2 frequencies;
- (5) Domain of the borrowing.

Of these factors, (1), (2), and (4) were found to have significant correlations with T2 frequencies.

The initial borrowings are found to have a bimodal density distribution in the T2 corpus. A Shapiro-Wilk test shows that this distribution is not due to chance. In fact, the distribution of new lexical borrowings is composed of two separate normal distributions. These results echo those of Baayen and Lieber (1997), who argue that a bimodal distribution of Dutch words with a particular prefix shows a difference between well-entrenched lexical items and nonce formations using the prefix. Correspondingly, I propose that new lexical borrowings in French are composed of two types of borrowings: the normal distribution with the lower mean represents nonce borrowings, while the one with the higher mean designates borrowings en route to lexical integration.

## **Silvie Cinkova**

Charles University in Prague

### **Semantic annotation of a dialog corpus**

With morphological tagging and lemmatization having become a standard with large corpora, additional linguistic markup has been gaining importance in manual creation of gold-standard data for machine learning in NLP tasks. Though mainly used for technical purposes, the annotation conventions are usually defined as well as applied by linguists, and therefore they reflect a lot of interesting linguistic reasoning. This is evident in parsed corpora (treebanks) that enhance surface-syntax annotation with semantic markup, as it is the case of the Prague Dependency Treebank 2.0 and other corpora drawing on its annotation scheme.

The annotation scheme of the Prague Dependency Treebank 2.0 [Hajič et al., 2006] is a successful implementation of the Functional Generative Description, a formal natural-language description framework, developed in Prague since the 1960's [Sgall, Hajičová and Panevová 1986]. As FGD always has combined the structuralist linguistic tradition with cur-

rent trends within computational linguistics, it is easy to implement in treebank annotation.

FGD stratifies the language in several levels, with the most important ones being the *morphological*, the *analytical* (surface-syntax) and the *tectogrammatical* (underlying-syntax) levels. The tectogrammatical (semantic) level is the topmost and most abstract level within FGD. The essential features of the tectogrammatical representation (TR), i.e. syntactic dependencies, semantic labeling, valency (predicate-argument structure), ellipsis resolution and coreference, along with topic-focus articulation annotation, reflect *the linguistic meaning* of each sentence. TR describes syntactic as well as semantic relations among autosemantic words within a sentence, with coreference (pronominal anaphora) markup exceeding the sentence boundaries.

The proposed paper presents the tectogrammatical annotation of an English corpus of spoken dialogs. The dialogs capture reminiscing over photographs. In comparison to written monologue texts, the dialog annotation deals with far more complex issues of textual discontinuity and exophora. The dialog-adjusted TR also includes experiments with the *wh-path* annotation, which was previously used in the TIBAQ question answering system [Hajičová (ed.) 1995], and which identifies the most relevant segments in responses to wh-questions.

#### References

- [Hajič et al. 2006] Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havlka, Marie Mikulová, Zdeněk Žabokrtský, Magda Ševčíková-Razimová, *Prague Dependency Treebank 2.0*, Linguistic Data Consortium, 2006, LDC Catalog No. LDC2006T01, 1-58563-370-4
- [Hajičová (ed) 1995] Eva Hajičová (ed.), *Text-And-Inference-Based Approach to Question Answering*, Prague, 1995
- [Sgall, Hajičová and Panevová 1986] Petr Sgall, Eva Hajičová, Jarmila Panevová, *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Dordrecht:Reidel Publishing Company and Prague: Academia, 1986

#### **Arlene Clachar**

University of Miami

#### **The Effect of Computer-Mediated Communication on the Acquisition of Registers of Written Academic Discourse by Creole-English-Speaking Children**

The paper reports the findings of a study which examined whether Creole- English-speaking children exhibit different challenges related to their choice of lexico-grammatical resources of registers of written academic discourse when they are required to compose their academic texts in regular computer word-processing communication (CW-PC)

versus in Computer-Mediated Communication (CMC).

The study was motivated by the need to understand the challenges related to the acquisition of lexico-grammatical features of written academic registers faced by Creole-English children whose linguistic repertoire is characterized by speech varieties along a creole continuum, typified by constant bidirectional shifting between Standard English (SE) and Creole English (CE), with the latter occurring predominantly in spoken discourse, and allowing the mixing of lexico-grammatical features of speech registers of CE with those of SE.

In addition, varieties of Creole English or English-based Creoles show considerable vocabulary overlap with Standard English, but diverge from SE with respect to the morphological and syntactic systems. This lexical overlap between CE and SE tends to mask grammatical differences between registers of speech typical of the CE and registers of written academic discourse in SE. As a result, the Creole-speaking children might have difficulties separating certain lexico-grammatical features typical of registers of speech in CE from those typical of registers of academic writing in SE. Therefore, it was hypothesized that these phenomena would cause a strong tendency on the part of CE children to draw on the lexico-grammatical features of speech registers in their written academic discourse in (CW-PC). Would we see the same tendency when these children compose similar texts on the Web in CMC?

Findings indicate that when CE children create Web texts, they are intuitively encouraged to think about the interactive properties of the texts, such as links to homepages, help sites, and promotional information. Thus, they are forced to focus on the structure of clauses and how these will shape the entire textual discourse. Such psycholinguistic and cognitive factors help CE children to separate the registers of written academic discourse from the registers of speech in CE, SE, and the mesolectal varieties that mix CE and SE. These factors contributed to a higher frequency of certain register features of written academic discourse, namely, nominalizations, nominal group structures, and non-paratactic conjunctions in CE-speaking children's Web texts in CMC than in their (CW-PC)

## **Georgie Columbus**

University of Alberta

**“Ah lovely stuff, eh?”<sup>2</sup>**

### **On invariant tag meanings and usage across three varieties of English**

Invariant tags, such as *huh* and *innit*, are discourse markers which often occur at the end

---

2 HEU 618, BNCWeb.

of an utterance to provide attitudinal and/or evidential information above that of the proposition. Many previous studies exist examining the meaning or usage of these tags in single varieties/dialects of English, such as New Zealand *eh* (e.g. Stubbe and Holmes, 1995; Meyerhoff, 1992), London teenage speech, including *innit* (Andersen, 1997; Berland, 1997), and US *hunh* (Norrick, 1995). However, while some have investigated sociolinguistic divisions within a dialect (viz. Andersen, 1997; Meyerhoff, 1992; Stubbe and Holmes, 1995), none yet have compared usage between varieties. Furthermore, differences in research methodology and aims prevent comparison of the prior results.

This study investigates the meaning/functions and relative frequencies of four invariant tags, *eh*, *yeah*, *no* and *na*, in New Zealand, Indian, and British English. The results show differences in the meanings available as well as the usage frequencies across both items and varieties. The findings suggest that varietal differences at the level above propositional understanding could cause problems for intercultural and global communication. This has implications for pedagogy and materials for ESOL and English for Specific/Business Purposes, in that global communication in English requires an awareness of these subtle differences at the varietal level into account.

### References

- Andersen, G. (1997). "I goes you hang it up in your shower innit? He goes yeah." *The use and development of invariant tags and follow-ups in London teenage speech*. Paper presented at the 1<sup>st</sup> UK Language Variation Workshop, Reading, United Kingdom.
- Berland, U. (1997). Invariant tags: pragmatic functions of *innit*, *okay*, *right* and *yeah* in London teenage conversations. Unpublished master's thesis, University of Bergen, Norway.
- Meyerhoff, M. (1992). 'We've all got to go one day, eh?': powerlessness and solidarity in the functions of a New Zealand tag. In K. Hall, M. Bucholtz and Moonwomon, B. (Eds.), *Locating power: Proceedings of the Second Annual Berkeley Women and Language Conference*. Berkeley, California: Berkeley Women and Language Group.
- Norrick, N.R. (1995). *Hunh*-tags and evidentiality in conversation. *Journal of Pragmatics*, 23, 687-692.
- Stubbe, M. and Holmes, J. (1995). *You know, eh* and other exasperating 'expressions': an analysis of social and stylistic variation in the use of pragmatic devices in a sample of New Zealand English. *Language and Communication*, 15, 63-88.

**Susan Conrad and Sarah Albers**

Portland State University

### A New Corpus of Student Academic Writing

The teaching of academic writing has been a challenge for ESL teachers and composition instructors for decades. Despite many different types of studies over the years – from information about writing tasks (Horowitz, 1986) to analyses of specific language features

in student writing (e.g. Cortes, 2004 on lexical bundles)—we still know remarkably little about the range of writing that university students are asked to do, and about the language characteristics of papers that receive high grades. At our university, instructors in the ESL program feel a particular lack of information about the writing required in regular classes.

In response to the general need for information about student academic writing and, especially, the local needs of our ESL program, we have begun a project to compile and analyze a new corpus of student academic writing. In this presentation, we report on the project from three perspectives:

- (1) The overall design for the corpus and goals of the project. The goals include having papers from every department in the university, and comparing the papers written in regular classes with those in ESL classes, as well as developing teaching materials for the ESL classes.
- (2) A description of the current version of the corpus and current investigations. Presently, the corpus has 392 papers from 26 departments at the university, for a total of approximately 608,000 words. A wide range of assignment types are included, from self-reflection pieces to original research papers. One focus of current investigations concerns ways to further improve the corpus design.
- (3) A discussion of particular challenges that we face in the project. Prime among these is how to make the corpus useful to our ESL program instructors who are intrigued and see its usefulness, but are not committed to learning how to use a corpus on their own.

#### References

- Horowitz, D. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly*, 20, 445-462.
- Cortes, V. (2004) Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23, 397-423.

### **Mike Conway**

National Institute of Informatics

### **Emotive Language and Disease Outbreak Reports**

This paper explores stylistic differences between news reports that describe infectious disease outbreak events, and news reports that describe other types of event, using the annotated BioCaster corpus as experimental data. The BioCaster corpus – as used in this work – consists of 500 English language news reports. Almost half of these news reports are directly concerned with infectious disease outbreaks, while the remaining reports con-

sist of more general news (business, sport, politics, non-infectious disease health reporting and so on). Four results are presented in this paper. First, we show that negative affect words appear more frequently in infectious disease outbreak articles, compared to other types of news articles. Second, we show that some standard non-lexical features common in statistical text analysis (word length, sentence length and punctuation density) fail to distinguish between infectious disease outbreak reports and other kinds of reports, and suggest reasons why this is the case. Third, we show that by using feature selection methods to produce more compact feature representations, we can gain better classification results (at a statistically significant level) than by using a benchmark “bag-of-words” style feature representation. Finally, we show that non-topical function words are not good discriminators for the infectious disease outbreak report classification task.

## **Christopher Cox**

University of Alberta

### **Probabilistic tagging of a corpus of Mennonite Low German: A case study using QTag**

The current paper presents a case study in training a probabilistic parts-of-speech tagger, QTag (Tufis and Mason 1998), on a corpus of written and spoken texts being developed at the University of Alberta for a local minority language, Mennonite Low German (*Plautdietsch*). While probabilistic tagging has itself received considerable attention in the corpus and computational-linguistic literature (cf. Church 1988; DeRose 1988; Dermatas and Kokkinakis 1995), less attention has been paid to date on investigating those factors which may determine the rate at which accuracy develops in iterative, interactive probabilistic tagging systems.

In the case of minority-language corpora, where no fixed standards for either orthography or parts-of-speech tagging may necessarily exist, and where financial resources for corpus development are often limited (cf. McEnery and Ostler 2000), achieving a high degree of tagging accuracy may represent a significant challenge. Understanding those combinations of factors relevant to maximizing accuracy and minimizing expenditure of time and effort in tagging, then, is important to the planning and development of such projects. The present study therefore considers several factors affecting the rate at which accuracy increases over rounds of training, including the size of the initial training data set, tag-set size, and the frequency with which parts of speech identified in the selected tag set occur in the training data. The study concludes with a discussion of the relationship of the proposed factors to the rate of increase in tagging accuracy over rounds of training, as well as of potential confounds to the application of probabilistic taggers to similar minority language data.

## References

- Church, Kenneth W. 1988. A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the Second Conference on Applied Natural Language Processing*, 136-143.
- Dermatas, Evangelos and George Kokkinakis. 1995. Automatic stochastic tagging of natural language texts. *Computational Linguistics* 21 (2).137-163.
- DeRose, Steven J. 1988. Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics* 14 (1).31-39.
- McEnery, Tony and Nick Ostler. 2000. A New Agenda for Corpus Linguistics – Working with all of the World's Languages. *Literary and Linguistic Computing* 15.403-149.
- Tufis, Dan and Oliver Mason. 1998. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. *Proceedings of the First International Conference on Language Resources & Evaluation (LREC), Granada (Spain), 28-30 May 1998*, 589-596.

## **Eniko Csomay and Viviana Cortes**

San Diego State University

### **Positioning lexical bundles in university class sessions**

Over the past decade a growing number of studies applied corpus-based methodologies to describe the lexico-grammatical characteristics of academic language use. Among these are studies focusing on lexical bundles, which are frequently occurring word combinations (Biber et al. 1999), and comprehensive linguistic descriptions of spoken and written registers (Biber, 2006) in the academic context. Other studies focus on the relationship between linguistic variation and discourse structure as they describe language change within university class sessions (Csomay 2005).

The present study investigates the relationship between the discourse functions of lexical bundles found in classroom teaching (Biber, Conrad, & Cortes, 2004) and their position as they appear in previously identified smaller units of analysis, called Vocabulary-Based Discourse Units (Biber et al. 2004).

84 lexical bundles, frequently occurring 4-word combinations identified in the lectures of the T2KSWAL corpus, are tracked in the discourse units (VBDUs) of 176 lectures. Among others, expressions such as 'has to do with', 'I would like to', 'if you look at', 'in the case of' are traced in tandem with their previously identified classification of structural correlates and discourse functions.

Preliminary findings show that a) the lexical bundles appearing in the first three units of classroom discourse (77% of the total bundles types) reflect the type of functions linked to the beginning of university lectures such as discourse organizers (to introduce a topic or to focus the forthcoming discourse), and b) the bundles missing from these units (the remaining 23%) are from two functional categories: stance markers, and referential bundles.

These initial findings provide lexical evidence to support previous empirical studies describing the linguistic characteristics of these units and earlier claims made on the communicative and instructional functions of the opening “phase” (Young 1994) in the class session.

## **Adriana Teresa Damascelli**

Università degli Studi di Torino

### **The use of hedging devices in American English: Identifying some trends in the FROWN corpus**

Social interaction presupposes the continuous flux of information between people. The information conveyed can be of many different kinds, namely expression of ideas and communication of knowledge, and is shaped on the basis of complex cognitive processes and linguistic activities which take place in writers and speakers. Beside effective discourse organisation which is needed to structure information and make it intelligible, dynamic mechanisms underlie social interaction. Usually, the information is mediated by writers or speakers' attitude towards what is being written or said. In some cases uncertainty and deference are expressed in order to prevent personal involvement and commitment; in other cases, claims are mitigated and made more or less vague in order to neutrally convey pieces of information.

In 1972 Lakoff introduced the term hedge to denote “words whose job it is to make things fuzzier or less fuzzy”. These words include some lexical verbs (e.g.: tend, appear), modals (e.g.: may, might), and some adverbs (e.g.: probably, perhaps). Also, studies in epistemic modality have shown that there are many different ways to realise hedging, e.g. through condition clauses, passive voices, and impersonal phrases.

Hedging has been studied and considered as playing a very important role in communication, especially in scientific and academic writing where different strategies are used to convey a range of different meanings. The appropriate use of them reflect efficient social interaction by showing the ability to express degrees of certainty and demanding rhetorical strategies required for particular social demands. However, as recently documented, it is necessary to further investigate hedging in order to better understand communicative strategies which not necessarily are used in scientific settings.

The aim of this study is to provide further insights into hedging in a corpus-based perspective. To this aim The Freiburg - Brown Corpus of American English will be investigated in order to identify some trends in American English.

## **Mark Davies**

Brigham Young University

### **The 360 Million Word *BYU Corpus of American English* (1990-2007)**

In this presentation I will discuss the BYU Corpus of American English, which is the first large-scale, publicly-available corpus of American English. (The American National Corpus is just 22 million words, has rather uneven composition, has not been updated in nearly three years, and may never reach the projected 100 million words.)

The BYU Corpus of American English is nearly done, and will be freely and publicly available via the web in late January 2008. Its major features are the following:

#### CONTENT

- 20 million words each years (1990-2007), for a total of 360 million words
- All texts are from American sources; e.g. TV, radio, magazines, newspapers, journals published in the US
- The corpus will be expanded at least twice each year (10 million words every 6 months; 2 million words in each of the five registers)
- The corpus is divided (overall, and for each year) into five equally-sized registers:
  - Spoken: Transcripts of unscripted conversation from more than 100 different TV and radio programs (Oprah, Good Morning America, Talk of the Nation, etc)
  - Fiction: Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, movie and TV scripts
  - Popular Magazines: Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (sports, entertainment, science, hobbies, etc)
  - Newspapers: Ten newspapers from across the US, with a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
  - Academic Journals: Nearly 100 different peer-reviewed journals. These were selected by Library of Congress call number to create a good balance across domains

#### FEATURES

- Interface similar to our interface for the British National Corpus and Time Magazine with several important improvements (which are now available for our web-based Corpus del Español)

- The corpus as been tagged by CLAWS, the same tagger that was used for the BNC
- Queries by word, phrase, alternates, substring, part of speech, lemma, and customized lists (e.g. user-created lists related to a particular semantic category, or a user-defined part of speech)
- Chart listings (totals for all matching forms in each register or year, 1990-present) and table listings (frequency for each matching form in each year or register)
- Full collocates searching (up to ten words left and right of node word)
- Comparisons between registers or time period (e.g. collocates of a given word that are more common in one register than another, or which appear only after 2003)
- Comparison of collocates of related words (e.g. one-step comparison of collocates of big and large, or with men and women, or comparison of collocates of chair in different registers)
- Incorporation of semantic information from WordNet and other thesauruses directly into query (e.g. search for all synonyms of sweet + type of food, or find the frequency of all verbs related to walk)

## **Boyd Davis and Charlene Pope**

University of North Carolina at Charlotte

### **Into the woods:**

#### **First steps in a collaborative corpus of medical conversations**

Carolina Conversations, recently funded by the National Library of Medicine, involves a multidisciplinary team representing the humanities, social sciences, and health sciences, in establishing a corpus of digital recordings and transcripts. The collection will house two cohorts: (1) 200 consented conversations with 72 older men and women of multiple ethnicities, with any of 12 chronic medical conditions and (2) a longitudinal set of 200 naturally-occurring conversations with 72 persons with Alzheimer's disease. Cohort 1 will be recorded twice a year, once with health professional graduate students, and once with community persons of concordant ethnicities and languages, in home or community settings. Cohort 2 will talk twice annually with gerontology and linguistics students. Common questions will unify interviews.

Approved users of the password-protected internet portal will be able to retrieve and conduct online analysis by gender, age, medical condition, ethnicity, and first language for speaker and/or interviewer, by provider, community, or facility site, and by single or

multiple interviewers over time. The key to storage, access, search and retrieval is the time-stamped alignment of the transcript with audio and/or video. Transcripts are created in Transcriber software. The mainframe package, ONZE Miner, links the database for the transcripts to the lexical database CELEX and to PRAAT phonetics software, as well as to parsers. Online modules will illustrate how to organize, retrieve, and analyze transcripts online with these tools. This digital Web-based corpus will allow researchers to examine issues such as patient-provider and client-community conversation styles and cultural effectiveness in healthcare.

Information about health literacy, health status, demographics, and cognitive function will identify participant conditions, requiring high standards for privacy and confidentiality. This presentation reports on necessary confidentiality regulations, protocols for review, multidisciplinary training needs, and audio, video and metadata standards.

## **Ferdinand de Haan and Sheila Dooley**

University of Arizona

### **The role of constructions in the assignment of modal meanings**

This paper is concerned with the application of ideas from Construction Grammar (CxG) in the area of modality. It is argued that constructions (in the sense of Goldberg 1995, 2006 and Croft 2001) play an important role in the disambiguation of modals. The presentation will show constructions from various Germanic languages.

For instance, it has been shown (Coates 1983:44-5, among others) that there is a strong correlation between Progressive and epistemic *must* in English. We argue, based on data from the Brown and Switchboard corpora, that the construction *must* plus Progressive is associated with epistemic meaning, even if the sentence is not progressive. That is, the Progressive form (V-ing) is now associated with epistemic meaning, and the original progressive meaning is suppressed.

(1) That's the only place I was sore, and I thought, well, I must not be doing them right ...

This has given rise to a non-compositional construction in which the meaning is not predictable from its constituent parts (giving support for arguments of CxG about compositionality, see e.g., Goldberg 2006:45ff). This also entails that in rare cases *must* can occur with deontic meaning when a Progressive form is used, but only if the reading is progressive. One such example from the Brown corpus is shown in (2):

(2) The article also said that a person had to be 18 years old or over, and must not be going to high school to attend these classes.

The presentation will focus on similar examples, such as the correlation between Perfect and epistemic modality, and between tense and modality. We will show different stages of constructions, from fully entrenched ones to emergent ones. Implications for the status of modals will be discussed as well.

## Paul Deane

Center for Automated Scoring and Natural Language Processing

### Creating Subcorpora to Explore the Topic Structure of Domain-Specific Text

Topic modeling is a relatively new statistical technique for investigating the structure of a corpus. Related to Latent Semantic Analysis (LSA), it creates an explicit model containing ‘topics’, groups of related, co-occurring words (Blei et al. 2003, Griffiths & Steyvers 2004, Griffiths et al. 2004). This paper will explore methods for automatically extracting a relevant subcorpus from a larger (billion-word) corpus and using them to the topic structure of particular domains.

We report on initial experiments using a large (billion-word) corpus, a collection of books and articles employed in the Educational Testing Service’s internal SourceFinder tool. We used a database capturing word-cooccurrence frequencies and related statistics to extract a subcorpus containing paragraphs with a high proportion of words related to a target vocabulary set. For instance, we selected documents containing a high proportion of words strongly related to the word *democracy*, and then examined how these words were assigned to topic clusters by LDA. The resulting clusters clearly reflected topical distinctions that could be used to map the immediate semantic neighborhood around the target word. Some of the topics extracted include:

*policy, trade, economic, foreign, market, markets, countries, investment, economy, policies ...*

*vote, voters, Kerry, election, candidates, campaign, candidate, voting, presidential, polls ...*

*Jefferson, Little, Thomas, liberty, equal, principle, Buchanan, further, principles, Madison ...*

Similarly, documents were selected containing words found in a single document (one addressing Ben Franklin as an inventor). With 200 topics, only one contained the word *Franklin*:

*Franklin, Ben, Benjamin, Philadelphia, mail, lightning, wrote, fire, first, invented, started, William, electricity ...*

These preliminary results strongly suggest that we can appropriately generate prompt-specific corpora so that they contain concentrations of vocabulary relevant to the prompt, and can then use LDA to generate specific topics, which can then be used to explore the fine structure of specific domains.

## **Marta Degani**

University of Verona

### **The Maori presence in New Zealand English: a lexical approach**

New Zealand English (NZE) is a language whose vocabulary has been strongly influenced by British, Scottish and Irish but what makes it uniquely different from any other English variety is the presence of words in *te reo maori*, i.e. the Maori language (Bell & Kuiper, 2000). The first borrowings from Maori came into NZE when the country was colonized (at the end of the 18<sup>th</sup> century): most of them were words for plants and animals, some were cultural terms. Since then, a rather large number of Maori words have found their way both in spoken and written NZE (Kennedy, 2001).

This paper concerns present-day written NZE and is motivated by two general observations: 1) Maori terms can be found in New Zealand books and newspapers without any translation into English, indicating that most New Zealanders know – or allegedly know – what they mean; 2) most of them are used exclusively in connection with Maori culture, which points at their no more than partial integration into NZE (Trudgill & Hannah).

For the present study three Maori words have been selected, namely *aroha* ('love'), *mana* ('power') and *marae* ('meeting ground'), on the basis of their frequency of occurrence in NZE and their significance in terms of cultural identity. The analysis (both quantitative and qualitative) will be carried out working on a corpus of on-line NZ newspapers, namely *The New Zealand Herald*, *The Dominion Post* and *The Press*. The approach will be synchronic and all issues from mid-2006 to mid-2007 will be considered in order to investigate the three Maori words in their contexts of use.

The study aims at pointing out:

- the extent to which the use of these words contributes to the maintenance and/or creation of cultural stereotypes
- the extent to which the use of these words is a reflection of assimilationist policies

## Selected references

- Bell A., K. Kuiper (eds.), 2000, *New Zealand English*, Wellington, Victoria University Press.
- Kennedy G., 2001, "Lexical Borrowings from Maori into New Zealand English", in B. Moore (ed.), *Who's Centric Now? The Present State of Post-Colonial Englishes*, Oxford, Oxford University Press, pp. 59-81.
- Trudgill P., J Hannah J., 2002, *International English. A Guide to the Varieties of Standard English*, London, Arnold.

**Manuel Delicado-Cantero**

Ohio State University

**Corpus attestations and linguistic explanations:  
the case of the history of Spanish prepositional finite clauses**

While in Modern Spanish argument finite clauses may optionally be introduced by a preposition, as in *Me acuerdo (de) que vino* (lit. I recall (of) that s/he came, 'I recall s/he came'), the situation was different in Old Spanish. In this paper I study this syntactic change with extensive data from large corpora, such as *Corpus del español (CdE)* and *CORDE*, and reveal the inadequacy of previous accounts when confronted with the reality of the examples.

Barra Jover (2002) maintains that in Old Spanish finite clauses were not nominal and therefore, could not be assigned Case by a preposition. This restriction disappeared once finite clauses acquired nominal features in the 16<sup>th</sup> century. However, I will show that a thorough research in the corpora challenges this analysis.

By locating the change in the 16<sup>th</sup> century, this account fails to explain why adverbial prepositional finite clauses, such as those introduced by *porque* (lit. for that..., 'because'), are attested in Old Spanish. The proposed solution, namely that those prepositions were not Case-assigners, is inconsistent with the data. *CdE* and *CORDE* provide evidence that those prepositions selected for morphologically inflected pronouns (Oblique Case), which falsifies Barra Jover's hypothesis.

Furthermore, Barra Jover attributes different syntactic structures to the prepositional finite clauses and to their prepositionless counterparts. However, my study of more than seventy verbs provides arguments for a common syntactic analysis of both alternatives, especially on the basis of mood selection phenomena.

In this paper I clearly show the benefits of large corpora for diachronic studies. In particular, I offer data which reveal that the emergence of prepositional finite clauses in Spanish is not related to Case and, in addition, I simplify the analysis by proposing a common syntactic structure for both prepositional and non-prepositional finite clauses in Spanish.

**Philip Dilts**

University of Alberta

**Good Nouns, Bad Nouns: What the corpus says and what native speakers think**

Many researchers have found that some words or constructions tend to co-occur with words representing a positive or negative semantic nuance. Stubbs (1995), for example, shows that the verb *cause* tends to appear with an object denoting an unpleasant concept (e.g., *cause for anxiety*), and points out that this “semantic preference” is not intuitively obvious to English speakers. Other researchers have explored the negative or positive associations of words taken out of context, their “semantic orientation” (Osgood et al 1957, Turney & Littman 2003). In this paper, we investigate how well a word’s semantic orientation correlates with its semantic preference.

We extracted a list of nouns with strong semantic preferences for positive or negative adjectives from the British National Corpus, using a collocational analysis modeled after Dilts and Newman (2006). We then submitted these nouns to native speakers, asking them to judge the nouns’ semantic orientations. Some initial results are summarized below.

		Semantic Preference of nouns	
		Positive	Negative
Semantic Orientation of nouns	Good	12	10
	Neutral	8	3
	Bad	1	12

The table shows interesting mismatches between semantic orientation and semantic preference. Of the 22 nouns that were judged as “good” out of context, 10 actually collocate with negative adjectives in context. On the other hand, the 13 nouns judged as “bad” do indeed have a semantic preference for collocating with negative adjectives. The pilot results point to intriguing differences in the way in which “good” and “bad” nouns collocate with adjectives: while “bad” nouns attract negative adjectives (further reinforcing their semantic orientation), “good” nouns attract both positive (reinforcing) adjectives, as well as negative (qualifying) adjectives which have a greater transformative effect on the semantics of the noun.

References

Dilts, P. & J. Newman. 2006. A Note on Quantifying ‘good’ and ‘bad’ prosodies. *Corpus Linguistics and Linguistic Theory* 2(2):233-242.  
 Osgood, C. E., G.J. Suci & P.H. Tannenbaum. 1957. *The Measurement of Meaning*. Urbana: University of Illinois Press.  
 Stubbs, M. 1995. Collocations and Semantic Profiles. *Functions of Language*, 2(11), 23-55.  
 Turney, P.D. & M.L. Littman. 2003. Measuring Praise and Criticism. *ACM Transactions on Information Systems*, 21(4), 315-346.

**Luciana Diniz**

Portland Community College

**Suggestions and Recommendations in Academic Speech**

This presentation will investigate the communicative functions of modals (e.g., *should*, *could*), as well as lexical verbs, such as *recommend* and *suggest* by faculty in academic spoken discourse. We find that these expressions are frequently used as hedging devices by professors in a number of situations. To perform the analysis, we use several speech events from MICASE (Michigan Corpus of Academic Spoken English), including academic lectures, office hours, colloquia, and seminars. With the help of Wordsmith Tools (Scott, 1999) and the tools from the MICASE website, we examine the collocates and n-grams in which the modals and lexical verbs are included, as well as the corresponding sound files, in order to further investigate acoustic cues that may have an impact on the communicative functions of the target words. The applications of the results have immediate pedagogical applications. For example, international teaching assistants can be presented with politeness strategies in which they can utilize for efficient communication among their own students. Also, the results will be likely to assist international students who aim at studying at US universities to be aware of professors' techniques to hedge and, consequently, balance power. The lack of understanding of potential implicit meanings conveyed by modals and certain lexical verbs such as the ones mentioned above may lead to unsuccessful performance in academia.

**Radoslaw Dylewski**

Adam Mickiewicz University, Poznan

**Selected words, phrases, and meanings  
of African (American) provenience in General American:  
A corpus-based study**

The incentive to embark on the present study has been the recurrent claim of Geneva Smitherman recently repeated in her 2006 book; this claim states the ongoing process of lexical *Africanization* of White American English which is increasingly evident in the thousands of “examples of Black linguistic crossover into mainstream English – from the ever-popular Black “high five” that can be seen everywhere in White America, to words like “phat” and “bling-bling,” now comfortably housed in standard dictionaries of American English” (Smitherman 2006: 7).

The said statement has been approached with caution since Smitherman seems to have an ‘emotional’ attitude toward AAVE; since the initial assumption was that this optimism behind Smitherman’s works might sometimes assume a too optimistic scenario – at least in some areas – the present paper aims at establishing the degree of recent influence of the AAVE lexicon on its General American counterpart. More specifically, the author scrutinizes a selection of words/phrases/meanings of African (American) origin which have been present in the language of European Americans at the turn of the 21<sup>st</sup> century in order to determine their frequency of appearance, distribution across spoken and written media as well as their presence in various text types and semantic fields.

The list of lexical items and denotations of AAVE provenience has been compiled on the basis of various sources ranging from subject literature (among other sources, Green 2002 and Smitherman 2006) to the *Dictionary of American Regional English* and Smitherman’s 2000 dictionary. The material chosen for the study is the second release of the *American National Corpus*, comprising to date approximately twenty-two million words of written and spoken American English, the corpus of texts of American English from 1990 to the present, as found in TIME magazine (<http://corpus.byu.edu/time>), and the *Corpus of Contemporary American English* yet to be released later this year.

#### References

- Dictionary of American Regional English* 1985– Frederic G. Cassidy, chief editor. Cambridge, Mass.: Belknap Press.
- Green, Lisa J. 2002 *African American English: A linguistic introduction*. New York: Cambridge University Press.
- Lee, Margaret G. 1999 “Out of the hood and into the news: Borrowed Black verbal expressions in a mainstream newspaper”. *American Speech* 74.4: 369-389.
- Smitherman, Geneva 2000 *Black Talk: Words and phrases from the Hood to the Amen Corner*. Revised Edition. Boston & New York: Houghton Mifflin.
- 2006 *Word from the mother: Language and African Americans*. New York and London: Routledge.

### **Andrés Enrique-Arias and Laura Carmago**

Universitat de les Illes Balears

#### ***Biblia Medieval: a parallel corpus of medieval Spanish***

Historical parallel corpora (i.e. translations of a single original composed at different time periods) are a useful tool for historical linguistics, insofar as they make it possible to study the evolution of the language at all levels of analysis with particular clarity. This paper reports on the development of the *Biblia Medieval* corpus, the first aligned parallel corpus of medieval Spanish. *Biblia Medieval* is a freely accessible online tool which enables linguists

to consult and compare side to side the existing medieval Spanish versions of the Bible, as well as to access the facsimiles of the originals.

This paper addresses some of the main philological and technical issues that were faced by the creators of the *Biblia Medieval* corpus. Some of the topics covered are: selection of the texts, norms of transcription, procedure to digitize the facsimile images of the manuscripts, and design of the database and interface for the searches. The paper includes a discussion of how a parallel corpus of medieval biblical texts will enrich our theoretical understanding of phenomena of change and variation in Spanish from a diachronic perspective. The corpus, as well as other related documents and tools, are housed at [www.bibliamedieval.es](http://www.bibliamedieval.es).

## **Eileen Fitzpatrick and Joan Bachenko**

Montclair State University

### **Testing Language-Based Indicators of Deception On a Corpus of Legal Narratives**

Experimental laboratory results, often performed with college student subjects, have identified several linguistic phenomena as indicative of speaker deception. These phenomena fall into three classes: (1) Linguistic devices used to avoid making a direct statement of fact, including linguistic hedges, qualified assertions, which leave open whether an act was performed, unexplained lapses of time, overzealous expressions, and rationalizations. (2) Preference for negative expressions in word choice, syntactic structure and semantics. (3) Inconsistencies with respect to verb and noun forms, including verb tense changes, thematic role changes, noun phrase and pronoun changes, where different forms denote the same referent.

Can these results be replicated on “real world” statements from suspects and witnesses? To test the accuracy of these linguistic cues with respect to identifying deception, we assembled a corpus of criminal statements, police interrogations, and civil testimony and defined twelve linguistic indicators of deception cited in the psychological and criminal justice literature that can be formally represented.

Trained annotators hand tagged the narratives for the deception cues and another group of annotators marked the truth value of all propositions that could be externally verified as true or false. A measure of the density of cues was then calculated, with high cue density taken to identify a passage as deceptive. This method correctly distinguishes 70% of the hand-tagged propositions as true or false, using classification and regression techniques, compared to a human subject average of 78% on the same task, and correctly identifies 93% of the false statements. This preliminary result suggests that linguistic cues can provide a reasonable guide to the sectioning of narratives into deceptive and non-deceptive statements.

**William H. Fletcher**

US Naval Academy

**Complementing the BNC with a Corpus from the Web**

Since 2003 I have deployed a database online for simple but powerful access to the British National Corpus. Despite its tremendous usefulness, however, the BNC has serious limitations: as a static collection of texts from 15 years ago and more, it neglects current and emerging usage; its “small” size excludes many items of interest; it represents the linguistic usage of a single nation. The proposed paper details a project which complements this carefully compiled corpus with one based on texts acquired from the Web and tagged automatically with the BNC’s POS-tagsets. To ensure semantic and geographic representativeness, search engine queries with 6000 terms from across the semantic spectrum yielded texts from the principal English-speaking countries which are incorporated in proportion to their population. Currently almost 800,000 unique webpages totaling over a billion words have been downloaded, and a database of roughly half the corpus is already searchable online. When linguistic annotation is complete and a shared user interface has been implemented, users will be able to carry out comparable searches on both corpora, to study regional and diachronic similarities and divergences in usage. This web corpus will be dynamic as it grows in response to actual user queries to my various web as corpus interfaces, but “snapshots” of each generation of the corpus will be preserved to ensure replicability of results. Potentially this tagged corpus will comprise many billion words, enabling researchers to study the “long tail” of low-frequency items and to monitor emergence, diffusion and ultimate fate of linguistic innovations.

**Maria Freddi**

University of Pavia

**Studying phraseology and translation through the corpus and the database**

For more than a decade now, both parallel and comparable corpora have become a standard resource in corpus-based translation studies, either assisting comparative analyses of source-texts and their translations in one or more languages (bi- and multilingual parallel corpus), or as an aid to studies of translationese (comparable corpus of original and translated texts in the same language), or as a valuable tool for contrastive studies (comparable corpus of texts originally drawn in different languages). The database, however, is a less used tool which can integrate traditional corpus output such as aligned concordances.

Drawing from corpus-based translation studies, theory of phraseology and audio-visual translation studies, the proposed paper aims to investigate film translation, dubbing in particular, with a view to identifying recurrent translation solutions (translational routines) of sociolinguistic and context-dependent pragmatic aspects of film dialogues, in particular phraseology.

To this purpose, a small parallel corpus has been compiled consisting of 12 both British and American contemporary general distribution films for the period 1995-2005, and their dubbed Italian version. The corpus has then been segmented on a turn-by-turn basis, each turn with its sociolinguistic and paralinguistic features, in single cells of a database. In addressing issues of corpus design, sampling and representativeness, data representation and extraction, the paper thus shows how the parallel corpus can help study film translation and how a sophisticated though conventional tool such as the relational database allows the representation of multimedia data, their comparison and grouping, and complex searching combinations. Some queries are given as examples of how the database can work as an ideal basis on which to ground empirical and quantitative research, thus giving a fresh contribution to the new discipline of audio-visual translation studies.

## **Eric Friginal**

Northern Arizona University

### **Grammatical Expression of Stance in Outsourced Call Center Discourse**

Outsourced customer contact centers from the United States (U.S.) to the Philippines have paved the way for the creation of jobs for Filipino professionals who are able to communicate in English and provide telephone-based customer services to American clients. The Philippines has become one of the major centers for U.S.-based customer service outsourcing – second only to India – because of its tradition of English education, affinity to the American culture, and overall cheap labor market. This study explores the use of grammatical expression of stance in outsourced call centers involving Filipino call-takers or “agents” and American customers engaged in various types of technical and service support transactions. “Stance” in this study is defined as the linguistic mechanisms used by speakers to convey their personal feelings and assessments (Biber, 2006) in spoken interactions. The specific goals of this study are (1) to establish the distribution of grammatical stance features in the call center discourse relative to other registers of conversation (American conversation from the Longman corpus and telephone interactions from the Switchboard corpus), and (2) to examine how the interactants in outsourced call center transactions use stance expressions based on role (as agent or caller), gender, and the categories of transactions or “accounts.” The data for analysis come from a corpus of out-

sourced call center texts collected in the Philippines (N of texts=500, with approximately 553,630 words). The research design follows a framework developed by Biber (2006) that incorporates three major lexico/syntactic features used for stance analyses. These features are grouped into: (1) modal and semi-modal verbs, (2) stance adverbs, and (3) complement clauses controlled by stance verbs, adjectives, or nouns. Results show significant variation in the use of the three categories of grammatical stance features across registers. Similarly, the internal social and/or demographic categories in the call center corpus such as role, gender, and the categories of accounts are found to affect the use and frequency distribution of stance features in outsourced call center transactions.

## Alfonso Gallegos Shibya

Universidad de Guadalajara

### The diachronic development of some verbs with copulative function in Spanish

Besides *ser* and *estar*, Spanish has also a series of verbs that work as copula (for example *andar*, *resultar*, *constituir*, *representar*, *volverse*, *quedarse*, *hacerse*, etc.). All these verbs are relatively slightly developed grammaticalization cases; for that reason the special treatment that the grammar tradition has given to *ser* and *estar* is perfectly justified.

Some studies allude to these verbs as ‘pseudo-copulativos’ (Fernandez Leborans 1999), but in this paper I will speak rather about the copulative function of such items. These verbs establish a nexus between subject and predicative nucleus adding in many cases some meaning related to *Aktionsarten* (perdurative in *Juan anda/sigue molesto*, ingressive in *Juan se volvió muy celoso*) or modality (essentially epistemic: *Juan parece/ luce enfermo*). In this work there will be tackled the diachronic development of copulative functions of some Spanish verbs like *parecer*, *andar*, *lucir*, *hacer*, *mostrarse* etc. using the **Corpus del español** of professor Mark Davies. The analysis is based on functional theories and the results demonstrate the following:

- (1) That such verbs have developed possibilities as copula in different stages of the history of Castilian/Spanish, for example *mostrarse* (‘to show his/herself’) since the 15th Century (“Commo el guerrero temeroso **se muestra generoso** en el campo”); *lucir* (‘to shine’) however, from the 17th Century (“y cual granada **luce sazonada** en el prado florido”)
- (2) That the different uses of the copulas (vid. Rude 1978) don’t arise simultaneously in the same verb, for instance *andar* (‘to walk’) is used to express TEMPORAL ATTRIBUTION since the 13th Century, but the meaning PERMANENT ATTRIBUTION since the 16th (“Como el anda falso y fingido con el señor, assi es su bienaventurança falsa y fingida”), and

- (3) That there is also a gradual incorporation of different types of syntactic constructions (pronoun, adjective, NP, etc.) as predicative elements.

Además de *ser* y *estar*, el español dispone de una serie de verbos que pueden desempeñar funciones copulativas (por ejemplo *andar*, *resultar*, *constituir*, *representar*, *volverse*, *que-darse*, *hacerse*, etc.) En todos estos casos se trata de una gramaticalización relativamente poco avanzada, por lo que el tratamiento especial que la tradición gramatical ha dado a *ser* y *estar* está perfectamente justificado.

En algunos estudios se hace referencia a estos verbos como ‘pseudo-copulativos’ (Fernández Leborans 1999), pero en esta ponencia hablaré más bien de la *función copulativa* de tales unidades. Estos verbos establecen un nexo entre el sujeto y el elemento predicativo añadiendo en muchas ocasiones algún significado relacionado con *Aktionsarten* (por ej. perdurativo en *Juan anda/sigue molesto*; ingresivo en *Juan se volvió muy celoso*, etc.) o modalidad (fundamentalmente epistémica: *Juan parece/luce enfermo*). En este trabajo se abordará el desarrollo de la función copulativa de algunos verbos como *parecer*, *andar*, *lucir*, *hacer*, *mostrarse* etc., utilizando para ello el *Corpus del español* del profesor Mark Davies. El análisis se fundamenta en teorías funcionales, y los resultados demuestran:

- (1) que tales verbos han desarrollado posibilidades copulativas en diferentes etapas de la historia del castellano/español, por ej. *mostrarse* desde el siglo XV (“Commo el guerrero temeroso **se muestra generoso** en el canpo”); *lucir*, en cambio, desde el XVII (“y cual granada **luce sazonada** en el prado florido”);
- (2) que los diferentes usos de la cópula (vid. Rude 1978) pueden no surgir simultáneamente en el mismo verbo, por ej. *andar* se emplea para expresar atribución temporal desde el siglo XIII, pero con el significado atribución permanente a partir del XVI (“Como el **anda falso** y fingido con el señor, assi es su bienaventurança falsa y fingida”), y
- (3) que existe además una incorporación paulatina de diferentes tipos de construcciones sintácticas (pronombre, adjetivo, SN, etc.) como elementos predicativos.

#### References

- Fernández Leborans, Ma. Jesús 1999: “La predicación: Las oraciones copulativas”. En Bosque, Ignacio & Demonte, Violeta (coords.) *Gramática descriptiva de la lengua española. 2. Las construcciones sintácticas elementales*. Madrid: Espasa-Calpe. 2358-2460.
- Leal, Fernando 2003: “Clases de palabras en español”, en Matute, Esmeralda & Leal, Fernando (coords.): *Introducción al estudio del español desde una perspectiva multidisciplinaria*. Guadalajara: Universidad de Guadalajara, 107-139.
- Pustet, Regina 2003: *Copulas. Universals in the categorization of the lexicon*. Oxford: University Press.
- Raible, Wolfgang. 1990. „Types of Tense and Aspect Systems“, en Bechert, Johannes & Bernini, Giuliano & Buridant, Claude (eds.): *Toward a Typology of European Languages*. Berlin [Empirical Approaches to Language Typology, Bd. 8] , 195-214.
- Rude, Noel 1978: “A continuum of meaning in the copula”, en J. Jeager *et al.* (eds.) *Proceedings of the Fourth Annual Meeting of the Berkeley Linguistics Society, Berkeley*, 202-10.

## **Dee Gardner**

Brigham Young University

### **Semantic frequency and the creation of pedagogical word lists: What can we learn from SemCor?**

This paper utilizes SemCor, one of the few semantically-tagged English corpora in existence, to address several key considerations in creating viable word lists for English language education, including homonymy, polysemy, and multi-word items. The actual data for the paper is derived from isolating the SemCor data involving word forms and their tagged senses, and then importing the results into Excel spreadsheets where comparisons of sense-tagged words can be made within and between the various registers that make up the corpus. The primary focus of the analysis is on sense frequencies and sense ranges. In addition to discussing the ramifications of the analysis relative to the creation of pedagogically-valid word lists, the researcher uses the data to briefly address some of the benefits and challenges of computer-processed corpora in terms of informing language educators and their learners about English vocabulary, and providing them with corpus-based tools, such as concordancing, to facilitate the acquisition of the same. It is felt that this line of inquiry is becoming increasingly important as attention in corpus linguistics seems to be shifting from grammar and discourse to vocabulary and semantics. Additionally, there is a growing recognition among many corpus linguists of the need to build useful bridges to language-learning applications and research.

## **Concepción Godev**

University of North Carolina at Charlotte

### **Word-Frequency and Vocabulary Acquisition: An Analysis of Elementary Spanish Textbooks**

Rehearsal, repetition, and frequency are all facets of time, which ultimately is the ever-present factor in language processing and language acquisition. The so-called vocabulary frequency is that aspect of vocabulary that has to do with the number of times, and therefore the duration of contact with a word, which facilitates its acquisition by enhancing the chances to draw learners' attention to it. First-year Spanish instructors often wonder why their students perform below expectations in vocabulary tests, an analysis of vocabulary frequency in textbooks may shed some light on the reasons behind students' low performance.

I will present an analysis of five leading first-year Spanish textbooks that will consist of two parts: 1) mapping their vocabulary lists as well as the vocabulary that appears in their reading and listening input against the vocabulary-frequency information provided in *A Frequency Dictionary of Spanish: Core Vocabulary for Learners* written by Mark Davies and published by Routledge in 2006 and 2) using estimations available in the current literature to infer logistical aspects that need to be considered in first-year Spanish such as amount of contact-time with the material to be learned, depth of vocabulary knowledge, and the relationship between the receptive and productive dimensions. The ultimate goal of this analysis is to develop an instrument that may be used for instructors to understand that textbooks are not always in tune with the reality of the complex processes involved in vocabulary acquisition, it often being the case that the expectations suggested by these textbooks may be unrealistic both in the area of reception and production, and within these two skills, expectations may be even more unrealistic for listening and speaking as they usually lag behind reading and writing development.

## **Grant Goodall**

University of California, San Diego

### **Spoken Spanish in Corpora and in Textbooks: Implications for Acquisition**

This paper presents two test cases showing that (i) there are some significant differences between the spoken Spanish represented in corpora and that represented in introductory Spanish textbooks, and (ii) the way in which the textbooks differ from the corpora is potentially detrimental to learner acquisition.

The first case involves present progressive verbal morphology. Most widely-used textbooks introduce this in the first third of the book, after the present tense but before any others, and continue to use it frequently thereafter. An analysis using oral texts in the Corpus del Español, however, shows that this verbal form is much less frequent in spoken Spanish than forms such as the subjunctive or present perfect, which typically appear in the final third of textbooks (and to which students therefore receive relatively little exposure).

The second case involves reflexive verbs. Despite the name, reflexive morphology in Spanish does not always result in a canonically reflexive interpretation, but may instead induce a change in the verb's aspectual class (among other things). Analysis of oral texts in the Corpus del Español shows that reflexive verbs that at least allow a canonical interpretation (e.g., *levantarse* 'to get up, to lift oneself') are relatively infrequent, whereas many "aspectual" reflexive verbs are extremely frequent (e.g. *irse* 'to go away'). Introductory textbooks, however, generally focus exclusively on the canonical interpretation and provide

almost no exposure to “aspectual” reflexives.

In both of these cases, there are potentially detrimental effects for acquisition. Flooding the input with unrealistically high numbers of progressive forms plausibly makes it harder for learners to comprehend the semantic distinction between simple present and progressive in Spanish (the progressive is much more restricted than in English), while shielding learners from the most frequent uses of the reflexive encourages them to misanalyze the function of reflexive morphology.

## **Athelia Graham**

Brigham Young University

### **Semantic Frequency: a new look at word frequency counts**

This study looks at the significance of semantics in word-frequency counts in response to a call for new word lists (Read, 2000; Gardner, 2007). Read claims that no corpus projects to date have produced any “definitive, stand-alone word-frequency lists” (pg. 226). Many researchers are wary of the fact that the concept of a word is never clearly defined in most studies that have done word frequency counts. It is clear from the research that one universally acceptable construct for the concept of a word does not exist. In fact, many past word frequency counts only examine word-forms without considering the semantics of words and the possible effects of polysemy and homonymy.

Ming-Tzu and Nation (2004) did some research on the Academic Word List (AWL) that somewhat responds to the criticisms of word-frequency lists. They evaluate the extent of polysemy and homonymy through out the AWL. However, words found in the AWL are often not a part of the highest frequency word-forms in English.

The present study focuses on high frequency words. It evaluates a randomized sample of 50 words that occur at least 1500 times in the British National Corpus (BNC). Another random sampling of 200 examples for each word, in context, was analyzed. 100 of these examples were from the written portion and the other 100 from the spoken portion. The meanings for each word were found on WordNet. Each context was double and sometimes triple rated. The results indicate that semantic frequency is indeed significant, in some cases, when evaluating a word’s frequency. The study shows that semantic frequency tends to be significant especially for the most high-frequency words, as well as within certain part of speech groups. Detailed evidence will be given to support both Read’s and Gardner’s claims about distorted representations of frequency in current lists.

## Bethany Ekle Gray

Northern Arizona University

### Comparing stance in qualitative and quantitative research reports

Contrasting research paradigms have differing goals, methodologies, and beliefs about the purpose and application of research. Thus, it is likely that linguistic variations exist that reflect these paradigmatic differences.

The current corpus-based study focuses on hedging features, which show doubt, uncertainty, and imprecision (Silver, 2003), and anticipate consequences of overstatement (Hyland, 1998). The study asks the following research question:

- (1) Do quantitative and qualitative research reports differ in terms of
  - (a) the types and frequencies of hedges used
  - (b) the frequency of sentences containing hedges
  - (c) the average number hedges within sentences
  - (d) the functional reasons behind such markers of hedging?

Two social science disciplines (education and sociology) were selected for study because all conduct research in both paradigms, minimizing the effect of discipline. A corpus was collected from journals in each discipline that published quantitative and qualitative research.

The corpus was tagged, and computer programs were written to analyze a subset of hedging features, which were selected from stance markers identified in Biber (2006) and Biber et al. (1999). Although Biber (2006) and Biber et al. (1999) do not label these stance markers as hedges specifically, they are identified as showing possibility or likelihood (rather than certainty), and thus can be considered hedges that lessen the strength of a statement.

The following features were analyzed: modals of possibility (*e.g., can, could, may, might*), epistemic stance adverbs showing doubt, limitation, and imprecision (*e.g., perhaps, possibly*), and *that*-clauses controlled by likelihood verbs (*e.g., think, believe, doubt*), by epistemic adjectives showing likelihood (*e.g., possible, probable*), and by epistemic nouns showing likelihood (*e.g., hypothesis, claim*).

Functional analyses were performed to explain the differing quantitative patterns revealed in the corpus.

The results from this study provide detailed information that can be used to help novice researchers effectively report research.

### References

- Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5, 97-116.
- Biber, D., Johansson, S. Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. : Longman.

- Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text*, 18(3), 349-382.
- Silver, M. (2003). The stance of stance: A critical look at ways stance is expressed and modeled in academic discourse. *Journal of English for Academic Purposes*, 2, 359-374.

## **Stefan Th. Gries**

University of California, Santa Barbara

### **Measures of dispersion in corpus data: a critical review and a suggestion**

The most frequently used statistic in linguistics in general and corpus linguistics in particular is the frequency of occurrence of some linguistic variable (or the frequency of co-occurrence of two or more linguistic variables). However, as has been pointed out repeatedly, frequencies of (co-)occurrence in isolation may sometimes be severely misleading given that they alone do not take into consideration the degree of dispersion of the relevant linguistic variable in question (cf. Leech, Rayson, and Wilson 2001 and Gries 2006). In order to handle such problems, several scholars suggested a variety of dispersion measures and adjusted frequency measures (cf. Oakes 1998 for an overview). Unfortunately, however, some measures are still largely unknown, still problematic in both theoretical and quantitative respects, and all are underutilized.

I pursue three objectives with this paper. First, I present an overview of a large number of dispersion measures and adjusted frequency measures – including measures that have not found their way into the literature yet – and summarily discuss some of their advantages and disadvantages. Second, I propose for discussion a conceptually very simple alternative measure, *DP* (for *deviation of proportions*), explain and exemplify its properties, and compare it to some more or less established measures on the basis of fictitious distributions from the literature, word frequencies from a stratified sample of all frequency bands from the British National Corpus Sampler, and co-occurrence data from the ICE-GB. I will conclude that *DP* is at least as discriminatory as most if not all other existing measures but conceptually simpler and even better in some respects. Third, I will briefly outline, and exemplify selected aspects of, a research program for future work in this in my opinion surprisingly under-researched area of corpus-linguistics.

### References

- Gries, Stefan Th. 2006. Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54.2:191-202.
- Leech, Geoffrey N., Rayson, Paul, and Andrew Wilson. 2001. *Word frequencies in written and spoken English: based on the British National Corpus*. Longman: London.
- Oakes, Michael. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

## Jack Grieve

Northern Arizona University

### Corpus-based constituency tests and the structural position of auxiliary verbs

Three theories on the structural position of English auxiliary verbs have been proposed: the verb group theory (Chomsky 1957, Biber et al 1999), the ternary structure theory (Chomsky 1972, 1981, Jackendoff, 1972), and the binary structure theory (Chomsky 1975, 1986, 1995, Katz and Postal 1964, Pullum and Wilson 1977). While proponents of these theories assign the same structure to, for example, transitive sentences (1), these theories differ in that they assign different attachment positions for auxiliaries (2-4).

- (1) [that dog] [chases [the bird]]
- (2) Verb group: Verb [that dog] [[will chase][the bird]]
- (3) Ternary structure: [that dog] will [chase [the bird]]
- (4) Binary structure: [that dog] [will [chase [the bird]]]

The disagreement over the position of auxiliaries is because standard constituency tests do not provide clear empirical evidence for any syntactic structure. Of primary importance, the coordination test seems to license both the verb group theory and the binary structure theory, as sentences 5 and 6 are both grammatical.

- (5) Verb group: [that dog] [[[will chase] and [might kill]] [the bird]]
- (6) Binary structure: [that dog] [will [[chase [the bird]] and [find [its nest]]]]

With no direct empirical evidence, generative grammar has thus been forced to adopt binary structure for purely theoretical reasons: it is consistent standard X-bar theory (Chomsky 1986, Radford 1988).

This presentation argues that relevant empirical evidence is available, but only if the standard coordination constituency test is augmented to consider the frequency of sentence-types as opposed to the grammaticality of sentence-types. Based on an analysis of numerous corpora, the relative frequency of sentences of type (5) and type (6) are compared. It is discovered that sentences of type (6) are far more common, and it is thus concluded that there is empirical evidence for the binary structure theory, for reasons of derivational economy (Kidwai 2000).

### References

- Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.
- Chomsky, N. (1957). *Syntactic Structures*. Berlin: Mouton de Gruyter.
- Chomsky, N. (1972). *Studies on Semantics in Generative Grammar*.

- Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*.  
Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa Lectures*. Berlin: Mouton de Gruyter.  
Chomsky, N. (1986). *Barriers. Linguistic Inquiry Monograph Thirteen*. Cambridge, MA: The MIT Press.  
Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: The MIT Press.  
Katz, J. and Postal, P. (1964). *An Integrated Theory of Linguistic Descriptions*. Massachusetts Institute of Technology.  
Kidwai, A. (2000). XP-adjunction in Universal Grammar. Oxford University Press.  
Pullum, G. & Wilson, D. (1977). Autonomous syntax and the analysis of auxiliaries? *Language* 53, 741-788.  
Radford, (1988). *Transformational Grammar*. Cambridge University Press.

## Angus B. Grieve-Smith

University of New Mexico

### Controlling for Fads in Historical Corpora

Historical corpora can tell us many things about the history of a language, but the picture they present can be distorted by fads, literary movements and other top-down trends in writing. We can avoid this distortion by formulating basic expectations of non-distorted corpora beforehand and testing corpora to ensure that they correspond to these expectations. Based on these tests, we can identify problematic texts and compensate for them. In order to properly control for these trends, it is indispensable to have an understanding of the social history of the texts in the corpus.

I came across this problem in compiling a corpus of French theatrical texts to study the history of negation in the language. Throughout the written history of the French language there has been a general shift from the use of the preverbal negator *ne* to the “embracing negation” constructions *ne ... pas* and *ne ... point*, and Kroch’s (1989) Constant Rate Hypothesis predicts that the use of the embracing negations will increase following an S-shaped logistic curve. Contrary to this prediction, the initial data from my corpus actually showed a small decrease in use of embracing negations during the late sixteenth and early seventeenth centuries, followed by a sharp increase in the mid to late seventeenth century.

Comparing the data to the predicted S-curve, I found that the lowest use of embracing negations was in those subgenres that were most influenced by the *Pléiade* movement, which sought to emulate Classical Latin texts (Degaine 2002), while the other subgenres continued to follow the S-curve. Plays written after the beginning of the seventeenth century are also consistent with the S-curve; this more independent standard can be attributed to the reaction to the *Pléiade* led by Malherbe (Brunot 1891).

**Robbie Haertel**

Brigham Young University

**MayanWiki:  
Facilitating Consensus Through an Openly Editable Corpus**

The writing system used by the ancient Maya civilization has intrigued researchers and aficionados for centuries. Now that it has mostly been deciphered, the emphasis in the field of Mayan epigraphy has shifted to a study of the system of phonological, morphological, and grammatical rules that once governed the language that the hieroglyphs encode. Linguistic study of this type could be facilitated by a publicly available, comprehensive, electronic corpus of texts to investigate phraseology, frequency information, and collocations, as is done in more widely studied languages such as English. Such a resource would assist not only Mayan epigraphers, but linguists, archeologists, anthropologists, students, and hobbyists. However, a corpus of the hieroglyphs presents special challenges. For one, new texts are continually discovered. More importantly, since Mayan linguistic epigraphy is in its infancy, there is considerable disagreement concerning such issues as phonology, morphology, etc.; even the question of whether the language was the direct ancestor of Colonial Ch'olti' or a more distant ancestor is disputed. Unfortunately, a privately run database reflects only the viewpoints of the maintainer and is difficult to maintain under these circumstances. Such a corpus does not necessarily serve the community as a whole; instead, a corpus with decentralized control is needed.

To address these issues, MayanWiki is a corpus of transcribed and transliterated hieroglyphic texts that is openly editable by all. At the heart of MayanWiki is a relational database that is capable of storing glyphic and linguistic data, including annotations indicating unknown and reconstructed readings. The types of searches mentioned previously can be performed for detailed linguistic analysis. The principle behind the wiki is to accelerate the convergence of readings to a consensus, which is encouraged through a policy of "conservative transcriptions, innovative explanations". Furthermore, because it is a wiki, new texts can be added as they become available. Once the database is fully populated by users, it will become a valuable tool allowing the textual data to be manipulated in ways that will facilitate scientific discovery of new and interesting linguistic patterns. Most importantly, it will continually evolve to reflect the latest research in the field.

## **Patrick Hanks**

Masaryk University

### **A Corpus-Based Pattern Dictionary for Mapping Meaning onto Use**

This paper presents a “Pattern Dictionary of English” (work in progress), being developed as an infrastructure resource for a wide variety of applications in NLP, language teaching, and lexicography. Traditional word sense disambiguation (WSD) has, in the words of Ide and Wilks (2005) “proven to be problematic for NLP”. Our project will provide a resource that side-steps the WSD problem by attaching meanings to patterns rather than to words in isolation. Patterns are, for the most part, mutually exclusive, whereas word senses are not.

The patterns are based on intensive statistically aided corpus pattern analysis (CPA). A distinction is made in CPA between normal usage and exploitations of normal usage (e.g. ellipsis, dynamic metaphor, puns, etc.). In parallel with discovering patterns, the project is developing a shallow ontology of “shimmering” lexical sets, which (unlike WordNet, the best resource currently available) is empirically well founded.

The differences between CPA (word-pattern based) and FrameNet (frame based) will be summarized. One such difference lies in the methodology: CPA is corpus-driven; while FrameNet is theory-driven.

Brief mention will also be made of the potential for beneficial interaction between CPA and other research, in particular:

- Word-class tagging
- Shallow parsing (context-sensitive)
- Pronominal anaphora resolution

#### References

Ide, Nancy, and Yorick Wilks. 2005. ‘Making Sense about Sense’ in Eneko Agirre and Philip Edmonds (eds.) Word Sense Disambiguation. Springer.

## **Cristina Hansen**

Universidad de La Laguna

### **Corpora of Spanish versus an educational text of astronomy**

The relative frequencies of words appearing in an educational text of astronomy are compared with the results of different corpora of Spanish. These are the following: Juilland &

Chang-Rodríguez, 1964; Alameda & Cuetos, 1995; Justicia, 1995; LEXESP 2000; Davies 2002; Almela et al. 2005; REAL ACADEMIA ESPAÑOLA: Banco de datos (CREA) 2007.

Qualitatively there seem to be no differences between corpora, with the only exception being the corpus of Justicia. This is understandable as this corpus was built after lemmatisation and from texts written by students. All corpora, except the one of Justicia, show a similar distribution of relative frequencies for the words studied, regardless their size or selection of texts.

Quantitatively there seem to be bigger differences in words which could be considered necessary to build any discourse, that is, these words are not specific of any particular subject. We are talking of articles, prepositions, demonstrative and possessive adjectives, etc.

The comparison of corpora allows us to compute the average relative frequency and the standard deviation for each word. This average is compared with the results from our educational text. Many of the words which could be considered specific of the studied subject show higher quantitative differences with regard to the average.

## **Huaqing Hong and Paul Doyle**

Nanyang Technological University

### **Annotation, Indexing and Querying a Multilingual, Multimodal Classroom Discourse Corpus**

The application of computer corpus methodology to the study of classroom discourse is comparatively a very new research area (Seedhouse 2004 & 2006; Al-Garawi 2005; Walsh 2006). On the one hand, the shortage of large scale collection of data and extensive analysis makes it difficult to generalize the educational research results on the basis of traditional pen-and-paper approach or observation-based approach (Biber *et al.*, 2004) to classroom discourse. On the other hand, the lack of sophisticated annotation, indexing and querying tools also makes the attempt to do any large scale of analysis an expensive and labor-intensive task. In this respect, we need a comprehensive methodology, which can effectively adapt the computer technology of corpus tools to facilitate the description, analysis and evaluation of classroom discourse.

It is noted that the exercise of analysis of classroom interactional data as speech genres and discourses, we need both empiricist approaches to interaction or behavior (thus linguistic or paralinguistic features) and functional-communicative approaches of classroom language as conversation (Rojo, 2000). With the results from the on-going Singapore Corpus for Research in Education (SCoRE) project (Hong 2005), this paper reports how corpus linguistic methodology can be turned into a synergy of these two

approaches to pedagogical research as well as linguistic analysis of classroom practice. In doing so, first we discuss the feasibilities of applying a computer corpus approach to classroom interactions, and identify a set of evaluation criteria that suits best for the approach. Next, we demonstrate how to annotate, index and query the corpus of such a kind. A list of problems encountered and the answers to the challenges in various stages of corpus processing will also be reported. To conclude, we present the significance and implications of this study.

### References

- Al-Garawi, B. 2005. A Review of Two Approaches to L2 Classroom Interaction. *The Annual Review of Education, Communication, and Language Sciences*. Vol. 2.
- Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. 2004. *Representing language in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. ETS TOEFL Monograph Services MS-25. Princeton, NJ: Educational Testing Service.
- Hong, Huaqing. 2005. SCORE: A Multimodal Corpus Database of Education Discourse. *Proceedings of International Conference of Corpus Linguistics (ISSN 1747-9398), Birmingham, July 14-17, 2005*.
- Rojo, R. H. R. 2000. *A discursive approach to classroom interactions as speech genres: from heteroglossia and social languages to authoritative discourse*. Retrieved on July 05, 2007, at: <http://www.fae.unicamp.br/br2000/trabs/1050.doc>.
- Seedhouse, P. 2004. *The interactional architecture of the language classroom: a conversation analysis perspective*. UK: Blackwell Publishing.
- Seedhouse, P. 2006. Classroom Interactions. *Revue Française de Linguistique Appliquée*. Vol.11(2), 111-122.
- Walsh, S. 2006. *Investigating Classroom Discourse*. Routledge: London & New York.

## **Katherine Horwinski Healy**

Louisiana State University

### **Creole African American Vernacular English: Origins of a Dialect**

Free People of Color, particularly Colored Creoles of eighteenth and nineteenth century Louisiana, have attracted significant academic attention within the last few decades, and progress has been made in identifying elements of their social and linguistic history. Historically French speaking, the Creoles of Color were ethnically and culturally mixed, making it difficult to know when they began learning English and from whom. As they acquired English, a distinct dialect developed called Creole African American Vernacular English (CAAVE). In order to trace the origins of this dialect, I have collected a corpus of letters from the 19th century. They belong to three distinct groups of free people of color: members of a formerly prosperous Colored Creole planter family post Civil War era (48 documents), freed slaves sent to colonize Liberia (30 documents), and two plantation owners (33 documents).

Letters written by each group have been transcribed, remaining faithful to the original grammar, orthography, capitalization, and line breaks. They are being analyzed for linguistic features which may help trace the origins of non-standard features currently found in Creole English to other ethnic groups with whom the Creoles would have been in contact, and could have influenced early CAAVE. Special attention is given to specific morpho-syntactic features identified in the language of other ethnic communities from the same time period and region, and a general comparison is presented (Bailey 2001; Dubois & Horvath 2003; Schneider & Montgomery 2001; Van Herk & Poplack 2003). Uses of non-standard auxiliaries (have/has/had, was/were, is/are and deletion of any of these) are coded, as well as use of inflexible copula *be*, omission of plural or possessive *-s*, *a/be + verb + ing*, multiple modals, double negatives and a variety of lexical or formulaic items such as *ain't*, *for to*, *y'all*, *fixin' to*, *be done*, demonstrative *Them*, and irregular preterits (i.e. *knowed*). The presence or absence of these features, as well as frequency and linguistic constraints dictating their use, allows comparison between the Creoles of Color and other ethnic groups of the time, and paints a clearer picture of the social and cultural contact between communities.

#### References

- Bailey, Guy. "The relationship between African American and White Vernaculars in the American South." Lanehart, Sonja L. (ed.). *Sociocultural and Historical Contexts of African American English*. Philadelphia: John Benjamins Publishing Company, 2001. 53-92.
- Dubois, Sylvie & Barbara Horvath. "The English Vernacular of the Creoles of Louisiana." *Language Variation and Change* 15 (2003): 253-186.
- Schneider, Edgar & Michael Montgomery. "On the Trail of Early Nonstandard Grammar: An Electronic Corpus of Southern U.S. Antebellum Overseer's Letters." *American Speech* 74.4 (2001): 388-410.
- Van Herk, Gerard & Shana Poplack. "Rewriting the past: Bare verbs in the Ottawa Repository of Early African American Correspondence." *Journal of Pidgin and Creole Languages* 18.2 (2003): 231-266.

### Jian Huang and C. Lee Giles

Pennsylvania State University

#### **An Efficient Framework for Large Scale Cross Document Coreference (CRC)**

As a reader, one can readily decipher the reference of the name *Clinton* in the phrase "Clinton campaign camp", link this person to the same Senator Clinton appearing in the other articles read in 2007 and distinguish it from that of President Clinton. The problem of coreferencing names appearances across multiple documents, or cross document coreference (CDC), necessitates a solution that combines various natural language processing, data mining and machine learning techniques.

We propose a framework to efficiently determine the identity of a person based on the context of a name appearance and the metadata of an article. This extracted information includes unary properties such as gender and title, as well as binary relationships with other entities such as co-occurrence with other names and affiliation with political parties. The framework works in a divide-and-conquer fashion: it first segments the name appearances into non-overlapping blocks for the purpose of significant reduction of similarity computation. Smaller candidate sets of name appearances within a block that potentially refer to the same person are then identified by the winnower, which also enforces the compatibility of the records within a candidate set. The similarity of a pair of records is determined by a committee of experts, where each expert specializes on a particular type of relationship and measures the similarity of the relationship by using various syntactic and/or semantic matching techniques. To ensure the consistency of coreference decisions, a relational density-based clustering method DBSCAN is used to delineate name clusters based on the pairwise similarities provided by the committee of experts.

We demonstrate the practicality of our framework with experiments on large (hundreds of megabytes) real world news and internet data sets. Human evaluation shows the capability of the system to accurately distinguish namesakes and cluster name variations of the same person across documents.

## **Susan Hunston**

University of Birmingham

### **You can't deny the fact that...: An Application of Corpus Linguistics**

This paper attempts to demonstrate how techniques of corpus linguistics (word searches, collocation and phraseology identification) might be used to supplement findings from discourse analysis and epistemology. It discusses the concept of 'epistemic status', and its linguistic realisations, relating these to recent work in economic history in a project entitled 'How Well Do Facts Travel?' (Morgan et al 2007). It then presents a series of corpus studies of the words 'fact' and 'facts' to suggest how studies of this kind might assist the work of the economist.

**Shin Ishikawa**

Kobe University

**A Statistical Analysis of CEEJUS,  
Corpus of English Essays by Japanese University Students**

**Aims**

Learners' corpus can greatly contribute to TEFL (Teaching English as a Foreign Language). Many scholars have examined the NS (native speakers) and NNS (non-native speakers) gap seen in their essays or speeches (Granger ed.,1998, Ringbom 1998, Flowerdew, 2000). However, most of the previous studies pay attention to European learners of English with a considerably high level of proficiency. Therefore, how learners at a relatively lower proficiency level use the vocabulary and the expressions in the essays has not been sufficiently analyzed. In the present study, based on the newly compiled corpus of English essays written by Japanese university students, we will probe the features of their characteristic vocabulary use.

**Procedure**

The presenter has compiled The Corpus of English Essays by Japanese College Students (CEEJUS), which collects the essays written by approximately 250 Japanese learners as well as detailed proficiency data about writers. Variables possibly giving influence on the lexical quality of the essays such as time, topic, quantity, and dictionary-use in essay writing, are strictly controlled. We calculated the lexical indices such as tokens, types, TTR, Guiraud index, mean word length, and so on, and statistically compared them with the parameters of writers' receptive English proficiency such as the score in the TOEIC(R) test and the vocabulary size.

**Results**

Firstly, correlation of lexical indices was examined (Table 1). Then, using the technique of structural equation modeling (SEM), we attempted to create a statistical model explaining the complicated relationships between lexical indices and proficiency indices (Fig. 1). Finally, using the technique of correspondence analysis, we identified key words used characteristically by Japanese learners at different proficiency levels (Table 2).

Table 1. Correlation of lexical indices

	Cha	Tokens	Types	STTR	MWL	Sent	MSL
Cha	1	0.94*	0.7*	0.15	0.18*	0.53*	0.07
Tokens	0.94*	1	0.67*	0.02	-0.12	0.54*	0.1
Types	0.7*	0.67*	1	0.58*	0.12	0.33*	0.09
STTR	0.15	0.02	0.58*	1	0.37*	0	0
MWL	0.18*	-0.12	0.12	0.37*	1	-0.07	-0.01
Sent	0.53*	0.54*	0.33*	0	-0.07	1	-0.75
MSL	0.07	0.1	0.09	0	-0.01	-0.75*	1
Sum of r	3.57	3.15	3.49	2.12	1.47	1.58	0.5

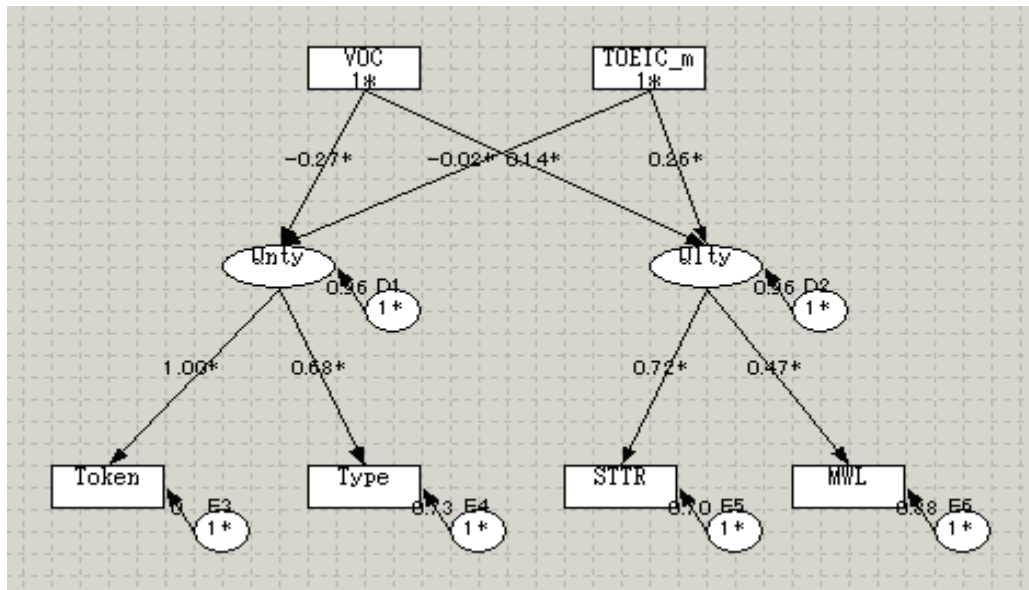


Fig 1. Structural Equation Modeling

Table 2. Key words at each proficiency levels

Category	Grade		Proficiency						
Var	1	2	11-15	16-20	21-25	26-30	31-35	36-40	41-45
N	72	137	3	14	58	63	54	16	1
1	Moreover	statement	wrong	did	No	Of	hour	band	Coffee
2	may	could	both	Besides	Such	Through	adults	move	Commonly
3	College	agree	Eventually	appeal	date	addition	rather	members	Starbucks
4	consider	talk	Unless	field	favorite	costs	themselves	kind	barista
5	serious	Thirdly	accept	retire	immediately	relationship	pain	regard	beans
6	university	following	avoided	Thanks	men	variety	selves	text	beverages
7	study	my	based	giving	mine	brought	Last	toward	bit
8	confused	depend	behaviors	talked	off	worry	dishes	word	fairly
9	stop	having	coming	daily	raised	Having	yen	him	issue
10	expenses	Secondly	energies	manners	Why	points	even	learning	livelihood
11	doing	daily	entertainment	Perhaps	cases	case	three	colleges	memorize
12	if	means	excessive	delicious	early	independent	My	store	relax
13	If	polite	furniture	probably	expected	trouble	Third	everyone	tastes
14	raised	myself	happen	man	places	human	these	instead	brings
15	until	There	hearing	life	easy	useful	always	knowing	coffee
16	knowledge	way	minority	his	too	club	feel	music	heard
17	What	actually	radical	bad	soon	When	free	pleasure	mentioned
18	lose	office	sleeping	teacher	After	way	aren	rely	attractive
19	experience	wanted	thought	around	done		careful	teachers	leisure
20	great	many	anxiety	cook	just		earlier	support	decided

## Conclusion

In this study, we found out that the writers' receptive English proficiency give quite limited influence on the lexical quality of the essays which they write. The striking gap between receptive proficiency and productive proficiency might be attributed to the peculiarity of Japanese learners of English as NNS.

## Gard Jensen and Christer Johansson

Bergen University

### Estimating the saliency of constructions using document frequencies from the web

We present results expanding on ideas from Goldberg (1995), and Gries & Stefanowitsch (2004), regarding the semantics of the English ditransitive construction (DTC). We propose the semantics of the DTC can be understood through significant and relevant patterns in very large sets of unannotated data.

No fundamental difference between annotated, edited corpora and other collections of usage events, such as the Internet accessed through a search engine need be posited.

Tummers et al (2005: 231) define a corpus as “a collection of non-elicited usage events.”

Searching for instantiated frequent patterns estimate relevance. Saliency of the patterns is estimated by an information theoretic measure: frequency adjusted mutual information (Lin, 1998; Church and Hanks 1990). We use document frequencies, and present an adjusted mutual information measure. We develop an analysis of variance to explain the choice of construction by factors.

Definition: **Mutual Information:  $\mu$**  =  $\log( \frac{p(\text{pattern})}{p(\text{words in pattern})} ) \approx \log(G \cdot df(\text{pattern})/df(\text{words in pattern}))$ ; where df is the document frequency, and the constant G is used to scale results. (Here, G = 1000,000, and log=log10).

Definition: **Saliency** =  $\mu \cdot \log(df(\text{pattern}))$ ; following in spirit the definitions of Lim (1998) and Modjeskaja et al. (2003).

The two constructions are:

1. “give X to him/her”
2. “give him/her X”

The factors are a) gender of the benefactor; b) abstract or concrete object; and c) the tense of the verb. We consider three near synonymous verbs (give, hand, donate).

We find interesting differences between constructions 1 and 2 regarding the verb and the gender of its object pronoun. This is consonant with the findings of Gries and Stefanowitsch (2004), but our results make their claims about distance of transfer less likely (cf. *ibid* p. 105ff). Specifically, we found patterning opposite their predicted preference.

## References

- Church, Kenneth Ward, and Patrick Hanks (1990). “Word association norms, mutual information, and lexicography.” *Computational linguistics* 16, 22-29.
- Goldberg, Adele E (1995). *Constructions*. Chicago: University of Chicago Press.
- Gries, Stefan Th., and Anatol Stefanowitsch (2004). “Extending collocation analysis: a corpus-based perspective on ‘alternations’” *International journal of corpus linguistics* 9, 97-129.
- Lin, Dekang. 1998. Extracting Collocations from Text Corpora. Workshop on Computational Terminology. pp. 57-63. Montreal, Canada.
- Modjeska, Natalia N., Katja Markert, and Malvina Nissim (2003). Using the web in machine learning for other-anaphora resolution. In *Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing Sapporo, Japan*, pages 176–183.
- Tummers, Jose, Kris Heylen, and Dirk Geeraerts (2005). Usage-based approaches in Cognitive Linguistics: a technical state of the art. *Corpus linguistics and linguistic theory* 1, 225-261.

**Patrick Juola**  
Duquesne University

### **Authorship Attribution: What Mixture-of-Experts Says We Don't Yet Know**

“Authorship attribution” or “stylometry” can be simply defined as the inferring of the author of a document or his/her properties through analysis of the text of the document itself. This paper summarizes some recent developments in empirical testing of authorship attribution, most notably the 2004 *Ad-hoc Authorship Attribution Competition* [1, 2]. Contest materials included thirteen problems, including a variety of lengths, styles, genres, and languages, mostly gathered from the Web but including some materials specifically gathered for this purpose. We reanalyze the results of this competition, and show that higher performance yet can be obtained via a mixture-of-experts and combining judgments.

A further goal of the AAAC was to encourage “researchers both to focus on the important differences between methods and to mix and match techniques to achieve the best practical results. [...] It is to be hoped that from this [mixing], researchers can identify the best inference techniques and the best models in order to assemble a sufficiently powerful and accurate system.” [3] We have taken up this challenge, applying a simple “mixture of experts” approach to try to find such an improvement. We have therefore combined the top three (five) performing methods and reanalyzed the published results using a simple unweighted plurality-wins rule.

The plurality of the top three experts scored 914% total percentage correct and the plurality of the top five scored 924% correct; in contrast, the top five individual experts scored 918%, 897%, 861%, 850%, and 804%, respectively. The plurality of the top three outscored all but the single best expert, while the collective expertise of the top five outscored any individual (as well as the expertise of the top 3).

The overall results have therefore not only borne out our predictions about the usefulness of multiple experts for such a problem, but also Juola’s original predictions about the usefulness of the AAAC framework for generating and testing new methods. It also further illustrates the need for greater theoretical analysis to understand the strengths and weaknesses of the individual methods applied.

#### References

- [1] Patrick Juola. Ad-hoc authorship attribution competition. In *Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities* (ALLC/ACH 2004), Göteborg, Sweden, June 2004.
- [2] Patrick Juola. Authorship attribution for electronic documents. In Martin Olivier and Sujeet Sheno, editors, *Advances in Digital Forensics II*, volume 222 of *International Federal for Information Processing*, pages 119–130. Springer, Boston, 2006.
- [3] Patrick Juola, John Sofko, and Patrick Brennan. A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21(2):169–178, 2006. Advance Access published on April 12, 2006; doi: doi:10.1093/lc/fql019.

## **Magdi Kandil**

Georgia State University

### **The Israeli-Palestinian Conflict in American, Arab, And British Media: Corpus-Based Critical Discourse Analysis**

Critical Discourse Analysis has often been successfully used to interpret the ideological underpinnings of a text. The methodology employed by CDA, however, has at times been severely criticized. One such criticism came from Stubbs (1997), who also proposed a number of steps that are likely to strengthen CDA methodology. One of his suggestions is that CDA analysts should conduct their analyses on relatively large corpora, using quantitative as well as qualitative techniques of analysis so that they can safely make some generalizations regarding a typical use of language. This study takes up this proposition by utilizing the computational techniques of Corpus Linguistics (CL) to investigate the similarities and differences between American, British, and Arabic media coverage of the Israeli-Palestinian conflict, one of the longest and most violent conflicts in modern history. The research is conducted on three news article corpora compiled from the CNN, BBC, and Al-Jazeera Arabic news websites over a period of twenty-seven months. The data is manipulated using a variety of corpus tools: the key-word and key key-word functions are used to identify the topics that tend to be emphasized in each news source, and the concordance and collocational tools are used to identify the semantic profiles of the different participants in the conflict. The ideological significance of the findings is discussed in light of CDA theory. It is argued that the use of corpus-based techniques can successfully address some of the most severe criticisms leveled against CDA.

## **Betsy Kerr**

University of Minnesota, Twin Cities

### **Semantic Anglicisms in Contemporary Metropolitan French**

A well-known category of borrowings from English to French is that of semantic borrowings, defined by Picone 1996 as follows: "This is when a preexisting French word, morpheme or locution shifts in meaning or becomes more extended or more restricted in meaning due to imitative language contact with English." In the field of language teaching, semantic Anglicisms are considered to constitute errors. However, a certain number of such borrowings are now widely used by native speakers, and more or less accepted by the general public and certain linguistic authorities, despite the continuing protests of

language purists. The most cited example is Fr. *réaliser*, previously restricted to the literal sense ‘to make real’, but now commonly used for ‘to become aware’, thus apparently displacing the Fr. (reflexive) verbal expression *se rendre compte*.

Semantic Anglicisms are treated in Picone’s very thorough 1996 study of Anglicisms and neologisms in contemporary French. Rifelj 1996, a much smaller-scale study, addresses the trend toward greater acceptance of such borrowings, citing a couple dozen examples.

The present study attempts to provide both an updated and a more detailed description of the current usage of semantic Anglicisms that have entered the language or significantly increased in frequency over the last half-century. Such Anglicisms will be identified with the help of such sources as Picone 1996 and Rifelj 1996, as well as didactic texts and purist commentaries, and through reading of current texts. Data on current usage will be gathered through the use of corpora such as corpora of recent journalistic texts (e.g. Chambers corpus), recent corpora of spoken French (e.g. Corpus de référence de français parlé), the Leeds Internet Corpus, and ordinary Web browsers. The study will seek to answer the questions: What is the extent of usage of the given Anglicisms? Is their use restricted to various media and registers? Are there any discernible trends over the last two decades, such as spread across registers of more established borrowings, or an increase in the number of new semantic borrowings? Dictionary entries will also be consulted as an additional indicator of status of a borrowing.

## **Ekaterina Lapshinova-Koltunski**

University of Stuttgart

### **Semi-automatic Classification and Extraction of Predicates from German Text Corpora**

The present research aims at semi-automatic classification of automatically extracted lexical data based on their subcategorization. Our results should contribute to the creation of subcategorization lexicons or the enhancement of existing ones. We use non-probabilistic methods of extraction (the lexical data are created to serve symbolic grammars).

The lexical information retrieved with acquisition tools can be stored in machine-readable lexicons and updated dynamically. Most existing linguistic studies concentrate exclusively on verbal predicates and there is an absence of extraction tools for certain classes of lexical data.

We focus on four types of predicates (verbs, nouns, adjectives and MWEs) retrieving them with their sentential complements (that-, wh- or if-subclauses) although our methods can be applied to other complements as well. We use German verb-final sentences as

an extraction context, because in this word order model we have a regular sequence of elements: the subcategorized subclause follows the verb, nominal or adjectival predicates tend to precede it.

Our extraction work starts from sentence-tokenized German newspaper corpora (a total of ca. 950M words) which are pos-tagged and lemmatized. Regular expressions for data extraction rely on the Stuttgart CorpusWorkBench. The extraction steps proceed from the general to the specific. General queries are underspecified and aim at sentences consisting of a main clause and a subordinate clause. Specific queries are used for further refinement and specification. They are applied to sentences extracted with a general query, and contain additional constraints for predicates which precede the verb. The extracted subordinate clause can be subcategorized by the verb, by the element preceding the verb or their combination.

Our experiments show that different kinds of predicates have their own subcategorization and contextual properties which should be considered in lexicon acquisition. This calls for tools to identify such cases by means of data extraction from corpora. We propose precision-oriented semi-automatic extraction which can operate on a tokenized, tagged and lemmatized text and should be elaborated with linguistic knowledge. In the future, we are going to extend the kinds of extracted complements beyond subclauses, and use more data, to achieve substantial coverage.

## Sander Lestrade

Radboud University Nijmegen

### Finnish case alternating adpositions: A corpus study

In Finnish, the adpositions *lähe*- ‘near’, *keske*- ‘in the middle’, and *ympäri*- ‘(a)round’, can assign both genitive and partitive case to their objects, all other adpositions exclusively assign either of these cases. In this paper, I will show how we can account for this adpositional case alternation. I will argue that the use of the partitive case results in an (abstract or fuzzy) extension of the meaning expressed by the genitive PP construction. Also, I will show how word order interacts with case assignment. Both findings are formalized in a Bidirectional Optimality Theoretic framework.

I will corroborate my analysis with corpus findings, using on-line newspaper corpora (CSC). It will be shown that the case alternating behavior of these adpositions is not attested across the board. Although the three adpositions in principle could assign either case to their objects, they tend to assign the same case to the same object over and over again. Only a subgroup of 25 adpositional objects is assigned both genitive and partitive case by the same adposition(s). I found a significant difference in frequency of these

different object classes, the mean frequency of the class of case alternating objects being significantly higher than that of nouns that appeared in one case only. I will argue that this difference in frequency is due to the difference in goodness of fit of the objects.

**Jianguo Li, Kirk Baker, and Chris Brew**

Ohio State University

### **A Corpus Study of Levin's Verb Classification**

Levin groups English verbs into classes that share both common semantics and syntactic alternations. This classification has been shown to be important in a wide range of language tasks, such as lexical resource construction and semantic parsing. The objective of this paper is to investigate to what extent Levin's methodology and classification are empirically supported by the corpus data. In particular, we evaluate the utility of different types of feature space to support automatic assignment of verbs into Levin classification. Automatic verb classifications are needed because manual creation usually requires many years of expert efforts.

- (1) Levin classifies verbs according to their syntactic alternations. In automatic verb classification, which type of features is more useful? We perform experiments using the Bayesian Multinomial Regression (BMR) with three types of feature space extracted from the Gigaword Corpus: subcategorization frame (SCF), neighboring word (NW) and enriched neighboring word (ENW: a mixture of words and part-of-speech tags). Our results demonstrate that using lexical information (NW and ENW) achieves a significantly higher accuracy than using syntactic information (SCF).
- (2) Levin uses a set of expert-defined SCFs (e.g. V-NP-NP, V-NP-PP(to)) for her classification. Are these expert-defined SCFs more helpful? We repeat the experiments using two different sets of SCFs. The first is obtained by extracting all SCFs from the Gigaword Corpus and the second by matching each SCF from the first set to one of the expert-defined SCFs. Our experiments show no difference in classification accuracy between using these two sets of SCFs.
- (3) Levin classifies verbs based on whether a given verb participates in a particular alternation, with regard to frequency. How much do we benefit from frequency information? We use SCF as features, with or without the frequency counts from the corpus. The results show that including frequency information significantly improve the accuracy.
- (4) It is implied that all the distinctions made in Levin's classification are of equal validity. Does the corpus data support some classes more strongly than others?

Our experiments reveal that some Levin classes are a lot more difficult to be classified than others.

Our study demonstrates that Levin's methodology in verb classification is not well empirically supported by the corpus data processed with the state-of-the-art NLP tools. We provide an analysis of the reasons for this result. Our experiments also show that lexical information is more useful in automatically deriving Levin-styled verb classification.

## **Kerstin Lindmark**

Stockholm University

### **A corpus-based investigation of cognate prepositions in English and Swedish**

The overall aim of my translation corpus project is to capture phenomena that cause problems in translation. Within this framework I am currently studying the translation of prepositions from English into Swedish, using the English-Swedish Parallel Corpus, my corpus of translations made by translation students, and elicitation tests in which the subjects are translation students, translation trainees in industry, experienced translators, and non-translators (people who use both Swedish and English professionally but lack translation experience). Prepositions in general are a notorious source of confusion, in part because of the highly lexicalised nature of prepositional phrases. What is more, in the two languages in question, prepositions are in many cases cognates, and while they do share several semantic features, there are differences in their use. The prototypical meaning of the cognate prepositions is often, but not always, clearly spatial. Examples are "over – över", "under – under", "for – för", "in – i".

In this paper, I explore a set of interlinguistically cognate prepositions in an attempt to make a corpus-based, WordNet-like categorisation of their senses in usage. Rather than using dictionary definitions as a primary point of departure (as in the Preposition Project), the English prepositions are investigated on the basis of their usage as shown in the BNC, and their Swedish counterparts are investigated using the available monolingual Swedish corpora, mainly Språkbanken and the Stockholm-Umeå Corpus. The language-specific patterns found are then linked together in order to obtain a crosslinguistic categorisation. Whereas the prototypical, spatial meanings can be expected to show a substantial degree of overlap in usage, secondary and extended meanings are expected to differ to various extents.

A subsequent step in my translation corpus project will be to compare the usage of prepositions in translated and non-translated texts, using this categorisation of meanings as a reference.

## References

- Ejerhed, E. and Källgren, G. 1997. "Stockholm Umeå Corpus version 1.0, SUC 1.0". Department of Linguistics, Umeå University.
- ESPC web site. <http://www.englund.lu.se/content/view/66/127/>. Visited 28 September 2007
- Språkbanken web site. <http://spraakbanken.gu.se>. Visited 28 September 2007.
- The Preposition Project website. <http://www.cres.com/prepositions.html> Visited 28 September 2007

**Deryle Lonsdale**

Brigham Young University

**Developing a Corpus for a Morphologically Rich, Endangered Language**

The Salish language family consists of some two dozen Native American languages spoken in the Pacific Northwest and in southwestern Canada. Many of these critically endangered languages have only a few dozen or fewer (mostly elderly) native speakers.

This paper discusses efforts to collect, annotate, and deploy a corpus of texts for Puget Salish. The challenges facing such an effort are appreciably different from challenges most often facing developers of corpora for more commonly spoken languages. The texts include a dictionary, several audio recordings transcribed in the Americanist orthography and translated with sentence glosses, plus various written materials developed by teachers of the language. The paper describes innovations to standard corpus linguistic methods required for this project; we expect these may apply to corpora for other languages that lack abundant existing language resources. We focus on the need to develop novel annotation methods.

For example, simply digitizing and standardizing the Americanist-script texts is non-trivial. OCR is also challenging.

Salish languages have amazingly complex morphological structure, and a meaningful corpus requires that the words be annotated for their inflectional and derivational morphemes including the root, affixes, clitics, and reduplication patterns. We discuss how finite-state tools have been developed to compute morphological parses for words. We present the results of trying two different syntactic parsing strategies for describing constituency. We have converted the results from these layers of analysis into relational database format, supporting queries over encoded textual data at the lexical, morphological, and syntactic levels.

The audio recordings form an essential part of the data. We discuss ongoing attempts to link textual corpus resources with speech data in digital and analog recordings, and the best practices followed for annotation of such information.

The dictionary forms the backbone of the project. We discuss how the lexical data,

which was rehabilitated from legacy formats, is used in all aspects of the corpus resource development for this language.

Though a comprehensive corpus for this language will never approach the size of well-known corpora for more traditional languages, the lessons learned from this work is likely to be of interest to corpus linguists everywhere.

## **Deryle Lonsdale and Yvon Le Bras**

Brigham Young University

### **Compiling a new French frequency dictionary**

We are in the process of compiling and developing, under a publisher's contract, a new frequency dictionary of French. It is based entirely on a 22-million word corpus and targets learners who want to focus on the most frequent 5,000 word types in the language. The dictionary will provide a valuable pedagogical tool to teachers and learners of the French language. Various frequency dictionaries for French do exist, but they suffer from significant limitations. Most importantly, all are based exclusively on written or spoken French, but not on a combination of these modalities. The existing dictionaries are also outdated given current corpus analysis capabilities, do not take into consideration the type/token distinction, only focus on one variety of French, are in exclusively electronic form, or are only accessible via costly subscriptions.

We have largely completed the task of collecting and cleaning texts for the corpus. It consists of both spoken and written language across several different genres, and leverages pre-existing collections as well as documents we have culled from the Web.

Now we are in the process of annotating the corpus for lexical and morphological information (part of speech tags, named entities, lemmas, etc.). The work is nontrivial since it involves reconciling the various tagsets that have previously been implemented for French. It also involved experimenting with various taggers for French from a wide spectrum of approaches from rule-based to statistical to hybrid approaches. The same is also true for the lemmatizers we have used for the task of reducing tokens to types.

This paper, then sketches how the French frequency dictionary is being compiled and how it compares to previous frequency dictionaries of French. It will explain the structure of the corpus used to define the scope of the dictionary, the tools used to annotate and extract information from the corpus, and how we will generate the various frequency, paradigmatic, syntagmatic, and thematic word lists presented in the dictionary.

## **Jose Lopes Moreira and Tony Berber Sardinha**

Pontifical Catholic University of Sao Paulo

### **The Reading Class Builder: A tool for creating corpus-based teaching materials**

In this paper, we present the Reading Class Builder (RCB), a Windows-based application that allows users to generate reading materials for teaching English as a Foreign Language. The tool was designed to meet the needs of teachers who want to use corpus-based materials in their classes but who are not familiar with corpus processing software and/or do not have much time for lesson planning. The tool allows users to generate materials quickly with its Planning Wizard, which guides users through the process in a few steps. Firstly, the user selects the size of the material, in terms of the number of items. Secondly, the user chooses a text (either from its built-in text bank or from another source) to be the reading text for the teaching material. Thirdly, the tool analyzes the text in a range of ways and displays the frequency of words, the text's keywords, a list of cognates, parts of speech, bundles, and the text's lexical density. Finally, the user selects a set of words to be the focal points of the material and a corpus (again either from the tool's internal corpora, which includes de BNC, or a user-defined corpus), and then over 15 exercises are instantly prepared, including concordance-based data-driven activities, guessing, matching, fill in the blanks, language awareness, and critical reading items. These activities are created based on a template that can be modified by the user. Visual Basic 6 was the main programming language for developing the software. This presentation will show how the program works and will also present an evaluation of the tool's functions, in terms of precision and recall. A trial was conducted with Brazilian teachers who used the software in their classrooms; results will be reported.

## **Hui-Chuan Lu, Yun-Hui Chen and Chia-Chi Tien**

National Cheng Kung University

### **Corpus Creation: CATE, CPEC & CAEC**

Compared with other languages such as French and German which have nearly twenty corpora available, the development of Spanish corpus is limited, and most of them are of native speakers' natural language. This paper aims to build cross-linguistics corpora and to provide the knowledge of its application to linguistics analysis for future research of the same interest.

This paper examines: (1) CATE: Corpus de Aprendices Taiwanese de Español, (2)

CPEC: Corpus Paralelo de Español-Chino, (3) CAEC: Corpus de Aprendices Españoles de Chino. Their computational technology of corpus creation depends on the assistance of engineering programmers, as well as the application of WordSmith and Paracon. The characteristics of these corpora are: (1) CATE: began from year 2005 and consists of 1500 texts (300,000 words) from different levels of Taiwanese Spanish learners, with annotations of error and correction; (2) CPEC: from 2006 on, based on the translated parallel Spanish-Chinese texts (11,734 words in Spanish and 21,076 words in Chinese) with POS tagging; (3) CAEC: starting from 2007, in the first year, we plan to compile texts which feature Chinese writings by Hispanic learners of Chinese, and to annotate the learners' errors with possible factors of second language acquisition.

By integrating the research from these three corpora, we would be able to develop an interlanguage continuum of developmental stages through analyzing the written production of different grammatical points, and also explore the variables of interlanguage in second language acquisition in order to reach the final goal of effective learning.

**Peter McClanahan, Eric Ringger, Robbie Haertel, Kevin Seppi, George Busby, Deryle Lonsdale**

Brigham Young University

### **Accelerating Corpus Annotation through Active Learning**

In the construction of annotated corpora, we are constrained by fixed budgets for expert annotation. Although a fully annotated corpus is required, we can afford only to label a subset. The focus of the current work is part-of-speech and morphological tagging, although other annotation tasks can also benefit from the techniques discussed. In addition to a labeled corpus, we also aim to produce a statistical tagger that can accurately tag future texts. A Maximum Entropy Markov Model (MEMM) tagger is trained from a labeled subset of the target corpus and employed to tag automatically the remainder of the corpus.

This paper addresses the question of where in the corpus to focus manual tagging efforts in order to deliver both an annotation of high quality and an accurate model. We demonstrate that by applying active learning techniques, a state of the art tagging model can be trained on as little as one-half of the amount of data required by more traditional annotation schemes to achieve the same levels of accuracy. We focus our active learning experiments on two techniques, namely Query by Uncertainty (QBU) and Query by Committee (QBC) and report on experiments with several baselines and variations of both QBC and QBU. Query by Uncertainty directs an annotator's attention to those data which appear to have greatest sequence entropy according to a given model. By contrast, Query by Committee requests annotation on those data where an ensemble of models

achieve least agreement.

Experiments on English prose from the Wall Street Journal, English poetry from the British National Corpus, and the Syriac New Testament test these approaches. The results allow us to make recommendations for both languages and for the two types of English text. The results also raise questions that are leading to further inquiry.

**Tony McEnery**

Lancaster University

### **Corpus Linguistics and the Humanities**

Language pervades the humanities. Textual analyses by historians, the analysis of sacred texts and the interpretation of literary material occur across the humanities in forms that are more or less familiar to linguists. It comes as a surprise, therefore, to find that the dialogue between linguistics and other disciplines for which language is key is minimal and at times non-existent. In this talk I will reflect upon the potential contribution of linguistics in general, and corpus linguistics in particular, to the humanities. In doing so I will note the important role that the humanities had to play in the founding of corpus linguistics and consider the potential, and real, contributions of corpus linguistics to a range of humanities disciplines, including the classics, history, literature and religious studies. I will conclude by noting that if corpus linguistics has much to offer the humanities, it is also true to say that the humanities has much to offer corpus linguistics.

**Alfonso Medina**

Instituto de Ingeniería

### **Towards a Quantitative Characterization of Corpora at the Morphological Level: the use of Morphological Profiles to Measure Diachronic Change**

The set of most prominent affixes of a wide variety of languages seems to be more intimate to those languages than any set of basic lexical items including cognates such as body parts, heavenly phenomena, personal pronouns, very basic numerals, etc. As widely known, measuring similarity among those basic sets of lexical cognates permits, among other things, estimation of how far back in time two or more languages were in fact the same language; usually in terms of millennia. Measuring similarity among sets of promi-

ment affixes and sequences of them, rather than among lexical items, allows for comparison of diachronic stages within much shorter periods of time.

Furthermore, there exist very diverse unsupervised methods for segmenting graphical words; and some of them can be applied to compile sets of affixes and sequences of them, *i.e.* morphological profiles, that can be used to intimately characterize corpora of a wide variety of languages. Comparison of different profiles of one given language (from different diachronic states) can be used for obtaining a general measurement of variation at the morphological level.

In this presentation, quantitative data for three centuries of the Spanish language spoken in Mexico will be presented (XVI, XVIII and XX centuries) with the intent of corroborating (or not) intuitions put forward by philologists.

**Ji Meng**

Imperial College London

### **Statistical Modelling of Empirical Data in Corpus Stylistics**

The study of corpus stylistics as an emerging field of research within literary and comparative linguistics exemplifies the use of corpus data and methodologies in pursuing scholarly inquiries into the nature of the idiosyncratic use of language by individuals. It shares certain similarities with closely related yet more established areas, *i.e.* authorship attribution, literary or forensic stylometry, which see an increasing use of statistical techniques in the description and analysis of quantitative linguistic data. In the place of discussing issues such as corpus text compilation and extraction of corpus frequency information, the present paper endeavors to test the potential and validity of more sophisticated statistical techniques in the exploration of parallel corpus texts. The statistical methodology presented here, which has proved to be revealing in a corpus-based study of the stylistic differences between two modern Chinese versions of Cervantes' *Don Quijote*, may well have a wider application in corpus stylistics, for it has been quite successful in making causal claims on the dependence or independence of the language style of translated texts from that of the original, as well as the possible stylistic influence from previous translations on more recent ones in the case of retranslation. The statistical tests introduced in the current study as well as their subsequent visualization on graphs have been performed in the freely-accessible and versatile analytical environment, R, whose increasing application in corpus linguistics seems to have been rarely tested in corpus-based literary stylistics.

## Reference

- Baayen, H, (forthcoming) *Analyzing Linguistics Data: An Introduction to Statistics*, CUP;  
 URL: <http://www.mpi.nl/world/persons/private/baayen/publications/baayenCUPstats.pdf>
- Biber, D. et al, (1998) *Corpus Linguistics: Investigating Language Structure and Use*, CUP
- McEnery, T & Wilson, A (2001) *Corpus Linguistics*, Edinburgh University Press (2<sup>nd</sup> edition)
- Hoover, D (2001) "Statistical stylistics and author attribution: an empirical investigation", in *Literary and Linguistics Computing*, 16(4): 421-44
- Ji, M (forthcoming) "Quantifying style in two modern Chinese versions of Don Quijote", in *Meta*, 53 (4)

**Viola Miglio**

University of California, Santa Barbara

**Online Databases and Language Change: the Case of Spanish "dizque"**

This paper explores the semantics and usage of *dizque*, an adverb used as an evidential strategy in Latin American Spanish and explores its development from the XII to the XX centuries, focusing on Iberian and Mexican Spanish and concentrating on changes during the Colonial period. While this peculiar adverbial is current in Latin American usage, it seems to have been lost in Iberian Spanish (Magaña 2005). The research questions explored in this paper are: 1) when does *dizque* emerge as an evidential strategy and what is its evolution from finite verb + complementizer to adverbial element; 2) when does the adverbial change from a purely "hearsay" marker to a marker of disbelief distancing the speaker from the information s/he is communicating; 3) how useful are online databases such as Mark Davies's *Corpus del Español* (henceforth CDE) and the Real Academia's *Corpus diacrónico del español* (CORDE) to carry out this type of historical analysis. This is evaluated by comparing the data obtained from CDE and CORDE to printed collections such as *Documentos lingüísticos de la Nueva España* (Company 1994).

The paper shows that *dizque* emerges very early as an evidential strategy marking information source in the XIII century, and it declines steadily from the XVII century onwards BOTH in Spain and in Mexico, according to CDE and CORDE. I show that while it is true that the form has all but disappeared in Spain, the reason for the relatively small number of *dizque*'s in 20<sup>th</sup> century Lat. Am. Spanish samples from the online databases is that this adverbial has changed registers, becoming a predominantly spoken or colloquial strategy, and therefore not present in the CDE and CORDE documents. The use of *dizque* as marking disbelief is secondary compared to the evidential meaning and is not found until the XVII century (both in Iberian and Colonial Spanish). I also conclude that online databases such as CDE and CORDE are an invaluable tool to establish trends in language change, and to understand diatopic variation.

**Jan Minagawa**

Temple University Japan

**Developing writer stance in intermediate level Japanese university student writing in academic and disciplinary writing courses**

Individual voice (Atkinson, 2001; Ramanathan & Atkinson, 1999), viewed as an L1 English cultural practice, may present a barrier for high writing assessment for second language writers. Ethnographic research on classroom second language acquisition emphasizes the importance of social context on language use, and genre pedagogy approaches have reported success in development of sociocognition for graduate level L2 English academic writers in foreign language university settings. Some Japanese universities are promoting academic genre and disciplinary studies in English for incoming students who may study at English-speaking universities or for higher educational or professional purposes.

New Rhetoric approaches (Samraj, 2002), describing social action occurring on multiple contextual layers, as well as views of culture as small cultures and overlapping, dynamic cultures (Holliday, 1999; Atkinson, 2004), suggest problem areas for development of EFL writer social identity and academic writer's voice. Japanese students may exhibit problems with polysemantic use of modal auxiliaries and appropriate use of stance devices which may result from both intentional and unintentional resistance (Wertsch, 1998) and from instructional sources. Recent academic and disciplinary textbooks and courses include exercises on use of modals for inference and logical reasoning and also provide modeled genre contexts in academic writing assignments.

This study queries whether 'adorable' and innocent, or inelegant, voices in EFL student writing may result from problems with academic stance devices. An analysis of development of epistemic and other stance complement clause constructions commonly found in academic writing, as suggested in Biber, Johansson, Leech, Conrad & Finegan (1999) and MDM analysis of spoken and written university language (Biber, 2006), should provide insight into EFL university student use of stance and suggest better instructional approaches to develop an appropriate academic writer's voice.

References

- Atkinson, D. (2001). Reflections and refractions on the JSLW special issue on voice. *Journal of Second Language Writing, 10*, 201-124.
- Atkinson, D. (2004). 'Contrasting rhetorics/contrasting cultures: why contrastive rhetoric needs a better conceptualization of culture'. *Journal of English for Academic Purposes, 3*(4), 277-290.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins Publishing Company.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Essex: Pearson Education Limited.
- Holliday, A. (1999). Small cultures. *Applied Linguistics, 20*(2), 237-264.

- Samraj, B. (2002). Texts and contextual layers. In Johns, A. M. (Ed.). *Genre in the classroom: Multiple perspectives*. Mahwah NJ: Lawrence Erlbaum Associates, Publishers.
- Ramanathan, V., & Atkinson, D. (1999). Individualism, academic writing and ESL writers. *Journal of Second Language Writing*, 8(1), 45-75.
- Riazi, A. (1997). Acquiring disciplinary literacy: A social-cognitive analysis of text production and learning among Iranian graduate students of education. *Journal of Second Language Writing*, 6(2), 105-137.
- Wertsch, J. V. (1998). *Mind as action*. Oxford: Oxford University Press.

## Ilka Mindt

Universität Würzburg

### Colloquialization: An Alteration in Written English

*Colloquialization* is an alteration which can be observed in 20<sup>th</sup>-century English. Mair describes colloquialization as a “significant stylistic shift in twentieth-century English” (2006: 187). This shift is observable in the written medium. Mair argues that the shift works in two ways:

- (1) “away from a written norm which is elaborated to maximal distance from speech” towards “a written norm that is closer to spoken usage” (2006:187),
- (2) “away from a written norm which cultivates formality” towards “a norm which is tolerant of informality” (2006:187).

Based on this account of colloquialization, the paper will investigate colloquialization more closely on a theoretical basis and demonstrate that it is indeed a current alteration in the written English language of the 20<sup>th</sup> century.

The theoretical part of this paper focuses on an explication of the two ways the stylistic shift works. Special attention is given to the varieties ‘medium’ (spoken vs. written) and ‘attitude’ (formal vs. informal) as described by Quirk et al (1985) and by Biber (1988, 1995).

A linguistic analysis of three linguistic markers will demonstrate how colloquialization can be detected. The data are taken from the British National Corpus and the Time magazine. The three linguistic markers under investigation are

- (1) the passive,
- (2) whether the conjunction *that* in object clauses is realized by *that* or by zero, and
- (3) contractions.

The paper finishes with some further characteristics of colloquialization.

## References

- Biber, Douglas (1988) *Variation across speech and writing*. Cambridge: CUP
- Biber, Douglas (1995) *Dimensions of register variation. A cross-linguistic comparison*. Cambridge: CUP
- Mair, Christian (2006) *Twentieth-Century English. History, Variation, and Standardization*. Cambridge: Cambridge University Press.
- Quirk, Randolph/Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (1985) *A Comprehensive Grammar of the English Language*. London: Longman

**Cristina Mota**

Instituto Superior Tecnico & New York University

**Journalistic Corpus Similarity over Time**

Kilgarriff (2001) proposed a method to measure corpus similarity based on the distance between word frequency lists. His aim was to compare language varieties. In our study, we used his method to assess corpus similarity over a short period of time both within topic and cross-topic. The corpus samples were drawn from CETEMPúblico (Santos & Rocha, 2001), a 180 million word Portuguese journalistic corpus. The corpus spans 8 years (from 1991 to 1998), and is comprised of article extracts marked with the year, semester and newspaper section of publication. We conducted forward, backward and centered in time experiments (i.e., taking as reference the oldest and newest texts, and one text in the middle of the time interval, respectively, and comparing it with all remaining texts published in different semesters). We will show that (i) the similarity between two texts within the same topic generally decreases as the time gap between them increases, being more significant for some topics; (ii) in some cases, the texts over time become as different as two texts from different topics.

Since the ultimate goal of our work is to understand how the changes in corpus similarity affect the performance of a named entity tagger, we also decomposed the word frequency lists into two different lists (one containing capitalized words and the other containing lower case words), and measured the similarity based of those lists instead. The former similarity aims at comparing the corpora from the viewpoint of the named entities content, whereas the latter one approximately compares the surrounding contexts of the named entities. The results show that the similarity values based on these lists also generally decrease over time, even though the way it decreases and compares with the all words based similarity depends on the topic.

## References

- Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 1 (6), 1-37.
- Santos, D., & Rocha, P. (2001). Evaluating CETEMPúblico, a Free Resource for Portuguese. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, (pp. 00-00). Toulouse.

**Hilary Nesi**

Coventry University

**The design of a web-based interface for the BAWE corpus**

The British Academic Written English (BAWE) corpus, <http://www.coventry.ac.uk/bawe>, contains about 3000 university student assignments of a good standard, at all levels from first year undergraduate to masters degree, and in many disciplines. Whereas transcripts from the comparable British Academic Spoken English (BASE) corpus, <http://www.coventry.ac.uk/base>, are downloadable from the website, BAWE corpus holdings cannot be released in their entirety to non-registered researchers, because of the fear that the resource will be treated as an 'essay bank' and will facilitate plagiarism. A web-based interface solves this problem by allowing free access to limited quantities of text relating to specific corpus queries. This paper introduces and demonstrates the beta version of our BAWE interface <http://www.coventry.ac.uk/bawe-search>, and discusses some key issues surrounding its creation and development to meet the needs of both expert and non-expert users. The interface provides collocation, pattern and frequency information, with optional filters for contextual properties (discipline, text type, year of study) and textual properties (section, position in text, position in paragraph). Free access should encourage use of the corpus data by users from within and outside the corpus linguistics research community, and we would like to attract to the site applied linguists, writing tutors and students seeking to develop their own writing skills. This leads to tensions between the conflicting requirements of usability and meaningfulness, however. We have experimented with Sketch Engine, a corpus query system developed by Adam Kilgarriff and Pavel Rychly. This is powerful but relatively opaque, and although our own interface, which attempts to simplify the search process (for example by creating drop-down menus of search categories) may support non-expert users, we wonder whether it might also encourage more superficial, perhaps pointless, queries, and lead users away from open-ended investigations. The final interface design will be arrived at through consultation and trialling with all kinds of potential end-users, so we will welcome feedback at the conference.

**Francisco Ocampo**  
University of Minnesota

**A diachronic process that gives birth to a Spanish discourse particle:  
The case of “claro”**

In contemporary spoken Rioplatense Spanish, the lexical item *claro* may have an adjectival use as in (1):

- (1) (The numbers and letter indicate the location of the example in a Rioplatense Spanish corpus)  
C: este:, yo d descubro cuando - grabo algo se oyen los ruidos de - así a aleja-  
dos de esos que uno no no queridos, con una nitidez impresionante.  
F: e he he  
C: sí:, muy **claros**, fieles y y y en cambio lo que usted quiere grabar no sale.  
6a12

or it may be utilized as a discourse particle, as in (2):

- (2) M: Le paga p siempre para ayudarlas a ellas que que tanto le habían hecho por él cuando vino de Italia.  
L: **claro**, porque ellas a él lo han criado, se puede decir. 1b20

In (2) *claro* conveys a token agreement, and is used as a tool by speaker L to secure the turn, as revealed by the overlapping between M and L's turns.

Previous work (Ocampo 2006), based on a corpus of 20 hours of transcribed Rioplatense Spanish casual conversations, has attempted to establish the meaning, and further evolution of this form in contemporary Spoken Spanish. It hypothesizes that the meaning of *claro* has changed from 'luminous', hence 'distinctly perceived' to 'evident'<sup>3</sup>, and that the discursive use of *claro* stems from this later meaning. This meaning has been discursively exploited to communicate messages of agreement/understanding, and to acknowledge receipt. In later stages, these two messages have been employed to allow *claro* to function as a self-selection particle, as in (2). Ocampo 2006's work focuses on *claro* in present spoken discourse, and tries to characterize its development as an instance of discoursivization, a movement distinct from classical grammaticalization.

The present work, on the other hand, traces the process of evolution that gave birth to the discourse particle use, first attested at the beginning of the 19<sup>th</sup> Century in a Spanish adaptation of Molière's play *Le médecin malgré soi* (1648). The data utilized for the analy-

<sup>3</sup> This term is not to be taken as designating the semantic notion of evidentiality, which refers to the source of information. Although both concepts are epistemically related, 'evident' here simply indicates that what is presented is easily perceived by the senses or by the mind.

sis is taken from Davies (2002). Since the discursive use of the adjective begins with the meaning of 'evident', in my analysis I follow the development of this meaning. I start from the meanings present in Latin, but I focus on Spanish data from 13<sup>th</sup> to 18<sup>th</sup> Centuries. The data show a modest increase of the meaning 'evident' in *claro* adjective, from 2% in 14<sup>th</sup> Century to 15.78% in 18<sup>th</sup> Century. However, the most important rise of this meaning does not appear in adjectives but in constructions such as *claro está*, *claro es que*, *es claro*. On the one hand, *claro* carries the meaning 'evident' in the majority of these constructions (90%). On the other, these constructions rise dramatically over the centuries: In 13<sup>th</sup> Century, *claro* adjective constitutes 93.18% of the collected data, constructions with *claro* only 2.27%, whereas in the 18<sup>th</sup> Century there are 39.56% *claro* adjectives, versus 47.67% of constructions with *claro*. Therefore, I hypothesize that these constructions are the main factor responsible for the increase use of *claro* with the meaning of 'evident', which in turn gives rise to its use as a discourse particle.

### References

- Abraham, Werner. 1991. The grammaticization of the German modal particles. *Approaches to Grammaticalization*, ed. by Elizabeth Closs Traugott and Bernd Heine. Amsterdam/Philadelphia: John Benjamins. Vol. 2, 331-380.
- Academia Quinque Germanicarum. 1907. *Thesaurus Linguae Latinae*. Lipsia: Teubner.
- Aijmer, Karin. 1997. I think: An English modal particle. *Modality in Germanic languages. Historical and comparative perspective*, ed. by Toril Swan and Olaf Jansen Westvik. Berlin/ New York: Mouton de Gruyter. 1-47.
- Auwera, Johan van der. 2002. More thoughts on degrammaticalization. *New Reflections on Grammaticalization*, ed. by Ilse Wischer and Gabrielle Diewald. Amsterdam/Philadelphia: John Benjamins. 19-29.
- Blánquez Fraile, Agustín. 1961. *Diccionario Latino-Español*. Barcelona: Editorial Ramón Sopena.
- Brinton, Laurel. 1996. *Pragmatic markers in English: Grammaticalization and Discourse Function*. Berlin: Mouton de Gruyter.
- Company Company, Concepción. (Forthcoming). Subjectivization of verbs into discourse markers: Semantic-pragmatic change only? *Modalization and Pragmaticalization*, ed. by Nicole Delbecque and B. Cornillie. Amsterdam/Philadelphia: John Benjamins.
- Contini-Morava, Ellen. 1995. Introduction: On linguistic sign theory. *Meaning as Explanation. Advances in Linguistic Sign Theory*, ed. by Ellen Contini-Morava and Barbara Sussman Goldberg, Berlin/New York: Mouton de Gruyter. 1-39.
- Davies, Mark. 2002. *Corpus del Español*. <http://www.corpusdelespañol.org>.
- Erman, Britt and Ulla-Britt Kotsinas. 1993. Pragmaticalization: The case of *ba'* and *you know*. *Studier I Modern Språkvetenskap*. Acta Universitatis Stokholmiensis. New Series 10. 76-93.
- Facciolatus, Jacobi, and Ægidii Forcellinus. 1828. *The Universal Latin Lexicon*, vol 1, Edition by James Bailey. London: Baldwin and Cradock.
- Fischer, Olga, and Anette Rosenbach. 2000. Introduction. *Pathways of change. Grammaticalization in English*, ed. by Olga Fischer, Anette Rosenbach, and Dieter Stein. Amsterdam/Philadelphia: John Benjamins. 1-37.
- Giacalone Ramat, Anna, and Paul Hopper, eds. 1998 *The Limits of Grammaticalization*. Amsterdam/Philadelphia: John Benjamins.
- Givón, Talmy. 1979. *On understanding grammar*. New York/San Francisco/London: Academic Press.
- Glare, P., ed. 1984. *Oxford Latin Dictionary*. Oxford: Clarendon Press.
- Günther, Susanne, and Katrin Mutz. 2004. Grammaticalization vs. pragmaticalization? The development of pragmatic markers in German and Italian. *What makes grammaticalization? A look from its fringes and*

- its components*, ed. by Walter Bisang, Nikolaus Himmelmann, and Bjern Wiemer. Berlin/New York: Mouton de Gruyter. 77-107.
- Haspelmath, Martin. 2004. On directionality in language change with particular reference to grammaticalization. *Up and down the cline - The nature of grammaticalization*, ed. by Olga Fischer, Muriel Norde, and Harry Perridon. Amsterdam/Philadelphia: John Benjamins. 14-44.
- Heine, Bernd, Ulrike Claudi, and Friederike Hünemeyer. 1991. *Grammaticalization*. Chicago and London: The University of Chicago Press.
- Herring, Susan. 1991. The grammaticalization of rhetorical questions in Tamil. *Approaches to Grammaticalization*, ed. by Elizabeth Closs Traugott and Bernd Heine. Amsterdam/ Philadelphia: John Benjamins. Vol. 1, 253-284.
- Hopper, Paul, and Elizabeth Closs Traugott. 1993. *Grammaticalization*. Cambridge: Cambridge University Press.
- Kuriłowicz, Jerzy. 1975. The evolution of grammatical categories. *Esquisses Linguistiques II*, Munich: Fink. 38-54.
- Lehmann, Christian. 1985. Grammaticalization: Synchronic variation and diachronic change. *Lingua e stile* 20:303-318.
- Lenker, Ursula. 2000. *Soþlice* and *witodlice* discourse markers in Old English. *Pathways of Change. Grammaticalization in English*, ed. by Olga Fischer, Anette Rosenbach, and Dieter Stein. Amsterdam/Philadelphia: John Benjamins. 229-249.
- Martín Zorroaquin, María Antonia and José Portolés Lázaro. 1999. Los marcadores del discurso. *Gramática Descriptiva de la Lengua Española*, ed. by Ignacio Bosque and Violeta Demonte. Madrid: Espasa. 4051-4213.
- Menéndez Pidal, Ramón. (1950): *Orígenes del español. Estado lingüístico de la Península Ibérica hasta el siglo XI*. Madrid, Espasa-Calpe.
- Mithun, Marianne. 1989. The grammaticization of coordination. *Clause Combining in Grammar and Discourse*, ed. by John Haiman and Sandra Thompson. Amsterdam/Philadelphia: John Benjamins. 331-359.
- Moliner, María. 1994. *Diccionario de uso del español*. Madrid, Gredos.
- Ocampo, Francisco. 2006. Movement towards discourse is not grammaticalization: The evolution of *claro* from adjective to discourse particle in spoken Spanish. *Selected Proceedings from the 9<sup>th</sup> Hispanic Linguistics Symposium*, ed. by Nuria Sagarra and Almeida Jacqueline Toribio, Somerville, MA: Cascadia Proceedings Project, 308-319.
- Onodera, Noriko. 1995. Diachronic analysis of Japanese discourse markers. *Historical Pragmatics*, ed. by Andrea Jucker. Amsterdam: John Benjamins. 393-437.
- Pinto de Lima, José. 2002. Grammaticalization, subjectification and the origin of phatic markers. *New Reflections on Grammaticalization*, ed. by Ilse Wischer and Gabrielle Diewald. Amsterdam/Philadelphia: John Benjamins. 363-378.
- Real Academia Española. 1956. *Diccionario de la Lengua Española*. Madrid: Espasa-Calpe.
- Saussure, Ferdinand de. (1973). *Curso de Lingüística General*, Translation, prologue and notes from Amado Alonso. Buenos Aires: Editorial Losada.
- Schwenter, Scott. 1996. Some reflections on *o sea*: A discourse marker in Spanish. *Journal of Pragmatics*. 25 (1996) 855-874.
- Silva-Corvalán, Carmen. 1999. *Ahora*: From Temporal to Discourse Deixis. *Essays in Hispanic Linguistics Dedicated to Paul M. Lloyd*, ed. by Robert Blake, Diana Ranson, and Roger Wright. Newark, Delaware: Juan de la Cuesta. 67-81.
- Trask, R. 1993. *A Dictionary of Grammatical Terms in Linguistics*. London: Routledge.
- Traugott, Elizabeth Closs. 1986. From polysemy to internal semantic reconstruction. *Berkeley Linguistics Society*. 12:539-550.
- Traugott, Elizabeth Closs. 1995. The role of the development of discourse markers in a theory of grammaticalization. Paper presented at the ICHL XII, Manchester. Version of 11/97. In <http://www.stanford.edu/~traugott/ect-paperonline.html>.

- Traugott, Elizabeth Closs, and Richard Dasher. 2002. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.
- Travis, Catherine. 2005. *Discourse Markers in Colombian Spanish: A study in polisemy*. Berlin/New York: Mouton de Gruiter.
- Wischer, Ilse. 2000. Grammaticalization vs. Lexicalization. 'Methinks' there is some confusion. *Pathways of change. Grammaticalization in English*, ed. by Olga Fischer, Anette Rosenbach, and Dieter Stein. Amsterdam/Philadelphia: John Benjamins. 355-370.

## **Matthew Brook O'Donnell, Catherine Smith, Robert Sanderson, Clare Llewellyn and John Harrison**

University of Liverpool

### **Using an XML database for large corpora: Introducing Cheshire3**

The advantages of using a relational database to store and analyse language corpora have been demonstrated by a number of projects and tools, such as the VIEW interface to the BNC. The speed, extensibility and wide variety of searches possible using this paradigm and SQL queries are unquestionable. However, there is a fundamental mismatch between the irregular hierarchical structure of text and the two dimensional tabular representations of the relational model. This paper discusses the benefits of using an XML database, specifically Cheshire3, to store and analyze large corpora.

Cheshire3 is record-based (in corpus terms a record is equal to an entire text) and allows an unlimited number of indexes to be constructed which extract data from specific locations within these records. These indexes then enable fast discovery and analysis of the records and their linguistic features in the contexts of both the corpus as a whole and the individual matching records. For instance, a frequency list indicates not only how many times a word occurs in the corpus but the number of texts (records) in which it is found and the number of occurrences in each record. Comparisons between textual structures (paragraphs, sentences) or physical properties of a text (pages, lines) may also be accomplished using the same record based analysis. For example, one might wish to compare the frequency of a word within the first paragraph of a text to the overall frequency in that text, as well as the number of occurrences in the corpus.

These advantages are illustrated through a web-based corpus query tool that provides the standard frequency list, KWIC listing and collocation analysis alongside some new text level features.

**Matthew Brook O'Donnell, Mike Scott and Michaela Mahlberg**

University of Liverpool

### **Exploring Text-initial Concgrams in a Newspaper Corpus**

Recent analysis of a newspaper corpus has confirmed the assertion that certain lexical items have a tendency to occur at particular points in a text, i.e. the beginning or end of texts, paragraphs or sentences (Hoey 2005; Scott & Tribble 2006). Lexical items cannot then be assumed to be evenly distributed across texts and corpus sampling and statistical procedures should reflect this fact (Sinclair 1991; Stubbs 1996). These positional associations are found for single words, phrases and clusters (N-Grams/lexical bundles or clusters). For example, *yesterday* and *announced* are text-initial words in newspaper stories as are the clusters *announced yesterday*, *yesterday announced* and *it was announced yesterday*. This paper explores the question of whether and to what extent such text-initial associations extend to non-contiguous patterns or units such as those identified by the recently proposed Concgram procedure (Warren & Greaves 2006). Using a corpus of 113000 texts (52 million words) from the *Guardian* paper that has been divided so that all the first sentences of texts are grouped together, as are the first sentences of paragraphs and all non text or paragraph-initial sentences, we calculate the Concgrams for each of these subcorpora. Implications are explored for both corpus linguistic theory, by relating Concgrams to clusters and other functional units, and for theories of the structure of news text.

#### References

- Cheng, W., C. Greaves and M. Warren (2006), 'From n-gram to skipgram to concgram, *International Journal of Corpus Linguistics*, 11 (4): 411-433.
- Hoey, M. (2005), *Lexical Priming: A new theory of words and language* (London: Routledge)
- Scott, M. and C. Tribble (2006), *Textual Patterns: Key words and corpus analysis in language education* (Amsterdam: John Benjamins)
- Sinclair, J. (1991), *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Stubbs, M. (1996), *Text and corpus analysis*. Oxford: Blackwell.

**Donald Parry**

Brigham Young University

### **Verbal Imperative Variations in Qumran Legal Texts and Other Registers**

The principal goal of this paper is to quantify linguistic variations of the positive (second-person volitionals) and negative imperatives (two different negative particles plus imper-

fect verbs) in a variety of texts from Qumran (the Dead Sea Scrolls) and the Hebrew Bible. By quantifying the imperatival variations, I aim to set forth estimations of the variations by quantity or numbers through figures, and to a lesser extent I will attribute qualities or characterizations to the variations.

I will focus on the imperatives of Qumran legal texts, more specifically, the Temple Scroll (11Q19), Rule of the Community (1QS), and Damascus Document (CD). I have placed these three texts into a module or *register*, a collection of texts belonging to the same genre or text type. For purposes of comparison, I have also prepared four other registers from the Qumran caves and three from the Hebrew Bible, each of which forms a genre or text type. My rationale for selecting the texts that belong to these registers is to provide several points of comparison, because the discourse purposes for each register is different from the others.

### **Iria Pastor-Gómez**

University of Santiago de Compostela

#### **Nominal Modifiers in Present Day English Noun Phrase Structure**

Building on earlier approaches to Noun Phrase (NP) structure such as Varantola (1984) or Raumolin-Brunberg (1991), a pioneering study by Biber & Clark (2002) has examined the historical shift that has taken place in English whereby adjectival (e.g. “an *expensive* car”) and clausal (e.g. “the desk *that is nearby the window*”) modifiers have been gradually encroached upon by nominal structures (e.g. “*subway* travellers”) and prepositional phrases (e.g. “production of *Japanese abalone*”).

In my proposed presentation I aim to supplement Biber & Clark’s preliminary study with a comprehensive analysis of the development of English nominal premodifiers, as examined in two matching corpora of written British and American English, namely FLOB (sampling year: 1991) and FROWN (sampling year:1992).

To date, the results obtained from the corpus analysis reveal that premodifying nouns display a wide variety of different structural patterns, ranging from NP heads premodified by nominal compounds (e.g. “*copy-right* system”) to structures consisting of an Adjective + Noun coordinate (e.g. “*secular or entertainment* style”). Further, message-oriented genres, such as Science or Press: reportage, contain high percentages of nominal premodifiers. By contrast, premodifying nouns are much less frequent in subjective genres such as Fiction or Essays. Likewise, fictional genres opt for simpler, more institutionalized patterns (e.g. “car wash”) than scientific texts (e.g. “magnetic tape electronic data processing system”).

My presentation will put forward an explanation for the above mentioned tenden-

cies in terms of the role of nominal modifiers as devices making for compactness, in an age in which the pressure to communicate information economically is constantly on the increase.

### References

- Biber, Douglas, and Victoria Clark. 2002. "Historical shifts in modification patterns with complex noun phrase structures", in Teresa Fanego et al. (eds.) *English Historical Syntax and Morphology*. Amsterdam: Benjamins. 43-66.
- Raumoling-Brunberg, Helena. 1991. *The Noun Phrase in early sixteenth-century English: a study based on Sir Thomas More's writings*. Helsinki: Societ  Neophilologique.
- Varantola, Krista. 1984. *On noun phrase structures in engineering English*. Turku: Turun Yliopisto.

## Pamela Pearson

Georgia State University

### **A corpus-based study of mandative subjunctive triggers in published research articles**

This corpus-based investigation of the mandative subjunctive (MS) (Quirk & Greenbaum, 1973) explored the structure as it co-occurred in published research articles with 10 lexical items previously identified as MS triggers (Overgaard, 1995; Peters, 1998; Crawford & Albakry, 2004): ask, demand, direct, insist, order, propose, recommend, request, require and suggest. Through a variety of quantitative (frequency, form and pattern) and qualitative (function) analyses, the present study examined the use of these lexical items in a corpus of published research articles ( $\approx 1.75$  million words) by authors in four academic disciplines (Applied Linguistics, Biology, Business and Electrical & Computer Engineering) with the purpose of identifying the discourse factors that contribute to the triggering of the MS structure in academic writing as its main objective.

Results of the quantitative analyses revealed that when functioning as verbs, the majority of these lexical items infrequently co-occurred with that-clauses in the corpus; and, of the 1337 instances that did co-occur with a that-clause, only 50 (3.74%) contained a subjunctive verb form. In the qualitative analysis, 24 of the MS structures (48%) were found to be used to call for a particular action (Peters, 2004) or propose a potential course of action (Biber et al., 1999); one significant exception was require, which co-occurred with an MS structure 17 times (34% of the total instances) to report a necessary condition.

The findings of this study indicate that use of these lexical items does not alone trigger the mandative subjunctive in that-complement clauses in academic writing; rather, the factors involved in triggering the structure were determined to be multiple and complex. Thus, although the structure does indeed co-occur with certain verbs, this study has found the rules governing the triggering of the MS structure to be unpredictable.

## **Phuong Dzung Pho**

Monash University

### **Linguistic realizations of rhetorical structure:**

#### **A corpus-based study of research articles in applied linguistics and educational technology**

Previous studies of rhetorical structure of research articles have tended to focus on individual sections of the article especially the introduction section with more attention being given to 'hard' sciences (e.g. biochemistry) than to 'soft' sciences (e.g. education). The literature on the linguistic features of research articles is also not complete in that most studies observed the distribution patterns of only one or two linguistic features at the section level rather than the move level. This study thus aims at examining not only the rhetorical structure of the article as a whole from the abstract through to the conclusion section but also the linguistic realizations of moves. A corpus of 40 research articles in applied linguistics and educational technology was xml tagged for moves and parsed for a range of linguistic features to investigate what features are prototypical of each move. The analysis showed that a combination of features such as grammatical subjects, verb tenses, voice, stance words and self-reference words can help distinguish moves. Variations across the two disciplines were also reported. The findings of the study have some pedagogical implications for academic writing courses for graduate students in general and for students from non-English backgrounds in particular.

## **Kornwipa Poonpon**

Northern Arizona University

### **The Use of Linking adverbials in EFL Learners' Time-Constrained Spoken Monologue**

Linking adverbials can help language learners improve their quality of speaking in terms of clarity and coherence. Previous research has demonstrated that the more the speaker use linking adverbials in utterances, the more his text avoids ambiguity of the listener's interpretation (Ball, 1992; Ellis et al, 2005; Thompson, 1994). While a number of studies examine the use of linking adverbials in different types of spoken discourse, studies in time-constrained spoken monologue are rare. This presentation will discuss the results of a study that investigated the distributions of different semantic functions and positions of linking adverbials used by Thai EFL learners. A learner corpus (21,897 words) was collected from 44 university students who volunteered to take a TOEFL-like speaking test, including three dependent and three integrated speaking tasks. Each speaking task

was manually coded for frequency and types of linking adverbials, based on the list from *Longman Grammar of Spoken and Written English* (LGSWE) (Biber et al., 1999). Then, comparisons were made by task type. Since there was little research on linking adverbials in spoken monologue, the results of the study were mainly discussed on a basis of the explanation of linking adverbials in the LGSWE to draw out unique patterns of linking adverbials used in spoken monologue. While the results confirmed the use of linking adverbials in other spoken (and written) discourses, some evidence revealed distinctive patterns which reflected on characteristics of time-constrained spoken monologue. This information is useful for language teachers in teaching English for communicative purposes, especially in such time-constrained situations as oral testing.

## **Yufang Qian**

Lancaster University

### **Discursive construction of terrorism in *People's Daily* and *The Sun* before and after 9.11**

The focus of this study is on how terrorism is linguistically represented in Chinese and English popular newspaper before and after “9.11” and how this representation contributes to the construction of discourses of terrorism in different social systems and doctrines. However, it is important to point out the dynamic and fluid nature of these representations and the identities that they help construct. As a historically situated practice, the construction of social meaning through discourse is a process, not solely a product (Hodge and Kress 1993, Lemke 1995). On the other hand, the social meanings constructed are also seen as the product of an interaction between the texts and how they are interpreted or ‘read’ from different positions. The social meanings analyzed are not all explicitly reflected in the linguistic choices that appear in the texts but also in the implicit connections the texts and the readers make from the linguistic evidence available to them. Corpus methodologies have a huge potential for use in discourse studies. Examples of authentic data are gathered, in order to support the researcher’s argument or, perhaps even more importantly, as counter-evidence to make them think again. Automatic text analysis tools can throw into relief the non-obvious in a single text, shedding light on what may be hidden thoughts, hidden perhaps even from authors themselves (Partington 2003:7). Hunston (2002) observes that corpora are a useful tool for the critical linguist, because they identify repetitions which can be used to identify implicit meaning. Because data in corpora are de-contextualised, the researcher is encouraged to spell out the steps that lie between what is observed and the interpretation placed on those observations. The socio-political agenda of this research is to raise awareness of how discourse on ter-

rorism, is being manipulated to the extent of long-term being biased and uneven so that such a phenomenon can be formed in both newspapers. This study is based on *People's Daily Terrorism Corpus* and *The Sun Terrorism Corpus*. Corpus keywords, concordance, collocation analysis and critical discourse analysis are integrated.

## **Randi Reppen**

Northern Arizona University

### **'Students must': A corpus based look at directives university language**

In the past, studies that have investigated directives have tended to focus on conversation (e.g., Ervyn-Tripp 1976). These studies have contributed to our understanding of how directives are used to accomplish different purposes. More recent studies have begun to look at directives in academic writing (e.g., Hyland 2002). This presentation will expand the work done on spoken and written directives, by focusing on directives in the context of American university classes. University professors regularly use spoken and written directives to guide students through both in-class and out-of-class tasks. Professors use directives during the beginning of classes to provide an overview of the class structure, or of general issues related to the course (classroom management). Course syllabi and assignment descriptions are written modes that are often filled with directives which explain assignments, announce readings, or provide information about tests or exams (course management).

Through a corpus-based approach, this study provides an empirical investigation of directives found in a corpus of university language. The corpus, used in this study of directives, is composed of over 39,000 words of spoken classroom management, and over 50,000 words of written course management. It is a sub corpus of the almost 3 million word TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) corpus (Biber et. al. 2004). Some of the questions that can be addressed through this sub corpus include: How spoken and written directives differ? Are there linguistic forms of directives that are more frequent in spoken or written language?

In addition to being among the first empirical studies of spoken and written directives in university language, it will contribute to a more fine grained description of the types of directives that exist in American university language. This work will help researchers and teachers to better understand the types of directives used in American university classes.

**Sally Rice and John Newman**

University of Alberta

**Beyond the lemma: Inflection-specific constructions in English**

Recent research has called into question the concept of the lemma or phrase structure category as a relevant unit for morphosyntactic analysis. Attention has begun to shift in some quarters towards constructions based around actual inflected forms of a word. Research in this vein, which is corpus-based and has tended to take a cognitive/functional perspective, includes findings reported by Bybee & Hopper (2001), Thompson & Hopper (2001), Tao (2003), Knowles & Zuraidah (2004), Newman & Rice (2004, 2006), Rice & Newman (2005, 2006). Most of these studies are centered largely around the collocational, constructional, and grammaticalization behavior of inflected forms of individual verbs. We have posited the idea that inflectional categories and syntactic constructions are “islands” which strand particular lexical items. Under this inflectional island hypothesis, lexico-semantic properties tend to inhere in specific morphosyntactic inflections of a lexical item (especially in register-specific contexts). These properties may not extend across all the inflections to characterize the lemma as a whole and so we have proposed the idea of a WIC or word-in-context as the most ecologically veridical starting point for lexico-syntactic research.

In this paper, we take on another commonplace inflectional category in English— that of pronominal inflection— and briefly describe some of the distributional skews and idiosyncracies affecting case, possession, and number marking. Relying mainly on the British National Corpus, we use individual pronominal forms as key words and investigate some of the different collocational preferences of each. We find that little analytic advantage attaches to positing lemma-level representations in general. This is especially so in the case of pronouns where the lemma status is by no means clear-cut. This particular case study joins a set of others which advocates not only language- and construction-specific analyses, but also inflection-specific ones.

References

- Bybee, J. & Hopper, P. [eds.] (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: Benjamins.
- Knowles, G. & Zuraidah, M.D. (2004). The notion of a “lemma.” *International Journal of Corpus Linguistics* 9:69-81.
- Newman, J. & Rice, S. (2006a). Transitivity schemas of English EAT and DRINK in the BNC. In Gries, S. & Stefanowitsch, A. [eds.], *Corpora in Cognitive Linguistics*, Vol. 2, *The Syntax-Lexis Interface*. Amsterdam: Benjamins.
- . (2006b). English adjectival inflectin: A radical Radical Construction Grammar approach. Paper presentation at the 8th CSDL (UC, San Diego), 3-5 Nov 2006.
- . (2004). Patterns of usage for English SIT, STAND, and LIE: A cognitively-inspired exploration in corpus linguistics. *Cognitive Linguistics* 15: 351-396.

- Rice, S. & Newman, J. (2005). Inflectional islands. Paper presentation at the 9th ICLC (Yonsei University; Korea), 17-22 July 2005.
- Tao, H. (2003). A usage-based approach to argument structure: 'Remember' and 'forget' in spoken English. *International Journal of Corpus Linguistics* 8:1, 75-95.
- Thompson, S. & Hopper, P. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In Bybee, J. & Hopper, P. [eds.], *Frequency and the Emergence of Linguistic Structure*, 27-60. Amsterdam: Benjamins.
- Stefanowitsch, A. & Gries, S. Th. (2003). Collocations: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8:2, 209-243.
- Thompson, S. & Hopper, P. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In Bybee, J. & Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure*, 27-60. Amsterdam/Philadelphia: John Benjamins.

## Joseph Richardson

LDS Foundation

### A Corpus-Based Model for the Work of Editors

Corpus linguistics has frequently been promoted as a tool for translators and language learners. A similar but relatively unexplored application of corpus linguistics is in editorial tasks. Editors work with language in the confluence of writer, reader, and corporate interests. They mediate in the relation between institution, writer, text, and reader to enable the accurate and efficient production and transmission of knowledge. Editorial work includes analytic, interpretive, creative, and synthetic functions. These functions include pattern identification and analysis tasks and illuminate possible applications of corpus tools. Editorial work is also grounded in a consideration of audience and purpose. Consequently, the work of editors provides an opportunity to study the "principles underlying the selection, storage, retrieval and reproduction of knowledge" (Stubbs 1996: 237), particularly in relation to corpora, and to develop linguistic theory and test the application of corpus tools. The work of editors illuminates the relation between "language, knowledge and social institutions" (Stubbs 1996: 237) and helps address the question of how language relates to the world and action in it.

This paper reviews a survey of editors in a large multinational corporation, who help produce materials in a variety of media for audiences in over 100 languages. It includes an analysis of the ways in which these editors use corpora to resolve conflicts over style and usage, refine the content of style guides, and facilitate the efficient use of language. It also suggests how editors can use corpora to determine why some texts are more effective than others. This study suggests limitations of corpora for editorial practice and outlines ways that corpus tagging can be tailored to facilitate editorial practice.

## **Helena Riha and Kirk Baker**

Ohio State University

### **Tracking Sociohistorical Trends in the Use of Roman Letters in Chinese Newswires**

Chinese orthography is showing the effects of contact with English through the use of roman letters in written discourse, especially in news writing. While Chinese writing delimits syllables and morphemes, English writing delimits phonemes and words. Chinese script mixing is interesting linguistically because it requires overcoming a mismatch between a syllabic writing system and a phonemic one. We examine how this difference is dealt with to discover types of linguistic units written with roman letters in Chinese. We analyzed two sections of the Chinese Gigaword Corpus\* representing Chinese societies differing in their contact with English: Taiwan (Central News Agency/CNA) and mainland China (Xinhua/XIN). Roman letter strings occur 4 times more often in CNA, but their use has grown steadily in both societies from 1991 to 2005, although more quickly in Taiwan (slope test;  $p < 0.0001$ ). Overall, letters are common enough in both societies that they may indicate a move toward a mixed writing system (CNA: 113.5 letter strings/100k Chinese characters; XIN: 26.9).

Categorizing letter strings by script type, we find that All\_Caps are most frequent and are usually abbreviations (WTO) (proportion of All\_Caps to Mixed\_Case: XIN (0.94), CNA (0.77)). Examples also appear in which an upper case letter is combined with Chinese characters. Many mixed and lower case 'spelled-out' items are nonce borrowings appearing as hapax legomena (Barracuda180). Mixed and lower case items tend to be more frequent in Taiwan, indicating greater familiarity with English (two proportions test;  $p < 0.0001$ ).

We suggest that the high frequency of upper case letters used as abbreviations and the relative lack of 'spelled-out' items reflects a preference for letters used in the manner of Chinese characters: each letter is discrete visually, pronounced as a syllable, and may have an independent meaning.

\*Third Edition, LDC2007T38.

## **Eric Ringger**

Brigham Young University

### **Compiling and Annotating a Corpus for Syriac**

This paper describes a joint British-American project to create an electronic corpus of Syriac literature. Syriac is a Semitic language spoken mostly by Christians and which has

its own orthographic scripts (more than one). Our goal is to develop an open-source, web-enabled text corpus with annotations that will help scholars perform textual, historical, religious, and linguistic analyses for Syriac. We describe the phased approach we are using to build up the corpus and develop the tools necessary to annotate its content.

Defining the scope of the corpus itself is an issue we are addressing. We are deliberating whether to collect the whole corpus of Syriac literature, focus on only one era (for example the early period), or stop at some stage (say the 14<sup>th</sup> Century). Our discussions include to what extent manuscripts should be included. The paper mentions ongoing efforts to use in production mode an experimental Syriac OCR engine, the only one we know of capable of handling the language.

Our annotation of the textual content involves, first, part-of-speech tagging. Morphological analysis can be executed standalone or for helping to determine parts of speech; we discuss the morphology engine that we have been developing for an autosegmental treatment of parsing wordforms.

Lexical information for the morphological engine relies on an XML encoding of useful entries from pre-existing dictionary resources for Syriac. We describe how we follow current standards for lexical content markup, and how this information serves as a crucial resource for corpus processing.

The discussion is framed by the overall organization of the project: the different types of data, the tools used to analyze and annotate information, the encoding standards, and the project workflow.

Finally, we discuss issues about user interface tools, data visualization, and other questions about deployment of the corpus and related lexical data. We give short examples of the type of linguistic questions that the corpus will eventually help researchers to explore.

## **Chandrika Rogers**

Western Carolina University

### **Articles in Registers of Indian English: A Corpus-based Study**

Indian English has for long been the subject of inquiry from theoretical and historical sociolinguistic perspectives but there is no comparable empirical research tradition on the subject. Kachru's observes that 'There is as yet no large-scale study of spoken or written South Asian English. Nor has any serious attempt been made to distinguish the features in terms of the proficiency scale, the register-specificity of the features and the distribution of grammatical features with reference to the regions' (1994:518) These observations remains true today, and existing studies on Indian English typically focus on how it differs from

other “standard” varieties of the language, such as standard British or American English.. The current paper is part of a larger project that constitutes the first ever large-scale, empirical study of variation and change in the Englishes spoken and written in India. The corpus for the project is a corpus of Contemporary Indian English with 13 written and 10 spoken registers.

Research on the structure of Indian English has thus far focused on a selection of near-stereotypical grammatical features such as stative progressives, prepositions, and use of the perfect aspect instead of the past tense; these features have frequently been discussed as characteristic of Indian English grammar.

The current paper discusses article use in registers of Indian English. Article use in Indian English has been previously studied, but not in appreciable detail. Further, previous studies on article use in Indian English have not adopted a register perspective, but have tended to view Indian English as homogeneous. The current analysis includes the following six different analyses:

- Articles before ordinal numbers
- Articles in noun phrases
- Articles in the following determiners/quantifiers:
  - a lot of*
  - a little*
  - a few*
  - a number of*

## **Ute Römer**

University of Michigan

### **A neo-Firthian approach to academic writing: Uncovering local patterns and local meanings in the discourse of linguistics**

In a lecture on “[d]escriptive linguistics and the study of English” delivered about 50 years ago, John R. Firth noted that “descriptive linguistics is at its best when dealing with [...] restricted languages” (Firth 1956a/1968: 106). A restricted language can be defined as the language of a particular domain (such as science of politics) that serves “a circumscribed field of experience or action and can be said to have its own grammar and dictionary.” (Firth 1956b/1968: 87)

Continuing John Sinclair’s “search for units of meaning” (Sinclair 1996) and using new-generation corpus tools that enable us to explore restricted languages semi-automatically (*Collocate*, Barlow 2004; *ConcGram*, Greaves 2005; *kfNgram*, Fletcher 2007), the aim of this paper is to uncover the phraseological profile of a particular sub-type of academic

writing and to see how meanings are created in a 3.5-million word corpus of linguistic book reviews written in English, as compared to larger corpora of less specialised (or less restricted) languages. Suggesting to connect the corpus-based variationist tradition (e.g. Biber et al. 1999) with the corpus-driven pattern grammar approach (e.g. Hunston & Francis 2000), we will discuss the development of a local lexical grammar of book review language, i.e. a particular type of functional grammar that matches frequent phraseological items and their contextual functions, and that thus helps to provide insights into the creation of meaning in the discourse of linguistics.

### References

- Barlow, Michael. 2004. *Collocate 1.0: Locating collocations and terminology*. Houston, TX: Athelstan.
- Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan (1999). *Longman Grammar of Spoken and Written English*. London: Longman.
- Firth, John R. 1956a. Descriptive linguistics and the study of English. In: Frank Robert Palmer (ed.) 1968. *Selected Papers of J. R. Firth 1952-59*. Bloomington: Indiana University Press. 96-113
- Firth, John R. 1956b. Linguistics and translation. In: Frank Robert Palmer (ed.) 1968. *Selected Papers of J. R. Firth 1952-59*. Bloomington: Indiana University Press. 84-95.
- Fletcher, William H. *KfNgram*. Annapolis, MD: USNA.
- Greaves, Chris. 2005. *ConcGram Concordancer with ConcGram Analysis*. HongKong: HKUST.
- Hunston, Susan & Gill Francis (2000). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.
- Sinclair, John McH. 1996. The search for units of meaning. *Textus IX*(1): 75-106.

## Ute Römer and Stefanie Wulff

University of Michigan

### Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive

Recent corpus studies have shown that language learners are aware of systematic associations between verbs and their preferred argument structures to the same extent that native speakers are (e.g. Gries and Wulff 2005). Given evidence for similarly systematic associations in native speaker data at the lexis-morphology interface (e.g. Römer 2005), the question arises if, and to what extent, learners of English are also sensitive to lexical dependencies at the level of morphology, and how their verb-aspect associations compare to those of native speakers.

In order to address this question, the present study focuses on the potential associations between verbs and progressive aspect in German learners' academic writing, which constitutes a particularly interesting case since progressive aspect marking has no equivalent in German and is therefore often considered to cause problems for German learners,

even at an advanced level. On the basis of the German component of the *International Corpus of Learner English* and the *Cologne-Hanover Advanced Learner Corpus*, learners' significantly preferred verb-aspect pairs are identified using an adaptation of collocation analysis (Stefanowitsch and Gries 2003). The results are complemented with corresponding analyses of the *Michigan Corpus of Upper-level Student Papers* on the one hand and published research articles from the *Hyland Corpus* and the *British National Corpus* on the other hand.

In a nutshell, the findings suggest that advanced German learners of English exhibit clear lexical preferences in the use of progressives; furthermore, the comparative analyses suggest that verb-aspect preferences gradually shift as a function of writers' mastery of text type-specific conventions rather than language proficiency at large, which adds to the growing evidence in favor of English as a *lingua franca* in academic writing settings that ultimately blurs a distinction between native and non-native speakers.

### References

- Gries, Stefan Th. & Stefanie Wulff. Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3:182-200.
- Römer, Ute. 2005. *Progressives, Patterns, Pedagogy. A Corpus-driven Approach to English Progressive Forms, Functions, Contexts and Didactics*. Amsterdam: John Benjamins.
- Stefanowitsch, Anatol and Stefan Th. Gries. Collocations: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2:209-43.

## Juhani Rudanko

University of Tampere

### Recent Change in Core Grammar: a Case Study Based on Corpus Evidence

Consider sentence (1):

- (1) The miserable wife submitted to be fed, looked with forlorn wonder at the children . . . (CLMET, 1867, Ward, *Marcella*)

Sentence (1) illustrates how the verb *submit* selected *to* infinitive complements about 150 years ago. However, as far as sentential complements in present-day English are concerned, *to -ing* complements are generally more common with *submit* than *to* infinitives.

The paper examines the complement selection properties of *submit* in the nineteenth and the first two decades of the twentieth century, on the one hand, and in present-day English, on the other, on the basis of electronic corpora. For historical data, the Corpus of Late Modern English Texts, the CLMET, and the Corpus of English Novels, the CEN, both

developed at the Catholic University of Leuven, are used, and for present-day English, major segments of the full Bank of English Corpus are drawn on. The objective is to trace and to document the change affecting *submit*, paying attention to both British and American English, and to examine the factors that may have influenced the change. Such factors may be syntactic, semantic, or even phonological (the *horror aequi* principle) in nature, and the aim is to understand what may have promoted or inhibited the change in question. The broader objective is to shed further light on a central aspect of what Rohdenburg (2006) has called the Great Complement Shift and on the evolution of the system of English predicate complementation in recent times.

#### References

Rohdenburg, Günter. 2006. "The Role of Functional Constraints in the Evolution of the English Complementation System," Christiane Dalton-Puffer, Dieter Kastovsky, Nikolaus Ritt and Herbert Schendl, eds., *Syntax, Style and Grammatical Norms: English from 1500-2000*. Bern: Peter Lang, 143-166.

### **Moises Almela Sanchez**

Universidad de Murcia

#### **Collocates for Word Sense Disambiguation by means of a Discriminant Function Analysis Model: A Corpus-Based Approach**

This paper explores the potential of collocates for word sense disambiguation (WSD). The WSD prototype pursues two priorities: first, minimize computational costs, and second, deal with different degrees of sense granularity. Computationally, this model has the advantage of involving relatively low-dimensional feature space, because it runs on raw contextual data (concordances). We use discriminant function analysis as it allows us to compute distances between each occurrence and each semantic class; for each meaning, we determine the location of the point (group centroids) that represents the means for all variables (collocational data) and for each case we then compute the distances (of the respective case) from each of the group centroids. Finally, we classify cases as belonging to the group (meaning) to which it is closest. The transition from coarse-grained senses to finer-grained ones can be achieved by means of reiteration of the same algorithm on different levels of contextual differentiation.

**Robert Sanderson, Matthew Brook O'Donnell and Clare Llewellyn**

University of Liverpool

**'Milk, bread and toothpaste':  
Adapting Data Mining techniques for the analysis of collocation  
at varying levels of discourse**

The analysis collocation is one of the fundamental components of corpus linguistics and the pervasiveness of the phenomenon in language has been repeatedly demonstrated. It was established early on that using a window of 4 or 5 words either side of a node item tends to produce the optimum results in capturing the main association patterns without introducing too great a level of 'noise' from textually frequent items.

Association Rule Mining (ARM) is a technique developed in the field of data mining in order to discover frequently co-occurring items in commercial transaction databases. For example, it is used to determine which sets of products are bought at the same time at a supermarket in order to inform store layout, sales and advertising strategies. This paper describes how ARM can be applied to textual data and specifically to the discovery of collocates at clause, sentence, paragraph and text level. This reapplication of the technique enables researchers to automatically and efficiently locate lexical associations, at different levels of textual granularity from the word level up to the entire text, without having to manually sift through the noise generated by increasing the collocation window size. We illustrate how the use of varying transaction sizes and statistical thresholds can produce comparable and in some cases more specific results to those produced using standard collocational analysis. For example, compared to the number of times each word appears separately in a random selection of newspaper articles, the words 'nuclear', 'biological' and 'chemical' very frequently co-occur at a sentence level.

**Monte Shelley and James Rosenvall**

Brigham Young University

**An Introduction to WordCruncher 7.1**

WordCruncher is a text retrieval and analysis program that enables users to develop or use a simple text or very large multilingual UNICODE corpora. It supports the addition of tags (such as part of speech, definitions, lemma, etc), graphics, and hyperlinks to text or multimedia files.

WordCruncher allows users to do simple word searches or more complex word,

phrase, and tag searches. WordCruncher also includes many analytical reports, including collocation, vocabulary dispersion, frequency distribution, vocabulary usage, and various other reports. Text and reports can be copied and pasted into a word processor.

We will demonstrate WordCruncher with the BNC, Dead Sea Scrolls and Popol Vuh.

## Rosário Silva

Linguatca

### Colouring COMPARA: Contrastive and monolingual colour studies in English and Portuguese

In this paper we present the Portuguese and English colour studies we made using COMPARA (Frankenberg-Garcia & Santos, 2003), the largest edited parallel corpus in the world (as far as we know), as well as some findings concerning contrastive analysis.

Being an everpresent element in our world and in our lives, colour was our first choice in exploring COMPARA semantically. Moreover, this apparently simple subject has already given rise to a lot of linguistic argumentation over the status of language vs. cognition and language vs. world.

We start by discussing the annotation process regarding colour, namely, what is colour, what semantic categories were defined and what colour groups were created. Briefly, we marked as colour all straightforward words conveying colour, regardless of word class (*blue, reddened, blackly*, etc.), words that convey colour but are not in themselves a colour (*blond, brunette, blushed*, etc.), words that have colours in them but have gone beyond the mere colour reference (*greyhound, Whitehall*, etc.), and the word *colour* itself and its derivations. (The absence of colour was not marked.) We then defined five semantic categories (`colour`, `colour:race`, `colour:human`, `colour:wine` and `colour:original`) and classified our “colourful” words accordingly, having in mind the context in which they appeared. The words that fell into the category (pure) `colour` were later grouped into seventeen groups, to allow us finer-grained comparisons between authors and languages (Blue, Red, Yellow, Green, Orange, Brown, Beige, Black, White, Grey, Pink, Purple, Gold, Silver, Other, Multiple and Unspecified). (Silva, Inácio and Santos, 2008)

Having done this, we were able to discover which were the favourite colours of English-speaking authors, who contributes the most and the least to these colour preferences, what colour categories dominate each author’s writing, etc., etc. Since preferences by the Portuguese authors in COMPARA were also investigated, we were able to contrast the use of colour in the two languages.

In addition, and taking advantage of the fact that the Portuguese part of COMPARA is syntactically analysed (automatically by the PALAVRAS parser (Bick, 2000) and then

manually revised and documented, see Santos & Inácio, 2006), we studied colour-related syntactical patterns in Portuguese, and also described the kinds of lexemes associated to colour in the texts.

Finally, we discuss a few cases where colour is not translated or is changed, ending with some comments on translation practice in the two directions (English to Portuguese translation and Portuguese to English translation). Our study may present some literary surprises, in the sense that our initial expectations of more coloured authors were not confirmed.

### References

- Bick, Eckhard. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press, 2000.
- Frankenberg-Garcia, Ana & Diana Santos. "Introducing COMPARA, the Portuguese-English parallel translation corpus", in Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), *Corpora in Translation Education*, Manchester: St. Jerome Publishing, 2003, pp. 71-87.
- Santos, Diana & Susana Inácio. "Annotating COMPARA, a grammar-aware parallel corpus". In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)* (Genoa, Italy, 22-28 May 2006), pp. 1216-1221.
- Silva, Rosário, Susana Inácio & Diana Santos. "Documentação da anotação relativa à cor no COMPARA". Continually updated. First version: 27 November 2007. Current version: 8 February 2008. <http://www.linguateca.pt/COMPARA/DocAnotacaoCorCOMPARA.pdf>

### **Marc Silver and Sara Radighieri**

University of Modena and Reggio Emilia

#### **The representation of time and space across historical discourse: a comparative analysis of evaluative effects**

The paper explores ways in which specialized corpora can be used to study language variation within and across disciplines. It is based on an analysis of corpora consisting of research articles in history of ideas, art history and social history written between 1999 and 2001 (approx. 1,000,000 tokens per subcorpus). By choosing three disciplinary areas which at least nominally reference 'history' as a super-ordinate disciplinary category, we wish to analyze temporal and spatial forms of representation characterizing each and explore how these forms may offer insight into the different strategies assumed by the writer in constructing her/his discourse. The fact that these otherwise symmetrical disciplinary areas focus their attention on dishomogeneous types of objects – ideas, works of art, people and events – offers a rich textual environment for exploring differences in interpretation and evaluative positioning.

Although it is now commonly held that the historian's role is not to find out 'what actually happened' but to help construct new visions of the past, deconstructing and reconstructing 'reality' through the argumentational weaving of other texts and voices (Silver and Bondi, 2004; Silver, 2006), little in the way of a systematic linguistic investigation has yet been attempted (Burke, 1991; Coffin, 2003). This paper, which hopes to be an initial contribution in this sense, focuses on some of the typical lexical and grammar patterns and sequences (Hunston and Francis, 2000; Groom, 2005; Hunston, 2006) while situating relevant temporal (e.g. adverbial patterns) and spatial (e.g. metaphorizing sequences) representation which characterize and help differentiate these disciplines.

Within the study of academic discourse, it is hoped that our findings will prove useful in demonstrating both the centrality of temporal versus spatial coordinates in constructing point of view and the innovative potential of concordancing when applied to lexical sequences.

### References

- Burke, P. (1991) *New Perspectives on Historical Writing*. – Polity Press: Oxford, U.K.
- Coffin, C. (2003) Reconstituting the past- settlement or invasion? The role of JUDGEMENT analysis. In: J. R. Martin, R. Wodak *Re/Reading the Past: Critical and Functional Perspectives on Time and Value*, 219-246. Benjamins: Amsterdam.
- Groom, N. (2005) Pattern and meaning across genres and disciplines: An exploratory study. – *Journal of English for Academic Purposes*, 4: 257-277.
- Hunston, S., Francis, G. (2000) *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. – Benjamins: Amsterdam.
- Hunston, S. (2006) Phraseology and system: a contribution to the debate. – In: G. Thompson and S. Hunston (eds.) *System and Corpus. Exploring connections*, 55-80. Equinox: London, Oakville.
- Silver, M., Bondi, M. (2004) Weaving Voices: a Study of Article Openings in Historical Discourse. In Del Lungo Camiciotti, G. & Tognini Bonelli, E. (eds.) *Academic Discourse – New Insights into Evaluation*. Bern: Peter Lang.
- Silver, M. (2006) *Language Across Disciplines: Towards a critical reading of contemporary academic discourse*. – Brown Walker Press: Boca Raton.

**Eric J.M. Smith**

University of Toronto

### Using a Query Language as an Annotation Tool

The motivation for the research described here was a study of the agreement morphology of two ancient languages: Sumerian (southern Iraq) and Elamite (southwestern Iran). There are available electronic corpora for these languages, but these largely provide only transliterated texts with little or no linguistic annotation.

Lacking a suitable annotated Sumerian corpus, Oxford's Electronic Text Corpus of

Sumerian Literature (ETCSL) (Black et al 1998-2006) was chosen as the base corpus for the agreement study. Within the ETCSL the linguistic annotation associated with each Sumerian word consists only of a lemma, a part-of-speech tag, and occasionally an indication of a bound morpheme.

However, even using a sophisticated query language such as LPath (Bird et al 2006), queries of the sort necessary to extract morphological information needed for the study turned out to be impossibly cumbersome, since such queries must refer to orthographic strings that only imperfectly represent the language's morphology. Nonetheless, by using a collection of relatively simple LPath queries it is possible to isolate particular linguistic entities (e.g. *genitive-case-noun*, *ergative-case-noun*, *finite-verb-complex*). Once these entities have been defined, they serve as building blocks for defining higher-level linguistic entities (e.g. *subordinate-clause*, *sentence*). As a first pass, these defined entities can serve as macros within an LPath query that are expanded at run-time. For performance reasons, it is also possible to create a restructured representation of the corpus that reflects the hierarchy that has been revealed by the query process. In either case, the result is an annotation-like structure that was created entirely by queries.

Not only is this query-based approach more practical than manual annotation, but it also avoids the problem where manual annotation requires decisions that presuppose a better understanding of the language's morphology than we actually have. The approach is not specific to Sumerian and the ETCSL, but should be a useful technique for quick and dirty annotation of corpora for other low-resource languages. To test the approach's generality, it is also being used to annotate a corpus of Elamite texts that has been assembled specifically for studying agreement.

## References

- Steven Bird, Yi Chen, Susan Davidson, Haejoong Lee, and Yifeng Zheng. Designing and evaluating an Xpath dialect for linguistic queries. In 22nd International Conference on Data Engineering (ICDE), pages 52–61, Atlanta, April 2006.
- J.A. Black, G. Cunningham, J. Ebeling, E. Flückiger-Hawker, E. Robson, J. Taylor, and G. Zólyomi. The Electronic Text Corpus of Sumerian Literature. <http://www-etcs.orient.ox.ac.uk/>, 1998-2006.

## Laurel Stvan

University of Texas at Arlington

### *Fat and Health Literacy:* Two Revealing Terms in CADOH (Corpus of American Discourses on Health)

This paper presents results of a project that is examining the ways that American discussions of health issues are framed within three types of discourses — those of health care

providers, of patients, and of the popular media. With the goal of examining the linguistically encoded evidence of health beliefs as they appear in the voices of American popular media, a pilot corpus of materials on the topic of health issues was compiled. The data consists of 60 text files spanning the years 1995-2007. Using this corpus, two terms were investigated: one (*health literacy*) is currently widely used in the public health materials as well as media coverage of health, and the other (*fat*) is frequent in both mass media and vernacular usage.

Files for the popular media dataset come from blogs, blog comments, newspaper articles, letters to the editor, magazine articles, transcripts of a panel discussion on women and health, and transcripts of local and national news broadcasts of health-related stories. AntConc was used to collect the contexts and then pinpoint where in the discourse the terms were used and how they were defined. The examination includes descriptions of collocations and the relative frequency of the terms compared to both general corpus of American English and a specialized corpus of medical texts. In addition, the ways in which definitions are overtly presented or their component information is presupposed is described.

The study shows that both of these ostensibly clear terms show evidence of lexical conflation, whereby distinct senses of polysemous terms are mistakenly used interchangeably. Identifying the separate intended meanings of *fat* and *health literacy* can facilitate more effective discussions for all the participants in an exchange of healthcare information — medical providers, public health administrators, and reporters, as well as medical consumers.

## Hongyin Tao

University of California, Los Angeles

**“I’ve never seen anything like it”:  
An inquiry into the grammatical clustering involving *never*  
in spoken American English**

In most standard English grammars, *never* is simply treated as just another *no*-type of negative token that can be added to, and hence negate, a basic declarative sentence. An examination of natural language corpus data shows that there are preferred patterns in spoken discourse which manifest as bundles of certain types of clausal elements. Using the Santa Barbara Corpus of Spoken American English as database, I show that *never* in spoken discourse collocates strongly with the first person singular pronominal subject, verbs of cognition (e.g. *know*, *think*, *hear*, and *see*), and marked aspects and voices with a modal expression, as exemplified by the utterance in the title of the paper: “I’ve never seen

anything like it”.

Given that there are no intrinsic semantic constraints for a negative token such as *never* to form such grammatical constellations, I argue that the explanation for this has to come from the pragmatics of interactive discourse and its interaction with the lexical semantics of the negator. Specifically, such an explanation involves the speaker role (first person, main speaker) and the subjective nature of the utterances associated with such a role, the extremity of the semantic connotations of the negative token, and the interactive nature of the communicative context. Each of these factors can be shown to be responsible for some elements in the constellation, and together they give rise to the observed grammatical patterning in discourse. Thus, this paper provides just another example demonstrating the discourse basis of grammar and the necessity to integrate corpus-based findings with interactive linguistics.

**Laura Teddiman**

University of Alberta

### **Comparative and Superlative Adjectives and Textual Frequency**

This study explores the textual frequencies of the broad classes of comparative and superlative adjectives, as they are distributed in English. The primary source of linguistic data is the British National Corpus (BNC). Comparative and superlative adjectives are unmarked with respect to each other, according to the hierarchy described by Greenberg and as reported on in Croft (1990: 92), although both are marked with respect to positive (degree) adjectives. Through the use of a large corpus such as the BNC, it is possible to test whether comparative and superlative adjectives in English are unmarked with respect to each other, or if there are finer gradations than would otherwise be supposed. Comparative adjectives (tagged as [ajc]) occur with an overall frequency of 1,959.32/million, while superlative adjectives (tagged as [ajs]) occur with an overall frequency of 915.84/million. This is a difference of 1043.48/million words, indicating that in this data, comparative adjectives occur more frequently than superlative adjectives by a fair margin. Similarly, the frequencies reported for individual words in the superlative category drop off sharply after the top item (reported by occurrence per 1,000,000) <best 260.61, latest 64.62>, whereas the decrease proceeds more slowly in the comparative category <further 214.53, better 208.12, higher 154.99>. Identities of the most frequent adjectives in both categories are considered, as is a brief discussion of text type. Results from this study suggest that superlative adjectives in English may be marked with respect to comparative adjectives when textual frequencies are employed.

## References

Croft, W. (1990). Markedness in typology. Chapter 4 of W. Croft, *Typology and Universals*, pp. 63-94. Cambridge: Cambridge University Press.

## Elke Teich

Institut f. Sprach- und Literaturwissenschaft

### Characterizing genre: The case of scientific texts

In this paper, we report on a project investigating the registerial properties of English scientific texts.\* The goal of the project is to investigate the emergence of registers at the boundaries of computer science and other disciplines (e.g., bioinformatics, computational linguistics). To this end, we have compiled a corpus, called the Darmstadt Scientific Text Corpus (DaSciTex), drawing on 23 sources (scientific journals) covering 9 scientific disciplines, each of them forming a subcorpus. Altogether the corpus contains around 2 million running words.

Before starting any register analyses, we want to make sure that the text collection in the DaSciTex corpus is sufficiently distinctive in terms of genre. While scientific texts do vary in terms of register (field, tenor and mode), they are commonly considered a genre in its own right within the class of factual writing (cf. Halliday & Martin, 1993; Swales, 1990). The shared properties of scientific texts commonly acknowledged are information density, technicality and abstractness - properties that make scientific texts functional for experts but hard to consume for laypersons. At a shallow linguistic level, these properties are reflected as follows:

- information density is reflected in a relatively high lexical density (LD, measured as the number of lexical words per clause);
- technicality is reflected by a relatively low type-token ratio (TTR);
- abstractness is reflected by a relatively high ratio of nouns (compared to other parts-of-speech).

If it is true that information density, technicality and abstractness are characteristic of scientific texts, then the DaSciTex corpus should be (a) significantly different from corpora containing exponents of other genres or from genre-mixed corpora wrt LD, TTR and noun ratio; and (b) it should be internally coherent wrt LD, TTR and noun ratio (i.e, the values of these features should not vary significantly).

We conducted several experiments testing for LD, TTR and noun ratio comparing the DaSciTex corpus with FLOB using techniques such as feature selection, clustering, and classification. The results clearly show that the DaSciTex corpus is both distinct and coherent with regard to these features.

## References

- Halliday MAK & Martin J.R., 1993. *Writing Science. Literary and Discursive Power*. Falmer Press, London.
- Swales J., 1990. *Genre Analysis: English in Academic and Research Settings*. CUP, Cambridge.

\*Supported by Deutsche Forschungsgemeinschaft (DFG).

## Yukio Tono

Tokyo University of Foreign Studies

### **Towards a Multi-layered & Multimodal Annotation Model of Learner Corpora**

One of the difficulties of tagging L2 learner data is the fact that the texts contain various kinds of performance errors. Automatic taggers tend to show a much lower rate of accuracy in assigning part-of speech or syntactic information for professional written texts, not to mention the feasibility of tagging learners' errors in students' writings. Learners' errors could be of many kinds, ranging from spelling errors to word order errors or pragmatic errors. Traditional error tagging has been done in such a way that an annotator assigns one of the possible interpretations of errors to the given case, and usually do not show more than one way to correct the errors. However, error corrections could be very problematical, for, in reality, there could be more than one way to correct the errors and it is totally up to the annotator which correction she or he will choose. Thus, there is a growing need of annotation schemes and tools which make it possible to assign more than one type of interpretation to the given position of the text. Embedded annotations are not appropriate for this task, nor is a flat or linear annotation model. Instead, it would be necessary to apply the stand-off multiple-layered annotation models to tag the learner corpora (Ludeling et al, 2005).

In this paper, we will present an on-going collaborative research project of annotating L2 learner corpora for complex extra-textual information as well as textual information especially required for learner corpora, such as error types and suggested corrections. We will firstly review the existing approaches to error-tagging and their potential limitations and problems. Next, we will report on how to customize MMAX2 (Müller & Strube, 2006) to deal with various aspects of L2 learner information, including very complex learner and task variables, as well as complicated error annotations inside the text. Demonstration on how to annotate, query and generate report with this tool will be followed by interactive Q & A from the audience. In so doing, we hope this error tagging model can fulfill various requirements of error-tagging in learner corpora, and will eventually benefit researchers in the community.

## References

- Anke Lüdeling, Maik Walter, Emil Kroymann, Peter Adolphs (2005): Multi-level error annotation in learner corpora. Proceedings of Corpus Linguistics 2005, Birmingham. (<http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc>)
- Müller, C. & Strube, M. (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee (Eds.): *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, pp.197-214. (*English Corpus Linguistics*, Vol.3).

**Gunnel Tottie**

University of Zurich

**Did the boys leave or not? Negation and quantifier scope: a corpus study**

Philosophers and theoretical linguists have given a lot of attention to sentences with universal quantifiers and negation like (1), discussing the possible interpretations in (2) and (3). In (2) negation has scope over the universal quantifier *all* (Neg-Q), whereas in (3), the negative has scope only over the verb (Neg-V).

- |   |       |
|---|-------|
| (1) All the boys didn't leave                                   |       |
| (2) Not all the boys left (some stayed)                         | Neg-Q |
| (3) All the boys 'not-left' (none of the boys left, all stayed) | Neg-V |

Scholars basing their arguments purely on introspection have argued that the Neg-V reading is the only logical one.

There has been little empirical research, but especially Carden (1973) has argued that there are different dialects preferring Neg-Q or Neg-V, based on elicitation experiments. However, a pioneering unpublished study by Taglicht from the 1980's is based on corpus evidence from the small Brown, LOB and London-Lund corpora and points to some interesting facts. Taglicht reminds us that sentences containing *all* plus negation can in fact have three different interpretations, as shown by (4)-(7).

- |  |              |
|--|--------------|
| (1) All the bills did not amount to fifty dollars                |              |
| (2) Not all the bills amounted to fifty dollars (but some did)   | Neg-Q        |
| (3) All the bills 'not-amounted' to fifty dollars (not one did)  | Neg-V        |
| (4) All the bills taken together did not amount to fifty dollars | Coll[ective] |

The third reading is a "collective" one, meaning that the total number of bills did not amount to fifty dollars. All three readings occurred in Taglicht's material, but the Neg-V

reading was definitely the least frequent one. However, Taglicht's corpora were too small to yield conclusive results; in particular, he found no instances of universal quantifier/negative interaction in spoken language at all.

We carried out a large-scale study based on the British National Corpus, which yielded some interesting findings, including many examples from speech. Neg-Q and Coll readings by far outnumbered Neg-V readings. We also found that results differed sharply when *all* was used as a predeterminer (as in *All the boys...*) and when it was a NP head (as in *All is not lost*), something that has not been previously observed in the literature. Moreover, we noticed that a large proportion of examples were formulaic-collective, as in *All the money in the world could not save him*. Our results clearly demonstrate the value of very large corpora.

## **Carrie Hsin-wen Tseng and Shelley Ching-yu Hsieh**

National Cheng Kung University

### **Stock Jargon in Discourses: A Corpus-based, Comparative Study of Mandarin Chinese and English**

Stock is a popular investment instrument in Taiwan. When reading newspaper, stock issues are always one of the main topics. The stock market jargons are also applied to our daily conversations, e.g., *wo gen ta laoshi bu duipan* 'I-and-he-always-not-suit-stock quotation board; we don't get along with each other'. This study examines Mandarin Chinese and English data relating to stock market jargon that are applied as a metaphor in our daily conversation, such as *duipan* 'suit-stock quotation board; getting along'. The data are collected from Corpus of newspapers in Taiwan and Academia Sinica Balanced Corpus of Modern Chinese.

Theoretical background: Lakoff and Johnson's Contemporary Theory of Metaphor (1980) proposes that people rely on conceptual mappings in interpreting conceptual metaphors. Ahrens's Conceptual Mapping Model (2002) based on Contemporary Theory of Metaphor (Lakoff and Johnson 1980, Lakoff 1993) shares the notion that metaphors have systematic source to target domain pairing. Our research questions are: (1) What are those stock market jargons applied to general discourses? (2) How does Conceptual Mapping model work in the Mandarin Chinese and English data? (3) Do Mandarin and English share similar conceptual metaphors? Why?

The results show that (a). Most common stock jargons are applied to general discourses, e.g., *lianai jiu gen touzi yiyang, zongshi gai wei ziji she ge tingsundian ba* 'Love is the same as investment, you should set a goal of profit and loss.' (b). Overall, the Conceptual Mapping Model shows similar results in Mandarin and English, however, some differ-

ences are significant. (c). Different languages share similar conceptual metaphors, but they differ in what is mapped linguistically. Corpora provide great linguistic data to examine stock market jargons.

### References

- Ahrens, Kathleen. (2002). "When Love is Not Digested: Underlying Reasons for Source to Target Domain Pairings in the Contemporary Theory of Metaphor" In Yuchau E. Hsiao (ed.). *Proceedings of the First Cognitive Linguistics Conference*. Cheng-Chi University. 273-302.
- Lakoff, G. & Mark J. (1980). *Metaphors We Live By*. Chicago and London: The University of Chicago Press.
- Lakoff, George. (1993). "The Contemporary Theory of Metaphor." In Andrew Ortony (ed.). *Metaphor and Thought*. Second Edition. Cambridge: Cambridge University Press. 202-251.

## Jeff Turley

Brigham Young University

### **Reflexive Pronoun as Discourse Focus Marker: The Case of Spanish *morir* vs. *morirse***

The Spanish verb *morir* 'to die' frequently appears with a reflexive pronoun (RP), the function of which persists in defying complete exposition. While two factors seem to check or prohibit the use of RPs, namely (a) hyper-formal style, e.g. journalistic or scientific prose, and (b) the explicit mention of the cause of death, e.g. *(\*se) murió de malaria* 'she died from malaria' (*Corpus del español*), and while several factors are known to favor its appearance, namely (c) informal conversational register, (d) death as metaphor, e.g. *\*(me) muero de sueño* 'I'm dying of sleepiness' (*Corpus del español*), as well as (e) the presence of the so-called "ethical" dative pronoun, e.g. *\*(se) le murió el canario* 'her canary died' (*Corpus del español*), in all other contexts the two forms seem to be "virtually interchangeable" (Moreira Rodríguez and Butt 1996:239). Previous writers have proposed that the interpretation of reflexive intransitive verbs arises from the interplay of the idiosyncratic lexical semantics of each predicate and the RP, which focuses attention on the subject (Bull 1952, García 1975, Maldonado 1999). While this analysis is probably sound, it fails to account for the apparently random appearance of the RP with *morir* mentioned above. It is proposed that instead of the RP occurring in free distribution with zero reflexive marking in these problematic cases, the RP has a tendency to function as a discourse focus marker; specifically it signals that the event of dying is new(er) information. A series of syntactic and discourse structures are postulated as evidence of focus on predicates containing *morir*. Data supporting this claim has been culled from the *Corpus del español*, as well as from the "Chiara" Corpus of Dominican Spanish.

## References

- Bull, William E. 1952. "The Intransitive Reflexive: 'Ir' and 'Irse.'" *The Modern Language Journal* 36.8: 382-386.
- Chiara, Daniel, and Angela Chiara. 1999. *Corpus of Dominican Spanish*. MA Project, Brigham Young University.
- Davies, Mark. *Corpus del español*. <http://www.corpusdelespanol.org>, last accessed 3 October 2007.
- García, Erica C. 1975. *The Role of Theory in Linguistic Analysis: The Spanish Pronoun System*. North-Holland Linguistic Series, 19. Amsterdam: North-Holland.
- Maldonado, Ricardo. 1999. *A media voz: Problemas conceptuales del clítico SE en español*. Ciudad de México: Instituto de Investigaciones Filológicas, UNAM.
- Moreira Rodríguez, Antonia, and John Butt. 1996. *Se de matización and the Semantics of Spanish Pronominal Verbs*. London: King's College, Department of Spanish and Spanish-American Studies.

## **Geertje van Bergen and Peter de Swart**

Radboud University Nijmegen

### **Scrambling in Spoken Dutch**

Scrambling is a word order phenomenon referring to the placement of a direct object with respect to an adverbial. In Dutch, direct objects can either precede adverbs (the scrambled order) or follow them (the unscrambled or basic order). Theoretical analyses of scrambling in Germanic languages in general and Dutch in particular have often attributed the scrambling behaviour of objects to their definiteness value. In this view, definites scramble obligatorily in order to escape semantic restrictions (Diesing 1992, Diesing and Jelinek 1995) or alternatively are expected to show a strong preference for scrambling when they are anaphoric (De Hoop 2003).

In this paper we present a corpus study of the scrambling behaviour of direct objects in spoken Dutch based on 4000 sentences involving a direct object and an adverbial. These sentences are extracted from the syntactically annotated part of the Spoken Dutch Corpus (Corpus Gesproken Nederlands; <http://lands.let.ru.nl/cgn>). We will report the influence of a number of different factors on scrambling in spoken Dutch: animacy, pronominality, definiteness, anaphoricity, weight of the object/adverb, type of adverb, and type of verb. In particular, we will show that definiteness does not have the effect on scrambling that it has been ascribed in the theoretical literature. Both definites and indefinites show a strong tendency not to scramble and the anaphoricity of the object does not alter this behaviour.

**Karen Vogel and Gerald Delahunty**

Colorado State University

### **A Corpus Analysis of Dative Clitic Doubling in Spanish**

In this presentation, we report on an empirical study using the Corpus del Español, of the factors that favor and inhibit dative clitic doubling in contexts in which it is optional.

Spanish allows for the expression of the indirect object of transitive verbs in three forms: 1) as a dative clitic, 2) as an indirect object NP, or 3) as a construction in which both the dative clitic and the IO NP are expressed. Constructions of the last type are referred to as clitic-doubled constructions (CL-D), which are problematic for theoretical and pedagogical descriptions, as the object argument in these structures appears to be expressed twice. While most research on dative CL-D has been carried out from a theoretical perspective within the generative grammar framework, some studies have attempted to explain the pragmatic aspects of CL-D expression, i.e., what discourse factors motivate speakers to employ or to not employ CL-D in contexts in which it is optional. Earlier researchers (Blickford 1985; Silva-Corvalán 1981; Solé/Solé 1977) suggested that inherent qualities of the NP (such as animacy, definiteness, specificity, or number, among others) predicted the use or omission of dative CL-D. Recent quantitative analyses of CL-D (Weissenrieder 1995, Koontz-Garboden 2002) suggest that the status of the IO in discourse, specifically its degree of topicality, is related to dative CL-D: more topical NPs favor CL-D.

The present study expands the quantitative research of Weissenrieder (1995) and Koontz-Garboden (2002) to a larger corpus (the Corpus Del Español), which includes various genres and dialects. Using the Corpus Del Español, this paper investigated the relationship between inherent IO NP qualities and the expression of clitic-doubled indirect objects, specifically addressing the NP features of humanness, animacy, specificity, definiteness, number, and an overall topicality score. The results of the analysis indicate that the features of humanness and/or animacy, combined with singular number, strongly favor CL-D. Specificity and definiteness were not highly correlated with CL-D, calling into question the hypothesized relationship between the anaphoric topicality status of NP referents and CL-D. The study also examined the relationship between genre and dative CL-D, and found that genre was a significant factor in the presence or absence of dative CL-D, indicating that stylistic preferences may override NP features with regard to preference for CL-D.

**Xingfu Wang, Eric Ringger, Guohui Liu, Shiping Liu**  
Chongqing University and Brigham Young University

### **Analysis of Canonical Chinese Antonym Co-occurrence**

Chinese antonym pairs act differently than antonyms in English because with Chinese parataxis they can be combined as compounds without any connecting word. Chinese opposites in a pair often co-occur in a sentence, and in a phenomenon unique to Chinese most of them are good bases for making 4-character idioms. The opposites in a pair can usually be separated by an arbitrary number of words to form new larger collocations, in such a way that the same number of characters precedes or follows each element of the antonym pair to make the whole collocation symmetric. Thus new phrases or idioms are made, although most of them retain a trace of the original antonym pair meaning. A Chinese scholar (Tan 1989) claimed that there are 371 canonical antonym compounds that are used often in Chinese. Our research focuses on two parts using this list: firstly, we analyzed Tan's antonym pair list using the modern CCL corpus (Corpus of the Center of Chinese Linguistics, PKU) and found that Tan's list should be revised in light of the data from the corpus. Secondly, we identify the patterns involving the interpolated elements. Some phrases incorporating Chinese antonym pairs have the expanded form  $x+m+!x+n$  or  $m+x+n+!x$ , where  $!x$  denotes the opposite of  $x$ , and  $m$  and  $n$  usually have the same length in characters. We have analyzed typical interpolations ( $m$  or  $n$ ) of length one in Chinese spatial antonym pairs and identified the patterns involved in the separation and linkage of the antonym pairs. Analyses using the CCL corpus reveal that Chinese antonym pairs contain richer lexical and syntactic information than was found by Jones for English (Jones 2006) due to the unique characteristics of Chinese. The features of the interpolated characters and their semantic relations are predictive of the patterns for new collocations involving antonym pairs in Chinese and are also predictive of constructions unlikely to emerge in usage.

**Martin Warren**

The Hong Kong Polytechnic University

### **A corpus-driven study of phraseological variation**

This paper describes a new way of identifying and categorising word associations which captures all of the permutations of up to five associated words. In other words, the search engine finds instances of the associated words (e.g. AB) irrespective of constituency varia-

tion (e.g. A\*B, A\*\*B) and positional variation (e.g. BA, B\*A) generated by the association of between two and five words. The products are termed 'congrams' and they differ from n-grams (a.k.a. clusters or bundles) and skipgrams (gapped n-grams) in that they include all of the possible forms of phraseological variation. As a result, congram search results reveal more fully the phraseological tendency (Sinclair, 1987) in language.

Fully automated searches (i.e. the user does not need to nominate any search items) on a 5 million-word corpus of written and spoken English have found that the majority of congrams are made up of non-contiguous word associations. Contiguous word associations are also found in congram searches, but, since many collocational patterns never occur contiguously, searches which focus on contiguous collocations present an incomplete picture of the word associations that exist. Examples of the main types of congram are presented and a framework for analysing them is explained.

#### Reference

Sinclair, J. McH. 1987. Collocation: A Progress Report. In R. Steele and T. Threadgold (eds.) *Language Topics: Essays in Honour of Michael Halliday*. Amsterdam: John Benjamins. 319-331.

### **Arthur H. Wendorf**

University of Georgia

#### ***De Ahí, Por Consiguiente, Por Ende, Por Lo Tanto and Por Tanto: A Distributional Diachronic and Synchronic Analysis***

The Spanish discourse markers (DMs) *de ahí*, *por consiguiente*, *por ende*, *por lo tanto* and *por tanto* have traditionally, especially in the case of the latter two, all been grouped together as near synonyms, and all of them indicate that what follows them in a given discourse is a consequence of the argument that precedes them. This study questions their synonymy with the help of corpora based analyses. First, a diachronic comparison of the distribution of these five DMs between the thirteenth and twentieth centuries is made using the Corpus del Español provided by Mark Davies. Results demonstrate that the relative frequencies of these DMs has fluctuated significantly during the given time period. Next, synchronic comparisons are made according to country, with all officially Spanish speaking countries, other than the Equatorial Guinea, and the United States being represented, register, contrasting the frequencies of use in books, newspapers, magazines and oral registers, and position within a sentence, comparing sentence initial, sentence medial and sentence final positions. The synchronic comparisons are made using the Corpus de Referencia del Español Actual provided by the Real Academia Española. Results indicate that in modern Spanish the relative frequency of use of these DMs varies significantly

according to country and register, though not according to position within a sentence. It is concluded that, while these DMs may serve similar semantic functions, their historical and current distributions indicate that they are not pragmatically synonymous.

## **Hongmei Wu**

University of Arizona

### **Corpus consultation in drafting and revising: A case study of a biomedical ESL graduate student**

The pedagogical benefits of introducing corpora to advanced ESL writers as a reference tool have been suggested in many corpus-based EAP/ESP studies. It has been argued that using corpora may help promote learner autonomy and data-driven, inductive learning as opposed to teacher-centered, rule-governed and deductive learning. However, empirical studies of these potential benefits are few, especially in the area of teaching academic writing to ESL graduate students, with only a few notable exceptions (e.g. Chambers & O'Sullivan, 2004; Coniam, 2004; Lee & Swales, 2006). To have a close look at how corpus can be used to assist academic writing, the present study investigates an ESL graduate student's experience of learning to consult a corpus when preparing a manuscript for publication. 40 journal articles in three key journals in his discipline of biomedical studies were compiled into a corpus, and the student was taught to search the corpus for a number of linguistic forms, such as tense in different sections, citation attribution, linguistic expressions for academic praise and criticism, and hedging devices in the corpus. He was then asked to refer to the corpus when drafting and revising his paper. His first two drafts, along with comments from his lab mate and advisor, were collected and analyzed. Interviews were also conducted with the student and his lab mate and advisor. Initial analysis and interviews indicate that corpus can be a very convenient tool for finding the appropriate hedging and evaluative expressions and that by displaying word concordances in limited contexts, it draws attention to linguistic forms (e.g. tenses) that tend to be overlooked when reading for information. Practical constraints such as time investment on the part of the student were also discussed.

## Pengcheng Wu and Huaqing Hong

Nanyang Technological University

### An Efficient Query Package for Richly Annotated Discourse Corpus

The technique of indexes has been used in corpus search engine research and achieved great success (Lou, 2004; Lai, 2006; Mark 2005). However, the previous applications focus on the query at the lexical or syntax level, and don't work with the richly annotated discourse corpus.

Our work aims at the highly efficient index and query on the richly annotated corpus. Richly annotating not only mark up the information an individual level, but also represent the relationships between information from different levels (Wouter, 2006). The stand-off annotation strategy provides a simple principle in which a new level of annotation can be easily added without damaging the existing ones (Müller 2005). Nevertheless, it would also bring difficulties to query because it is quite often a query requires concurrent markup merged. (Peter, 2006).

To deal with this problem, we investigate how to effectively combine the indexes of multiple levels. We first classify different annotation levels based on the annotation schemes. Then we index the annotation information of the same level via reverse indexes (Sergey, 1998), and index the annotation information of the different levels as well as their relations via forward indexes. To make the query further efficient, we create a cache to store the most frequent indices. We have built a component to map heavy annotation of different levels into our specified model, and setup a website for the search purpose.

In our on-going \*\*\*\* project (omitted for blind review), we have annotated a large number of data at different linguistic and discourse levels, which is to be delivered with this query package. We evaluate our solution on a pseudo corpus comprised of millions of marked information. The experimental results are promising; in most cases, the query results can be returned in just one second.

### References

- Christoph, Müller. 2005. A flexible stand-off data model with query language for multi-level annotation. *Annual Meeting of the ACL*, 2005.
- Stefanie D. Lukas F. Ulf L. & Anke L. 2004. Challenges in Modelling a Richly Annotated Diachronic Corpus of German. Retrieved on Aug. 05, 2007, at: <http://citeseer.ist.psu.edu/659987.html>
- Michael,C.& Oren E. 2005 A Search Engine for Natural Language Applications. *International World Wide Web Conference Committee (IW3C2)*.
- Dan,C. & Cristina, B. 2004. Hierarchical XML Layers Representation for Heavily Annotated Corpora. *Proceedings of the LREC 2004 Workshop*.
- Hong, Huaqing. 2005. SCORE: A Multimodal Corpus Database of Education Discourse. *Proceedings of International Conference of Corpus Linguistics (ISSN 1747-9398), Birmingham, July 14-17, 2005*.
- Lou,B. 2004. Xaira: an XML Aware Indexing and Retrieval Architecture. *Digital Resources for the Humanities and Arts. DRHA2007*.
- Lai, C. 2006. A Formal Framework for Linguistic Tree Query. Retrieved on Aug. 05, 2007, at: <http://eprints>.

[unimelb.edu.au/archive/00001594/](http://unimelb.edu.au/archive/00001594/).

- Mark D. 2005. The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics* 10: 301-28.
- Wouter, A. Valentin, J. David, A. Maarten, de, R. Peter, B. & Arjen, V. 2006. Representing and Querying Multi-dimensional Markup for Question Answering. *11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Peter, S. 2006. Querying XML documents with multi-dimensional markup. *11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Fabio, Rinaldi. Gerold, Schneider. Kaarel, K. Michael, H. & Martin, R. 2006. An environment for relation mining over richly annotated corpora: the case of GENIA. *Second International Symposium on Semantic Mining in Biomedicine (SMBM) Jena, Germany*.
- Nancy, I. Laurent R. & Eric de la C. 2003. International Standard for a Linguistic Annotation Framework. *Proceedings of the HLT-NAACL 2003 workshop on Software engineering and architecture of language technology systems - Volume 8*
- Steven, B. & Mark, L. 2001. A formal framework for linguistic annotation. *Speech Communication archive Volume 33, Issue 1-2 (January 2001)*
- Paul, R & James W. 2006. Annotated web as corpus. *the 11th Conference of the European Chapter of the Association for Computational Linguistics*.
- Sergey, B. & Lawrence, P. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the seventh international conference on World Wide Web 7*

## Stefanie Wulff

University of Michigan

### A multifactorial approach to *that*-deletion in English complement constructions

This paper investigates the variable presence of the complementizer *that* in English subject, object, and adjectival complement constructions exemplified in (1) to (3).

- (1) She thought (that) her colleague was in New York. (direct object)
- (2) The problem is (that) we don't have internet access here. (subject complement)
- (3) I'm glad (that) I found my purse later that night. (adjectival complement)

Recent studies presented semantic (Dor 2005), cognitive-functional (Kaltenboeck 2006), discourse-functional (Thompson and Mulac 1991), and frequency-driven (Tagliamonte and Smith 2005) accounts for this phenomenon. There is general agreement that more than one factor is responsible for the absence or presence of *that*, and that the relative weight of these factors varies depending on the type of complementation construction. To date, however, multifactorial analyses are scarce and are restricted to the analysis of individual complement constructions.

The present paper presents a comparative analysis of the factors determining the function of *that* in different complement constructions. It is the first study that (i) takes

more than one complementation construction into consideration and (ii) subjects the data to a multifactorial analysis. All instances of subject, object, and adjectival complementation constructions were extracted from the British component of the *International Corpus of English* and coded for the various factors proposed in the literature (including register; formality; the structural complexity of the subject in the matrix and complement clauses; the presence or absence of material intervening between the clauses; the coreferentiality of the subjects of the clauses; and the corpus frequency of the main verb) and fed into a logistic regression model. The results expose related, yet distinctive construction-specific meanings of *that*, which are argued to be accounted for best as different instantiations of the abstract feature of 'distance' as suggested by Kaltenboeck (2006).

#### References

- Dor, D. 2005. Toward a semantic account of *that*-deletion in English. *Linguistics* 43.2:345-82.
- Kaltenboeck, G. 2006. '... *That* is the question': complementizer omission in extraposed *that*-clauses. *English Language and Linguistics* 10.2:371-96.
- Tagliamonte, S. and J. Smith. 2005. *No momentary fancy!* The zero 'complementizer' in English dialects. *English Language and Linguistics* 9.2:289-309.
- Thompson, S.A. and A. Mulac. 1991. The discourse conditions for the use of the complementizer *that* in conversational English. *Journal of Pragmatics* 15:237-51.

### Shozo Yokoyama

University of Miyazaki

#### A Cross-sectional Analysis of Lexical Bundles in Written Medical Discourse

This paper investigates the characteristics of written academic discourse by means of a corpus-based quantitative analysis. It examines those multi-word expressions usually referred to as lexical 'chunks', 'clusters' or 'bundles'. Bundles have begun to attract considerable attention in corpus studies in EAP (e.g. Hyland 2007; Cortes 2004; Biber et al. 2004). However, the issue of differences according to field or genre-specific collocations and set phrases still remains uncertain. This paper will explore the structures and functions of 4- or 5- word bundles in one representative academic discourse: medicine. The corpus-data used for this investigation is comprised of scientific articles of more than 5 million words in the fields of medicine and nursing, taken from over one hundred kinds of international journals electronically published on the Web, covering the period 2001-2006. The data was accumulated and categorized according to the following specialized subfields: genome biomedicine, clinical surgery, nursing and public health. All articles were separated electronically and POS-tagged according to the following rhetorical sections: Introduction, Methods, Results and Discussion (IMRD). This was done in order to determine how rhe-

torical functions are realized by target bundles in each category, employing the taxonomy and methodology proposed in Hyland (2007). The analysis shows that the sets of lexical bundles found in those different sub-disciplines and rhetorical sections play a greater part in establishing cohesive and persuasive discourse, and performing genre-specific lexico-rhetorical functions, both for readers and writers of academic articles. This research is partly supported by the Grant-in-Aid for Scientific Research by Ministry of Education, Culture, Sports, Science & Technology, Japan and the Project Research 2007 funded by Kumamoto University in Japan.

## **Tatiana Zdorenko**

University of Alberta

### **Subject omission in Russian: A study of the Russian National corpus**

This study investigated subject omission in spoken and written corpora of Russian in order to produce a quantitative comparison of omission in different genres and morphosyntactic environments. Previous theoretical studies of Russian described subject omission using isolated constructed sentences, and most corpus studies analyzed written literary language. However, subject omission in Russian is a discourse phenomenon, and the present study took a more appropriate approach, namely to investigate subject omission in coherent spontaneous text, focusing on spoken data.

The analysis showed that subjects were not omitted to the same extent in all genres and registers. The percentage of omitted subjects was the highest in the corpus of informal spontaneous conversations, and omitted subjects were practically absent in the written corpus, even in the most informal register.

In a more detailed analysis of the spoken corpus, the relation between null subjects and the notion of subject topicality was examined. The comparison of the frequency of null subjects in different person contexts provided support for the Topicalization Hierarchy of person. More null subjects were found in the first and second person contexts than in the third person contexts. In contrast, in written Russian there was no significant effect of person on the proportion of null subjects.

Finally, an analysis of omitted subjects used with specific verb types was built on previous cross-linguistic studies of grammaticalized collocations such as *I dunno* or *y'know*. The analysis was aimed at isolating potential discourse markers, i.e. verbs with a particular pragmatic function that grammaticalized either in a subjectless form or with a certain pronominal subject. It was concluded that *znat'* 'know' and *ponimat'* 'understand' were the likely candidates for grammaticalization as discourse markers.

## Presenters

Adami, Elisabetta	Tubing the Web: a corpus-based study on video-communication	University of Verona	Italy
Al Nafjan, Eman	The Nora Corpus: a study of Arab EFL written discourse	King Saud bin Abdulaziz University for Health Sciences	Saudi Arabia
Albakry, Mohammed	Social Taboo, Evaluation, and Identity Construction Online	Middle Tennessee State University	US
Alexander, David	El coche ese vs el coche de marras: The postnominal demonstrative and de marras in Diachronic and Synchronic Corpora	Ohio State University	US
Amaral, Patricia and Chad Howe	Using corpora to quantify contexts: The case of the Portuguese Perfect	University of Coimbra	Portugal
Andrade, Aroldo	A corpus-based research on the history of clitic climbing in European Portuguese	State University of Campinas (Unicamp)	Brazil
Artstein, Ron and Massimo Poesio	The Arrau corpus of anaphoric relations	University of Southern California, University of Essex, and Università di Trento	US, UK, Italy
Baayen, Harald	Co-occurrence below and above the word level: exploring language at the intersection of corpus linguistics, psycholinguistics and statistics	University of Alberta	Canada
Baker, Paul and Costa Gabrielatos	Representations of Islam in the British and American press 1999–2005	Lancaster University	UK
Baker, Paul and Costa Gabrielatos	Using collocational profiling to investigate the construction of refugees, asylum seekers and immigrants in the UK press	Lancaster University	UK
Barrett, Del	Parents, victims, suspects, victims...: The 'framing' of the McCanns in the British tabloid press.	King's College London	UK
Bartsch, Sabine	Intermodal cohesion and coherence in multisemiotic text	Technische Universität Darmstadt	Germany

AAFL 2008

Belladelli, Anna	American slang in mainstream magazine writing	University of Verona, Italy	Italy
Berber Sardinha, Tony	The CEPRIL Metaphor Candidate Identifier: A program for identifying metaphor in corpora	Pontifical Catholic University of Sao Paulo	Brazil
Biber, Douglas	Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles	Northern Arizona University	US
Bloom, Ken	Learning Appraisal Extraction Patterns	Illinois Institute of Technology	US
Boulton, Alex	Evaluating corpus use in language learning: State of play and future directions	CRAPEL-ATILF/CNRS, Nancy Université.	France
Brinton, Laurel	Historical Pragmatics and Corpus Linguistics: Problems and Strategies	University of British Columbia	Canada
Camargo, Laura	The Project for the Sociolinguistic Study of the Spanish Language of Spain and the Americas (PRESEEA)	Universitat de les Illes Balears	Spain
Chao Castro, Milagros	The OED as a Corpus: Looking for Dual-Form Adverbs	University of Santiago de Compostela	Spain
Chapman, Don	Using Corpora in an English Usage Class	Brigham Young University	US
Chen, Ya-Jie and Hui-Chuan Lu, Kailing Liu	Corpus-based Study of New Motherhood in Charlotte Perkins Gilman's <i>Herland</i>	National Cheng Kung University	Taiwan
Chesley, Paula	Towards Predicting New Words from Newer Words: Lexical Borrowings in French	University of Minnesota	US
Cinkova, Silvie	Semantic annotation of a dialog corpus	Charles University in Prague	Czech Republic
Clachar, Arlene	The Effect of Computer-Mediated Communication on the Acquisition of Registers of Written Academic Discourse by Creole-Speaking Children	University of Miami	US
Columbus, Georgie	"Ah lovely stuff, eh?" On invariant tag meanings and usage across three varieties of English	University of Alberta	Canada

Conrad, Susan and Sarah Albers	A New Corpus of Student Academic Writing	Portland State University	US
Conway, Mike	Emotive Language and Disease Outbreak Reports	National Institute of Informatics	Japan
Cox, Christopher	Probabilistic tagging of a corpus of Mennonite Low German: A case study using QTag	University of Alberta	Canada
Csomay, Eniko and Viviana Cortes	'Positioning lexical bundles in university class sessions'	San Diego State University	US
Damascelli, Adriana Teresa	The use of hedging devices in American English. Identifying some trends in the FROWN corpus	Università degli Studi di Torino	Italy
Davies, Mark	The 360 million word BYU Corpus of American English (1990–2007)	Brigham Young University	US
Davis, Boyd and Charlene Pope	Into the woods: First steps in a collaborative corpus of medical conversations	University of North Carolina at Charlotte	US
de Haan, Ferdinand and Sheila Dooley	The role of constructions in the assignment of modal meanings	University of Arizona	US
Deane, Paul	Creating Subcorpora to Explore the Topic Structure of Domain-Specific Text	Center for Automated Scoring and Natural Language Processing	US
Degani, Marta	The Maori presence in New Zealand English: a lexical approach	University of Verona	Italy
Delicado-Cantero, Manuel	Corpus attestations and linguistic explanations: the case of the history of Spanish prepositional finite clauses	Ohio State University	US
Dilts, Philip	Good Nouns, Bad Nouns: What the corpus says and what native speakers think	University of Alberta	Canada
Diniz, Luciana	Suggestions and Recommendations in Academic Speech	Portland Community College	US
Dylewski, Radoslaw	Selected words, phrases, and meanings of African (American) provenience in General American: A corpus-based study.	Adam Mickiewicz University, Poznan	Poland
Enrique-Arias, Andrés and Carmago, Laura	<i>Biblia Medieval</i> : a parallel corpus of medieval Spanish	Universitat de les Illes Balears	Spain

AACL 2008

Fitzpatrick, Eileen and Joan Bachenko	Testing Language-Based Indicators of Deception on a Corpus of Legal Narratives	Montclair State University	US
Fletcher, William H.	Complementing the BNC with a Corpus from the Web	US Naval Academy	US
Freddi, Maria	Studying phraseology and translation through the corpus and the database	University of Pavia	Italy
Friginal, Eric	Grammatical Expression of Stance in Outsourced Call Center Discourse	Northern Arizona University	US
Gallegos Shibya, Alfonso	The diachronic development of some verbs with copulative function in Spanish	Universidad de Guadalajara	Mexico
Gardner, Dee	Semantic frequency and the creation of pedagogical word lists: What can we learn from SemCor?	Brigham Young University	US
Godev, Concepción	Word-Frequency and Vocabulary Acquisition: An Analysis of Elementary Spanish Textbooks	University of North Carolina at Charlotte	US
Goodall, Grant	Spoken Spanish in Corpora and in Textbooks: Implications for Acquisition	University of California, San Diego	US
Graham, Athelia	Semantic Frequency: a new look at word frequency counts	Brigham Young University	US
Gray, Bethany	Comparing stance in qualitative and quantitative research reports	Northern Arizona University	US
Gries, Stefan Th.	Measures of dispersion in corpus data: a critical review and a suggestion	University of California, Santa Barbara	US
Grieve, Jack	Corpus-based constituency tests and the structural position of auxiliary verbs	Northern Arizona University	US
Grieve-Smith, Angus B.	Controlling for Fads in Historical Corpora	University of New Mexico	US
Haertel, Robbie	MayanWiki: Facilitating Consensus Through an Openly Editable Corpus	Brigham Young University	US
Hanks, Patrick	A Corpus-Driven Pattern Dictionary for Mapping Meaning onto Use	Masaryk University	Czech Republic
Hansen, Cristina	Corpora of Spanish versus an educational text of astronomy	Universidad de La Laguna	Spain

Hong, Huaqing and Paul Doyle	Annotation, Indexing and Querying a Multilingual, Multimodal Classroom Discourse Corpus	Nanyang Technological University	Singapore
Horwinski Healy, Katherine	Creole African American Vernacular English: Origins of a Dialect	Louisiana State University	US
Huang, Jian and C. Lee Giles	An Efficient Framework for Large Scale Cross Document Coreference (CRC)	Pennsylvania State University	US
Hunston, Susan	You can't deny the fact that...: An Application of Corpus Linguistics	University of Birmingham	UK
Ishikawa, Shin	A Statistical Analysis of CEEJUS, Corpus of English Essays by Japanese University Students	Kobe University	Japan
Jenset, Gard and Christer Johansson	Estimating the saliency of constructions using document frequencies from the web	Bergen University	Norway
Juola, Patrick	Authorship Attribution : What Mixture-of-Experts Says We Don't Yet Know	Duquesne University	US
Kandil, Magdi	The Israeli-Palestinian Conflict in American, Arab, and British Media: Corpus-Based Critical Discourse Analysis	Georgia State University	US
Kerr, Betsy	Semantic Anglicisms in Contemporary Metropolitan French	University of Minnesota, Twin Cities	US
Lapshinova-Koltunski, Ekaterina	Semi-automatic Classification and Extraction of Predicates from German Text Corpora	University of Stuttgart	Germany
Lestrade, Sander	Finnish case alternating adpositions: A corpus study	Radboud University Nijmegen	Netherlands
Li, Jianguo and Kirk Baker, Chris Brew	A Corpus Study of Levin's Verb Classification	Ohio State University	US
Lindmark, Kerstin	A corpus-based investigation of cognate prepositions in English and Swedish	Stockholm University	Sweden
Lonsdale, Deryle	Developing a Corpus for a Morphologically Rich, Endangered Language	Brigham Young University	US
Lonsdale, Deryle and Yvon Le Bras	Compiling a new French frequency dictionary	Brigham Young University	US

AACL 2008

Lopes Moreira, Jose and Tony Berber Sardinha	The Reading Class Builder: A tool for creating corpus-based teaching materials	Pontifical Catholic University of Sao Paulo	Brazil
Lu, Hui-Chuan and Yun-Hui Chen, Chia-Chi Tien	Corpus Creation: CATE, CPEC & CAHC	National Cheng Kung University	Taiwan
McClanahan, Peter and Eric Ringger, Robbie Haertel, Kevin Seppi, George Busby, Deryle Lonsdale	Accelerating Corpus Annotation through Active Learning	Brigham Young University	US
McEnergy, Tony	Corpus Linguistics and the Humanities	Lancaster University	UK
Medina, Alfonso	Towards a Quantitative Characterization of Corpora at the Morphological Level: the use of Morphological Profiles to Measure Diachronic Change	Instituto de Ingeniería	Mexico
Meng, Ji	Statistical Modelling of Empirical Data in Corpus Stylistics	Imperial College London	UK
Miglio, Viola	Online Databases and Language Change: the Case of Spanish “dizque”	University of California, Santa Barbara	US
Minagawa, Jan	Developing writer stance in intermediate level Japanese university writing in academic and disciplinary courses	Temple University Japan	Japan
Mindt, Ilka	Colloquialization: An Alteration in Written English	Universität Würzburg	Germany
Mota, Cristina	Journalistic Corpus Similarity over Time	Instituto Superior Tecnico & New York University	Portugal
Nesi, Hilary	The design of a web-based interface for the BAWE corpus’	Coventry University	UK
Ocampo, Francisco	A diachronic process that gives birth to a Spanish discourse particle: The case of “claro”	University of Minnesota	US
O’Donnell, Matthew Brook and Catherine Smith, Robert Sanderson, Clare Llewellyn and John Harrison	Using an XML database for large corpora: Introducing Cheshire3	University of Liverpool	UK

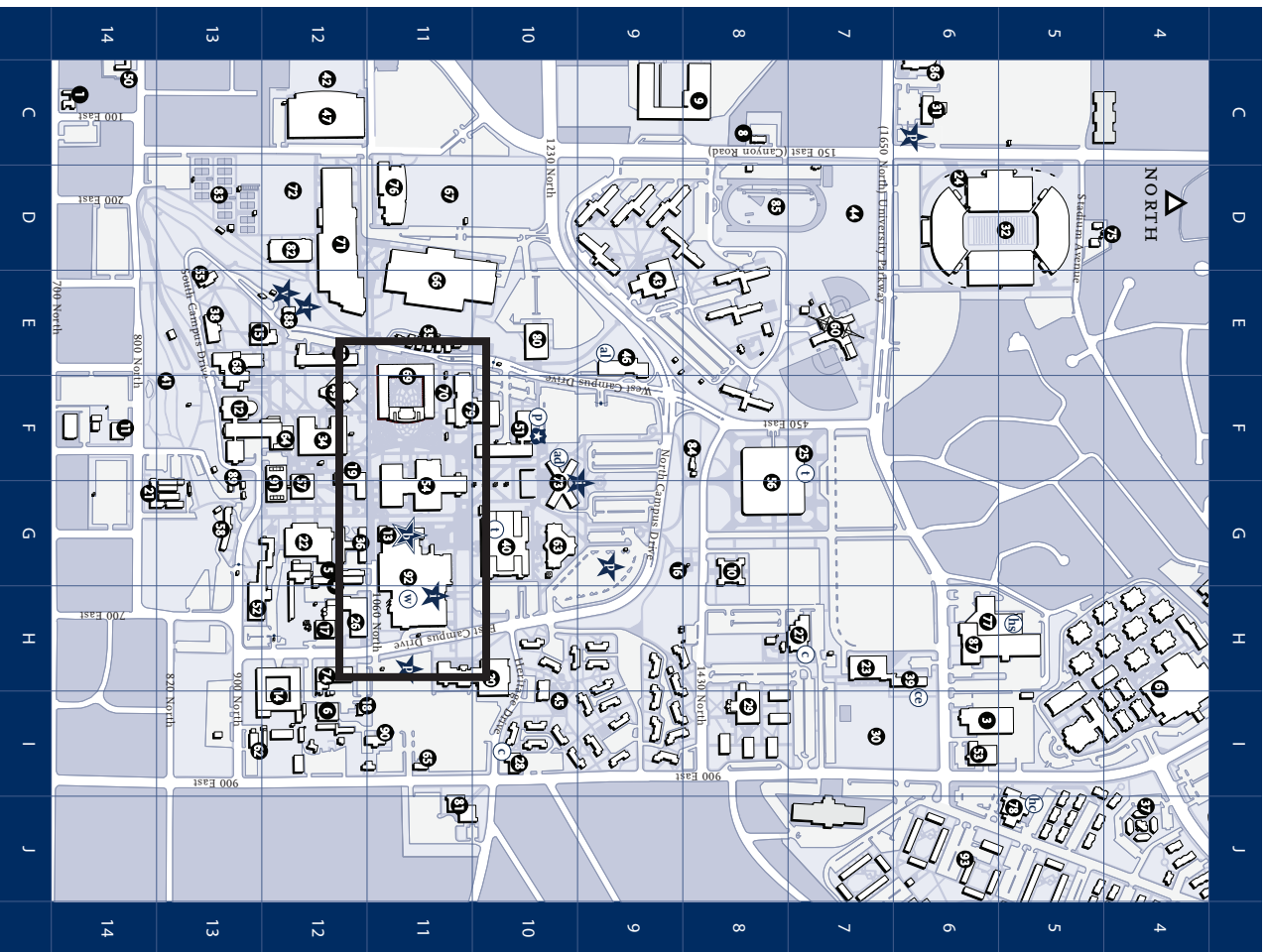
O'Donnell, Matthew Brook and Mike Scott, Michaela Mahlberg	Exploring Text-initial Concgrams in a Newspaper Corpus	University of Liverpool	UK
Parry, Donald	Verbal Imperative Variations in Qumran Legal Texts and Other Registers	Brigham Young University	US
Pastor-Gómez, Iria	On the development of Nouns as Internal Dependents in the Contemporary English Noun Phrase	University of Santiago de Compostela	Spain
Pearson, Pamela	A corpus-based study of mandative subjunctive triggers in published research articles	Georgia State University	US
Pho, Phuong Dzung	Linguistic realizations of rhetorical structure: A corpus-based study of research articles in applied linguistics and educational technology	Monash University	Australia
Poonpon, Kornwipa	The Use of Linking Adverbials in EFL Learners' Time-Constrained Spoken Monologue	Northern Arizona University	US
Qian, Yufang	Discursive construction of terrorism in <i>Peoples Daily</i> and <i>The Sun</i> before and after 9.11	Lancaster University	UK
Reppen, Randi	"Students must": A corpus based look at directives in university language	Northern Arizona University	US
Rice, Sally and John Newman	Beyond the lemma: Inflection-specific constructions in English	University of Alberta	Canada
Richardson, Joseph	A Corpus-Based Model for the Work of Editors	LDS Foundation	US
Riha, Helena and Kirk Baker	Tracking Sociohistorical Trends in the Use of Roman Letters in Chinese Newswires	Ohio State University	US
Ringger, Eric	Compiling and Annotating a Corpus for Syriac	Brigham Young University	US
Rogers, Chandrika	Articles in Registers of Indian English: A Corpus-based Study	Western Carolina University	US

Römer, Ute	A neo-Firthian approach to academic writing: Uncovering local patterns and local meanings in the discourse of linguistics	University of Michigan	US
Römer, Ute and Stefanie Wulff	Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive	University of Michigan	US
Rudanko, Juhani	Recent Change in Core Grammar: a Case Study Based on Corpus Evidence	University of Tampere	Finland
Sanchez Almela, Moises	Collocates for Word Sense Disambiguation by means of a Discriminant Function Analysis Model: A Corpus-Based Approach	Universidad de Murcia	Spain
Sanderson, Robert and Matthew Brook O'Donnell, Clare Llewellyn	Milk, bread and toothpaste: Adapting Data Mining techniques for the analysis of collocation at varying levels of discourse	University of Liverpool	UK
Shelley, Monte and James Rosenvall	An Introduction to WordCruncher 7.1	Brigham Young University	US
Silva, Rosário	Colouring COMPARA: contrastive and monolingual colour studies in English and Portuguese	Linguateca	Portugal
Silver, Marc and Sara Radighieri	The representation of time and space across historical discourse: a comparative analysis of evaluative effects	University of Modena and Reggio Emilia	Italy
Smith, Eric J.M.	Using a Query Language as an Annotation Tool	University of Toronto	Canada
Stvan, Laurel	<i>Fat and Health Literacy: Two Revealing Terms in CADOH (Corpus of American Discourses on Health)</i>	University of Texas at Arlington	US
Tao, Hongyin	I've never seen anything like it: An inquiry into the grammatical clustering involving never in spoken American English	University of California, Los Angeles	US
Teddiman, Laura	Comparative and Superlative Adjectives and Textual Frequency	University of Alberta	Canada
Teich, Elke	Characterizing genre: The case of scientific texts	Institut f. Sprach- und Literaturwissenschaft	Germany

Tono, Yukio	Towards a Multi-layered & Multimodal Annotation Model of Learner Corpora	Tokyo University of Foreign Studies	Japan
Tottie, Gunnel	Did the boys leave or not? Negation and quantifier scope: a corpus study	University of Zurich	Switzerland
Tseng Hsin-wen, Carrie and Shelley Ching-yu Hsieh	Stock Market Jargon Metaphors in General Discourses: A Corpus-based Study of Mandarin Chinese	National Cheng Kung University	Taiwan
Turley, Jeff	Reflexive Pronoun as Discourse Focus Marker: The Case of Spanish <i>morir</i> vs. <i>morirse</i>	Brigham Young University	US
van Bergen, Geertje and Peter de Swart	Scrambling in Spoken Dutch	Radboud University Nijmegen	Netherlands
Vogel, Karen and Gerald Delahunty	A Corpus Analysis of Dative Clitic Doubling in Spanish	Colorado State University	US
Wang, Xingfu and Eric Ringger, Guohui Liu and Shiping Liu	Analysis of Canonical Chinese Antonym Co-occurrence	Chongqing University and Brigham Young University	China / US
Warren, Martin	A corpus-driven study of phraseological variation	The Hong Kong Polytechnic University	Hong Kong
Wendorf, Arthur H.	<i>De Ahí, Por Consiguiente, Por Ende, Por Lo Tanto</i> and <i>Por Tanto</i> : A Distributional Diachronic and Synchronic Analysis	University of Georgia	US
Wu, Hongmei	Corpus consultation in drafting and revising: A case study of a biomedical ESL graduate student	University of Arizona	US
Wu, Pengcheng and Huaqing Hong	An Efficient Query Package for Richly Annotated Discourse Corpus	Nanyang Technological University	Singapore
Wulff, Stefanie	A multifactorial approach to that-deletion in English complement constructions	University of Michigan	US
Yokoyama, Shozo	A Cross-sectional Analysis of Lexical Bundles in Written Medical Discourse	University of Miyazaki	Japan
Zdorenko, Tatiana	Subject omission in Russian: A study of the Russian National corpus	University of Alberta	Canada

# BRIGHAM YOUNG UNIVERSITY

## CAMPUS FACILITIES

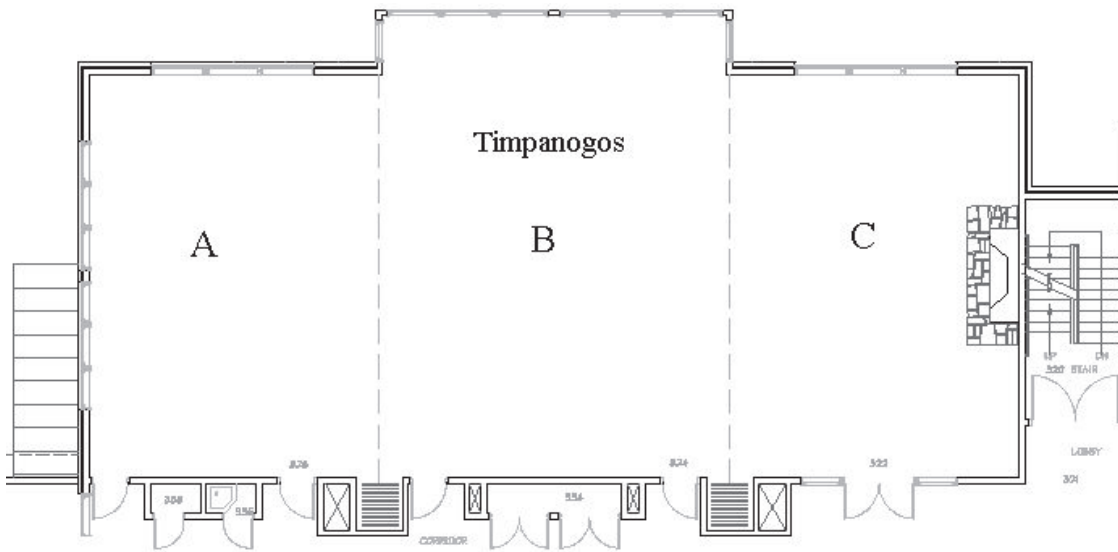


1	ALLN	Allen Hall (Museum of Peoples and Cultures)	C/14	48	INRA	Intramural Recreation Area (West Stadium)	A, B, 4, 5, 6
2	FARM	Animal Science Farm	C/1	49	SWKT	Kimball Tower, Spencer W. Kimball Hall, Ananda Knight Building, Kesse Knight Building	F/12, B/C/14, F/10
3	AXMB	Auxiliary Maintenance Building	W/56	50	AKH	Knights Hall	B/C/14
4		B-21 to B-32 (Service Buildings)	G/12	51	J48	Knights Building, Kesse Knight Building	F/10
5		B-34, B-38, B-41, B-51 (MMC Temporary Buildings)	G/12	52	KMB	Knights Barragon Building	G/H/12, 13
6	866	B-66 Classroom/Lab Building	W/12	53	AXLB	Laundry Building, Auxiliary Services	W/6
7	867	B-67 Service Building	C/2	54	HBL	Lee Library, Harold R. Lee Library	FG/10, 11
8	872	B-72 Building (LDS Philanthropies)	C/8	55	MS18	Maeser Building, Karl G. Maeser Building	E/13
9	877	B-77 Service Building (Former USJC Building)	C/8, 9	56	MC	Marriott Center, J. Willard Marriott Center	FG/7, 8
10	MLBM	Bean Life Science Museum, Monroe L. Benson Agriculture and Food Institute, Ezra Taft Benson Building	C/8, 9	57	MANB	Marrin Building, Thomas L. Marrin Building	FG/12
11	B-49	Benson Agriculture and Food Institute, Ezra Taft Benson Building	F/14	58	MB	McDonald Building, Howard S. Miller Park (Baseball/Softball Complex)	G/13
12	BRSN	Benson Building, Ezra Taft Benson Building	F/12, 13	59	MCPB	Miller Park (Baseball/Softball Complex)	E/12
13	WSC	Bookstore, BYU	G/11	60	MIP	Miller Park (Baseball/Softball Complex)	E/12
14	BRWB	Brewster Building, Sam F. Brimhall Building, George H. Brimhall Building	H/1, 12	61	MTC	Missionary Training Center	H/4, 5
15	BRMB	Brimhall Building, George H. Brimhall Building	E/12, 13	62	PPWV	Motor Pool Vehicle Shelter	I/12, 13
16	BELL	Centennial Carlton Tower	G/8	63	MOA	Museum of Art	G/10
17	PPCH	Central Heating and Cooling Plants	W/12	64	MCB	Nichols Building, Joseph K. Nichols Building	F/12
18	CMB	Chemicals Management Building	W/12	65	OLVH	Oliver House (Performing Arts Management)	I/11
19	HNCB	Clark Building, Heald R. Clark Building, J. Reuben (Law School) Clark Building, Benjamin (Plant Science Lab) Clark Building	FG/12, H/10, 11	66	RB	Richards Building, Stephen L. Richards Building	DE/10, 11
20	JRCB	Clark Building, J. Reuben (Law School) Clark Building	H/10, 11	67	RFB	Richards Building, Stephen L. Richards Building	DE/10, 11
21	CLFB	Clark Building, Benjamin (Plant Science Lab) Clark Building	FG/13, 14	68	JSB	Smith Building, Joseph F. Smith Building	E/13
22	CB	Cable Engineering Building, W. W. Clyde Conference Center, BYU	G/12	69	JFSB	Smith Building, Joseph F. Smith Building	E/11
23	CONF	Conference Center, BYU	H/7	70	JFSG	Smith Building Parking Garage, Joseph F. Smith Building	E/11
24		Cogner Room, Laykell Edwards Stadium	D/6	71	SFH	Smith Fieldhouse, George Albert Smith Fieldhouse	DE/12
25	CRB	Cable Technology Building, Roland A. Creaney (Dairy Products Laboratory) Creaney on North East	H/12	72	SFLD	Smith Fieldhouse, South Field (Varsity Soccer)	D/12
26	CRB	Cable Technology Building, Roland A. Creaney (Dairy Products Laboratory)	H/12	73	ASB	Smoot Administration Bldg, Abraham O. Smoot Administration Bldg	FG/10
27	DPL	Creaney (Dairy Products Laboratory)	H/7	74	SNLB	Snell Building, William H. Snell Building	H/12
28	CONE	Creaney on North East	I/10	75	STEH	Stadium East and West (STWH) Houses	D/5
29	DT	Deseret Towers and Morris Center (MORC) Desert Towers Recreation Area	H/8	76	SAB	Student Athlete Building	D/11
30	DIRA	Deseret Towers Recreation Area	I/6, 7	77	SA5B	Student Auxiliary Services Building	H/5, 6
31	ESM	Earth Science Museum	I/6, 7	78	SHC	Student Health Center	I/5
32	LVKS	Edwards Stadium, Laykell Edwards Stadium	C/6	79	TMCB	Talmage Math Sciences/Computer Building	F/10, 11
33	ELLB	Ellsworth Building, Leo B. Elsworth Building	B/1	80	TN8B	Tanner Building, N. Eldon Tanner Building	E/10
34	ESC	Eyring Science Center, Carl F. Eyring Faculty Office Building	F/12	81	TL8B	Taylor Building, John (Comprehensive Clinic) Taylor Building	J/11
35	F08	Faculty Office Building	E/11	82	TCB	Tennis Courts Building	D/12
36	FB	Fratcher Building, Harvey L. Fratcher Building	G/12	83	TCF	Tennis Courts, Outdoor	C/D/13
37	FLSR	Foreign Language Student Residence	J/4	84	TOHH	Thomas House, Risk Management and Safety	F/8
38	HGB	Grant Building, Heber J. Grant Building	E/13	85	TRAK	Track and Field Complex	D/7, 8
39	HCEB	Harman Bldg., Caroline Henaway (Cont. Ed.) Harman Bldg., Caroline Henaway (Cont. Ed.)	H/6	86	UPC	University Parkway Center	B/6, 6
40	H4C	Harris Fine Arts Center, Franklin S. Harris Alabertum and Botany Pond, Brandt F. Heavens Field	G/10, F/13, 14	87	UPB	University Press Building	H/5, 6
41		Harrison Alabertum and Botany Pond, Brandt F. Heavens Field	G/10, F/13, 14	88	WCR	Widest Center	E/12
42	HAWF	Heavens Field	B/C/12	89	WAH	Wahle House (MAXXIST) Walls Building, Daniel H. Walls Building	G/13
43	HL	Helaman Hills and Cannon Center (CANCC)	DE/F/8, 9	90	ROTC	Wahle House (MAXXIST) Walls Building, Daniel H. Walls Building	I/11
44	HLBA	Helaman Hills and Cannon Center (CANCC)	D/7	91	WDB	Widest Building, John A. Widest Building	FG/12
45	HR	Heritage Halls and Central Building (HRCN)	H/9, 10	92	WSC	Wilkinson Student Center, Ernest L. Wilkinson Student Center	G/H/11
46	HC	Hinckley Alumni and Visitors Center, Gordon B. Under Construction	E/9	93	WT	Wynton Terrace and Adm. Bldg.	JK/L/3, 4, 5, 6, 7
47	IPF	Indoor Practice Facility	C/12	94	WP	Wynton Terrace and Adm. Bldg., Viewview Park and Central Building	A, B, 2, 3, 4

A complete, interactive campus map is available online at <http://map.byu.edu>.

# Apsen Grove Conference Center

## Third Floor Timpanogos Room



## Second Floor Tree Rooms

