AACL 2009 Abstract Book

Table of Contents

Helpful Notes

This electronic abstract book contains internal links, connecting the program to the abstracts and the contact information of authors. Click on the title of a talk in the program, and it will take you to the abstract for that talk. Click on the title in an abstract, and it will take you to the relevant location in the program. Click on the name of an author in the abstract, and it will take you to that person's entry in the contact information pages. The abstract book also contains numerous bookmarks to help you navigate. The plenary abstracts are listed first in chronological order, followed by the rest of the abstracts in alphabetical order, by first author.

October 1, 2009

Dear AACL 2009 Participant

On behalf of the Organizing Committee, let me welcome you to AACL 2009! I am particularly happy that we have this opportunity to host the conference in Edmonton – the first time that AACL conferences have been held outside the USA. Naturally, I hope that holding the conference in Canada will provide new and fruitful opportunities for academic collaboration, especially international collaboration, among participants.

The Department of Linguistics at the University of Alberta has played a key role in hosting a number of international conferences in recent years: *The International Conference on the Mental Lexicon* (2002, 2008) and the *Conference on Conceptual Structure, Discourse and Language*, CSDL (2004). Hosting AACL 2009 is fully in keeping with the Department's tradition of promoting empirically oriented research. The Department is also home to the Canadian component of the International Corpus of English (ICE-CANADA), making it a natural candidate for hosting a corpus-oriented conference in Canada.

We have six plenary speakers for AACL 2009 and close to eighty regular presentations over the two and a half days of the conference. The papers cover the full range of topics that one expects at AACL conferences. The two pre-conference workshops, on XML and ELAN, were planned to provide more practical training in techniques of corpus construction – something that is fundamental to the field of corpus linguistics.

I hope you enjoy the conference and I hope you can explore more of Edmonton than the Conference Centre while you are visiting (before the weather becomes…er…cooler).

John Newman
Chair, Organizing Committee AACL 2009
&
Professor and Chair, Department of Linguistics
University of Alberta

American Association for Corpus Linguistics AACL 2009

Lister Conference Centre, University of Alberta, 87 Avenue & 116 Street
Edmonton, Alberta, Canada
October 8-11, 2009

Thursday Oct 8

8.30 - 10.00	**Registration**
	WORKSHOPS Maple Leaf Room
10:00 - 12:30	**Stefan Sinclair** *Practical XML for Linguists*
12:30 - 13:30	*catered lunch*
14:00 - 16:30	**Chris Cox** *Time-aligned transcription in ELAN*
17:30 - 20:30	**Registration & Welcome Reception** *Light refreshments and a cash bar serving beer and wine* Maple Leaf Room

AACL 2009 Organizing Committee:
John Newman (Chair), Harald Baayen, Elizabeth French, Georgie Columbus, Christopher Cox, Diane McKen, Sally Rice, Oliver Rossier
Abstracts Booklet: Lindsay Griener
Technician: Al Denington

AACL 2009 is supported by grants from:
Social Sciences and Humanities Research Council of Canada
Office of the VP (Research), University of Alberta
Faculty of Arts, University of Alberta
Department of Linguistics, University of Alberta

AACL 2009 Scientific Committee: Sally Rice and Harald Baayen (co-chairs)
Laurence Anthony, Michael Barlow, Douglas Biber, Laurel Brinton, Paul Deane, Stefan Dollinger, Gerard Van Herk, Susan Hunston, Kaori Kabata, Tyler Kendall, Ian Lancashire, Suzanne Kemmer, Greg Kondrak, Katja Markert, Yoshio Ono, Paul Rayson, Randi Reppen, Sali Tagliamonte, Hongyin Tao, Chris Westbury, Stefanie Wulff, Richard Xiao

The Organizing Committee is grateful to the following for their support of AACL 2009: John Benjamins, Mouton de Gruyter, Routledge, Department of Psychology (University of Alberta), Arts Resource Centre (University of Alberta)

The Graduate Student Networking Event is supported by a grant from the Graduate Students' Association, University of Alberta.
Sincere thanks to the University of Alberta graduate students who have assisted in AACL 2009 in numerous ways.

Friday Oct 9

DAY ONE
MORNING

8:00 - onward **Registration- available throughout the day**

8:30 - 9:00 **Opening Remarks**
John Newman
Maple Leaf Room

9:00 - 10:00 **PLENARY- Brian MacWhinney**
"TalkBank- Reintegrating the disciplines"
Maple Leaf Room, Chair: Rice

	Session 1	**Session 2**	**Session 3**
	Prairie Room	Maple Leaf Room	Glacier Room
	Register and Discourse	**Learner Corpora and**	**Semantics**
	Chair: Baayen	Chair: Kryuchkova	Chair: Shaoul
10:00-10:20	Jankowski: *Grammatical and register variation and change: A multi-corpora perspective on the English genitive*	Barlow & Calude: *Individual differences in the use of cleft constructions in speech*	
10:20 - 10:40	Pezik: *On the uniqueness of casual spoken discourse – insights from the National Corpus of Polish*	Römer & O'Donnell: *Positional variation of phrase-frames in a new corpus of proficient student writing*	Taboada: *Comparable corpora for cross-linguistic sentiment analysis*
10:40-11:10		*coffee break*	

	Session 1	**Session 2**	**Session 3**
	Chair: Baayen	Chair: Barlow	Chair: Shaoul
11:10-11:30	Geeraert: "I haven't drank in weeks"*: The use of past tense forms as past participles in English corpora*	Johansson & Geisler: *Syntactic aspects of Learner English*	
11:30- 11:50	Columbus: *Irish like as an invariant tag: evidence from ICE-Ireland*	Römer: *Phraseological items in apprentice academic writing: Does nativeness matter?*	Akiyama: *The constructional meaning of Infinitival Relative Clauses in English: A corpus-based Approach*
11:50- 12:10	Gales: "Get out before we get you!": *A corpus-based analysis of stance in threatening communications*	Tono, Nonura, Murakami, Kaneta, & Mochizuki: *Unsupervised learning of criterial features of L2 acquisition stages using parallel learner corpora*	
12:10 - 13:30		*lunch*	
Friday Afternoon	**Session 1**	**Session 2**	**Session 3**
	Prairie Room	Maple Leaf Room	Glacier Room
	Morphosyntax	**Learner Corpora**	**Semiotics and Registers**
	Chair: Teddiman	Chair: Dilts	Chair: Wulff
13:30- 13:50	Miglio & Gries: *Narrative function(s) of tense switching: a corpus-based application to medieval Icelandic sagas*	Neary-Sundquist: *The role of task type in a learner corpus*	Bednarek: *Emotional practices and character identity in American popular culture*
13:50 - 14:10	Gajdos: *Rethinking the German three-way system of spatial demonstrative adverbs: evidence from electronic corpora*	Friginal, Baker, & Pearson: *Linguistic Characteristics of Non-Native Speaker Writing in English: A Corpus-Based Analysis*	Kuo: *The representation of the elderly in Taiwanese newspapers: A corpus-based study*
14:10- 14:30	Tomišić: *A Corpus-based analysis of Slovene Reflexive Verbs (Verbs with SE)*		Caple: *Using corpus linguistics to explore the 'reading' of multi-semiotic play*

14:30 - 14:50		*coffee break*	
	Session 1	**Session 2**	**Session 3**
	Chair: Rice	Chair: Tono	Chair: Bednarek
14:50- 15:10	Snoek: *Variation in –im suffix usage in a Tok Pisin corpus*	Sundquist & Neary-Sundquist: *A corpus-based evaluation of vocabulary in the second language classroom*	Aull: *How textbook genres make readers, disciplines, nations: A qualitative and quantitative corpus approach*
15:10 - 15:30	Säily: *Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations*	Grant: *Developing and Analysing the Engineering Lecture Corpus (ELC)*	Wulff: *A multifactorial analysis of (un)attended demonstratives in academic writing*
15:30- 15:50	Ju: *A Corpus-based Study of the Korean Quantifiers Cokum and Com*	Iberri-Shea: *University Student Public Speech: A Corpus-Based Study of Student Produced Language*	
15:50- 16:10		*coffee break*	
	Chair: Teddiman	Chair: Grant	Chair: Newman
16:10-16:30	Caldwell: *A Corpus Analysis of Japanese Mimetics*	Huang: *The Acquisition of Grammatical Knowledge and Usage Using a Corpus-Aided Discovery Approach*	
16:30-16:50	Takeda: *Japanese adjective 'sugoi' and adverb 'sugoku' in conversations*	Lee: *Like the experts? A multi-corpus approach to the features of Chinese research writing in English*	Becher: *New applications of translation corpora: Investigating language contact and language change*
16:50 - 17:10	Unser-Schutz: *Developing a Text-Based Corpus of the Language of Japanese Comics (Manga)*		Miglio: *A corpus analysis of Scandinavian legal concepts in Anglo-Saxon Laws*
17:15-18:15		**PLENARY- Shana Poplack** "Corpora as tools for the study of linguistic change" Maple Leaf Room, Chair: Rice	
18:30- 22:00		**Conference Dinner** Faculty Club	
18:30		Cocktails- Cash Bar	
19:30		Buffet Dinner	

DAY TWO

Saturday , Oct 10 **MORNING**

9:00-10:00	**PLENARY- Mark Liberman** "The Journal of Experimental Linguistics" Maple Leaf Room, Chair: Baayen		
	Session 1 Prairie Room **English Morphosyntax**	**Session 2** Aurora Room **Corpora and Linguistic Theories**	**Session 3** Glacier Room **New Corpora**
	Chair: Brinton	Chair: Römer	Chair: Davies
10:00-10:20	Kendall, Van Herk & Bresnan: *The Dative Alternation in African American English: Researching Syntactic Variation and Change in a Conglomerated Sociolinguistic Corpus*	Meyer: *Apposition from the Perspective of Construction Grammar*	Raso & Mello: *The C-ORAL-BRASIL corpus*
10:20 - 10:40	Teddiman: *Subject Ellipsis by Text Type: An Investigation using ICE-GB*	Bergh: *Complex extraction in English*	Berber Sardinha: *The Brazilian Corpus*
10:40-11:00	Stvan: *Where Are They Bare? The Frequency and Distribution of Bare Nouns in American English*	Duffley: *What grammaticality judgements and decontextualized examples indicate: reflections in the light of corpus data*	
11:00-11:30		*coffee break*	
	Chair: Van Herk		Chair: Cox
11:30- 11:50	Tagliamonte & Waters: *A tale of two cities: Comparing sociolinguistic patterns in England and Canada*	Barlow: *Demonstration of Collocates Software*	Ussishkin, Francom & Woudstra: *Electronic corpora for two Semitic languages*
11:50- 12:10	Brinton: *The development of <that said>*	Barlow: *Demonstration of Collocates Software*	Scheffler: *"VOICE Awards": A German Human-Machine Dialog Corpus*
12:10- 13:30		*lunch*	

Saturday Afternoon	**Session 1**	**Session 2**	**Session 3**
	Prairie Room	Aurora Room	Glacier Room
	English Morphosyntax	**Psycholinguistics, Speech and Register**	**Corpus Tools**
	Chair: Meyer	Chair: Gries	Chair: Liberman
13:30- 13:50	Chartrand,Kunichika & Takeuchi: Analysis of Modal Auxiliaries in Two Consecutive Phrases Extracted from the British National Corpus	Dilts, Libben & Baayen: *A Corpus Analysis of Frequency Effects on Eye-Movements in Sentence Context*	Alarcón & Sierra: *CORCODE: A Corpus of Definitional Contexts as a Lexicography Resource*
13:50 - 14:10	Kemmer & Barlow: A contrastive analysis of causal expressions	Tucker & Tremblay: *Frequency and multi-word sequences: A psycholinguistic comparison of two corpora*	Kendall: *The Value of Relational Databases for Time-Aligned Annotation*
14:10- 14:30	Gotscharek, Neumann, Reffle, Ringlstetter, & Schulz: *Constructing a lexicon from a historical corpus*	Shaoul, Westbury & Baayen: *Agreeing with Google: We are Sensitive to the Relative Corpus Frequency of Phrases*	O'Donnell: *The Adjusted Frequency List: Evaluating a method to produce cluster-sensitive frequency counts*
14:30- 14:50		*coffee break*	

	Session 1 Chair: Duffley	**Session 2** Chair: Tucker	**Session 3** Chair: Columbus
14:50- 15:10	Hsieh & Chung: *“Do” and “Make”: A Corpus-based Study*	Vaughn, Pierrehumbert & Rohde: *Using character n-grams to classify native language in a non-native English corpus of transcribed speech*	Nesi, Ahmad & Ibrahim: *Pragmatic annotation in an international corpus of engineering lectures*
15:10 - 15:30	Aberra: *English light predicate construction with an indefinite deverbal complement*	Weinberger & Kunath: *Towards a Typology of English Accents*	Anthony, Chujo & Oghigian: *A novel, web-based, parallel concordancer for use in the ESL/EFL classroom*
15:30- 15:50	Rodríguez-Puente: *“And new meanings turned up:” The development of new meanings of English phrasal verbs with up: Evidence from the Helsinki Corpus and ARCHER1*	Albers: *A corpus-based study of how doctors construct diagnoses of osteopenia or osteoporosis with their patients*	Li & Fang: *Age Tagging and Word Frequency for Learner’s Dictionaries*
15:50-16:10	Liu: *Is it a* chief, main, major, primary, or principal *concern? A corpus-based behavioral profile study of the near-synonyms and its implications*	Bruce, Friginal, Pearson & Pickering: *Developing a highly specialized corpus of spoken English: AAC discourse in the workplace*	Gardner & Davies: *A frequency dictionary of contemporary American English*

16:10-17:00 *coffee break*

17:00-18:00 **PLENARY- Mark Davies**

"Creating the first reliable monitor corpus of English: problems, solutions, and insights"
Maple Leaf Room, Chair: Newman

18:30-20:30 **GRADUATE STUDENT NETWORKING EVENT**

Chair: Snoek

This Graduate Student Networking Event is open to all (and only!) graduate students and post-docs at AACL 2009. The event includes a talk by **Dekang Lin - from Google** (*Unsupervised Acquisition of Lexical Knowledge from N-Grams*), to be followed by a Q and A session on being a computational/corpus linguist outside of academia. Finally, there will be a chance to network with others in your field over refreshments and drinks. Graduate students who are not registered at AACL will need to pay a $10 door charge. We hope you will join us. – the Linguistics Graduate Students’ Association.

Alumni Lounge, 2-900 Students' Union Building (SUB)

Sunday, Oct 11 **DAY THREE**

9:00-10:00 **PLENARY- Stefan Th. Gries**
"Corpus linguistics and theoretical linguistics: A love-hate relationship?"
Maple Leaf Room, Chair: Newman

	Session 1 Prairie Room **Connectives**	**Session 2** Aurora Room **Unsupervised Learning, Part-of-Speech Tagging**	**Session 3** Glacier Room **Corpus Development**
	Chair: Waters	Chair: Vaughn	Chair: Kendall
10:00- 10:20	Soyeon Kim: *Emergent Patterns of Conjunctive Adverbial "Though" in Academic Spoken English: A Corpus-based Study*	Fung: *Towards the unsupervised discovery of syntactic categories for typologically-varied languages*	Van Herk, Childs & Thorburn: *Safe Harbour: Ethics and accessibility in sociolinguistic corpus building*
10:20 -10:40	Jia: *A Corpus-based study of the connectives 'danshi' 'keshi' and 'ran'er' in Mandarin Chinese*	Sun: *Part-of-speech tagging for a Southern Min Corpus*	Lonsdale: *Beyond the dictionary: creating a long-term corpus resource*
10:40-11:00	Sangbok Kim: *Interclausal Semantic Functions of the –E Connectives in Korean*	Barreda: *The use of Trigrams in Classification of texts based on Authorship*	Cox: *Corpus linguistics and language documentation: Challenges for collaboration*

11:00-11:30 *coffee break*

11:30-12:30 **PLENARY - Geoffrey Rockwell and Alexandre Sevigny**
"The extraordinary effectiveness of words"
Maple Leaf Room, Chair: Newman

12:30-12:40 **CLOSING REMARKS**

Workshops

Practical XML for Linguists: This workshop will provide a brief overview of XML and focus on practical applications for linguistics, including the encoding, searching and transformation of data. Basic familiarity with XML is recommended.

Time-aligned transcription in ELAN: This workshop will introduce participants to the use of ELAN, a tool developed by the Max Planck Institute (Nijmegen) for the production of XML-based, time-aligned transcripts of audio and video resources. No previous experience with ELAN or related software is required.

TalkBank – Reintegrating the Disciplines

Brian MacWhinney
Carnegie Mellon University

Human communication is a single, complex, unified process. However, specific aspects of this process are studied in detail by 20 largely separate academic traditions, ranging from information theory and phonetics to conversation analysis and psycholinguistics. Although this fractionated approach has led to many advances, it has failed to illuminate links between the disciplines. To achieve this, the TalkBank Project has constructed the world's largest available database of spoken language interactions, all contributed by individual scientific work groups from across the globe. Much of this database is now available in a form that links transcripts directly to audio and video media for playback on the desktop or over the web. TalkBank communities are now particularly active in the areas of child language development, aphasia, conversation analysis, legal discourse, gesture-language linkages, classroom discourse, and phonological analysis. For all of these groups, TalkBank is providing increasingly powerful interoperable tools for analyzing phonetics (Praat, Phon), gesture (ELAN), and morphosyntax (GRASP), and for supporting collaborative commentary on target corpora across the web. Using these tools, researchers will eventually be able to study a set of richly annotated multimedia transcriptions from a wide variety of analytic perspectives.

Corpora as tools for the study of linguistic change

Shana Poplack
University of Ottawa

Spoken language is characterized by inherent variability and the competing variants, particularly when salient and stigmatized, are often construed as signs of change. Establishing whether a feature in fact represents an innovation requires a real-time benchmark predating the current situation. Written texts, the standard fare of historical linguists, are of limited use, since the vernacular features of interest are either grossly underrepresented or not attested at all. The most appropriate benchmark is an oral precursor, but such precursors necessarily have a time depth too shallow to chart the full course of change.

In this paper I describe the construction of a series of diachronic corpora, each of which represents at least one aspect of the speech of earlier times, and the novel uses we make of them in our efforts to trace the locus and trajectory of change over long periods. Each has complementary sources of error, but interpreted in conjunction with synchronic age distributions, they are capable of distinguishing stable variation from change with a high degree of accuracy. Results converge in showing, for a series of morphosyntactic variables, that change tends to be very slow, infrequent and highly circumscribed.

The Journal of Experimental Linguistics

Mark Liberman
University of Pennsylvania

The Journal of Experimental Linguistics is a new Open Access co-journal now being developed within the Linguistic Society of America's eLanguage initiative. It will be a linguistic "journal of reproducible results", that is, a journal of reproducible computational experiments on topics related to speech and language. These experiments may involve the analysis of previously published corpora, or of experiment-specific data published for the occasion. Other articles will include computational simulations, implementations of diagnostic techniques or task scoring methods, methodological tutorials, and documentation of relevant software.

In all cases, articles will be accompanied by executable recipes for recreating all figures, tables, numbers and other results. These recipes will be in the form of source code that runs in some generally available computational environment. Although JEL will be centered in the discipline of linguistics, we aim to publish research from the widest possible range of disciplines that engage speech and language experimentally, from electrical engineering and computer science to education, psychology, biology, and speech pathology.

In this interdisciplinary context, "reproducible research" is especially useful in helping experimental and analytical techniques to cross over from one subfield to another. Publication will be in online digital form only, with articles appearing as they complete the review process. There will be a rigorous but rapid process of peer review, supplemented by a vigorously promoted system for adding refereed remarks and replies after publication.

The publication of "executable articles" obviously raises some new conceptual, editorial and technical issues, which are also on the agenda for our colleagues in other disciplines. This talk will discuss our initial plans for dealing with these problems.

Creating the first reliable monitor corpus of English: problems, solutions, and insights

Mark Davies
Brigham Young University

Monitor corpora allow researchers to look at change in real time, by using data from continually-expanding corpora. It has been claimed that there are large, reliable monitor corpora for English, but we will argue that this is not the case. We will also show that -- in spite of their initial attraction -- text archives like online newspapers and magazines (or even the Web, via Google) won't work as monitor corpora either.

We will show how the Corpus of Contemporary American English (400+ million words, 1990-2009) is a very useful monitor corpus, and in fact the first and only really usable monitor corpus of English. This is due in large part to its design -- 20 million words each year, divided exactly the same each year (20% spoken, 20% fiction, 20% popular magazine, 20% newspaper, and 20% academic journals). We will provide a number of examples of how data from the corpus can be used to answer a number of interesting questions about ongoing change and variation in American English -- including morphological change, syntactic change, semantic change, and lexical change.

Corpus linguistics and theoretical linguistics: a love-hate relationship?

Stefan Th. Gries
University of California, Santa Barbara

Most accounts of the development of corpus linguistics categorize the field into different stages on the basis of the kinds of corpus-linguistic resources that were available. A different perspective looks at the field in terms of its relationship to what's going on in theoretical linguistics, and the picture that seems to arise from that is that of a love-hate relationship. In this talk, I am going to discuss:

- reasons why corpus linguists as a whole can only benefit from embracing and exploring theoretical aspects more than many more traditional corpus linguists have done so far;
- suggestions as to which linguistic theory, or more broadly framework, appears to be the most promising point of connection;
- commonalities between corpus linguistics on the one hand and different newer developments in a particular linguistic framework on the other hand ;
- methodological developments which not only render corpus-linguistic work more appropriate, but also allows to connect it more intimately to particular theoretical work.

The extraordinary effectiveness of words

Geoffrey Rockwell[1] and Alexandre Sevigny[2]
[1]University of Alberta, [2]McMaster University

Words are at the centre of grammar, culture and information. They are the building blocks of many theoretical paradigms, whether it is how critical-cultural theory examines the construction of power relations in society, how certain linguists understand grammatical structures in the brain, how communications studies content or how the digital humanist analyzes texts in the tradition of concording.

In this talk, we will discuss the power and effectiveness of words from two perspectives. We will survey how words are theorized and used in humanities computing and how communication studies and applied linguistics uses words in content analysis. We will highlight the differences between these approaches and propose a few thoughts on how the study of communication and literature can benefit through a shared examination of the effectiveness of words.

English light predicate construction with an indefinite deverbal complement

Daniel Aberra
University of Alberta

This paper presents a constructional analysis of the English light predicate construction using the BNC, and OED corpora. From the partially schematic structure of the construction i.e., [V_{LIGHT} a $N_{DEVERBAL}$], it is observed that the construction has three components. The light verb is mostly instantiated by verbs such as *give*, *take*, *make*, *get*, and *have* followed by the indefinite article 'a' and the open class of deverbal nouns. The British National Corpus examples illustrate this construction.

> ...***give a vote*** *of confidence to ...*
> ...***get a climb****, or go down for a landing...*
> ...***make a go*** *of it, will not be discouraged...*
> ...***take a look*** *around you...*
> ...***have a go*** *at doing the part of Sharon sometime...*

Based on the synchronic properties of the construction drawn from the BNC, the study will discuss about the conventionality of the unit, and based on the diachronic OED data it will answer the question when the construction emerged as a unit.

Some of the BNC-based observations include but not limited to preference for *infinitive*, and second person singular subject by the construction. From the open class of deverbal nouns *look*, *go*, *drink*, *profit*, *move*, *point*, *chat*, *copy*, *start*, *talk*, and *note* occur frequently in the construction, and from the top 10 deverbals only *drink*, *lead*, *look*, *move*, *name*, *talk*, and *view* occur with more than one light verbs. Finally, the English light predicate construction in its present form and function is traced back to the 14th century where it emerged with all three elements intact according to the OED record.

References:

BNC. http://corpus.byu.edu/bnc/x.asp.
OED.http://corpus.byu.edu/oed/x.asp.

The Constructional Meaning of Infinitival Relative Clauses in English: A Corpus-based Approach

Takanobu Akiyama
Nihon University

Infinitival relative clauses (henceforth IRCs) can be simply defined as *to*-infinitive clauses which modify preceding nouns instead of being their complement (e.g. *I have something to eat* [IRC]). It is reasonable to hypothesize, however, that there are some restrictions on the use of IRCs, i.e. particular syntactic and semantic circumstantial patterns which IRCs enter into. Such restricted environments would help us discern the constructional meaning of IRCs. To a great extent, the constraints on the occurrence of IRCs can be described in terms of certain grammatical and lexico-semantic 'triggers' preceding the NP containing the IRC. The aim of this paper is to investigate the circumstances in which IRCs occur and to explore the constructional meaning of this grammatical structure, on the basis of the data extracted from the British National Corpus. In particular, this paper will focus on the following four objectives: a) to cast light upon the nature of main verbs which co-occur with IRCs, b) to clarify the nature of the modifiers which precede the antecedent nouns of IRCs, c) to consider a general constructional meaning or schema for IRCs, and d) to investigate what sorts of *to*-infinitive clauses can be used as IRCs.

The organization of this paper is as follows. Section 1 gives a brief survey of previous analyses of IRCs, in particular, by Chomsky and Lasnik (1977) and Suzuki and Yasui (1994). Here the validity of introspective analyses of the target construction will be questioned, and the necessity of empirical corpus-based studies will be suggested. In section 2, the main verbs that take IRCs are scrutinized to detect the first trigger of the construction. Our corpus survey will clarify that five types of higher clauses precede an IRC. Section 3 focuses on modifiers operating on antecedent nouns of IRCs. Based on the corpus evidence, it is found that the modifiers tend to express evaluative adjectival meaning (e.g. *good*, *best*, *right*, etc.). Section 4 draws upon the results of the earlier sections and explores the constructional meaning of the IRC in English, attempting to develop a schema for this construction.

Kerswill, P. (2002c) Models of linguistic change and diffusion: new evidence from dialect levelling in British English, *Reading Working Papers in Linguistics 6. 187*-216.

Kerswill, P (2006a). Migration and language, In Klaus Mattheier, Ulrich Ammon & Peter Trudgill (eds.) *Sociolinguistics/Soziolinguistik. An international handbook of the science of language andsociety*, 2nd edition, Vol 3. Berlin: De Gruyter.

Kerswill, P (2006b), RP, Standard English and the standard/non-standard *relationship*, in David Britain (ed.) Language in the British Isles. Cambridge: Cambridge University.

Kerswill, P, & Williams, A. (2000), Creating a new town koiné: children and language change in Milton Keynes. *Language in Society* 29: 65-115.

Labov, W. (1966), The social stratification of English in new York city. Washington: Centre for Applied Linguistics.

Labov, W. (1989), Exact description of the speech community: Short a in Philadelphia, in R. Fasold and D. Schiffrin (eds.) (1989), *Language Change and Variation*, Amsterdam/Philadelphia: John Benjamins, p.1-57.

Trudgill, P. (1974), *The Social Differentiation of English in Norwich*, Cambridge: Cambridge University Press.

Trudgill, P. (1981), Linguistic accommodation: sociolinguistic observations on a sociopsychological theory, in *Papers from the Parasession on Language and Behavior*, May 1-2, 1981, pp. 218-237, Chicago: Chicago Linguistic Society.

Trudgill, P. (1986) *Dialects in contact*, Oxford: Basil Blackwell Ltd.

Trudgill, P. (1999) The dialects of England , Oxford: Blackwell.

Trudgill, P. (2003) A Glossary of Sociolinguistics, Oxford University Press.

Trudgill, P & Chambers, J (1980), *Dialectology***,** Cambridge; New York: Cambridge University Press.

CORCODE: A Corpus of Definitional Contexts as a Lexicography Resource

Rodrigo Alarcón and Gerardo Sierra
Universidad Nacional Autónoma de México

The development of computational tools to help on lexicography and terminography tasks is indeed a growing interest in the NLP field. Some tools have been developed for the extraction of terms from corpora. In addition, some studies are focusing on developing methods for acquiring definitional knowledge about terms. This last task is a complex process that requires a deep study on how the terms are commonly defined in texts.

In this sense, we present a corpus of *definitional contexts* (DCs), i.e., textual fragments where an author defines a term. Such a corpus is a repository that contains many different ways a term is defined and represents a useful tool for lexicography, terminography and information extraction work.

From a lexicography and terminography point of view, this kind of corpus will be an aid for the building of semasiological and onomasiological dictionaries, both on general and specialised language, since it will provide definitions which are the basic elements for a *lexical knowledge base* (LKB) and a primary source of linguistic knowledge to be considered. DCs corpora are an important resource for attaining the high degree of completeness required in a LKB.

In the information extraction field, it could be a starting point for developing extraction algorithms for a rule-based system. This means that it is helpful for analysing the structure of terms, definitions and characteristic elements that constitute DCs, and will help to formulate rules for extracting them from annotated corpora.

This paper describes a corpus of definitional contexts, from the definition of the term to the description of the applications of the corpus in lexicography. The intended applications for the work, and how these affected the corpus design, are described. In addition, the methodology for corpus building and the corpus structure are outlined, and the applications of the work are presented.

A corpus-based study of how doctors construct diagnoses of osteopenia or osteoporosis with their patients

Sarah Albers
Verilogue, Inc.

Using a corpus of 596 doctor-patient conversations (968,000 words), this study explores the language of diagnosis and treatment recommendations associated with osteopenia and osteoporosis. Osteopenia and osteoporosis are medical conditions that refer to the erosion of bone density, estimated to affect approximately 44 million Americans, most of whom are post-menopausal women. Osteopenia is the less severe form of, or precursor to, osteoporosis. Either of these conditions places the patient at greater risk for breaking their bones if they fall or even while doing everyday activities. Fractures of the spine or hip can lead to loss of height, stooping posture, pain, or mortality (Cline, et al., 2005).

Bone loss is detected with an X-ray bone density test, known as the DEXA scan. In 1994, the World Health Organization put forth guidelines about which DEXA scores should correspond with a diagnosis of osteoporosis. However, these guidelines are not without controversy, as some skeptics claim the diagnostic boundaries were arbitrarily defined and covertly influenced by pharmaceutical companies as a way to drive prescriptions of anti-osteoporosis medications to otherwise health individuals (Alonso-Coello, et al., 2008).

It is true that in the last few years there has been a considerable amount of marketing and development of new osteoporosis treatments, not to mention tactical pushes by some pharmaceutical companies to start treatment for patients at the osteopenia stage. However, in consultation with patients, physicians display a range of attitudes regarding the urgency to treat bone loss at its different stages (Griffiths, Green, & Tsouroufli, 2005).

This presentation will focus on the collocational profiles, frequent lexical bundles, and metaphorical/analogical framing associated with osteopenia and osteoporosis (Stefanowitsch, 2005; Biber, Conrad, & Cortes, 2004; Skelton, Wearn, & Hobbs, 2002) and will show how the naturally-occurring discourse between doctors and patients plays into both sides of the debate about when to treat bone loss.

References:

Alonso-Coello, P., Lopez Garcia Franco, A., Guyatt, G., & Moynihan, R. (2008). Drugs for pre-osteoporosis: Prevention or disease-mongering? *British Medical Journal* 336, pp. 126-129

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): pp. 371-405

Cline, R., Farley, J., Hansen, R., Schommer, J. (2005). Osteoporosis beliefs and antiresorptive medication use. *Maturitas* 50: pp. 196-208.

Griffiths, F., Green, E., & Tsouroufli, M. (2005). The nature of medical evidence and its inherent uncertainty for the clinical consultation: Qualitative study. *British Medical Journal Online* doi:10.1136/bmj.38336.482720.8F, pp. 1-7

Skelton, J., Wearn, A., Hobbs, F. (2002). A concordance-based study of metaphoric expressions used by general practitioners and patients in consultation. *British Journal of General Practice* 52: pp.114-118

Stefanowitsch, A. (2005). The function of metaphor: Developing a corpus-based perspective. *International Journal of Corpus Linguistics* 10(2): pp. 161-198

A novel, web-based, parallel concordancer for use in the ESL/EFL classroom

Laurence Anthony[1], Kiyomi Chujo[2], and Kathryn Oghigian[1]

[1]Waseda University, [2]Nihon University, Japan

In this paper, we describe a novel, web-based, parallel concordancer designed specifically for use in the classroom by second and foreign language learners of English. Currently, the use of parallel concordancers in the classroom has been a largely unexplored area, particularly with beginner level students. As a result, there are few guidelines on the design of these tools or the functions they should offer. In addition, the number of available parallel concordance tools is surprisingly few, and we know of no web-based tools offering this functionality. In the preparation of a revised English course for Japanese university students of English, we have created a new web-based, parallel concordancer that is built on a standard LAMP (Linux, Apache, MySQL, PHP) framework. The concordancer is both powerful and intuitive for teachers and learners to use. In addition, the concordancer is designed on a similar architecture to the Google search engine, allowing it to work comfortably on very large corpora of hundreds of millions of words. To enable the smooth processing of both English and Japanese texts, the concordancer is built to Unicode standards, and its internal token definition settings also employ Unicode character classes meaning that no cumbersome user-defined settings are necessary. Preliminary results show that the new software is considerably easier to use than standard desktop parallel concordance programs, and because it is web-based, it can be accessed out of class to allow more time for hypothesis-verification and production activities. Course-wide introduction of the software is planned for the fall of the 2009 academic year.

How textbook genres make readers, disciplines, nations: A qualitative and quantitative corpus approach

Laura Aull
University of Michigan

Biber, Connor, and Upton have recently called for simultaneously qualitative and quantitative approaches to large bodies of texts, a call that resonates with Vijay Bhatia's notion of both "text-external" and "text external" approaches to genres.

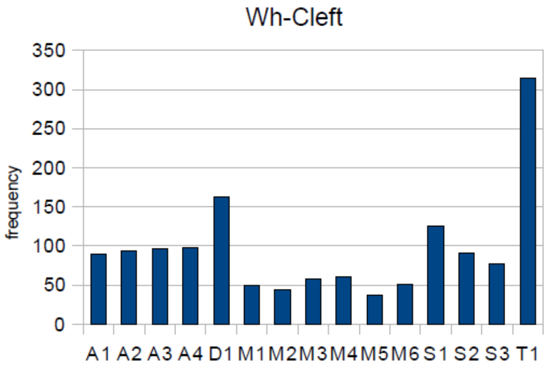
In this paper, I take a text internal and external approach to university humanities textbooks, which have not received the same attention as textbooks in the sciences (e.g. Kuhn, Myers). I specifically examine the editorial apparatus of American literature anthologies, textbooks that have dictated US canon and cultural representation over the last 40 years. To do so, I have created a corpus of what I view as two underexamined genres in the construction of "American literature": the prefaces and period overviews to the two leading anthologies in the US. These two genres – the preface, which narrates the position and authority of the editors and the anthology, and the historical period overview, which tells the story of particular national periods – influence discipline, nation, culture, and knowledge formation. Despite extensive recent scholarship on "American canon" construction and the dynamic nature of textual genres (namely, that texts enable and constrain particular social and rhetorical actions [Swales; Miller; Bawarshi]), these editorial/pedagogical texts, whose power derives from their relative invisibility, remain unchallenged. Classrooms and scholarship repeatedly cast them as "apolitical" texts, though they narrate the work(s) of cultural events, social groups, writers and (student) readers, and how those have changed over time.

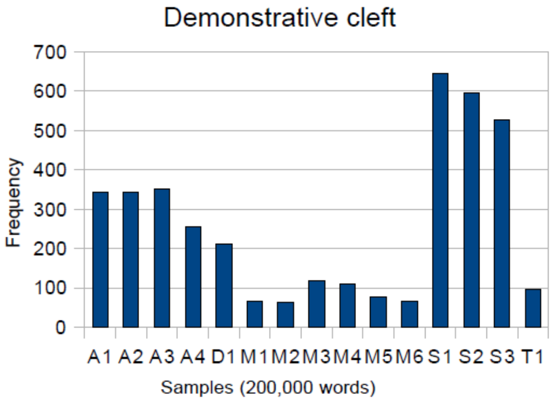
A recursive analysis that draws on historical detail and Antconc results reveals language patterns, such as gender imbalance and an Anglo-American focus, that contradict the "multicultural" claims of anthology editors. This is a timely examination with necessarily multi-disciplinary lenses and methods, and one which begs deeply epistemological questions for pedagogy and the field of American literature, such as whether supposedly "multicultural" textbooks are at all efficacious pedagogical tools for cultural and literature representation.

Individual differences in the use of cleft constructions in speech

Michael Barlow and Andreea Calude
Auckland University

Cleft constructions constitute a grammatical means for focusing specific information and directing the attention of the hearer/reader to the most salient part of the message. English is extremely rich in clefts, particularly in comparison to other languages where the highlighting or focus is achieved by different means. The present study adds to the considerable literature on English clefts (Calude 2009, Collins 2004, Hedberg 2000, Lambrecht 2001) by examining individual differences in the use of clefts in the speech of five White House Press Secretaries over a period of a year or more. The amount of data for each press secretary varies from 200,000 to 1.2 million words and allows comparisons of inter- and intra-speaker variability. The results show (a) remarkable stability in the frequency of use of demonstrative clefts, wh-clefts etc. in the samples of speech for each speaker over time and (b) marked differences in the use of clefts from one speaker to the next.





The graphs show the frequency of wh-clefts and demonstrative clefts in each sample. (The initial letter indicates the name of the speaker.) It can be seen that many but not all the speakers favour one of cleft types. We discuss these and other results in more detail in the presentation.

The use of Trigrams in Classification of texts based on Authorship

Santiago Barreda
University of Alberta

This study explores the use of trigrams as a basis for the automatic classification of texts (Biber et al. 2004). More specifically, the study investigates the effects of systematically varying two distinct parameters, frequency and dispersion, in an attempt to determine the optimal settings of these parameters to yield the best classificatory results.

To this end 20,000 word excerpts from 5 novels from each of 4 authors (Joseph Conrad, Jack London, Bram Stoker and H. G. Wells) were used. Trigrams were included or excluded on the basis of two distinct measures: by including only those trigrams with a minimum within-text frequency and a minimum level of dispersion. Frequency is the number of times a trigram appears within a text while dispersion refers to the number of texts a trigram is found in, regardless of its within-text frequency. A list was created of all trigrams appearing in each text along with their within-text frequencies. After this, matrices were created including all trigrams with minimum frequency and dispersion levels for every combination of dispersion level (1…20 with 1 being all trigrams found at least in one book and 20 being all trigrams found at least all books) and within-text frequency (minimum frequencies used were 1…6). Classification analyses were then performed on each of the resultant matrices (Baayen 2008).

Results indicate that frequency and dispersion are independent but not necessarily disjoint properties and, as a result, manipulation of each property separately results in more control on the result of the analysis. Furthermore, while increasing the minimum frequency and dispersion a trigram must have for inclusion in the analysis generally results in increased performance, a low frequency threshold benefits most from a high dispersion minimum while a high frequency threshold benefits most from a low dispersion minimum.

References:

Baayen, Harald. *Analyzing linguistic data : a practical introduction to statistics using R*. Cambridge, NY: Cambridge University Press, 2008.

Biber, Douglas, Susan Conrad, and Viviana Cortes. (2004). If you look at . . .: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371-405.

New applications of translation corpora: Investigating language contact and language change

Viktor Becher
University of Hamburg

Corpora in which source language texts are aligned with their translations into one or more target languages are not only useful for translation studies, but also for the study of language contact and language change. Particularly useful in this respect are diachronic translation corpora, as the analysis of translation choices in different time periods not only allows the investigator to observe diachronic changes in language use, but also enables him/her to explain different developments in terms of causal relatedness. It will be argued that the kind of evidence translation corpora can offer in this respect cannot be obtained from monolingual corpora. In support of this argument, two case studies will be presented that were carried out on a corpus of English-German translations and comparable (non-translated) German texts from the years 1978 - 2002. The studies not only evidence remarkable changes in German language use that have occurred in this short time-span (such as e.g. a trend towards paratactic syntax in certain contexts), but also suggest that some of these recent developments are due to the influence of English on German. Most interestingly, it seems that one gateway through which English patterns of language use enter the German language are the investigated translations from English.

Emotional practices and character identity in American popular culture

Monika Bednarek
University of Technology, Sydney

The notion of character identity in fiction and film has been relatively neglected in stylistics and narratology (Toolan 2001: 80, Rimmon-Kenan 2002: 29) while television dialogue remains largely uninvestigated within corpus linguistics and linguistics as a whole (with only a few notable exceptions). However, character identity is particularly important with respect to contemporary television series that are often character-driven. This paper is a contribution to the analysis of character identity in television dialogue, using a corpus stylistic (Wynne 2005) approach and focussing on 'emotive' (Ameka 1992) interjections in an American television 'dramedy' (Ross 2004). I report on the analysis of the distribution and use of interjections such as *oh my god*, *for X's sake*, *damn/damn it* with respect to character identity and relations between characters. Such interjections are clearly part of 'surge features' that work as 'implicit cues' to characterisation (Culpeper 2001: 190). As shown, they are part of the construal of interpersonal identities; they are related to external factors (e.g. being a 'family friendly' show), and they partially reflect cultural stereotypes (e.g. a conception of men as less 'emotional' than women). Interestingly, there is also a link between characters that are in some way associated with each other (for instance as relatives), and their usage of interjections. While I argue that interjections can be usefully investigated in a corpus stylistic analysis it is also acknowledged that this methodology needs to be complemented with a more in-depth approach, since interjections are always produced in reaction to a context (e.g. *oh my god* can be associated with different emotions). Thus, a three-pronged approach to character identity is proposed that combines large-scale and small-scale corpus linguistics with qualitative discourse analysis.

The Brazilian Corpus

Tony Berber Sardinha
Sao Paulo Catholic University

In this paper we present the Corpus Brasileiro (Brazilian Corpus), which, when ready in 2010, will comprise one billion words of contemporary Brazilian Portuguese, from a wide range of registers, written and spoken. Currently there is a gap among online Portuguese collections for corpora with the dimensions and variety such as the Corpus Brasileiro. The largest online corpora of Portuguese are the CETEM Público (at www.linguateca.pt), with 180 million words of newspaper text, and the Corpus do Português, with 45 million words of speech and writing (www.corpusdoportugues.org/). The structure of the corpus follows the architecture proposed by Mark Davies, which consists of importing textual data into relational databases and then querying the databases via PHP. Our search engine offers access to both frequency information and concordances for user generated searches, as well facilities for exporting search results to txt and Excel, which enables users to post-edit the data to fit their purposes. Searching the whole corpus is extremely fast: a simple search will typically take about half a second. Users do not have access to the whole texts, since this would impinge on copyright laws. The need for corpora as large as one billion words derives from the fact that a corpus is a sample of a large population (language), and in the case of general language, the size of the population is unknown; hence, the larger the sample, the closer it will be to the population, thus being a more representative sample of the range of variation in language. The corpus can have a significant social impact, as it will make it possible for everyone to search the vast quantities of text and talk and find out for themselves how Brazilian Portuguese is typically used in diverse situations.

Complex Extraction in English

Gunnar Bergh
Mid-Sweden University

Focusing on complexity aspects in the field of unbounded dependencies, the present paper reports on a corpus study of long extraction (or long movement, cf. Haegeman 1991:342 ff) in the four main historical subperiods of English. Among the structures highlighted, the following types are of primary relevance, realizing the notion of complex extraction:

(i) *XP* + *XP* [matrix clause] […Δ…Δ…]
(ii) *XP* [matrix clause] + [link clause] […Δ…]

These two extraction formulae can be illustrated through the following Early Modern English examples:

(i’) And here I brynge but one doctour, *whose testimony in the balance of any trewe christen man’s herte* [me thynketh] [Δ sholde weye downe Martyn Luther Δ] (JOHN FISHER’S SERMONS)
(ii’) then he layd heinously to her charge, y^{e} thing *y^{t}* [her self could not deny,] [that al y^{e} world wist] [Δ was true] (KING RICHARD)

Long extraction has been the focus of a great deal of work in the field of generative grammar in the last few decades. This has led to a situation where today we seem to know a great deal about such constructions from a theoretical point of view and with regard to present-day English (e.g. Chomsky 1986, Rizzi 2000), but where we know considerably less about their practical properties, and in particular their usage in earlier varieties of the language. Thus, it is the aim of the present study to try to redress some of this imbalance by looking into the usage patterns of complex extraction as they appear in a 1.4-million-word corpus of historical text, analysing specifically frequency data as a function of both structural type and level of complexity.

References:

Chomsky, N. 1986. *Barriers*. Cambridge, MA: MIT Press.
Haegeman, Liliane. 1991. *Introduction to Government and Binding Theory*. Oxford: Blackwell.
Rizzi, Luigi. 2000. *Comparative Syntax and Language Acquisition*. London: Routledge.

The development of *that said*

Laurel Brinton
University of British Columbia

In Present-day English, the absolute construction *that said* functions as a "contrastive-concessive" conjunct (Quirk et al. 1985:636) with hedging/boosting pragmatic usages (cf. Beeching fc.), as in:

> Many noted the presence of theft and vandalism. **That said,** many seemed relaxed at the possibility of theft (Strathy Corpus)

That said has been characterized as a modern "voguism" and is popularly seen as deriving from the full predicate "(with) that having been said" (Safire 2002).

This paper begins with a study of *that said* and related constructions in corpora of Present-day English, such as the BNC, COCA, and Strathy. *That said* is nearly three times more frequent in American than in British English and is most common in spoken English.

The paper then turns to the historical development of parenthetical *that said*, querying whether there is evidence for its evolution from a full verbal predicate. Visser (1972:1266) sees *being* or *having been* as modern additions. Corpus evidence also provides no support for derivation from a full predicate: absolute forms of {*that, this*} *said* arise in Middle English, while full passive constructions ({*this/that*} {*being, having been*} *said*) are late and extremely rare. The more common active construction (*having said {this, that}*) appears in the Early Modern period. This finding is in concert with studies of the origin of other—supposedly reduced clausal—parentheticals such as *I mean* (Brinton 2008).

This paper also traces the rise of non-propositional, concessive meaning via processes of (inter)subjectification, a change dating from the late 19th/early 20th century, as in:

> The impression that the Standard Co. of New Jersey has any trade relations with the Soviet government is incorrect.... " **<u>This</u> <u>said</u>**, President Teagle sailed on his 55th trip to Europe. (TIME 1927)

The period poses a challenge for the empirical study given the scarcity of Late Modern English corpora.

Developing a highly specialized corpus of spoken English: AAC discourse in the workplace

Carrie Bruce, Eric Friginal, Pamela Pearson, and Lucy Pickering
Georgia State University

For the more than 3,500,000 individuals with severe communication impairment in the US alone (Beukelman & Mirenda, 2005), augmentative and alternative communication (AAC) devices are presumed to facilitate interaction, thus promoting fuller participation in social situations (e.g., school or work). However, AAC devices have been found ineffective at providing quick access to the context-specific language needed in work settings (Bryen, Potts, & Carey, 2007). Therefore, this specialized corpus of AAC user (n = 4) interactions and those of their job-equivalent non-AAC using counterparts (n = 4) in seven different places of work was created to examine the lexicogrammatical features and communicative patterns used by each pair -- AAC and non-AAC -- for the purpose of determining: 1) the workplace-specific language needs of AAC users, and 2) whether there is a gap in existing lexical and grammatical prediction capabilities of AAC devices and what the user may actually require on the job.

Data collection methods mirrored those used in the development of the Language in the Workplace Project (Holmes, 2000). Speech samples were (a) gathered from five consecutive work days to ensure a wide range of routine and novel topics, (b) captured via wearable digital voice recorders, and (c) transcribed following an enhanced orthographic transcription scheme (Crowdy, 1994; BNC). The result is a 180-hour+ corpus involving eight focal participants and their interactions with over 60 interlocutors in seven separate workplaces, totaling nearly 500,000 words in all.

This session will 1) address never before documented issues encountered in the collection, transcription, and annotation of this highly specialized corpus of AAC workplace discourse, and 2) provide implications for natural language processing (NLP) with respect to AAC device prediction systems and the creation of more efficient pre-programmed pages.

A Corpus Analysis of Japanese Mimetics

Joshua Caldwell
Brigham Young University

Sound symbolism, onomatopoeia, ideophones and mimetics are commonly considered fringe linguistic phenomenon, and thus traditionally left unexamined or ignored. However recent research shows that in many African, Asian and Native American languages they are an integral component of the language (Nuckolls 1999, 239-240). One such language is Japanese. “Mimetics in Japanese are extremely productive and ubiquitous” (Tsujimura 2005, 146), and “are indispensable for enriching colloquial as well as literary expression in both spoken and written Japanese” (Baba 2003, 1862).

Much of the research on Japanese mimetics relates to which part of speech the mimetic occurs as. “A single mimetic word can appear as a noun, an adjective, an adverb, or a verb” (Tsujimura & Deguchi 2007, 340). The favorite example of this phenomenon is *iraira*, as seen below.

part of speech	Sample Sentence
Noun	Kodomo-no seiseki-ga waruku iraira-ga tamatta Since my child’s grades have been bad, my irritation has accumulated
Adverb	Ano hito-wa itsumo iraira(-to) hanasu That person always speaks in an irritated manner
Verb	Otto no kudaranai hanasi-ni iraira shita. I got irritated by my husband’s silly talk

Table 1-Various forms of iraira (Tsujimura 2005, 144)

However other mimetics can sometimes only appear as a single category. Examples of this include seisei suru (to feel refreshed), utouto suru (to doze), sappari suru (to feel refreshed) (Tamori 1980, 167). It is assumed by many scholars that mimetic “words function essentially as adverbs” (Inose 2007, 98) and are “more closely related to the Japanese culture than standard adverbs” (Yang 1984, 147). However, no data based study exists that verifies the frequency of which mimetics occur most often in which part of speech (grammatical category). In this paper I will describe a study performed to investigate this issue. Approximately 1700 memitics will be analyzed using the data from the Kotonoha (http://www.kotonoha.gr.jp) and JpWaC (http://corpus.leeds.ac.uk/) Corpora, to measure the overall occurrence of mimetic per parts of speech.

Using corpus linguistics to explore the 'reading' of multi-semiotic play

Helen Caple
University of Sydney/University of New South Wales, Australia

This paper explores the complex cultural and institutional practice of ‘playing with the reader’ in the print-media, as exemplified through a particular news story genre in the Australian broadsheet newspaper, *The Sydney Morning Herald,* collected in a corpus of 1000 texts. In this genre, news stories are presented as short, witty news bites with a heading, a dominant photograph and a caption. The heading and image enter into a verbal-visual (multi-semiotic) play that relies on the manipulation of common idiomatic expressions or allusions to other texts, while the caption elaborates on the story's news value. This type of discourse constitutes a semiotic practice that is tied to an Anglo culture with a long history of this kind of play in media discourse. In producing and perpetuating this practice newspapers construe ‘an obliging reader’ (Kitis & Milapides, 1996: 585) who fulfils certain expectations on the part of the newspaper as far as different types of knowledge (linguistic, cultural, multi-semiotic) are concerned. Making use of Sinclair's open-choice and idiom-principle I will highlight how we can use corpus linguistic methodology and theory to explore the interpretation of this multi-semiotic play by the reader.

Reference :

Kitis, E. and Milapides, M. 1996. ‘Read it and believe it: how metaphor constructs ideology in news discourse. A case study’. *Journal of Pragmatics 28*: 557-590.

Analysis of Modal Auxiliaries in Two Consecutive Phrases Extracted from the British National Corpus

Robert Chartrand[1], Hidenobu Kunichika[1], and Akira Takeuchi[2]
[1]Kurume University, [2]Kyushu Institute of Technology

This paper investigates the use of modal auxiliaries in two consecutive phrases extracted from the British National Corpus. One of the ways that modals are being used is by colloquial pairs. These pairs of modals are often used in consecutive phrases but they are not well understood or researched. We discuss the reasons for analyzing this particular aspect of the English language, the statistical analysis conducted to determine the more frequent uses and provide some concrete examples for evaluation. Further emphasis is placed on the meanings of the pair use of modal auxiliaries and report on a survey examining their collocational patterns and corresponding semantic meanings. In most cases of using a corpus for looking at language use, only a simple concordance is carried out for simple pattern matching. Moreover it is reasonable to assume that some expressions are more common than others. By using concordances, learners can investigate co-occurrence patterns or learn common expressions. It is our belief that some combinations of modals are more commonly used than others, and people use each of these pairs of modals to express a specific meaning. It is not well known exactly, however, which combinations of modals are more popular. A purpose of our study is to investigate which colloquial expressions are more common and to identify their meanings. This kind of information is very useful for non-native learners of English to expand their expressive abilities. Moreover, we investigated the meanings of the fifteen most common pairs of modals and analyzed the appropriateness of the results. The way in which modals are interpreted in sequence is important in teaching or learning English as a Second Language and learners and educators could make use of these results to gain a better understanding of modal auxiliaries and to facilitate the process of learning English.

Irish *like* as an invariant tag: evidence from ICE-Ireland

Georgie Columbus
University of Alberta

Discourse markers, such as *like*, *you know*, and *innit*, are the smallest units of attitudinal meaning available to English with the exception of prosody and intonation. As such, they have been researched thoroughly for their effects on sociocultural meaning (e.g. Schiffrin, 1987, Stubbe and Holmes, 1995). With the advancement of corpus linguistic techniques, the study of discourse markers has been made more methodical, and allowed for larger samples to be utilized.

One particular type of discourse marker, the invariant tag (e.g., *innit, eh, yeah*) has been much investigated in English varieties for their social and semantic/pragmatic functions. *Like* has also been the focus of attention in discourse marker studies, particularly for its social and quotative uses (e.g., D'Arcy, 2005, Tagliamonte and D'Arcy, 2004, Miller and Weinert, 1995). The recent completion and distribution of the International Corpus of English for Irish English (Kirk et al. 2007), however, has revealed strong evidence for Irish *like* as an invariant tag (Kallen, p.c.). The aim of this study is to describe the pragmatic/semantic functions of Irish *like*, and investigate its frequency compared to other invariant tags in Irish English. Comparisons for the tag functions of *like* will also be made, particularly with reference to previous research on a common invariant tag in Englishes, *eh*. The results of this study add to the ongoing description of variations in Englishes.

References:

D'Arcy, A. (2005). *Taking a new perspective on discourse 'like'.* Paper presented at the University of Canterbury 2005 Seminar Series.

Kirk, John M. , Kallen, Jeffrey L., Lowry, Orla, Rooney, Anne, and Margaret Mannion. (2007). *International Corpus of English: Ireland Component. The ICE-Ireland Corpus*. CD-ROM

Miller, J. and R. Weinert. (1995). The function of LIKE in dialogue. *Journal of Pragmatics*, 23: 365-393.

Schiffrin, Deborah. 1987. *Discourse markers.* Cambridge: Cambridge University Press.

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.

Tagliamonte, S.A. & A. D'Arcy (2004). "He's like, she's like": The quotative system in Canadian youth. *Journal of Sociolinguistics*, 8 (4): 493-514.

Corpus linguistics and language documentation: Challenges for collaboration

Christopher Cox
University of Alberta

It has been suggested in the recent literature concerning both corpus linguistics (e.g. McEnery & Ostler, 2000) and language documentation (e.g. Johnson, 2004) that these two disciplines may share natural points of interaction. From the perspective of corpus linguistics, the diverse and consistently organized collections of digital language materials arising from contemporary language documentation (cf. Himmelmann, 1998; Woodbury, 2003) appear to present ideal sources of 'raw' language data for corpus construction. Conversely, from the perspective of language documentation, corpus construction and the techniques which corpus based methodologies offer for later analysis may present additional means by which portions of the documentary record might be put to descriptive and theoretical use, opening the products of language documentation to further linguistic treatment and to new forms of community access.

Although considerable benefit might thus be anticipated from close collaboration between these two disciplines, one may nevertheless submit that such interaction may not always be as simple to foster as might be hoped. The purpose of this paper is to consider points of commonality and difference in the goals and commitments maintained in corpus construction and language documentation. This discussion concentrates upon four specific areas in which corpus construction and language documentation practices might be contrasted:

1. The relationships which typically exist between project stakeholders;
2. The methods of linguistic sampling employed;
3. The editorial treatment of the resulting data and metadata;
4. The technologies conventionally employed in both disciplines.

Concerns within each of these areas are exemplified by corpus construction and language documentation efforts centred around a minority language of Canada. It is hoped that balanced consideration not only of the natural compatibilities between these two disciplines, but also of potential mismatches between their respective practices might provide serviceable information supporting ongoing efforts to bring these two disciplines into closer contact.

References:

Himmelmann, Nikolaus P. 1998. “Documentary and descriptive linguistics.” *Linguistics* 36: 161-195.
Johnson, Heidi. 2004. “Language documentation and archiving, or how to build a better corpus.” *Language Documentation and Description. Volume 2*, ed. by Peter Austin, 140- 153. London: School of Oriental and African Studies.
McEnery, Tony and Nick Ostler. 2000. “A new agenda for corpus linguistics – working with all of the world's languages.” *Literary and Linguistic Computing* 15.403-18.

Woodbury, Anthony. 2003. “Defining documentary linguistics.” *Language Documentation and Description. Volume 1*, ed. by Peter Austin, 35-51. London: School of Oriental and African Studies.

A Corpus Analysis of Frequency Effects on Eye-Movements in Sentence Context

Philip Dilts[1], Gary Libben[2], and R. Harald Baayen[3]
[1]University of Alberta, [2]University of Calgary

Corpora are extremely useful for psycholinguistic research. The Dundee Corpus (Kennedy et al. 2003), for example, contains eye-movements recorded while 10 participants read the same 20 newspaper editorials. This type of psycholinguistic corpus has allowed for large-scale explorations of lexical access in a natural reading context.

We use the Dundee Corpus to investigate the effect of word frequency on eye-movements in sentence context. Van Petten and Kutas (1990) showed that while word frequency affects the N400 Event-related potential (ERP) of words presented in isolation and at the beginning of sentences, this frequency effect gradually decreases as readers proceed through a sentence. They found that the frequency effect disappears completely by the end of the sentence, suggesting that top-down effects of context may supersede bottom-up effects of lexical properties when we access words during reading.

We analyzed the Dundee corpus, attempting to determine whether the effect of frequency on eye-movements diminishes as readers proceed through a sentence. After using several types of regression modeling, we find no support for attenuation of the frequency effect on fixation times, instead finding a small (non-significant) trend towards stronger frequency effects later in a sentence, even when sentence length, word length, and word position on screen and in the text are taken into account. This suggests that bottom-up lexical effects still play a role in lexical access, even when readers have more contextual information to help them identify a word.

References:

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee corpus. Poster presented at ECEM12: 12th European Conference on eye movements., Dundee, August 2003.

Van Petten, C. & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. Memory & Cognition 18(4), 380-393.

What grammaticality judgements and decontextualized examples indicate: reflections in the light of corpus data

Patrick Duffley
Université Laval

The recourse to grammaticality judgements and examples made up by linguists continues to characterize publications written within the generative grammar framework, and many studies in cognitive grammar as well (cf. Broccias 2004, Langacker 2005). Although cognitive grammar purports to be “usage-based” (Langacker 1987: 46), this feature is understood to refer to a model that incorporates not only general rules of production but also irregular and idiosyncratic phenomena in the form of conventional expressions learned as units. Generative grammar, on the other hand, seeks to construct a model of competence, the rules that allow an ideal speaker to generate the sentences of a language and only the sentences of that language, and justifies the recourse to grammaticality judgements theoretically. The latter are claimed to be both richer and less noisy than performance data (cf. Schütze 1996: 2-3): richer, because they provide data not found in attested usage, viz. “negative information, in the form of strings that are not part of the language”; less noisy, because they allow one to “distinguish slips, unfinished utterances, and so forth, from grammatical production.” Grammaticality judgement data has been severely criticized on both these counts however. Birdsong (1989: 72) provides an extensive list of intruding factors in grammaticality judgements which introduce so much noise into this data that he feels obliged to denounce “the hypocrisy of rejecting linguistic performance data as too noisy to study, while embracing metalinguistic performance data as proper input to theory”. Duffley (2002: 55) shows that having recourse to negative information is a consequence of adopting a theoretical view of a grammar as an algorithm for making structural grammaticality choices between strings that belong to a language and those that do not, which does not correspond at all to what speakers actually do with their language in a normal speech situation.

The present paper examines the discrepancy between grammaticality judgements and corpus data with five items that bring out various factors at work in grammaticality judgements. These factors are: (1) frequency, (2) experiential salience, (3) semantic harmony between word-meanings, (4) the natural tendency of human psychology to adopt certain attitudes, (5) the internal logic of linguistic constructions. The conclusion drawn is that using corpus data can help to neutralize such factors by exposing the analyst to the full range of linguistic usage and allowing him to infer the limits of the speaker’s unconscious linguistic knowledge. The paper thus vindicates Ruhl’s (1989: 125, 235) claim that “consciousness cannot hold all the possibilities at once, and thus must selectively partialize,” but “with abundant data, consciousness can infer the range and limits of the meaning [of a word].”

References:

Birdsong, David. 1989. *Metalinguistic performance and linguistic competence*. Berlin: Springer-Verlag.

Broccias, Cristiano. 2004. The cognitive basis of adjectival and adverbial resultative constructions. *Annual Review of Cognitive Linguistics* 2: 103-126.

Duffley, Patrick J. 2002. Linguistics as an empirical science: the status of grammaticality judgments in linguistic theory. *Lacus Forum 28*: 51-58.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Volume 1. Theoretical Prerequisites.* Stanford: Stanford University Press.

Langacker, Ronald W. 2005. Dynamicity, fictivity and scanning: the imaginative basis of logic and linguistic meaning. In Pecher, Diane and Rudolf A. Zwaan (eds), *Grounding cognition: the role of perception and action in memory, language and thinking.* Cambridge: Cambridge University Press, 164-197.

Ruhl, Charles. 1989. *On monosemy*. Albany: State University of New York Press.

Schütze, Carson T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Linguistic Characteristics of Non-Native Speaker Writing in English: A Corpus-Based Analysis

Eric Friginal, Amanda Baker, and Pamela Pearson
Georgia State University

There is an increasing number of corpus-based studies conducted to describe genres of academic writing and also to assess the quality of writing by foreign students enrolled in universities across the United States (U.S.) (e.g., Hinkel, 2002; Lee & Swales, 2006). For students coming from non-English speaking countries, second language (L2) writing ability is an important consideration measured by compulsory tests such as the Test of English as a Foreign Language (TOEFL) before these students gain formal admission to universities in the U.S. This paper presents the linguistic characteristics of academic essays written by non-native speakers (NNS) of English (N = 386) who are enrolled in eight universities in the U.S. The study utilizes corpus-based approaches in comparing the frequency distribution of a range of lexical and syntactic features of NNS academic writing from a corpus of written responses to two prompts from the writing section of the TOEFL (N text = 772, with approximately 254,000 words). The linguistic features of these L2 writing samples are compared across sub-groups of data based on writing quality, first language background of writers, and the use of prompts. Results show that the distribution of some linguistic features could be indicators of NNS writing quality (e.g., linking adverbials, stance markers, lexical verbs). In addition, variations exist in the linguistic characteristics of NNS essays based on the use specific writing prompts. It is argued that results of comparative data in the present study have pedagogical implications for the teaching of academic writing and materials production especially for NNS students in intensive English programs in the U.S.

Towards the unsupervised discovery of syntactic categories for typologically-varied languages

Simon Fung
University of Alberta

Successful unsupervised induction of syntactic categories not only enables part-of-speech tagging without corpus annotation, but also allows linguists to rigorously test criteria for distinctions between syntactic categories. It also enables the exploration of the nature of these distinctions in a more concrete way; syntactic categories may differ formally to different degrees. Previous efforts have compared distributions of word contexts (especially high frequency words in the contexts) to form syntactic categories, and have achieved encouraging results. However, although designed to be language-independent, they have so far been evaluated predominantly on English. Performance may be poorer for languages that mark syntactic relationships with morphology more than word order. Performance may also be poorer for languages that make less formal distinction between categories which nevertheless can be distinguished based on semantic criteria (e.g. nouns and verbs in Tagalog and Salishan languages).

This paper will compare the performances of existing unsupervised algorithms for syntactic category induction on Lushootseed, Totonac, and Tagalog, to their performances on English. These three languages will introduce significantly more morphological complexity, as well as more nuanced distinctions between nouns and verbs, compared to English. Since the amounts of available text vary among languages, corpus size will be controlled. From a technological perspective, good performance on all languages would be highly encouraging, though unexpected; meanwhile, variable performance would point to areas for future improvement, whether it be the integration of morphological analysis or semantic information. For linguistics, this evaluation will help pinpoint the most important types of criteria for syntactic categorization among structurally-varied languages, as well as concretely show varying degrees of difference between syntactic categories.

Rethinking the German three-way system of spatial demonstrative adverbs: evidence from electronic corpora

Johnathan Gajdos
University of Iowa

Spatial demonstrative adverbs allow for speakers to refer to objects in the context of the location relative to the speaker-hearer interaction; in German the primary spatial demonstrative adverbs are *hier* ‘here’, *dort* ‘there’, and *da* ‘here/there’. The assignment of a location identified by *da* is challenging. Since *da* can occur independently of speaker location it is therefore often understood as able to have a non-deictic function; the strength of this claim is called into question in this paper.

In an effort to analyze the use of *hier*, *da*, and *dort*, data were collected from a five syntactically-tagged electronic corpora of spoken and written German. The starting point of the corpus study was an examination of the occurrences of *hier*, *dort*, and *da*, both in combination with each other and in isolation. Previous literature has noted that *da* occurs frequently and early in L1 child speech, though many of those uses of da appear to serve a discourse-level function as opposed to a strictly deictic function. To examine German spatial deixis from the context of child L1 acquisition and language use, seven corpora from the CHILDES database were analyzed.

The data suggest that *da* does indeed occur with regularity from an early age; an examination of the frequency of the utterances of the target expressions was undertaken in conjunction with a consideration of the interaction between the child and the parent or other adults present in the corpus data. The child learner corpus data were also compared with that from the various non-learner corpora utilized in this study. The data from these corpus analyses provide evidence for a reexamination of the treatment of *da* as having a completely non-deictic interpretation and suggest that it is not actually devoid of deictic/locative information.

"Get out before we get you!" A corpus-based analysis of stance in threatening communications

Tammy Gales
University of California, Davis

Each year, countless numbers of threats are received at law enforcement agencies (Fitzgerald, 2007). Analysts receiving these threats perform any number of forensic analyses depending on the case context in which they are received—from a stylistic assessment of authorship to a content analysis of the threat level. Linguistics, as a relative newcomer to forensic sciences (Coulthard and Johnson, 2007), does not currently possess a unified framework for assessing threats, even though multiple methods, from quantitative statistical programs to qualitative stylistic analyses, are employed (Grant, *to appear*). Yet, despite this search for the most valid and reliable method, there is still one core element, the foundation upon which all others must be built, that is still lacking—a comprehensive linguistic description of the genre of threatening communications (Leonard and Shilling-Estes, *under review*).

As is well-established in corpus linguistics, genres and registers exhibit unique linguistic patterns and unique collocations of linguistic patterns (Biber *et al.*, 1998). Because threats are socially-constructed illocutionary acts that are the product of the society from which they derive (Eggins and Martin, 1997), linguistic patterns related to authorial 'stance,' or an author's culturally-organized "personal feelings, attitudes, value judgments, or assessments" about a theme, recipient, or proposition being presented (Biber *et al.*, 1999:966), can be used to aid analysts in assessing the level of authorial intent or commitment to the threatened act. Furthermore, patterns of variation within the genre can also aid analysts in identifying the *consistency* of authorship and *distinctiveness* of features within a threatening text—the two categorizations of expert opinion currently proffered in court (French and Harrison, 2007). Therefore, through a corpus-based analysis of over 300 authentic threats, this paper uncovers patterns of linguistic features of stance, such as personal pronouns and adverbial clauses, that manifest within and are unique to the genre of threats.

A Frequency Dictionary of Contemporary American English

Dee Gardner and Mark Davies
Brigham Young University

This presentation will discuss the design and progress of the first frequency dictionary of American English. The dictionary is based on the 360-million-word *Corpus of Contemporary American English* or *COCA* (1990-2007), which is the first large corpus of American English ever created, and the first to offer a balance of English usage between the registers of spoken, fiction, popular magazines, newspapers, and academic journals. The dictionary will be oriented towards English language learners, allowing them to focus on the content words (5,000) they are most likely to encounter in a wide range of registers, as well as the most frequent collocates associated with those words. For example, the entry for the word *break* would include (among other information) the following:

SUBJ: heart, hell, silence, bone, scandal, fighting, levee, dawn
OBJ: law, heart, record, rule, news, barrier, silence, ground
ADV: apart, abruptly, accidentally

This dictionary, approved as part of the Routledge series, should also be very useful for linguistic research, pedagogical materials development, and classroom-based language instruction.

I haven't drank in weeks: The use of past tense forms as past participles in English corpora

Kristina Geeraert
University of Alberta

Previous research on the English strong verb paradigm and the neutralization between these verb forms has largely focused on the past participle form being used as the past tense form, such as *she drunk ale* or *he rung the bell*, occasionally referred to as "Bybee verbs" (cf. Bybee 1995; Anderwald 2007). In this study, I consider a similar but less explored phenomenon of levelling in the reverse direction, i.e., the use of the past tense form as a past participle, such as *he has drank the wine* or *I've went to the store*. This phenomenon includes the verbs *drank*, *went*, *took*, *sang*, and *beat*. Interestingly, there is no common phonetic structure to these forms, unlike the V– nasal–velar pattern associated with the Bybee verbs. The Oxford English Dictionary (1989) makes reference to the past tense of these verbs being occasionally used as the past participle, for example *took* as a past participle is attested already in the 16th century. This study attempts to identify a variety of properties associated with the use of these verbs: the individual verbs used in this construction, the productivity of their usage, the linguistic context (*have took* vs*. has took* vs. *'ve took*), the genres in which they occur, the varieties of English, etc. I first investigate these properties within traditional corpora, such as the BNC and COCA. Since the phenomenon under study is relatively infrequent, large-scale corpora such as the web are needed to provide further data for this non-standard form, requiring the astute use of internet searches, and thus contributing to the current "web as corpus" (cf. Hundt, Nesselhauf, and Biewer 2007) debate.

References:

Anderwald, L. (2007). '*He rung the bell*' and '*she drunk ale*' – non-standard past tense forms intraditional British dialects and on the internet. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 271-286). Amsterdam, NL: Rodopi.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425-455.

Hundt, M., Nesselhauf, N. & Biewer, C. (2007). *Corpus linguistics and the web*. Amsterdam, NL: Rodopi.
Oxford English Dictionary Online, (2nd ed.). (1989). Retrieved on April 29, 2009 from http://dictionary.oed.com.login.ezproxy.library.ualberta.ca.

Constructing a Lexicon from a Historical Corpus

Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter, and Klaus Schultz
Ludwig-Maximilians-Universität München

We describe the construction of a historical lexicon for German to support Natural Language Processing. By such a historical lexicon we mean a collection of wordforms that represent variants of modern words occurring in proofread historical documents. Each lexicon entry is manually checked.We store at least one attestation and assign the corresponding modern lemma(s). In applications, as for example Information Retrieval Systems, a lexicon of this form can be used in online mode at query time or in offline mode at indexing time.

In order to collect the vocabulary of historical language, special corpora are needed. Most "historical corpora" found in the web are just large collections of document images. In contrast, we needed proofread symbolic texts. For German, appropriate sources are distributed over various places, often only known to a small and special scientific community. Because of that some efforts were necessary to compose a corpus of acceptable size. Our complete corpus contains 2,693,966 tokens and 288,709 types. The texts come from the period of 1500-1900.We give a thorough statistical analysis of the vocabulary found in this corpus.

A web based, collaborative lexicon tool was developed to support the construction of the historical lexicon. As a central background resource, we use a matching procedure for historical variants, which helps to considerably reduce the manual work. For a given historical token, we automatically receive one or several modern full forms. For non-pattern induced historical words, a historical linguist provides the matching between historical wordform and modern variants. It remains to assign the correct lemma(s) and part-of-speech to the full form(s). An automatic analyzer computes the lemma(s) - including part-of-speech info - which may underly the full form.

Several views of a Graphical User Interface control the steps to create a lexicon entry. They visualize different readings of a token, show the corpus attestations and provide access to resources as the lexical database, the set of patterns, and the texts of the background corpus. Two working modes are available: frequency lists of the tokens that remain to be analyzed and a full text mode providing information on the status of each token in a selected historical document.

Developing and Analysing the Engineering Lecture Corpus (ELC)

Lynn Grant
AUT University, Auckland, NZ

Students, in particular English as an Additional Language (EAL) ones, have much to gain from a better understanding of the academic spoken English used in lectures in their tertiary study. Focusing specifically on the area of engineering, previous research has looked at writing (Robinson & Blair, 1995; Archer, 2008), at the role of chunks, phrases and body language (Khuwaileh, 1999) and at understanding lectures (Miller, 2002, 2007). It is the latter two that are relevant to this study. Presently, three universities in different parts of the world (Coventry University, UK; Universiti Teknologi, Malaysia; AUT University, N.Z.) have been collaborating to develop and establish the Engineering Lecture Corpus (ELC). This study describes how we analysed the entire collection of lectures (at least 60 hours), in order to identify discourse features, engineering-related vocabulary as well as note similarities and differences in lecturing style in the three countries. University lecturers would benefit by increasing their awareness of both the discourse features and the delivery of their academic lectures. This awareness would in turn benefit both native speaker and EAL (English as an additional language) learners in these engineering lectures.

References:

Archer, A. (2008). 'The place is suffering': Enabling dialogue between students' discourses and academic literacy conventions in engineering. *English for Specific Purposes 27,* 255-266.

Khuwaileh, A.A. (1999). The role of chunks, phrases and body language in understanding co-ordinated academic lectures. *System* 27*,* 249-260.

Miller, L. (2007). Issues in lecturing in a second language: lecturer's behaviour and students' perceptions. *Studies in Higher Education 32* (6), 747-760.

Miller, L. (2002). Towards a model for lecturing in a second language. *Journal of English for Academic Purposes 1*, 145-162.

Robinson, C.M. & Blair, G.M. (1995). Writing skills training for engineering students in large classes. *Higher Education* 30, 99-114.

“Do” and “Make”: A Corpus-based Study

Yi-Chen Hsieh and Siaw-Fong Chung
National Chengchi University

Fu (2006) has found that two of the high-frequency verbs, *do* and *make*, are often misused by EFL language learners and has proposed that nominals following *do* are relatively negative (such as *disadvantage* and *harm*) and those following *make* are rather positive (such as *decision* and *judgment*). However, counterexamples could be seen in *make a mistake* and *do justice.* The study aims to explore the collocational behaviors of the two high-frequency words, *do* and *make* based on UKWac (cf. Bailey & Thompson, 2006), a large collection of web-based materials with more than two billion words, available through the Sketch Engine (Kilgarriff and Tugwell, 2002). We hypothesize that a lexical and corpus-based approach will provide quantitative information regarding the similarities and differences between *do* and *make*.

A total of 4,663,926 instances with any verb forms was examined for *do* and 2,838,486 for *make*. Our results show that both *do* and *make* can be followed by an event (such as *do homework,* and *make stride*) (cf. Saeed, 2003 for “event”). Besides, both *do* and *make* can be used metaphorical in ***make** a recommendation* or ***do** the talking*, in which both *recommendation* and *talking* are not physically actions. Since only *make* has a causative reading, it can denote a “change of state” (e.g., *Tom **makes** a cake.).* Furthermore, we also found that (See Appendix), most of the nominals following *make* + *from* are manufactured products (96% of the top 25 collocates examined are concrete.) but for *do* + *from* are abstract nouns (88% of the top 25 collocates examined are abstract). With respect to the teaching practice, such a lexical and corpus-based approach can expose English learners to distinguishing the use of two highly frequent verbs.

References:

Bailey S. & Thompson D. (2006). UKWAC: Building the UK's First Public Web Archive. *D-Lib Magazine,* 12 (1).

Fu, Z. (2006.) What to Follow “Make” and What to Follow “Do”--Corpus-Based Study on the De-lexical Use of “Make” and “Do” In Native Speakers’ and Chinese Students’ Writing. *US-China Education Review, 3*(5), 42-47.

Kilgarriff, A. & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In the Proceedings of the *ACL Workshop COLLOCATION: Computational Extraction, Analysis and Exploitation*. Toulouse. 32-38.

Saeed, J. (2003). *Semantics*. Blackwell, 2nd edition.

Appendix:

Table 1: Top 25 Collocates for the PREPOSISTION_FROM Relation of *Do* and *Make* based on UKWak

Do			*Make*		
Collocates	**Freq**	**Saliency1**	**Collocates**	**Freq.**	**Saliency**
scratch	21	27.65	plastic	291	46.15
perspective	43	25.82	steel	353	44.22
beginning	33	24.85	cotton	204	43.02
home	100	23.46	grape	131	42.73
start	37	21.8	wood	397	42.54
time	169	21.75	aluminium	154	40.74
outside	15	20.18	material	1062	40.49
comfort	17	18.9	ingredient	224	39.81
point	56	18.36	polyester	67	39.12
inside	11	18.31	fabric	193	38.84
position	42	18.18	leather	140	36.65
memory	16	13.27	nylon	64	36.31
computer	22	13.03	hardwood	58	36.31
distance	13	12.58	resin	85	36.08
age	20	12.28	alloy	99	35.99
moment	13	11.66	fibre	159	35.85
duty	12	10.87	polypropylene	39	35.33
day	32	10.04	wool	97	35.22
office	15	10.03	flour	91	34.57
side	17	9.37	clay	133	34.29
owner	11	9.33	pulp	53	33.47
location	11	9.18	rubber	88	32.62
source	12	7.72	scratch	67	31.89
page	16	7.67	milk	138	31.09

ACKNOWLEDGEMENT: The research discussed herein is under the funding of the National Science Council (NSC 97-2410-H-004-001).

The Acquisition of Grammatical Knowledge and Usage Using a Corpus-Aided Discovery Approach

Li-Shih Huang
University of Victoria

Recent technological improvements have enhanced language learners' and instructors’ access to corpora (e.g., Conrad, 2005; Hunston, 2005; Sinclair, 2004). Consequently, corpus-aided discovery learning (Huang, 2008) that actively engages learners in analyzing language is an increasingly workable option for English-as-anadditional language instructors (Gavioli, 2000). This presentation reports on the pedagogical applications and outcomes of incorporating a corpus-aided discovery learning approach that enhances the acquisition of English grammatical knowledge and usage in an undergraduate-level English language and usage course taken by students across disciplines. The study involves ten groups of students, with two to three participants in each group, over the course of 12 weeks, who engage in weekly learning activities that promote learners’ abilities to observe, analyze, and generalize about the linguistic features in their findings. Qualitative and quantitative analyses of participants’ background questionnaires, weekly audio-taped recordings of group discussion sessions, weekly individual written language logs, and exit surveys provide empirical evidence about the efficacy of the approach. The presenter will, on the basis of the empirical study, share pedagogical strategies and challenges involved in implementing the approach, which helps students to develop sharper awareness of the linguistics features of spoken and written texts in different contexts.

References:

Conrad, S. (2005). Corpus linguistics and L2 teaching. In E. Hinkel (Ed.). *Handbook of research in second language teaching and learning.* Mahwah, NJ: Lawrence Erlbaum Associates.

Gavioli, L. (2000). The learner as researcher: Introducing corpus concordancing in the classroom. In G. Aston (Ed.), *Learning with corpora* (pp. 108-137). Houston, TX: Athelstan/Bologna: CLUEB.

Huang, L.-S. (2008b). Using guided, corpus-aided discovery learning to generate active learning. *English Teaching Forum. 46*(4): 20-27.

Hunston, S. (2005). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Sinclair, J. (2004). *How to use corpora in language teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

University Student Public Speech: A Corpus-Based Study of Student Produced Language

Gina Iberri-Shea
United States Air Force Academy

Public speaking is an important part of academic life for many students. It is often taught as if uniform and homogeneous, but the extent to which it varies across the university has been previously unknown. This study explores the language variation in university student public speech across two academic disciplines: business administration and education. A corpus of university student public speech, made up of 102 classroom presentations (approximately 215,000 words), was designed, constructed, and analyzed using both quantitative and qualitative methods. Using multi-dimensional analysis, the co-occurrence of linguistic variables was functionally interpreted in order to provide a description of university student public speech, and to compare the language use to other university registers.

University student public speech was compared on four dimensions of variation previously found in university language (Biber, 2006). Significant differences were found for discipline and task on each of the four dimensions of variation. University student public speech scored extremely high on dimension 1, which describes an oral-literate continuum. In fact, these student presentations were more "oral" than any university register previously studied. Another interesting finding was that the university student public speech texts exhibited positive features associated with teacher-centered stance. This differs from previous work, where positive scores were strongly associated only with instructor-controlled registers. This presentation will first describe the corpus and methods, and then present the variation found in university student public speech on each of the four dimensions of university language, focusing on how university student public speech compares to and differs from other academic registers.

Grammatical and register variation and change: A multi-corpora perspective on the English genitive

Bridget Jankowski
University of British Columbia

(1) a. ...use it to house ***Canada's*** *first responsible government*... [Macleans/h/1956]
b. ...doing a work of national importance and polishing *the treasures* ***of Canada***. [Macleans/h/1956]

This paper presents a corpus-based variationist analysis of the *s*- and *of*-genitive in Canadian English, as in (1), using parallel, part-of-speech-tagged corpora. The corpora were compiled from two Canadian English written registers spanning 1906–2006: journalistic prose from *Maclean's* magazine and *Hansard* parliamentary transcripts.

Results show that the *s*-genitive is increasing, with journalistic prose leading as expected (Hinrichs & Szmrecsanyi 2007: 440). Szmrecsanyi & Hinrichs' conclusion (2008: 299 – 301, 307) that "economy" results in increased *s*-genitive use in journalism — a register-internal change rather than a change in the underlying grammar — is tested by replicating their factors of lexical density and frequency of the possessor NP head noun. However, subject animacy is a stronger predictor of *s*-genitive use. Additionally, the apparent spread of *s*-genitives into a subtype of inanimate subjects (places) indicates a possible ongoing grammatical change.

These results are then compared to a recent variationist study of *s*-and *of*-genitive in Canadian vernacular speech (Tagliamonte and Jarmasz 2008) in order to determine whether the shift toward increased *s*-genitive use in the written registers is originating, as would be expected, in the vernacular (Pintzuk 2003: 525). Intriguingly, though use of the *s*-genitive is far more common overall in vernacular speech than in the written registers, there are indications that journalistic prose is leading the spread of the *s*-genitive into the inanimate subject contexts. The unique perspective afforded by such multi-corpora, multi-register comparison permits the simultaneous study of variation and change at both internal grammatical and cross-register levels. I argue that such methodology is integral to further understanding of language variation and change across both spoken and written forms of language.

References:

Hinrichs, Lars & Benedikt Szmrecsanyi. 2007. "Recent changes in the function and frequency of standard English genitive constructions: a multivariate analysis of tagged corpora". *English Language and Linguistics* 11(3): 437–474.

Pintzuk, Susan. 2003. Variationist approaches to syntactic change. In *The Handbook of Historical Linguistics,* eds. Brian D. Joseph and Richard D. Janda. 509–528. Oxford: Blackwell.

Szmrecsanyi, Benedikt & Lars Hinrichs. 2008. Probabilistic determinants of genitive variation in spoken and written English: a multivariate comparison across time, space, and genres. In *The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present*, eds.

Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen. Amsterdam: Benjamins, 291–309.

Tagliamonte, Sali and Lidia Jarmasz. 2008. “Variation and change in the English genitive: A sociolinguistic perspective.” Paper presented at the annual meeting of the Linguistic Society of America, Chicago, IL, January 3–6, 2008.

A Corpus-based study of the connectives *'danshi' 'keshi'* and *'ran'er'* in Mandarin Chinese

Ruiting Jia
University of Alberta

This study investigates the uses of three Mandarin Chinese connectives '*danshi', 'keshi',* and *'ran'er'* which are all translated as *'but/however'* in English. Two Mandarin corpora are used: LCMC Mandarin written corpus (15 files/ 1,000,000 words) and Callhome Mandarin spoken corpus (120 files /300,767 words). Five types of data will be examined: (1) The frequency distribution of the three connectives in written and spoken texts; (2) collocational patterns with these three connectives in written and spoken texts; (3) word clustering with these three connectives in written and spoken texts; (4) the semantic connotation found with these three connectives in written and spoken contexts. Results show that there are interesting differences among these three connectives, which have not been recognized in dictionaries and grammar of Mandarin Chinese: (1) connective *danshi* and *keshi* are losing their contrastive function in current spoken Mandarin Chinese, and connective *ran'er* is more likely to occur with the negative message in written texts; (2) among three connectives, *danshi* is the most frequently used connective in both written and spoken texts.

References:

Wang, Chueh-chen & Huang, Lillian M (2006), *Grammaticalizaion of Connectives in Mandarin Chinese: A corpus-Based Study.* Language and Linguistics 7.4: 991-1016.

Li, Charles N. & Thompson, Sandra A (1981), *Mandarin Chinese: A Functional Reference Grammar.* Berkeley: University of California Press.

Lü, Shuxiang (1979), *Lü Shuxiang quan ji.* Liaoning Education Publishing Company

Syntactic Aspects of Learner English

Christine Johansson and Christer Geisler
Uppsala University

The present paper studies syntactic development in learner English. The data are drawn from a new corpus of learner English (see Geisler & Johansson 2007). The Uppsala Learner English Corpus (ULEC) is a collection of essays by Swedish junior and senior high school students aged between 14 and 19. We will focus on the structure of clauses and sentences in students' writing. Preliminary results show that many high school students use features of spoken language in their writing, such as multiple clausal coordination, sentence fragments, and discourse markers (*well*, *good-bye*). Another feature of learner English writing involves the use of subclauses as main clauses, as in (1).

(1) At first i did'nt wanted to sitt next to it. But i am very satisfied now when i had done it. Because the tiger was must bigger then me. So it could eat me with only one bite.

It is, however, possible to trace syntactic development in the Swedish students' writing. As a measure of syntactic complexity, we will use the concept of the T-unit, which is defined as an independent clause with all its dependent clauses (see Biber et al. 1998: 178). In the ULEC data, 70% of the dependent clauses are adverbial clauses (*when*, *since*, *because*, and *then* are the most common subordinators). In (1) above the T-units are short since the dependent (adverbial) clauses are treated as main clauses. Example (2), which is from a third-year high school student, looks different.

(2) He is going to work as a florist because no one would suspect that a simple florist would be a superhero (Spiderman).

The more advanced third-year students produce longer T-units, as in (2). In addition, they use nominal *that*-clauses and relative clauses to a greater extent.

References:

Biber, D., S. Conrad & R. Reppen. 1998. *Corpus Linguistics*. London: Longman.

Geisler, C. & C. Johansson. 2007. "*ULEC* - The Uppsala Learner Corpus". Paper presented at ASLA, 8-10 November, 2007, at Lund University.

Johansson, C. 2006. "Supplemental instruction och lärarstudenters intresse för grammatik i språkundervisningen och i examensarbeten". In Approaches to Teaching and Learning in Linguistic Research: Papers from the ASLA Symposium in Växjö,10-11 November, 2005, ed. J. Einarsson, E. Larsson Ringqvist & M. Lindgren. Uppsala: Svenska föreningen för tillämpad språkvetenskap, 101-111.

A Corpus-based Study of the Korean Quantifiers *Cokum* and *Com*

Hee Ju
University of California, Los Angeles

This study examines differences between *cokum* and *com*, two Korean quantifiers which may be translated in English as 'a little.' Using Sejong Korean corpus of 10 million words (Koo 2005) and Wordsmith tools (Scott 2008), this study looks at their distribution and lexico-grammatical associations over different registers and the sequential organization of conversation in the spoken data.

The existing studies seem to agree that *cokum* is restricted to referential meanings, whereas *com* has pragmatic functions in addition to referential meanings (Chu (2004; Im, 1995; Ku, 1998; Lim, 2003; Lee, 1992). However, no study has conducted any quantitative analysis of actual usage, and most analyses are restricted to sentential units.

The results of the study show that *com* is 4 times as frequent as *cokum* in total frequency and 2~6 times across each register. In terms of specific genres, it is found that both forms occur most frequently in genres that involve informal and interpersonal interaction such as novels, drama scripts, and TV talk-shows. Based on the fact that both forms show similar distribution across different genres yet *com* shows overwhelmingly higher frequency than *cokum*, we may apply Zipf's "law that the length of a word is inversely proportional to its frequency" to the finding (quotation from Bybee and Hopper 2001, p.1). Also, findings on lexico-grammatical association and sequential environments of these forms supports the hypothesis that *com* might have undergone grammaticization which involves formal reduction and functional change due to its high frequency (Bybee and Hopper 2001).

References:

Bybee, J. & Hopper, P. (2001). "Introduction to frequency and the emergence of linguistic structure." In J. Bybee and P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*, pp.1-24. Amsterdam/Philadelphia: John Benjamin Publishing Company.

Chu, K. H. (2004). A semantic & pragmatic study of 'Com' grammaticalization. *Kuke kyoyuk 115*: 433-453.

Goodwin, C. (1996). "Transparent Vision." In E. Ochs, E. A. Schegloff and S.Thompson (eds), *Interaction and Grammar*, Cambridge: Cambridge University Press, pp. 370-404.

Heritage, J. 1984. *Garfinkel and Ethnomethodology*, Cambridge: Polity Press. Im, Y. J. (1995). About "Com/Cokum" *Hanyang Emun* 13.

Kärkkänen, E. (2006). Stance taking in conversation: From subjectivity to intersubjectivity. *Text & Talk 26* (6): 699-731.

Kim, K-H. & Suh, K-H. (2002). Demonstratives as proximal indexicals: *Ku* and *Ce* in Korean conversation. *Japanese/ Korean Linguistics 10* , 192-205.

Koo, H. J. (2005). On Corpus-based studies of Spoken Language. *The Journal of Linguistic Science 32*: 1-20.

Ku, J. N. (1998). On pragmatic marker ‘Com’. *Hanguk ena munhak 41*: 411- 434.

Lee, H. G. (1991). The pragmatics of Korean pragmatic morpheme com. *Studies in the Linguistic Sciences 21.2*.

Lee, H. G. (1992). *The Pragmatics and Syntax of Pragmatic Morphemes in Korean*. Urbana-Champaign, IL: University of Illinois Doctoral Dissertation.

Lim, G. H. (2003). Pragmaticization of Korean Degree Verbs. *The Journal of Linguistic Science 24*: 283-302.

Schegloff, E. , Ochs, E. & Thompson, S. A. (1996). *Interaction and Grammar*, Cambridge: Cambridge University Press.

Son, S. M. D.(1988). The contextual meaning of ‘com’. *Kukehak nonjip 14*: 477-508.

Scott, M. (2008). WordSmith Tools version 5.0. Oxford: Oxford University Press.

Tao, H. (1999). The grammar of demonstratives in Mandarin. *Journal of Chinese Linguistics 27(1)*:69-103.

A contrastive analysis of causal expressions

Suzanne Kemmer[1] and Michael Barlow[2]
[1]Rice University, [2]Auckland University

Causation, a complex notion involving construal of action in the physical and human domain, can be expressed implicitly by juxtaposition or explicitly by the use of conjunctions, prepositions, nouns, verbs or causal constructions (Altenberg 1984, Degand 2001, Hoey 2005, Stukker 2005). These different causal devices code causality in different formal patterns within a discourse and they express differences in volitionality, animacy, directness of causation, etc. (Kemmer and Verhagen 1991) In this study, we use parallel and monolingual corpora to contrast the coding of causal relations in different languages. For instance, English *lead to* is often accompanied by modals and can be used with outcomes that are positive or negative as shown in the examples below. The Czech *vést k* has followed a similar grammaticalisation path from spatial to causal meaning and can be used in a similar way to English *lead to*.

1. His apology **led to** a bond between them stronger than any they ' d ever had before.
   Jeho pokora **mu pomohla vytvořit** mezi nimi pevné pouto .
   'His humbleness helped him to create a bond.'
2. "You had information that might have **led to** Lozada ' s arrest and you failed to come forward with it ."
   „ Měla jste informaci , která mohla **vést k** Lozadovu zatčení , a neoznámila jste ji .“

In this presentation we examine contrasts in causal expressions in English-Czech and English-Chinese and explore the notion that different patterns can be distinguished on the basis of the animacy of the arguments and nature of the predicates.

The Value of Relational Databases for Time-Aligned Annotation

Tyler Kendall
Northwestern/NC State University

Recent work in a number of linguistic disciplines has stressed the importance and utility of time-aligned annotation for linguistic corpora (cf. Bird and Liberman 2001) and increasingly projects are making use of time-alignment for areas of annotation such as transcription. At the same time, XML has emerged as a popular technology for structuring various sorts of corpus data, more so than any other data management technology (cf. Gries 2009; e.g., Simons, Fitzsimons, Langendoen, Lewis, Farrar, Lanham, Basham, and Gonzalez 2004). With few exceptions (namely, Davies 2005), relational database engines, such as MySQL or PostgreSQL, have not been used for the storage and manipulation of linguistic corpora.

In this talk, I discuss the approach to time-aligned transcription and annotation implemented for the Sociolinguistic Archive and Analysis Project (SLAAP; http://ncslaap.lib.ncsu.edu/). SLAAP makes use of a (MySQL) relational database to store and interface with all aspects of the annotation – from transcription, to analysts' research notes, to the extraction and coding of features for analysis (such as sociolinguistic variables). By making use of a number of open-source tools, SLAAP connects the time-aligned, databased annotation with the original speech recordings allowing for such features as dynamic, user-customizable transcript views, audio-playing capability, and "on-the-fly" phonetic analysis directly from the transcripts.

I argue that the database-driven approach has some advantages over XML- or text-based strategies and illustrate this through examples and discussions of features of the SLAAP software. While XML is clearly useful for data *sharing*, I demonstrate the benefits of relational databases for the *storage* and *manipulation* of corpus data. However, I also propose ways that databased corpora may be shared without the need for intermediate formats.

The Dative Alternation in African American English: Researching Syntactic Variation and Change in a Conglomerated Sociolinguistic Corpus

Tyler Kendall[1], Gerard Van Herk[2], and Joan Bresnan[3]
[1]Duke University/NC State University; [2]Memorial University of Newfoundland; [3]Stanford University

Recent research shows that the dative alternation in English, as in (1), is a productive arena for examining the relationship between group-level variation and the internalization of individuals' grammars.

(1)	a. *Who gave that wonderful watch* ***to you****?*	prepositional (*to*-)dative
	b. *Who gave* ***you*** *that wonderful watch?*	double object construction

Experimental methods (e.g., Bresnan and Ford submitted) and the analysis of large published corpora (e.g., Bresnan et al. 2007) have revealed subtle cross-dialect differences for this variable. The current paper seeks to improve our understanding of this feature and its bearings on experience-based models of grammar (e.g., Bybee 2001; Jurafsky 2003) by examining spoken and written African American English (AAE) data from a number of sources.

AAE has long been a central object of study in North American sociolinguistics (e.g., Wolfram 1969; Labov 1972) inspiring five times as many publications as any other ethnic or regional dialect (Schneider 1996:3). Yet, the data from these studies often remain out of public use (cf. Kendall 2008) and, beyond sociolinguistics, AAE is rarely used to address central linguistic questions. In fact, unlike the data used for many corpus-based analyses – and used thus far for studies of the dative alternation – there are no definitive or publicly available transcribed corpora of AAE available through organizations like the Linguistic Data Consortium.

In this project, we compile a number of less conventional corpora (cf. Beal, Corrigan, and Moisl 2007; e.g., SLAAP [Kendall 2007], OREAAC [Van Herk and Poplack 2003]) to conduct a historical and comparative analysis of the dative alternation in AAE. We discuss how our current findings may illuminate previous work on the dative alternation, but focus on our data compilation methods and the benefits to theoretical linguistic work of examining more diverse data sources.

Interclausal Semantic Functions of the –E Connectives in Korean

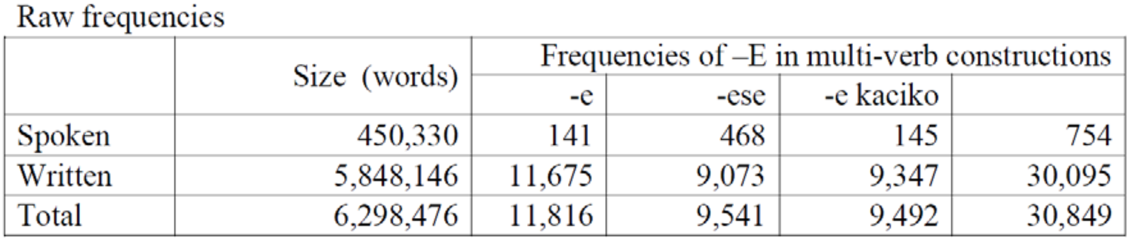
Sangbok Kim
University of California Los Angeles

There have been recent debates on the functions of the connectives –E (i.e., -*e*, -*ese* and –*e kaciko*), a group of suffixes attached to the first verb of a multi-verb construction in Korean, where two full lexical verbs occur in sequence with a shared subject. In previous studies where abstract theoretical ssumptions with individual/artificial sentences were the main resources for analysis, the various forms of -E were treated as clausal connectives in free variation appearing alternatively without being restricted in the same interclausal semantic relation (e.g., cause-result). At the same time, it is claimed that these variant forms have different functions in spoken and written registers. For example, -*e kaciko* is said to be more frequent in spoken discourse than in written discourse.

Analyzing 30,849 multi-verb constructions in the discourse context culled from a corpus of 6,298,476 words of the 21st C. Sejong Project Modern Korean Corpora built by Korean government, and drawing on 'interclausal semantic relations' in Role Reference Grammar (Van Valin & Wilkins 1993; Van Valin 2002, 2005), this study demonstrates that the –E forms do not function as clausal connectives in free variation. Instead, the findings of this study show that the three different forms of –E are clausal connectives with different semantic functions in discourse, reflecting the 'iconicity principle' (Haiman 1983). The shortest form of -e tends to connect two clauses that have the 'closest' semantic relations, the medial form –*ese* is used for a range of semantic relations, and the longest form –*e kaciko* is for the 'loosest' clausal relations. As for the use of the variant forms regarding spoken and written registers, it is shown that the –*ese* has a spoken tendency, at least compared to the other forms.

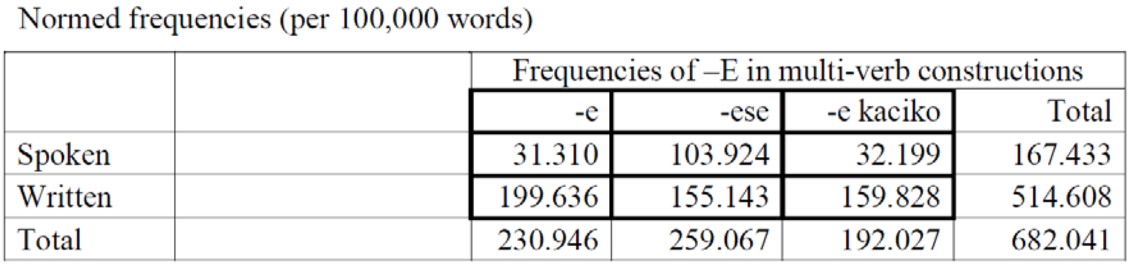
The findings of this study suggest that looking at actual usage by native speakers can bring new perspectives to the description of Korean multi-verb constructions. More importantly, they can contribute to discussing the existence of serial verb constructions in Korean. Recent cross-linguistic studies have purported that the multi-verb constructions are serial verb constructions.

Finding 1: Different functions of the –E in spoken and written registers



Raw frequencies

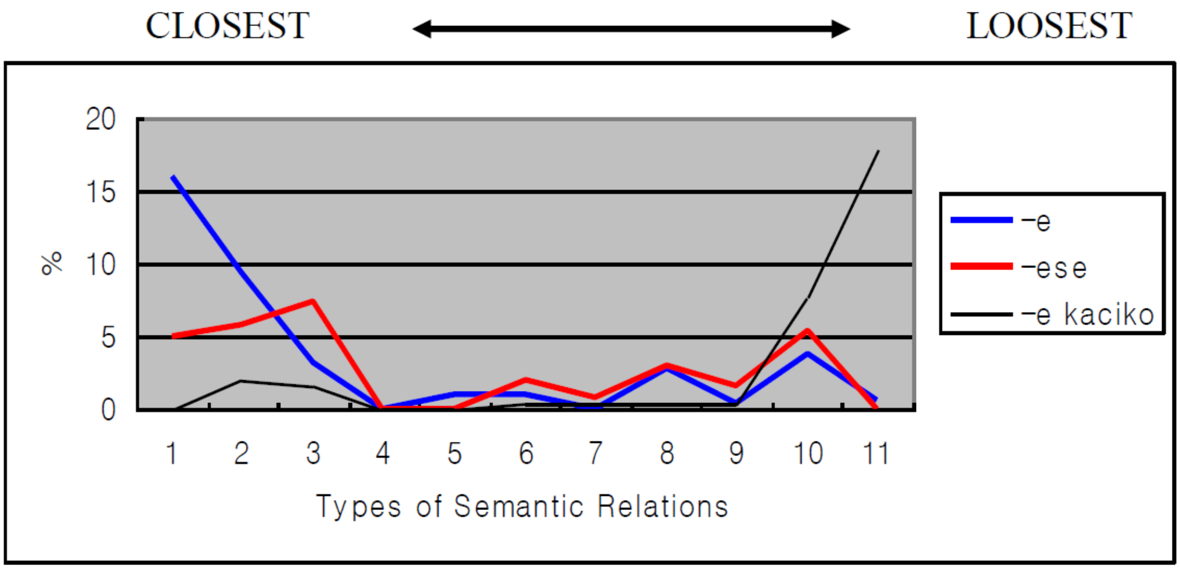
	Size (words)	Frequencies of –E in multi-verb constructions			
		-e	-ese	-e kaciko	
Spoken	450,330	141	468	145	754
Written	5,848,146	11,675	9,073	9,347	30,095
Total	6,298,476	11,816	9,541	9,492	30,849



Normed frequencies (per 100,000 words)

		Frequencies of –E in multi-verb constructions			
		-e	-ese	-e kaciko	Total
Spoken		31.310	103.924	32.199	167.433
Written		199.636	155.143	159.828	514.608
Total		230.946	259.067	192.027	682.041

Finding 2: Distributional patterns of the three different forms of –E



Explanation of 'Types of Semantic Relations':

1 = Manner-(Path)-Locomotion
2 = Manner-Action
3 = Tightly-bound actions in sequence
4 = Psycho-Action
5 = Purposive-Locomotion
6 = Causal-Resulting action/state
7 = Inchoative-Resulting states
8 = Non-overlapping immediately following actions in sequence
9 = Reason-Action/Event
10 = Non-overlapping-with-an-interval actions in sequence
11 = Comitative-Locomotion

Emergent Patterns of Conjunctive Adverbial *Though* in Academic Spoken English: A Corpus-based Study

So Yeon Kim
University of California, Los Angeles

This study investigates the use of *though* as a conjunctive adverbial in academic spoken discourse. Whereas much attention has been given to the usage of *though* as a concessive subordinator, it is only recently that the conjunctive adverbial *though* has become a topic of interest. A series of studies on the concessive patterns in conversation have provided a particularly detailed analysis of the conjunctive adverbial *though* in clause-final position (Couper-Kuhlen & Thompson, 2000; Barth-Weingarten, 2003). It is argued that this adverbial is essentially an abbreviated version of subordinator *though*-marked concessive clauses but downplays the conceding move prosodically and lexico-syntactically. As such, it is considered as an item belonging to the 'inner periphery of concession' on the adversativity-concession continuum. The current study expands on these previous studies by examining more diverse types of spoken discourse other than conversation. The analysis of over 250 tokens of *though* selected from the Michigan Corpus of Academic Spoken English (MICASE) reveals that the occurrence of the conjunctive adverbial *though* in clause-medial position is not uncommon in spoken discourse and even more so in monadic speech. In addition, the position of the medial *though* is not random but frequently marks the boundary between given information and new information. This study also argues that the effect of mitigating the degree of disagreement, which has been suggested as the primary function of the conjunctive adverbial *though* in the previous studies, should be explained by examining the sequential environment in which the form is placed, rather than by assuming the derivational relationship between the conjunctive adverbial *though* and the subordinator *though*. For example, it is frequently observed that the use of the conjunctive adverbial *though* in the disagreement context is accompanied by the employment of question formulations, hedges, and adverbs downgrading the contradictory force.

References:

Barth-Weingarten, D. (2003*). Concession in spoken English: On the realisation of a discourse-pragmatic relation.* Tübingen: Gunter Narr Verlag.

Couper-Kuhlen, E., & Thompson, S. A. (2000). Concessive patterns in conversation. In E. Couper-Kuhlen & B. Kortmann (Eds.), *Cause, condition, concession and contrast: Cognitive and discourse perspectives* (pp. 381-410). Berlin, New York: Mouton de Gruyter.

The Representation of the Elderly in Taiwanese Newspapers: A Corpus-Based Study

Sai-hua Kuo
National Tsing Hua University, Taiwan

Although there have been a number of studies on how elderly people are portrayed in magazine advertisements and television programs (e.g. Bell 1992, Miller et al. 1999, Roy and Harwood 1997), there is a dearth of current research on how they are represented in newspapers and most of these existing studies have been based on data from Western societies. This corpus-based research aims to discover how the elderly are presented and how that presentation may vary over time in Taiwanese newspapers. In addition, how male elders and female elders are portrayed differently in newspapers and how the changing representation of the elderly is related to the socio-cultural context will also be explored.

Data for this study are from three newspapers which circulate most widely in Taiwan, the *United Daily News*, the *China Times*, and the *Liberty Times*. A total number of 240 news stories sampled from four months, i.e. July, August, September, and October in the year of 2007. In addition, 157 news stories from the same period in the year of 1997 from the *China Times* are included for a chronological comparison.

My preliminary investigation has found that age-related news stories tend to appear in less prominent space, which usually devoted to local or regional news. Moreover, the number of age-related news has decreased significantly over time despite the growing population of the elderly. On the other hand, the chronological comparison has revealed that older adults were presented in a more positive light—as kind, active, happy, and healthy—than they had been in the past. These results indicate that although older adults in Taiwan's newspapers have been more positively portrayed over time, they are even more underrepresented. It is also found that although males were predominately the subjects of age-related news stories, females were pictured more favorably. Finally, the frequent reporting of family tragedies (e.g. murder, abuse, abandonment) caused by conflict between the elderly and their offspring further indicates the changing role and status of the elders in the changing Taiwanese society, in which the traditional Confucian ethical principle of filial piety is diminishing.

Like the experts? A multi-corpus approach to the features of Chinese research writing in English

David Y.W. Lee
City University of Hong Kong

In light of contemporary debates within the English language teaching profession about world Englishes and the role of English as a lingua franca, this paper aims to empirically uncover a number of discoursal and lexico-grammatical features that characterize Chinese academic research writing in English, with a view to contributing to the formal codification of its distinctive characteristics. A multiple-corpus-comparison approach, involving specialized corpora of research writing in English by mainland Chinese research-writing apprentices, "published experts", and "native-speaker apprentices", all in the same discipline, was used to determine a number of specific lexico-grammatical and discoursal choices that may either be addressed in EAP writing instruction in China if seen as problematic, or included in a new lingua franca 'canon' if seen as having wider, international acceptance. Previous studies of the English of Chinese writers have been either non-comparative in nature, or based largely on argumentative or narrative essays written specifically for exams, tests or proficiency-focused courses. Such essays are different in many respects from research writing, which requires the exercising of many text-rhetorical skills not relevant to other genres: for example, how to talk about data, questionnaires, methods and results; how to refer to tables and figures; and how to cite or use primary and secondary sources as part of argumentation. Other more general writing strategies, such as how to suggest or make recommendations, or how to hedge or emphasize claims, are also subtly different in scholarly articles compared to argumentative or narrative essays. This paper presents a research-writing-focused study that integrates quantitative analyses of genre-comparable corpora and qualitative investigations of texts and their generic structure, and is additionally informed by contrastive analyses of the two relevant languages. The paper presents a range of features of "non-nativeness" (not only errors, but also instances of characteristic under- and overrepresentation of words, phrases and structures) that, it is hoped, will contribute to a line of research leading to more empirically-based debates on what constitutes "China English" in the high-stakes genre of research writing.

Age Tagging and Word Frequency for Learner's Dictionaries

Hanhong Li and Alex C. Fang
City University of Hong Kong

The use of corpora for word frequency information is unquestioned in contemporary lexicography, particularly in learner's dictionaries. A review of the current practice shows that word frequency information is used for entry selection, sense ranking, and collocation identification as well as definition vocabulary. For example, *Longman Dictionary of Contemporary English 5th edition* (2009) specifies 3000 most frequent words (known as *Longman Communication 3000*) for spoken and written English respectively, selected according to the computerized analysis of the Longman Corpus Network. *Oxford Advanced Learner's Dictionary* (7th Edition 2007), as another example, has its own defining vocabulary which is composed of three thousand most frequent, widely used and familiar words based on the British National Corpus (BNC).

However, age information in linguistic corpora has not been adequately highlighted or exploited. Early experiments (Carroll and White 1973) have demonstrated that word retrieval in long-term memory is much more influenced by the age of acquisition than word frequency. For English learners in non-English speaking countries, it is necessary for them to know what words are used by native speakers at different ages besides those common words in frequential terms. Common words are not simply those with high frequency but also the ones with an even distribution in different age groups according to authentic use as recorded in corpora. If learner's dictionaries can incorporate a word profile based on frequency of occurrence and distribution across different age groups, it will be of much help to English learning and teaching as well as research in the formation and composition of core vocabulary. Moreover, it will open a new area in word frequency studies in lexicography.

Our research will take advantage of the age group information in BNC xml (2007). The age group information for the spoken part of BNC xml is indirectly tagged so that it is not that easy to extract words according to age group information. Some researchers even mistakenly took the age information for the respondents (the collectors of recording material) for the age of the speakers. In order to extract word units according to age group information in the spoken BNC xml, we modify its utterance tagging by replacing <u who="speaker's ID"> with a more detailed uniformed pattern <u who="ageGroup=X xml:id=X role=X sex=X soc=X dialect=X firstLang=X educ=X n=X">. In this case, it is much easier and accurate to collect data for further analysis. With the above modification, we will explore the word frequency distribution in different age groups, specify the word frequency with age information in dictionaries, and select the core vocabulary used by native speakers at all ages.

References:

British National Corpus XML Edition (cd-rom), 2007.

Carroll J. B., White, M. N. Word frequency and age-of-acquisition as determiners of picture-naming latency. *Quarterly Journal of Experimental Psychology*, 1973, 25: 85-95.

Longman Dictionary of Contemporary English 5th edition (cd-rom), 2009.

Oxford Advanced Learner's Dictionary 7th Edition (cd-rom), 2007.

Is it a *chief, main, major, primary,* or *principal* concern? A corpus-based behavioral profile study of the near-synonyms and its implications

Dilin Liu
University of Alabama

Using the Brigham Young University's Corpus of Contemporary American English as the source data and employing a corpus-based behavioral profile approach supplemented by a close reading of some concordance data, this study examines the internal structure of a set of five near-synonyms (*chief, main, major, primary,* and *principal*) by focusing on their distributional patterns, including, among others, their usage variance across the different types of nouns that they modify, their difference in the singular/plural and definite/indefinite uses, and their distribution variation across different registers. Based on a close analysis of the various data including pertinent statistical tests, the study has identified several important fine-grained semantic and usage differences among the five near-synonyms and produced a meaningful delineation of their internal structure. Furthermore, while the study has produced findings that support many of the dictionary definitions and descriptions about the adjectives, it has also yielded new findings, including those that challenge some of the existing understandings about the adjectives' meanings and usage patterns. The new findings can help develop a more adequate and accurate description of this near-synonym set, which in turn can assist language users in better grasping the use of the synonyms. Finally, the overall results of the study offer new evidence for the assumed correlation between near-synonyms' distributional and semantic similarity and demonstrate the effectiveness of the corpus-based behavioral profile approach in the study of near-synonyms and the dynamic nature of the micro procedures used in the approach.

Beyond the dictionary: creating a long-term corpus resource

Deryle Lonsdale
Brigham Young University

This paper discusses follow-on work that has been carried out in developing, deploying, and improving a 23-million word corpus of French. Our recently completed frequency dictionary for learners of French is based entirely on this balanced spoken/written corpus of recent language. We collected the corpus texts from a wide array of sources, cleaned and tokenized them, lemmatized and POS-tagged all of the words, and performed statistical analyses to find the core vocabulary to serve as the focus of the dictionary. Words were ranked and selected via a metric (“dispersion of proportions”) introduced at last year’s AACL meeting.

Though primarily a lexicographic effort, the work obviously involved integrating several different language resources and tools, therefore requiring extensive programming support. Now that the work is completed and the dictionary in print, we debated whether to re-deploy the corpus in a manner more amenable for further research.

Several considerations went into the decision to re-engineer the corpus. For example, only one person had done all of the programming and corpus tools integration work, and the methodology was not easily amenable to use by other users. In addition, the effort combined several different tools, command-line applications, various programming/scripting languages, and language resources. Finally, no similar corpus currently exists for French. For these reasons and other we decided to undertake the effort to standardize and store the corpus data in a more usable and flexible way.

In this paper we address how the corpus was thus re-engineered and give examples of the types of data represented. We also show how several annotations are being incrementally cleaned by targeted human intervention. We also discuss plans for representing further layers of linguistic annotation. Finally, we show examples of interesting linguistic phenomena that can be retrieved from the corpus, and how this is possible given current best practices in corpus creation, representation, and usage.

Apposition from the Perspective of Construction Grammar

Charles F. Meyer
University of Massachusetts Boston

Apposition has proven to be a problematic grammatical category, largely because treatments of it disagree about which constructions should be considered appositions. For instance, most studies consider a construction such as *Geoffrey Plimpton, police commissioner* as consisting of two units in apposition. However, if the two units are reversed, a reversal possible with some but not all appositions, a very different construction, *police commissioner Geoffrey Plimpton*, results—one that some studies favoring a more expansive view of apposition consider appositional (e.g. Meyer 1992), but that those favoring a more restricted view do not (e.g. Acuña 1996). These conflicting views are each problematical. The former resorts to a series of syntactic and semantic gradients to distinguish various 'degrees' of apposition. The result is the admission of constructions such as *a person like you* into the category of apposition, a construction so different from typical examples of apposition that it renders the notion of apposition almost meaningless. The latter view is so restrictive that any construction that does not contain juxtaposed noun phrases separated by an intonation boundary is not considered appositional.

In more recent work, Acuña (2006) attempts to bridge the gap between these two views. Working within the framework of construction grammar/cognitive linguistics, he views the many constructions considered appositional as occupying what he terms 'appositive space', a space where the differing types of appositions are related through notions such as 'family resemblance' and 'prototype'. In developing his argument, Acuña (2006) works primarily with invented examples or examples taken from books or articles discussing apposition. In my presentation, I will extend Acuña's ideas to the analysis of corpus data. My discussion has two parts. First, I very briefly describe the complicated nature of apposition. I then evaluate the two competing ways of describing the variable structure of appositions: viewing them as being on gradients between full and partial apposition, or regarding some as more prototypical than others. I argue that while both ways help explain the variable nature of apposition, discussing apposition from the perspective of construction grammar provides a more comprehensive view of the nature of apposition, since within construction grammar, constructions are viewed as '"vertical" structures' (Croft and Cruse 2004: 247); that is, structures that are defined not simply syntactically but semantically, phonologically, and pragmatically as well.

To illustrate this point, I use corpus data and frequencies taken from Meyer (1992) and ICE-GB to demonstrate that one of the more frequently occurring types of apposition, the so-called nominal apposition, has resemblances to what Fillmore, Kay, and O'Connor (1988: 505) label formal idioms: '…syntactic patterns dedicated to semantic and pragmatic purposes not knowable from their form alone'. For instance, in his study of the form of appositions in various registers of spoken and written English, Meyer (1992: 11) found that 55% consisted of two noun phrases. But a further examination of the particular form of the two NPs revealed some dominant structures:

(1) 44% of nominal appositions contained one unit that was a proper noun (e.g. *my friend, John Smith*)
(2) 63% the nominal appositions containing proper nouns occurred with a unit beginning with a definite NP (as in the example above) or an NP with an implied determiner (e.g. *Thomas Menino, mayor of Boston*). In 99% of the units whose determiner was implied, the implied determiner was definite (e.g. *Thomas Menino, [the] mayor of Boston*).
(3) The overwhelming number of appositions with proper nouns occurred in press reportage (Meyer 1992: 117), suggesting that this kind of apposition has a strong pragmatic function.

What these figures illustrate is that prototypical appositions have a fairly fixed form and a very specific pragmatic function. I will further demonstrate that the less frequently occurring types of appositions have a more varied range of forms and therefore a more diffuse relationship to prototypical appositions.

References:

Acuña Frariña, Juan Carlos (1996) *The Puzzle of Apposition: On So-Called Appositive Structures in English*. Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico.
----- (2006) ‘A Constructional Network in Appositive Space’. *Cognitive Linguistics* 17 (1). 1-37.
Croft, William and D. Alan Cruse (2004) *Cognitive Linguistics*. CUP.
Fillmore, Charles J., Paul Kay, and Mary Catherine O'Connor (1988) ‘Regularity and Idiomaticity in Grammatical Constructions: The Case of *Let Alone*’. *Language* 64 (3): 501-538.
Meyer, Charles F. (1992) *Apposition in Contemporary English*. CUP.

A Corpus Analysis of Scandinavian Legal Concepts in Anglo-Saxon Laws

Viola G. Miglio
University of California, Santa Barbara

The access to large amount of data from different periods provided by corpora such as *The Dictionary of Old English/Old English Corpus* (DOE Project, University of Toronto), with at least one copy of all existing OE documents (3037 documents in total), allows researchers to study the frequency of occurrence of the legal concepts derived from ON/OE contact and approach their semantic evolution in an exhaustive way. The present paper follows the development and frequency of words such as *utlah* 'outlaw' (13 examples in the corpus, from the XI onwards), and *lahslit* 'breaking of the law, crime', which starts appearing in Æthelred's Laws from the late 10th century (2 instances), and its use increases with the Danelaw's influence on English life from the year 1000 onwards (substituting the autoctonous *æwbryce* whose meaning specializes into 'violation of matrimony', i.e. adultery, 9 cases in the whole OE corpus). Specialisation is also found in pairs such as *frit/frid* (common Germanic, found early on, for instance in the 10th century Exeter Book, 671 examples in the DOE) and *grid/grit* (of Scandinavian origin, typically encountered from 11th century onwards, 177 instances in DOE): they could both be translated as 'peace', but the first is a generic term, whereas *grid* is typically more limited and refers to the peace afforded by a place of refuge, a sanctuary.

Studying OE legal terminology of Scandinavian origin using a corpus approach also facilitates highlighting subtler shifts of meaning: the OE word *mord* ('death, murder') acquires after Scandinavian contact the specific meaning of 'furtive, secret killing'. In Scandinavian society, where the enforcement of punishment was often left to the family of the victim, this was a particularly heinous crime, underlining the importance of openly admitting to a crime, rather than condemning the violent aspect of the deed, just as in ON one finds the difference between the heinous *mord* and the honourable *víg*. This paper shows that a corpus approach using DOE allows researchers to track early Scandinavian legal concepts as attested in OE texts, assess how their frequency increases with the contact and cultural influence from the Danelaw, as well as follow more subtle semantic changes in the usage of already existing OE words.

Narrative function(s) of tense switching: a corpus-based application to medieval Icelandic sagas

Viola G. Miglio and Stephan Th. Gries
University of California, Santa Barbara

Weinrich (1964) suggested that tense switching (TS) in spoken discourse in some Indo-European languages often serves the function of manipulating the addressee's allocation of attention. (Parts of) Texts requiring attentive and critical attitude are typically cast in non-past tenses, whereas descriptive (parts of) texts with mostly background information use past tenses. This association creates expectations in the recipient as to which tenses belong in which part of the narrative, and TS suggests a conscious manipulation of the text by an authorial hand.

In our paper, we study this association between tenses and informational states and its function in medieval Icelandic sagas using a corpus-based approach. We study the hypothesis that TS fulfills an alerting role, preparing the text recipient (a contemporary audience of the orally performed text) to take an active stance towards the narrative. We undertook a statistical analysis of the distribution of all verbs in present/past tense in /Hrafnkels saga freysgoða/ and their occurrence in narrative turning points. The data exhibit a statistically significant association and show that the distribution of tenses in medieval texts is not haphazard, as was traditionally maintained, but a performative device related to reader/audience response: TS does not simply encode temporal information but, much like ominous music in movies, alerts the reader of an impending momentous event.

The findings have several impliations. First, they are compatible to current theories of the origin of these texts (Würth 2000). Second, for the area of saga scholarship, we show that authors did not just use written sources, but that they were aware of the use of TS in oral narratives (and therefore used some oral sources), and wrote a text to be performed, a little explored aspect of the sagas. Third, they show that TS is not just found in medieval Romance (Fleischmann 1991), but is typologically more widespread. Finally, it exemplifies how corpus-based methods can contribute to the study of historical narratives.

The role of task type in a learner corpus

Colleen Neary-Sundquist
Purdue University

This study examines the use of cohesive devices (discourse markers and conjunctions) in a 24,000 word corpus of transcribed oral data from 40 learners of English. The corpus was created from four different tasks performed on an oral proficiency test. The results show that ESL learners at all levels used more conjunctions than discourse markers. Overall discourse marker and conjunction use both increase steadily with proficiency level, and the two higher proficiency levels were significantly different from the two lower levels in their use of cohesive devices (*p<.01*). These results demonstrate the efficacy of using a learner corpus to investigate how second language learners construct textual coherence.

When the data are examined by the task performed, interesting patterns concerning the interaction of task type and proficiency level can be observed. On a task in which the subjects were asked to give their opinion on a news item, discourse markers and conjunctions were used at almost the same rate. On a telephone message task, however, the learners' use of discourse markers was significantly different (*p<.05*) from other tasks, and they used far more conjunctions than discourse markers. These results highlight the importance of considering the type of task that the data is drawn from when creating a SLA corpus.

Pragmatic annotation in an international corpus of engineering lectures

Hilary Nesi, Ummul Ahmad, and Noor Mala Ibrahim
Universiti Teknologi Malaysia

Although a number of studies have examined lecture discourse from a pragmatic perspective (e.g. Bamford 2005, Crawford Camuciottoli 2008, Fortanet Gomez 2004, Morell 2004), pragmatic annotation of lectures tends to be limited to small corpora created by scholars for their own personal use. The two major spoken academic corpora in the public domain, BASE and MICASE, are not encoded for pragmatic features for various reasons: such annotation is costly in terms of time, and it is also interpretative, fixing utterance meanings in ways which corpus users must either accept or challenge. A section in the MICASE Handbook (Simpson & Leicher 2006), however, provides pragmatic information to supplement the standard MICASE mark-up. Building on this, and working with a corpus of lectures which exhibit a fairly limited range of pragmatic features, we are developing and testing a more robust system of annotation which facilitates the comparison of lecture delivery styles in different cultural settings.

For this study our corpus is the Engineering Lecture Corpus (ELC), a growing collection of English-medium university lectures on similar topics from different parts of the world (currently the UK, Hong Kong, Malaysia and New Zealand). So far, twenty-seven features have been identified and encoded; these include three types of referencing (to the external world and within and between lectures), enumeration, directives, and exemplification. Transcripts are aligned with the video component to enable searches for pragmatic features via SACODEYL, a search tool developed with funding from the European Commission. Differences in delivery style are startling, and on completion the corpus package should constitute a useful resource for engineering lecturers and students crossing continents for work or study.

The Adjusted Frequency List: Evaluating a method to produce cluster-sensitive frequency counts

Matthew O'Donnell
University of Michigan

Many corpus-based studies make use of frequency lists of various size chunks (called clusters, bundles or n-grams). Most software packages facilitate the creation of such lists, making it possible to compare units of difference length. However, each size unit is counted on its own terms. For example in a combined 1- and 2- gram list every instance of *know* is counted individually even if it is always preceded by *you*, thus *you know* and *know* have the same frequency in the list. But *know* has no independent occurrences to be counted. The issue also applies to larger units where collecting 3, 4 and 5 grams together will result in very similar and often identical counts for *at the end*, *the end of*, at *the end of*, *the end of the*, *at the end of the*, *the end of the day* and so on.

Frequency lists of items of various lengths are important in both computational and applied linguistics. Chunks of two or more words are key in the description and teaching of vocabulary. O'Keeffe *et al*. highlight the fact that 'many chunks are as frequent as or more frequent than the single-word items which appear in the core vocabulary' (2006: 46). They are also valuable for measuring the idiomatic/formulaic nature of text (Erman & Warren 2000).

The index-based method of constructing a frequency list proposed here adjusts the frequency of items of various lengths if they are part of a larger unit that occurs at or above a given frequency threshold. That is, if *you know* occurs 15 times in a corpus and *know* 20 times then the frequency of *know* will be adjusted to 5. The method outlined is 'cluster sensitive' because it boosts the rank of larger word sequences. A number of comparisons of unadjusted (standard) and adjusted frequency 1- to 5-gram lists are given using both spoken and written corpora (BNCBaby and MICASE) and developmental data from the CHILDES database.

References:

Erman, B. & Warren, B. 2000. 'The idiom principle and the open choice priniciple'. *Text* 20 (1): 29-62.
O'Keeffe, A, McCarthy, M. & Carter, R. 2006. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

On the uniqueness of casual spoken discourse – insights from the National Corpus of Polish

Piotr Pezik
University of Lodz

The National Corpus of Polish (NKJP) is a joint project carried out by a consortium of four Polish corpus teams, aimed at compiling a 1 billion words corpus of modern Polish, of which at least 300,000,000 will form a carefully balanced sub-corpus. A 30, 000,000 word spoken component will contain at least 1,500,000 words of casual spoken-conversational Polish.

In this paper I first describe the methodology of spoken data acquisition, transcription and annotation in the abovementioned project. I give an overview of the annotation guidelines used, and introduce some of the technical and legal aspects of recording authentic conversations.

Next, I focus on the characteristics of *Casual Spoken Discourse* (CSD) emerging from the data collected so far. One observation made from the data at hand is that CSD is very different from other registers of spoken language, such as the mediaconversational discourse also represented in NKJP. The samples analyzed so far reveal a number of idiolectic and familiolectic features of CSD, which are not to be found in other registers of spoken Polish. I also show that CSD is characterised by a high level of topical volatility and parallelisation. Given the density of CSD and the fact that speech transcriptions are particularly prone to semiotic impoverishment, I argue that for many aspects of spoken discourse analysis it is mandatory to link the transcriptions with the original recordings.

I conclude that due to its distinctive character, casual spoken discourse should be studied in its own right and that it deserves to be represented as a separate register in modern reference and national corpora.

The C-ORAL-BRASIL corpus

Tommaso Raso and Heliana Mello
Universidade Federal de Minas Gerais, Brazil

The C-ORAL-BRASIL is a spontaneous speech corpus of Brazilian Portuguese which follows the same architecture and segmentation criteria employed in the C-ORAL-ROM, a corpus which groups the four main European Romance languages: French, Italian, Portuguese and Spanish (Cresti & Moneglia 2005).The main features of the corpus along with its current state are:

1. 200 spontaneous speech texts amounting to 1,500 words each, mostly representative of the Mineiro dialect, divided into an informal half, already completed, and a formal one. The capture of data was carried with high sensitivity equipment that allow for prosodic analysis. 80% of the informal half represents private/family contexts and 20% public contexts. It is constituted by 1/3 dialogues, 1/3 conversations and 1/3 monologues;

2. texts are chosen so that they systematically represent diaphasic variation and, to a lower degree, also diastratic variation. A large situational variation is necessary to represent structural and illocutory complexities found in speech (eg., mason and engineer at construction site, women grocery shopping, woman shoe shopping, four peers discussing about a school chore, four friends playing a table game, individuals narrating life stories, talking about work, telling a fable to a child, etc.);

3. texts are segmented into utterances (based on the perception of terminal breaks which carry an illocution) and tone units (based on the perception of non-terminal breaks that in principle correspond to information units). The agreement among transcribers after a training period reached a 0.84 kappa (0.91 for terminal breaks). Half of the corpus has been already transcribed;

4. texts are aligned through the WinPitch software (Ph. Martin), which allows for the simultaneous analysis of text, sound, spectrogram and F0 curve;

5. transcriptions take into consideration the features of Brazilian Portuguese that are likely to be undergoing lexicalization and grammaticalization processes.

"And new meanings turned up:" The development of new meanings of English phrasal verbs with *up*: Evidence from the *Helsinki Corpus* and *ARCHER1*

Paula Rodríguez-Puente
University of Santiago de Compostela

English phrasal verbs or particle verbs are combinations of a verb and an adverbial particle which function, to varying degrees, as a single syntactic and semantic unit. Many scholars (Quirk et al. 1985; Biber et al. 1999) consider it a prerequisite that a verb-adverb combination must have an idiomatic meaning in order to be considered a phrasal verb. However, idiomatic, non-compositional meanings of phrasal verbs do not appear suddenly but rather possess a literal origin. Thus, as noted by Hiltunen (1983: 149) literal meanings of phrasal verbs were predominant in Old English, although some metaphorical uses had already started to develop at this early stage.

According to Thim (2006: 221-225), phrasal verbs can be divided into five different semantic groups, namely literal (*blow down*), partly literal (*send up*), pleonastic (*lift up*), figurative (*wrap up*) and non-compositional (*call down*). One of the aims of this paper is to explore how phrasal verbs with *up* adjust to these (or other) semantic types. I also intend to investigate how the particle *up*, whose original meaning is that of upward movement (cf. *OED* s.v. *up* adv. 1.a.), has become the "aktionsart particle par excellence" (Denison 1985: 37) or is used in combination with verbs for the formation of new phrasal verbs with an opaque, idiomatic meaning (e.g. *give up*). For my purposes, I will analyse the occurrences of phrasal verbs with *up* in the *Helsinki Corpus* and *ARCHER1*, two complementary multi-genre corpora of English which together cover the span 750-1990 and can, therefore, be used to study the development of linguistic phenomena through time.

References:

ARCHER1: *A representative Corpus of English Historical Registers 1650-1990*. Compiled by Douglas Biber and Edward Finegan. Northern Arizona University.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Denison, David. 1985. "The Origins of Completive *up* in English." *Neuphilologische Mitteilungen* 86: 37-61.

HC=Helsinki Corpus of English Texts 850-1710. 1991. Compiled by Matti Rissanen et al. Department of English, University of Helsinki.

Hiltunen, Risto. 1983. *The Decline of the Prefixes and the Beginnings of the English Phrasal Verb* (=*Annales Universitatis Turkuensis,* 160). Turku: Turun Yliopisto.

OED = Oxford English Dictionary. 1989. 2nd edition. Oxford: Oxford University Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London and New York: Longman.

Thim, Stefan. 2006. "Phrasal Verbs in Late Middle and Early Modern English: Combinations with *back*, *down*, *forth*, *out* and *up*." In Christiane Dalton-Puffer, Dieter Kastovsky, Nikolaus Ritt and Herbert Schendl (eds.) *Syntax, Style and Grammatical Norms: English* *from 1500-2000*. Bern: Peter Lang: 213-228.

Phraseological items in apprentice academic writing: Does nativeness matter?

Ute Römer
University of Michigan

Nowadays, a large and growing number of academic English texts (e.g. research articles, book reviews, dissertations) are written by non-native speakers of English. While the research world is becoming more and more Anglicized and large numbers of "non-Anglophones" (Swales 2004: 46) produce academic English alongside their native-speaker colleagues, it is yet unclear what status *nativeness* has in this context and what challenges novice or apprentice academic writers whose first language is not English have to face.

This talk will address the issue of nativeness and examine what the native/non-native-distinction means in the context of English academic writing. It will investigate how different the academic writing of native speakers and non-native speakers of English is and, based on comparisons of apprentice and expert performance data (in Bazerman's 1994: 131 terms), discuss whether nativeness has an effect on academic writing proficiency if other potentially influential factors like genre, discipline, and duration of university education are controlled.

The focus of the analyses reported on in this talk is on frequent phraseological items, e.g. word combinations such as *on the one hand* or *in the case of*, that are typical of academic writing, in comparable sets of successful apprentice academic writing by native speakers and non-native speakers of English in the disciplines of Linguistics and English (language and literature). Phraseological items (n-grams and phrase-frames of different lengths) have been extracted from the Cologne-Hanover Advanced Learner Corpus (CHALC, see Römer 2007) and from a subset of the Michigan Corpus of Upper-level Student Papers (MICUSP, see http://micusp.elicorpora.info). A collection of published expert academic writing (research articles from Linguistics journals) functions as a reference corpus and is regarded as a kind of target norm for our apprentice writers.

References:

Bazerman, C. (1994). Systems of genres and the enactment of social intentions. In: A. Freedman & P. Medway (eds.). *Genre and the New Rhetoric*. London: Taylor and Francis. 79-101.

Römer, U. (2007). Learner language and the norms in native corpora and EFL teaching materials: A case study of English conditionals. In: S. Volk-Birke & J. Lippert (eds.). *Anglistentag 2006 Halle. Proceedings*. Trier: Wissenschaftlicher Verlag Trier. 355-363.

Swales, J. M. (2004). *Research Genres. Exploration and Applications*. Cambridge: Cambridge University Press.

Positional variation of phrase-frames in a new corpus of proficient student writing

Ute Römer and Matthew O'Donnell
University of Michigan

The Michigan Corpus of Upper-level Student Papers (MICUSP) is a new corpus of proficient student academic writing compiled at the English Language Institute of the University of Michigan, Ann Arbor. It consists of more than 800 papers collected from final year undergraduates and first to third year graduate students across 17 disciplines. Each of the papers is marked up in TEI-compliant XML, maintaining the structural divisions (sections, headings and paragraphs) of the original paper.

Hoey (2005) expands the notion of colligation to include the possibility that words and phrases may carry with them particular associations for occurrence at a specific location in text (*textual colligation*). To test this notion, we extracted frequent and well-dispersed phrase-frames (e.g. *at the * of*; see Römer [forthcoming]) of different lengths, and their variants (e.g. *at the end/beginning/risk of*) from MICUSP. Using the XML annotation to identify where in a text an item is most commonly found, we analysed the positional variation of select phrase-frames to determine which items typically occur at the beginning or end of a text, paragraph or sentence.

For example, examining two frames: 1. *the * that* (* = *fact, idea, suggestion, impression*) and 2. *it * (*) that* (*= *is suggested, was shown, is certain, turns out, is evident/clear, is known*) we found that both frames have a similar distribution in terms whether they are found in the first or last sentence (~20% and 17% of instances) of the paragraph in which they occur. However, examining where in the sentence they occur, 52% of the instances of the *it * (*) that* frame occur in the first tenth of the sentence, compared to around 20% for *the * that*.

Our findings suggest that phrase-frames not only show restrictions in terms of lexical selection but also restrictions on where in a text unit they tend to occur.

Reference:

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Römer, U. (forthcoming). English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies* 20(2).

Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations

Tanja Säily
University of Helsinki

I study variation in the morphological productivity of the suffixes -*ity* and -*ness* in the *British National Corpus* (BNC) using nonparametric statistical methods to compute upper and lower bounds for type and hapax accumulation curves. The aims are both sociolinguistic and methodological.

First, a sociolinguistic study of the suffixes is carried out in the demographically sampled part of the spoken component of the corpus. This follows up on a historical investigation into the use of -*ness* and -*ity* in the 17th-century part of the *Corpus of Early English Correspondence*, which showed that the productivity of the 'learned' suffix -*ity* (but not of the native -*ness*) was significantly low in letters written by women, perhaps because of women's lack of education. In the BNC data, however, women seem to use both –*ness* and -*ity* less than men, which may reflect a cultural change in the way men and women speak (cf. Rayson et al. 1997).

The results of the above-mentioned studies were obtained using type accumulation curves; the confidence intervals for hapax accumulation curves turned out to be too wide for significant differences to emerge. As this could be due to an insufficient amount of data, hapax accumulation curves are now being tested in the full BNC (c. 100,000,000 words), with a comparison of the spoken and written components (cf. Plag et al. 1999). The results will show whether hapax-based productivity measures can be considered reliable in large corpora, or if the number of hapaxes in these is as much a matter of chance as in smaller corpora.

References:

Plag, I., C. Dalton-Puffer & R.H. Baayen (1999) Morphological productivity across speech and writing. *English Language and Linguistics* 3(2): 209–228.

Rayson, P., G. Leech & M. Hodges **(**1997) Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1): 133–152.

“VOICE Awards”: A German Human-Machine Dialog Corpus

Tatjana Scheffler
Deutches Forschungszentrum für Künstliche Intelligenz

We present the VOICE Awards-Corpus, a unique collection of interactions with spoken dialog systems (SDS). In the talk, we report on the creation and annotation of the corpus, as well as some initial usage experiments.

So far, spoken and multimodal corpora are relatively underrepresented, due mostly to the fact that they are harder to obtain, store, and process. At the same time, such corpora are essential for the study of spoken language and for training machine learning methods for natural language processing applications.

The new corpus presented here consists of dialogs between lay users and commercially deployed telephone-based SDSs, which were collected in the course of the evaluations for the 2005-2008 annual “VOICE Awards” contests. This collection represents the entire breadth of state-of-the-art commercially deployed SDSs in the German-speaking countries. Each year, about 10 lay users evaluate the usability of the submitted systems by calling them and solving assigned tasks. In addition, experts perform a closer evaluation of the systems with regard to error robustness and advanced features such as tolerance of barge-ins.

The corpus consists of audio recordings of the resulting over 1500 dialogs (from 120 systems), along with manual transcriptions. It is currently being hand-annotated for (1) dialog acts, (2) errors or sources of miscommunication, (3) task success, and (4) repetitions. An evaluation of the dialog act annotation schema has shown that annotators agree on segmentation 80% of the time, and an inter-annotator agreement of κ=0.89 for the dialog act types assigned (on matching segments only).
Dialog corpora are still a rare resource. The VOICE Awards-Corpus is a large, German, human-machine dialog corpus with extensive annotations, which represents the entire breadth of domains and technology of commercial SDSs. To our knowledge, no other similar resource exists. This corpus is a solid basis for future research in human-machine dialog.

Agreeing with Google: We are Sensitive to the Relative Corpus Frequency of Phrases

Cyrus Shaoul, Chris Westbury and R. Harald Baayen
University of Alberta

Much has been said about the power of the frequency of words to predict many psycholinguistic phenomena. The frequency of the written word is integral to most architectures of lexical representation and lexical processing. Does this sensitivity to orthographic frequency extend beyond single words and into multi-word phrases? To answer this question, we asked participants to judge the relative frequencies of two, three, four and five word phrases. The frequencies of these phrases were taken from the Google Web1T database. For the majority of items, participants performed at above-chance levels when judging which one was more frequent. We also found that for two and three word phrases, the relative Web1T frequencies predicted judgment accuracy. These results imply that phrasal frequency is available to us during language comprehension.

Variation in –*im* suffix usage in a Tok Pisin corpus

Conor Snoek
University of Alberta

This study focuses on the Tok Pisin suffix -*im* which prototypically derives transitive from intransitive verbs. The suffix may attach to a number of different forms in the language. Previous analyses (Verhaar 1995) have shown that the suffix may attach to some transitive verbs, e.g. *tromoi* and *tromoim* both occur with the meaning 'to throw'. According to the glossing in Mihalic (1971) and the analysis presented in Verhaar (1995), the two forms behave similarly with respect to their meaning and grammar. This leads to the question of when and why the suffix occurs and what factors govern its distribution.

These questions are investigated in the *Slone Wantok Corpus*, a collection of Ancestor stories (*Stori Tumbuna*) from a column of the Tok Pisin language newspaper *Wantok.* The available data allow for an investigation of three possible dimensions in suffix use: variation in the linguistic context of usage, variation along the temporal dimension (1972-1997), and regional variation (by province). While Mühlhäusler et al. (2003:4) have downplayed the role of regional of variation Tok Pisin in the past, they admit that with the increasing autonomy of the Papua New Guinean provinces in more recent times, regional differentiation cannot be ruled out anymore.

What emerges is a complex picture of suffix use. There is evidence to indicate that some regions of Papua New Guinea have integrated the suffix as a fully productive feature of the Tok Pisin morphology, while others have not. The study sheds light on this phenomenon by making use of textual and metadata available through the corpus.

References:

Mihalic, Francis. (1971) *The Jacaranda Dictionary and Grammar of Melanesian Pidgin.* Milton: Jacaranda Press.

Mühlhäusler, Peter. (1975) *Growth and Structure of the Lexicon of New Guinea Pidgin.* Canberra: Pacific Linguistics, C-52.

Mühlhäusler, Peter, Dutton, Thomas E. and Suzanne Romaine (2003) *Tok Pisin Texts: From the beginning to the present.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

*The Slone Wantok Corpus***.** Compiled by Thomas Sloan, Stuart Robinson and Esteban Gutierrez. URL = http://www.tokpisin.org/

Verhaar, John W.M. (1995) *Toward a Reference Grammar of Tok Pisin: An Experiment in Corpus Linguistics.* University of Hawai'i Press: Honolulu.

Where Are They Bare? The Frequency and Distribution of Bare Nouns in American English

Laurel Stvan
University of Texas at Arlington

Acquisition studies have examined the frequency of bare and articulated NPs to determine whether the input frequency of NP types in caregiver speech affects children's production of these forms cross-linguistically. Second language teaching (other than immersion) has a less deterministic expectation of input vs. output, though contrastive and error analysis does focus on particular omissions or alterations in L2 form. Studies of learner corpora show that article misuse, in particular, is the highest error type for Chinese and Korean learners of English. Because determiners are a functional category, however, their mastery is a low priority in classrooms emphasizing communicative practice. Conversely, the English bare singular form, a low frequency NP type, is less often available as input. The overall frequency is low enough that English bare singular count nouns as arguments have been described as "totally impossible" (Chiercia 1998); yet such bare singular uses are attested and when they occur, the marked forms have been claimed to be pragmatically meaningful (Stvan 1998). This paper establishes the current frequency of different bare forms in the production of native English speakers. Based on a dataset of the first 100 uses of the top 15 bare nouns culled from the online Corpus of Contemporary American English (Davies 2008), the results detail the frequency for both mass and bare singular count nouns in American English found in three grammatical positions (PP predicates, subjects, and direct objects), giving a baseline for comparison of NP distribution work in learner corpora.

References:

Chiercia G. 1998. Reference to Kinds Across Languages. Natural Language Semantics 6: 339–405.

Davies, M. (2008-) The Corpus of Contemporary American English (COCA): 385 million words, 1990-present. Available online at http://www.americancorpus.org.

Stvan, L. S. 1998. The Semantics and Pragmatics of Bare Singular Noun Phrases. Northwestern University, Ph.D. Dissertation.

Part-of-speech tagging for a Southern Min Corpus

Ching Chu (June) Sun
University of Alberta

This paper addresses the issue of part-of speech (POS) tagging for a spoken and written corpus of Southern Min, a Chinese dialect spoken in parts of China and Taiwan as well as in overseas Chinese communities. Written materials for this corpus were collected from online forums, blogs, and literary works. Spoken materials were recorded from spontaneous monologues, conversations, Internet broadcast, on-line chats, telephone conversations, speeches, talk shows, and television news broadcasts. Although a number of Mandarin Chinese corpora annotated for parts of speech have been created (cf.McEnery and Xiao 2004), no tagged Southern Min corpus is publicly available. The current tagsets used for written Mandarin Chinese are not necessarily suitable for Southern Min and a new tagset was developed for the Southern Min Corpus. In addition, POS tagging of spoken language corpora present significant problems (incomplete utterances, repetitions, false starts etc.). Parts of speech in the corpus, both written and spoken, were tagged with the semi-supervised tagging program HunPOS. The overall accuracy of the tagger for written and spoken Southern Min is quite respectable, ranging from 70% to 90%, with higher accuracies resulting from larger training models. I argue that POS-tagging can be efficient and accurate using the HunPOS program and discuss the problems and difficulties in coding parts of speech specific to Southern Min.

References:

Halácsy, P., Kornai, A., and Oravecz, Cs. 2007. Hunpos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Companion Volume, Proceedings of the Demo and Poster Sessions, pp. 209-212. Prague, Czech Republic. 2007. Association for Computational Linguistics. McEnery, Anthony and Xiao Zhonghua. 2004. The Lancaster Corpus of Mandarin Chinese:

A Corpus for Monolingual and Contrastive Language Study. *Proceedings of the Fourth International Conference on Language Resources and Evalation* (LREC) 2004, pp.1175-78. Available from: http://www.lancs.ac.uk/postgrad/xiaoz/papers/231.pdf

A corpus-based evaluation of vocabulary in the second language classroom

John Sundquist and Colleen Neary-Sundquist
Purdue University

This study compares the vocabulary presented in three first-year German textbooks with two list of the 1000 most frequent German words. The three textbooks, *Kontakte*, *Vorsprung*, and *Deutsch Heute,* were selected on the basis of their widespread use in first-year German courses. The source for the frequency list of German words are the DeReWo from the Institut für Deutsche Sprache at the University of Mannheim and the Projekt Deutscher Wortschatz from the University of Leipzig. The DeReWo word list was created from the DeReKo corpus of German texts, which is comprised of over two billion words from various text types.

The results show that there is a serious disconnect between the vocabulary presented in the typical German textbook and the words that a student is most likely to encounter in materials outside the classroom. The paper discusses the possible reasons for this disconnect in terms of the preference that textbooks show for concrete nouns that can be taught to beginning students by means of a pictorial display. Although the pedagogical reason for this preference is clear, the low frequency of the vocabulary items typically emphasized in first-year classrooms is problematic. This study highlights the importance of corpora for creating and improving pedagogical materials for the second language classroom.

Comparable corpora for cross-linguistic sentiment analysis

Maite Taboada
Simon Frasier University

In this paper I discuss the use of comparable corpora in two or more languages for two related purposes: linguistic analysis of evaluation and appraisal in text, and the development of a computational system to extract sentiment and opinion. The work described here is part of a large project that has produced a semantic orientation calculator (SO-CAL), a system that extracts opinion from text and calculates its polarity (positive or negative). The current system is geared towards informal on-line reviews of movies, books and consumer products.

SO-CAL was initially developed for English texts, and we are in the process of porting it to a new language, Spanish. In both cases, the initial stages of the process involve collecting a suitable corpus for linguistic analysis, development of resources (dictionaries in particular), and testing. Researchers in sentiment analysis (e.g., Bautin et al., 2008) have proposed that, instead of developing resources for each specific language, one could translate texts from other languages into English, and then use English-only resources. I show that this type of approach misses much of the language-specific and culture-specific aspects of sentiment, and that comparable corpora are much preferable to parallel translation corpora, whether the translations are human or automatic.

The paper describes the current comparable corpus of 1,600 reviews, the manual annotation process that involves assigning Appraisal categories (Martin and White, 2005), and provides an overview of the differences between the English and Spanish semantic orientation calculators.

References:

Bautin, Mikhail, Lohit Vijayarenu and Steven Skiena. (2008). International sentiment analysis for news and blogs. *Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media*. San Jose, CA.

Martin, James R. and Peter White. (2005). *The Language of Evaluation*. New York: Palgrave.

A tale of two cities:
Comparing sociolinguistic patterns in England and Canada

Sali A. Tagliamonte and Cathleen Waters
University of Toronto

Corpus linguistics typically focuses on data sets of tens of millions of words. However, sociolinguistic studies rely on much smaller corpora. The benefit is a data set that can model both linguistic and community-level patterns. In this paper, we conduct a comparative study of two corpora of contemporary cities: York in England (1.2 million words) and Toronto in Canada (2 million words). Both datasets are stratified by age, sex and education enabling a view of change in apparent time and by social embedding.

In order to establish similarities and/or differences between the two cities, we focus on a mix of different features. Some are undergoing change; others are stable. In addition, we target systems from different levels of grammar, including phonology, morpho-syntax, and discourse-pragmatics, e.g. variable (ing), the deontic modals, intensifying adverbs, among other features.

(1) We were *having* a good time out in what we were *doin'*. (YRK/E)
I keep *lookin'* at this 'cause it's so *interesting*. (TOR/NW)

(2) There *has to* be something more than this. (YRK/d)
The answer of course *has to* be 'yes'. (TOR/NS)

(3) I was *really* excited about it (YRK/R)
I was *really* excited (TOR/3U)

The variants within each community are remarkably similar (see 1-3). Moreover, there are striking similarities in their linguistic patterns (e.g. the [ɪ n] variant occurs most often with verbs; *really* is favoured with adjectives such as *excited*, etc. However, unpredicted differences emerge in the social distribution of forms.

Taken together, the results reveal that British and Canadian English share the same grammar, but the way each linguistic system is embedded in the speech community diverges substantially. We conclude that grammar coheres across time and space, but the society in which a language is spoken shapes variable features to suit local conditions.

Japanese adjective '*sugoi*' and adverb '*sugoku*' in conversations

Tomoko Takeda
San Francisco State University

The present study examines Japanese adjective *sugoi* ‘amazing; awesome’ and the adverb *sugoku* ‘very; terribly’ in spontaneous Japanese conversations. Like many adverbs, *Sugoku* is generally treated as a derived adverb (or nuclear adverbial) of *sugoi* (Martin, 2004; Tsujimura, 2007). *Sugoku* belongs to degree adverbs which “primarily modify words expressing states and express their degree (Teramura, 1987).

Syntactically, adjectives modify nouns only whereas adverbs adjectives, verbs, adverbs and even an entire clause（Tsujimura 2007:120）. The findings from the examination of *sugoi* and *sugoku* in ten spontaneous Japanese conversations between friends (approximately 190 minutes in total), however, contradict those syntactic functions. *Sugoi* in the data not only modifies nouns but adjectives, verbs and adverbs. Noun-modifying *sugoi* which one would assume to be its primary use is the least frequent (2 of 90, 2%) whereas adjective-modifying *sugoi* is the most frequent (54 of 90, 60%).

The further analysis of *sugoi* indicates that the majority (34 of 54, 63%) belongs to affective (objective) adjectives category which includes adjectives such as *kanashii* ‘sad; ruthful’ and *kowai* `scary; terrifying’.

Those findings in the present study suggest that *sugoi* is primarily used to express the degree of the speaker’s affective state. Given that the general infrequency of *sugoku* compared to *sugoi* and that *sugoi* can modify a wide range of word classes, it appears as if *sugoi* is replacing some functions of *sugoku.* I also propose to look at the result applying prototype theory following Hopper and Thompson (1983). That is, treating the adjective *sugoi* as one end and the adverb *sugoi* on the other end of continuum instead of discrete two categories, we can suggest that *sugoi* is moving (or extending its function) toward its counterpart adjective *sugoku*, at least in casual spoken Japanese.

Subject Ellipsis by Text Type: An Investigation using ICE-GB

Laura Teddiman
University of Alberta

Subject ellipsis is not traditionally considered to be a feature of English grammar, but although it does not occur often in formal writing, it has been observed to be relatively common in conversation (Biber, Johansson, Leech, Conrad, & Finegan, 1999: 1048; Nariyama, 2004) and in diary writing (e.g., Haegeman & Ihsane, 2001). The purpose of the current paper is to explore the distribution of subject ellipsis across different text types of English through the use of the International Corpus of English for Great Britain (ICE-GB) and to identify the verbs and elided pronouns associated with each.

Results show that subject ellipsis is most frequently attested in private conversation, unscripted monologues, and correspondence, and is not attested in more formal writing. In correspondence, *I* is the most frequently omitted pronoun, with *it* a distant second. In private conversation, *I* and *it* are both omitted at roughly the same rate, with small contributions from the other subject pronouns. In unscripted monologues, however, *he* is the most frequently omitted pronoun. This reflects data from live sportscasts, where announcers describe the continuing actions of a player who was identified earlier in the speech stream. The verbs attracted to sentence initial position vary as well, with more active verbs such as *plays* and *turns* being present in unscripted monologues and mental state verbs (e.g., *think*, *remember*) being more frequently attested in conversation. Other patterns in conversation are also identified. Conventional expressions such as *thank you* are more prevalent in correspondence and conversation.

This paper accomplishes two related goals. First, patterns in English subject ellipsis have been identified, and second, different patterns have been associated with different text types. These results highlight the importance of examining linguistic structure across genres.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Pearson Education Limited, Essex, UK.

Haegeman, L. & Ihsane, T. (2001). Adult null subjects in the non-pro-drop languages: Two diary dialects. *Language Acquisition, 9*(4), 329-346.

Nariyama, S. (2004). Subject ellipsis in English. *Journal of Pragmatics, 36*, 237-264.

A Corpus-Based Analysis of Slovene Reflexive Verbs (Verbs with *Se*)

Mojka Tomišić
University of Maribor, Slovenia

In the Slovene language reflexivity is expressed through the verb accompanied by the reflexive pronominal clitic *se* (e.g. *umiti se* – 'to wash oneself'). Slovene verbs with *se* can also express other meanings, e.g. reciprocity, non-volition, subjectlessness and passivity. Some verbs cannot exist without *se* and some others completely change their meaning when they become reflexive, which in some cases results in minimal phrases. Slovene linguistics has yet to produce a comprehensive study which would cover all aspects of the analysis of verbs with *se* and thus set relevant criteria for the correct placement of such verbs in the dictionary entry. A partial study of the verbs with *se* from the three most important Slovene dictionaries, namely from Pleteršnik's Slovene-German Dictionary (1894–1895), the Dictionary of the Slovene Standard Language (SSKJ) (1970–1991) and the dictionary section of the Slovene Orthography (SP) (2001), has indicated great differences in the entry placement of the verbs with *se*. Despite the guidelines given for their work, authors of these dictionaries often relied on their linguistic intuition and the frequency of the use of the forms with *se* and without *se*. For example, in Pleteršnik's Dictionary the verb *bližati se* ('to get closer') is included in the sample sentences, whereas in SSKJ it is the headword, and in SP it is the subhead word. Using the corpus method, I try to determine the changes that verbs with *se* from the mentioned dictionaries have undergone. On the basis of these findings, it is possible to determine which forms should be included in the future new monolingual dictionary of the Slovene language. The comparison of the selected verbs with *se* is implemented with the help of the computer program FidaPLUS Assistant, which has been developed for this purpose at the Faculty of Arts in Maribor.

Unsupervised learning of criterial features of L2 acquisition stages using parallel learner corpora

Yukio Tono, Mariko Nomura, Akira Murakami, Taku Kaneta, and Hajime Mochizuki
Tokyo University of Foreign Studies

This study aims to identify L2 learners' criterial features for the Interlanguage developmental stages by using unsupervised learning techniques over parallel learner corpora. Research into the nature of L2 learner language against the target language using learner corpora has shown many interesting characteristics of the Interlanguage, such as underuse vs. overuse phenomena, transitional error patterns, and the developmental patterns in various aspects of lexico-grammatical features (cf. Granger, 1998; Tono, 2004 among others). One of the recent research interests is to identify criterial features, features which can serve as criteria to indicate a particular proficiency level of L2 learners, using pseudo-longitudinal learner corpora. The present study is also one such attempt.

In this study, we prepared a parallel set of the JEFLL Corpus, a corpus of written compositions by more than 10,000 Japanese-speaking learners of English, ranging from the 1st-year junior high school to the 3rd-year senior high school. By "parallel set", we mean that the original essay is proofread and corrected by a native speaker and each sentence in the writings has its corrected counterpart. We prepared a parallel data for the entire JEFLL Corpus, and the differences between the original and corrected versions of the sentences were automatically identified by using DP-matching techniques in terms of three criteria: omission, addition and misformation. The output of DP-matching was then further tagged by POS, and further statistical analysis (e.g. Correspondence Analysis) was performed over the output in order to obtain unsupervised clustering of linguistic features with specific error types (omission, addition, and misformation) in association with different developmental stages (in our case, the school years).

The analysis of the data is still underway and will be reported in the presentation. Methodological implications of using DP-matching techniques with multivariate analyses will be discussed, together with some theoretical and pedagogical implications for future studies.

Frequency and multi word sequences: A psycholinguistic comparison of two corpora

Benjamin V. Tucker[1] and Antoine Tremblay[2]
[1]University of Alberta, [2]Georgetown University

In the field of psycholinguistics the use of corpora plays an important role. One of the functions of a corpus in a psycholinguistic experiment is to calculate the frequency of occurrence of a word, from a list of experimental stimuli, and then correlate the calculated frequency with listeners ability to process that word (i.e. measured in reaction time or error rate). In this paper we present a production experiment where 24 participants were asked to produce 432 four word sequences (e.g., *in the middle of*; *it could have been*). In this study we recorded the response latency (for all participants) and the acoustic duration (for 15 participants) of the productions. We then report how these measures correlate with frequency and probability of the strings and their parts. The frequency and probability measures in this experiment are derived from two corpora: the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA). We hypothesized that the frequency and probability values derived from COCA would better model the data than the BNC, owing to the fact that COCA is more reflective of the dialect spoken by the participants. We were in fact surprised that there were only minor differences in the results of the analysis and that the main trends agreed across the two corpora. For example, we found that as the frequency of the whole string increased the response latency decreased in both corpora (approximately 100ms using both corpora).

Developing a Text-Based Corpus of the Language of Japanese Comics (Manga)

Giancarla Unser-Schutz
Hitotsubashi University

Comics (manga) account for nearly 40% of all books and magazines published in Japan (Schodt, 19). With the vast majority of those published being narrative-driven titles, they represent an enormous amount of linguistic data, and have often been cited as sources of linguistic change (e.g., Endo, 2001 on personal pronouns). However, while there has been some research dealing with language in manga (Kinsui, 2007 on manga "role-language", Ueno, 2006 on gendered language in manga, etc.), most have been limited in scope. A comprehensive, corpus-based analysis of manga-language is essential to determining the linguistic influence they have on their readers.

However, while manga have been raised as a possible subject for corpus study in the past, their particularities have stalled such attempts (Maekawa, 8). As their linguistic elements appear embedded in the pictures (in speech-bubbles, etc.), any text-based corpus would need to develop a way to represent such data in a meaningful way. In this presentation, I will introduce a manga, pure-text corpus designed with such issues in mind. In particular, all text has been categorized into eight groups: lines, thoughts, narration, onomatopoeia, background text, background lines/thoughts, author comments and titles. By noting the category, page and frame number of each item, it becomes possible to keep track of their occurrences in-context.

The corpus presently consists of the first three volumes of six popular series, selected from (1) a series of surveys on Japanese high school students, and (2) sales ranking, for a total of approximately 422,000 characters over 3,220 pages. I will pay particular attention to the special issues faced in designing such a corpus, including demonstrating how it may be of use in examining linguistic phenomena. I will also touch upon plans to add additional series and a possible parallel English translation corpus of the included texts.

References:

Endo, Orie (2001). Onna no Ko no "Boku / Ore" wa Okasikunai. In O. Endo (Ed.), Onna to Kotoba: Onna wa Kawatta ka, Nihongo wa Kawatta ka, 30-39. Tokyo: Meiji-shoten.

Kinsui, Satoshi (2007). Kindai Nihon Manga no Gengo. In S. Kinsui (Ed.), Yakuwari-go Kenkyuu no Chihei, 97-108. Tokyo: Kurosio Shuppan.

Maekawa, Kikuo (2008). Kotonoha 'Gendai Nihongo Kaki-kotoba Kinkou Koupasu' no Kaihatsu. Nihongo no Kenkyuu, 4.1, 82-95.

Schodt, Frederik L (1996). Dreamland Japan: Writings on Modern Manga. Berkeley, CA: Stone Bridge.

Ueno, Junko (2006). Shojo and Adult Women: A Linguistic Analysis of Gender Identity in Manga (Japanese Comics). Women and Language, 29.1, 16-25.

Electronic corpora for two Semitic languages

Adam Ussishkin[1], Jerid Francom[2], and Dainon Woudstra[1]
[1]University of Arizona, [2]Wake Forest University

In addition to the clear role that electronic text resources play in computational applications, corpora are increasingly playing a larger role in the testing and development of linguistic theories. These resources have proved to be widely applicable for applications in areas such as formal linguistics, psycholinguistics and language acquisition; however, the majority of the resources document a relatively limited number of well-studied languages. A growing number of scholars working on languages that have limited or no corpus coverage have begun to develop resources for under-studied and under-documented languages in order to fill this existing gap. This paper describes such an effort for two Semitic languages: Maltese and Hebrew.

In this paper we document the creation of Maltese and Hebrew lexical corpora, developed in order to further psycholinguistic research on the mental organization of Semitic lexicons. The ultimate goal of this project is to create a sizable lexical corpus and a set of lexical calculation tools capable of producing a filtered set of lexical items for psycholinguistic experimentation. The sources for this effort include text extracted from the web and collaborative efforts from other scholars working in the area. We discuss both theoretical and practical issues in creating corpora in general and these corpora more specifically, elaborate the steps taken to bring this project to fruition, and report the statistical results of our efforts. Our discussion also includes an evaluation of these resources and their contribution to language research and the larger language community.

Safe Harbour: Ethics and accessibility in sociolinguistic corpus building

Gerard Van Herk[1], Rebecca Childs[2], and Jennifer Thorburn[1]
[1]Memorial University of Newfoundland, [2]Coastal Carolina University

Contemporary sociolinguistic corpus builders are often pulled in competing theoretical and methodological directions. Should corpora be broad (permitting large quantitative studies, with many speakers) or deep (ethnographically rich, with lengthy interviews)? Should they protect informant confidentiality (through sharply limiting access), or enrich a wide range of research (through data sharing)?

This paper describes our ongoing work on a sociolinguistic corpus of vernacular speech from Petty Harbour, a rapidly urbanizing fishing community in Newfoundland. We describe the specific challenges of working in such a community, but we focus on the procedures used to make the corpus amenable to both traditional large-scale quantitative analysis and contemporary smaller-scale ethnographic work while maintaining the ethically sound data sharing practices of corpus linguistics. To create this collection, interviews are transcribed and double verified prior to anonymization, in which identity-revealing information is removed from both sound files and transcripts, while preserving cues necessary to studying linguistic structure. The resulting anonymized files are accessed through an online password-protected interface, permitting authorized researchers from various disciplines and affiliations to make use of Petty Harbour interviews. Although our anonymization process puts some restrictions on phonological analysis, it ensures that the data meet both ethical and research standards.

Using character n-grams to classify native language in a non-native English corpus of transcribed speech

Charlotte Vaughn, Janet Pierrehumbert, and Hannah Rohde
Northwestern University

A central issue in second language acquisition research is the degree to which first language (L1) has an effect on the learning of a second language (L2). Recently, corpus-based computational methods have been employed to mine the language of L2 learners in order to detect such effects over large datasets. An important advancement in this area was the claim that a speaker's L1 phonology may have an effect on word choice in L2 [1], a hypothesis formulated from patterns observed in a corpus of writing of non-native English speakers.

This paper extends the analysis from L2 writing (as in [1]) to L2 naturalistic English speech, in which phonological effects might be stronger than in writing. The speech database comes from the Wildcat corpus of native- and foreign-accented speech [2], part of which was gathered using a referential communication task between dyads. Confirming the results in [1], a k-nearest neighbors classifier operating on character n-grams performs well above chance in predicting the native language of speakers (English, Korean, or Chinese).

The paper explores the relative contributions of specific word choices and phonological constraints to the character n-gram patterns. The classifier maintains high performance when highly frequent n-grams and words are removed, a strategy to control for function word statistics as a reflex of L1 background. Initial observations suggest that the effects of content words involve both specific lexical substitutions as well as systematic avoidance of words with problematic sequences. These effects are further examined through additional statistical classifiers. An advantage of this method for acquisition research is the non-reliance on linguistic errors to identify L1 effects on L2; rather, the approach allows for the analysis of more gradient effects of dispreference and avoidance. Overall, our results demonstrate how text-based corpus methods may be usefully applied to transcripts of naturalistic speech.

[1] Tsur, Oren and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 6-16, Prague, Czech Republic, June 2007.

[2] Van Engen, Kristin, Melissa Baese-Berk, Rachel E. Baker, Arim Choi, Midam Kim, and Ann R. Bradlow. In press. The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*.

Towards a Typology of English Accents

Steven Weinberger[1] and Stephen Kunath[2]
[1]George Mason University; [2]Georgetown University

The theoretical and practical value of studying human accented speech is of interest to linguists, language teachers, actors, speech recognition engineers, and computational linguists. It is also part of the research program behind the Speech Accent Archive (http://accent.gmu.edu). The Archive is a growing annotated corpus of English speech varieties that contains more than 1,100 samples of native and non-native speakers reading from the same English paragraph. The non-native speakers of English come from more than 250 language backgrounds and include a variety of different levels of English speech abilities. The native samples demonstrate the various dialects of English speech from around the world. All samples contain a complete digital audio version, and include a narrow phonetic transcription. Each speaker is located geographically, and crucial demographic parameters are supplied. For comparison purposes, the Archive also includes phonetic sound inventories from more than 200 world languages so that researchers can perform various contrastive analyses and accented speech studies.

Human listeners immediately and automatically notice that different speakers have discernable accents. For example, Chinese speakers of English sound different from Swahili speakers of English. But what exactly constitutes these differences? Just what specific information alerts a human listener that a Chinese speaker sounds different from a Swahili speaker of English? We have developed a device that computationally compares any two phonetic transcriptions from our corpus and distills from this analysis a set of phonological speech patterns (PSPs) for each speaker. Essentially, our task is to discover the precise features responsible for humans to categorize and assess any nonnative speaker of English.

This paper will discuss the architecture of the Speech Accent Archive, and discuss our computational device, the Speech Transcription Analysis Tool (STAT). We show how such a tool leads to a typology of English accents that can benefit researchers in the fields of theoretical and applied linguistics.

A multifactorial analysis of (un)attended demonstratives in academic writing

Stephanie Wulff
University of North Texas

The English demonstrative forms *this*, *these*, *that*, and *those* can function either as free-standing pronouns as illustrated in (1) or as determiners attending a head noun phrase as in (2).

(1) *This* is an example sentence.

(2) *This* sentence is an example.

Previous studies have suggested that various language-internal factors determine this variation, such as the abstractness, previous mention, and the grammatical function of the noun phrase (Geisler et al. 1985, Swales 2005). However, there is to date no systematic investigation of the relative impact of all these variables, their interactions, or the additional role that language-external factors such as genre or proficiency may play.

The present addresses these issues by presenting the results of a logistic regression analysis that takes all potential predictor variables simultaneously into account and not only identifies (combinations of) significant predictors but also weighs their relative impact. Next to the variables above we also consider the (frequency of the) matrix verb; the frequency of the noun phrase; whether the noun phrase occurs clause-initially or not; how often the referent of the noun phrase has been mentioned in the previous text, as well as the distance to the last mention of the referent. The analysis is based on more than 12,000 hits in the *Michigan Corpus of Upper Level Student Papers* (MICUSP). The results indicate that those variables representing discourse-organizational aspects generally outweigh other factors and testify to pronounced discipline-specific tendencies in the use of (un)attended *this*.

References:

Geisler, C., Kaufer, D. S. & Steinberg, E. R. (1985). The unattended anaphoric “this”. *Written Communication*, 2, 129-155.

Swales, J. M. (2005). Attended and unattended “this” in academic writing: A long and unfinished story.*ESP Malaysia*, 11, 1-15.

Contact Information

Aberra, Daniel	University of Alberta	daberra@ualberta.ca
Ahmad, Ummul	Universiti Teknologi Malaysia	m-ummul@utm.my
Akiyama, Takanobu	Nihon University	akiyama.takanobu@nihon-u.ac.jp
Alarcón, Rodrigo	Universidad Nacional Autónoma de México	ralarconm@ii.unam.mx
Albers, Sarah	Verilogue, Inc.	salbers@verilogue.com
Anthony, Laurence	Waseda University, Japan	anthony@waseda.jp
Aull, Laura	University of Michigan	laull@umich.edu
Baayen, Harald	University of Alberta	baayen@ualberta.ca
Baker, Amanda	Georgia State University	eslambx@langate.gsu.edu
Barlow, Michael	Auckland University	mi.barlow@auckland.ac.nz
Barreda, Santiago	University of Alberta	sbarreda@ualberta.ca
Becher, Viktor	University of Hamburg	viktor.becher@uni-hamburg.de
Bednarek, Monika	University of Technology, Sydney	Monika.Bednarek@uts.edu.au
Berber Sardinha, Tony	Sao Paulo Catholic University	tony@pucsp.br
Bergh, Gunnar	Mid-Sweden University	gunnar.bergh@miun.se
Bresnan, Joan	Stanford University	bresnan@stanford.edu
Brinton, Laurel	University of British Columbua	brinton@interchange.ubc.ca
Bruce, Carrie	Georgia Institute of Technology	carrie.bruce@coa.gatech.edu
Caldwell, Joshua	Brigham Young University	Josh.Caldwell@byu.edu
Calude, Andreea	Auckland University	acalude@gmail.com
Caple, Helen	University of Wollongong	helen.caple@usyd.edu.au
Chartrand, Robert	Kurume University	robertchartrand@mac.com
Childs, Rebecca	Coastal Carolina University	rchilds@coastal.edu
Chujo, Kiyomi	Nihon University, Japan	chuujou.kiyomi@nihon-u.ac.jp
Chung, Siaw-Fong	National Chengchi University	sfchung@nccu.edu.tw
Columbus, Georgina	University of Alberta	georgie.columbus@ualberta.ca
Cox, Christopher	University of Alberta	christopher.cox@ualberta.ca
Davies, Mark	Brigham Young University	mark_davies@byu.edu
Dilts, Philip	University of Alberta	pdilts@ualberta.ca
Duffley, Patrick	Université Laval	Patrick.Duffley@lli.ulaval.ca
Fang, Alex C.	City University of Hong Kong	acfang@cityu.edu.hk
Francom, Jerid	Wake Forest University	jeridf@u.arizona.edu
Friginal, Eric	Georgia State University	efriginal@gsu.edu
Fung, Simon	University of Alberta	simonmin@ualberta.ca
Gajdos, Johnathan	University of Iowa	johnathan-gajdos@uiowa.edu
Gales, Tammy	University of California, Davis	tgales@ucdavis.edu
Gardner, Dee	Brigham Young University	dee_gardner@byu.edu
Geeraert, Kristina	University of Alberta	geeraert@ualberta.ca
Geisler, Christer	Uppsala University	christer.geisler@engelska.uu.se
Gotscharek, Annette	Ludwig-Maximilians-Universität München	annette@cis.uni-muenchen.de
Grant, Lynn	Auckland University of Technology	lynn.grant@aut.ac.nz

Gries, Stefan Th.	University of California, Santa Barbara	stgries@linguistics.ucsb.edu
Hsieh, Yi-Chen	National Chengchi University	96551016@nccu.edu.tw
Huang, Li-Shih	Victoria University, CA	lshuang@uvic.ca
Iberri-Shea, Gina	United States Air Force Academy	Gina.Iberri-Shea@nau.edu
Ibrahim, Noor Mala	Universiti Teknologi Malaysia	noormala65@hotmail.com
Jankowski, Bridget	University of British Columbia	bljankowski@gmail.com
Jia, Ruiting	University of Alberta	ruiting@ualberta.ca
Johansson, Christine	Uppsala University	christine.johansson@engelska.uu.se
Ju, Hee	University of California, Los Angeles	hjugrace@yahoo.com
Kaneta, Taku	Tokyo University of Foreign Studies	kaneta.taku.919@gmail.com
Kemmer, Suzanne	Rice University	kemmer@rice.edu
Kendall, Tyler	Northwestern/NC State University	tsk3@duke.edu
Kim, Sangbok	University of California, Los Angeles	sbkim@ucla.edu
Kim, So Yeon	University of California, Los Angeles	ssoyeony@gmail.com
Kunath, Stephen	Georgetown University	skunath@gmu.edu
Kunichika, Hidenobu	Kyushu Institute of Technology	kunitika@ai.kyutech.ac.jp
Kuo, Sai-hua	National Tsing Hua Unversity, Taiwan	shkuo@mx.nthu.edu.tw
Lee, David Y.W.	City University of Hong Kong	davidlee@cityu.edu.hk
Li, John	City University of Hong Kong	johnlihanhong@yahoo.com.cn
Libben, Gary	University of Calgary	glibben@ucalgary.ca
Liberman, Mark	University of Pennsylvania	myl@cis.upenn.edu
Lin, Dekang	Google.com	lindek@google.com
Liu, Dilin	University of Alabama	dliu@as.ua.edu
Lonsdale, Deryle	BYU	lonz@byu.edu
MacWhinney, Brian	Carnegie Mellon University	macw@cmu.edu
Mello, Heliana	Universidade Federal de Minas Gerais, Brazil	heliana.mello@gmail.com
Meyer, Charles	University of Massachusetts Boston	meyer@cs.umb.edu
Miglio, Viola G	University of California, Santa Barbara	miglio@spanport.ucsb.edu
Mochizuki, Hajime	Tokyo University of Foreign Studies	motizuki@tufs.ac.jp
Murakami, Akira	Tokyo University of Foreign Studies	mr_optimistic39@yahoo.co.jp
Neary-Sundquist, Colleen	Purdue University	cnearysu@purdue.edu
Nesi, Hilary Ibrahim,	Coventry University, UK	h.nesi@coventry.ac.uk
Neumann, Andreas	Ludwig-Maximilians-Universität München	andi@cis.uni-muenchen.de

Newman, John	University of Alberta	john.newman@ualberta.ca
Nomura, Mariko	Tokyo University of Foreign Studies	m.nomura311@gmail.com
O'Donnell, Matthew	University of Michigan	mbod@umich.edu
Oghigian, Kathryn	Waseda University, Japan	oghigian@gol.com
Pearson, Pamela	Georgia State University	elspapx@langate.gau.edu
Pezik, Piotr	University of Lodz	piotr.pezik@gmail.com
Pickering, Lucy	Georgia State University	esllup@langate.gsu.edu
Pierrehumbert, Janet	Northwestern University	jbp@babel.ling.northwestern.edu
Poplack, Shana	University of Ottawa	spoplack@uOttawa.ca
Raso, Tommaso;	Universidade Federal de Minas Gerais, Brazil	tommaso.raso@gmail.com
Reffle, Ulrich	Ludwig-Maximilians-Universität München	uli@cis.uni-muenchen.de
Rice, Sally	University of Alberta	sally.rice@ualberta.ca
Ringlstetter, Christoph	Ludwig-Maximilians-Universität München	kristof@cis.uni-muenchen.de
Rockwell, Geoffrey	University of Alberta	geoffrey.rockwell@ualberta.ca
Rodríguez-Puente, Paula	University of Santiago de Compostela	paula.rodriguez.puente@usc.es
Rohde, Hannah	Northwestern University	hannah@northwestern.edu
Römer, Ute	University of Michigan	uroemer@umich.edu
Säily, Tanja	University of Helsinki	tanja.saily@helsinki.fi
Scheffler, Tatjana	Deutsches Forschungszentrum für Künstliche Intelligenz	tatjana.scheffler@dfki.de
Schulz, Klaus	Ludwig-Maximilians-Universität München	schulz@cis.uni-muenchen.de
Sevigny, Alexandre	McMaster University	sevigny@mcmaster.ca
Shaoul, Cyrus	University of Alberta	cyrus.shaoul@ualberta.ca
Sierra, Gerardo	Universidad Nacional Autónoma de México	gsierram@ii.unam.mx
Snoek, Conor	University of Alberta	snoek@ualberta.ca
Stvan, Laurel	University of Texas at Arlington	stvan@uta.edu
Sun, June	University of Alberta	chingchu@ualberta.ca
Sundquist, John	Purdue University	sundquist@purdue.edu
Taboada, Maite	Simon Fraser University	mtaboada@sfu.ca
Tagliamonte, Sali A.	University of Toronto	sali.tagliamonte@utoronto.ca
Takeda, Tomoko	San Francisco State University	takeda@sfsu.edu
Takeuchi, Akira	Kyushu Institute of Technology	takeuchi@minnie.ai.kyutech.ac.jp
Teddiman, Laura	University of Alberta	teddiman@ualberta.ca
Thorburn, Jennifer	Memorial University of Newfoundland	jennifer.thorburn@gmail.com
Tomišić, Mojca	University of Maribor, Slovenia	mojca.tomisic@uni-mb.si
Tono, Yukio	Tokyo University of Foreign Studies	y.tono@tufs.ac.jp
Tremblay, Antoine	Georgetown University	trea26@gmail.com

Tucker, Ben	University of Alberta	benjamin.v.tucker@gmail.com
Unser-Schutz, Giancarla	Hitotsubashi University	giancarlaunserschutz@yahoo.co.jp
Ussishkin, Adam	University of Arizona	ussishki@u.arizona.edu
Van Herk, Gerard	Memorial University of Newfoundland	gvanherk@mun.ca
Vaughn, Charlotte	Northwestern University	crvaughn@u.northwestern.edu
Waters, Cathleen	University of Toronto	cathleen.waters@utoronto.ca
Weinberger, Steven	George Mason University	weinberg@gmu.edu
Westbury, Chris	University of Alberta	chrisw@ualberta.ca
Woudstra, Dainon	University of Arizona	dainon@u.arizona.edu
Wulff, Stefanie	University of North Texas	stefaniewulff@gmail.com