14th American Association for Corpus Linguistics (AAACL) Conference

September 20-22, 2018 in Atlanta, GA
GSU Student Center East, 55 Gilmer St. SE, Atlanta, GA, 30303
ACKNOWLEDGEMENTS

We acknowledge the generous support of the AACL 2018 sponsors:

Center for Research on the Challenges of Acquiring Language and Literacy, Georgia State University
College of Arts and Science, Georgia State University
Department of Applied Linguistics and ESL, Georgia State University
Duolingo Inc., Pittsburgh
John Benjamins Publishing
University of Michigan Press
University Research Services and Administration, Georgia State University

We would also like to thank the members of the AACL 2018 Program Committee for their help with conference proposal reviewing:


We also very much appreciate the help of the members of the AACL 2018 Student Team:

Noor Abbady, Fatemeh Bordbarjavid, Analynn Bustamante, Yusuf Cengiz, Diana Hammond, Sanghee Kang, Minkyung Kim, William Lake, Xian (Jan) Li, Bineta Mbodj, Kátia Monteiro Vanderbilt, Tamanna Mostafa, Yunjung (Eunice) Nam, Haoshan (Sally) Ren, Andy Schneider, Justin Taylor, Yu (Terry) Tian, Rurik Tywoniw, Rachelle Udell, Qian Wan, Selahattin Yilmaz, Genggeng Zhang
# TABLE OF CONTENTS

## PRE-CONFERENCE WORKSHOPS

**WORKSHOP 1:** An Introduction to Natural Language Processing Tools: The Curious Case of the Lexicon ............................................. 1

**WORKSHOP 2:** Lexical Multi-Dimensional Analysis.......................................................................................................................... 2

**WORKSHOP 3:** Complexity in Writing Development: Untangling Two Approaches to Measuring Grammatical Complexity ......................................................... 3

## PLENARY SPEAKERS

Berber Sardinha, Tony - Catholic University of Sao Paulo, Brazil - Corpus Linguistics and Historiography ............................................. 4

Gray, Bethany - Iowa State University, Ames, Iowa - The Role of Register Analysis in Academic Writing Research.............................................. 5

Gries, Stefan Th. - University of California, Santa Barbara and Justus Liebig University - Towards a Unified Tupleization of Corpus Linguistics ........................................................................ 6

## PAPER PRESENTATIONS

Alanazi, Zaha - Iowa State University - Lexical complexity in ESL writing: Does the task make a difference? .................................................................................................................. 7

Alanazi, Zaha - Iowa State University - Topical organizational Patterns as indicators of ESL writing proficiency.......................................................................................................................... 7

Anthony, Laurence, Waseda University and Römer, Ute, Georgia State University - Identification, analysis, and application of n-grams, p-frames, and other multi-word units ........................................................................ 7

Appel, Randy, Waseda University and Szeib, Andrzej, Trent University - Linking adverbials in L2 English academic writing: L1 related differences ........................................................................ 8

Araujo, Rafael Fonseca de, Berber Sardinha, Tony and Delfino, Maria Claudia - Sao Paulo Catholic University - Representations of American identity: a multidimensional analysis of Google Books trigrams around North, Central, South and Latin America ........................................................................ 8

Balasubramanian, Chandrika - Sultan Qaboos University - Now we are having lot to think about, isn’t it? Indian English: What happens next? ...................................................................................... 8

Becker, Kimberly and Feng, Hui Hsien - Iowa State University - Stance in Unpublished Student Writing: An Exploration of MICUSP .................................................................................. 9

Belcher, Diane, Friginal, Eric and Yilmaz, Selahattin - Georgia State University - Exploring the Academic Blogging Practices of English as A Lingua Franca (ELF) Scholars ......................................................................... 9

Berber Sardinha, Tony, Sao Paulo Catholic University and Veirano, Pinto Marcia, Sao Paulo Federal University - A multi-dimensional text typology of American television programs ........................................................................ 10

Bertoli, Patricia and Shepherd, Tania - University of the State of Rio de Janeiro - Corpus-informed analyses of research papers in English as a lingua franca: the case of concessive structures ........................................................................ 11
Biber, Doug, Egbert, Jesse and Keller, Daniel - Northern Arizona University - **Reconceptualizing register as a continuous construct** ......................................................................................................................... 11

Black, Amanda - Northern Arizona University - **Non-native and native acceptability of variable slots in fixed sequences: A corpus-based approach** .............................................................................................................. 12

Brennan-Seely, Jane - Mary Immaculate College, University of Limerick - **A Corpus-Aided Comparative Analysis of Native and Non-Native Speaking English Language Teachers’ Approach to In-Class Spoken Feedback** ........................................................................................................................................ 12

Brezina, Vaclav, Gablasova, Dana and Gomide, Andressa Rodrigues - Lancaster University - **#LancsBox and BNC lab: A new-generation corpus analysis tools** .................................................................................................................. 12

Bristow, Mackenzie - Emory University - **An Exploratory Multidimensional Analysis of International Strategic Plans published by U.S. Universities** ......................................................................................................................... 13

Buell, Duncan, University of South Carolina, Picoral, Adriana,University of Arizona, Reppen, Randi Northern Arizona University, Staples, Shelley, University of Arizona, Novikov, Aleksey, University of Arizona, Liu, Qianti, University of South Carolina and Lan, Ge, Purdue University - **Annotating learner data for lexico-grammatical patterns: A comparison of software tools** .................................................................................................................. 13

Charles, Maggie - Language Centre, Oxford University - **Corpus-assisted editing for doctoral students: More than just concordancing** ........................................................................................................................................ 14

Cotas, Elena, Gray, Bethany and Smith, Jordan - Iowa State University - **Combining Multi-Dimensional Analysis with Rhetorical Move Analysis to Research Writing Across Disciplines** ........................................................................................................ 14

Craig, Betsy - Furman University - **Getting the Grade in First-Year Writing: Does nominal density matter?** .............................................................................................................................................. 15

Crawford, Bill, Northern Arizona University, McDonough, Kim, Concordia University and Brun-Mercer, Nicole, California State University-Fullerton - **Identifying linguistic markers of L2 collaboration: A lexicogrammatical approach** .................................................................................................................. 15

Crossley, Scott, Georgia State University, Bustamante, Analynn, Georgia State University, and Bradfield, Franklin, Georgia Institute of Technology - **GAMET: A tool for automatically assessing grammar and mechanic errors in learner corpora** .......................................................................................... 16

Csomay, Eniko - San Diego State University - **Writing for STEM in (country) Georgia: A corpus-based study of academic word use in student writing** ............................................................................................................. 16

Cunningham, Clark - Georgia State University College of Law - **Teaching lawyers about using corpus linguistics** .............................................................................................................................................. 16

Cupery, Maria - Purdue University - **Price-Based Lexical Variation in Turkish Online Restaurant Reviews** 17

Davies, Mark - Brigham Young University - **Using the new 14 billion word iWeb corpus for teaching and research** .............................................................................................................................................. 17

Deshors, Sandra C. - Michigan State University - **Does the passé composé influence L2 learners’ use of English past tenses? A semantic exploration of the Present Perfect in French-English interlanguage** .... 18

Doolan, Stephen - Texas A&M University - Corpus Christi - **A corpus-based analysis of undergraduate, source-based L1 student writing** .................................................................................................................. 18
Durrant, Philip - University of Exeter - **Lexical development in English school children’s writing from six to sixteen** .................................................................................................................. 18

Dutra, Deise, Universidade Federal de Minas Gerais, Costa, Danilo, Universidade Federal dos Vales do Jequitinhonha e Mucuri, Macedo, Luciana, Universidade Federal de Minas Gerais and Queiroz, Jessica, Universidade Federal de Minas Gerais - **Adjective as nominal premodifiers in Chemistry and Applied Linguistics Corpora** .................................................. 19

Edalatishams, Idee - Iowa State University - **Development of a Corpus of TA Classroom Speech** .............. 19

Egbert, Jesse and Biber, Douglas - Northern Arizona University - **Incorporating text dispersion into keyword analyses** .................................................................................................................. 19

Fahy, Matthew, Northern Arizona University, Egbert, Jesse, Northern Arizona University, Biber, Douglas, Northern Arizona University and Szmrrecsanyi, Benedikt, Katholieke Universiteit Leuven - **Comparing logistic regression, conditional inference trees and random forests applied to ternary variables** ............ 20

Garner, James - Georgia State University - **The Longitudinal Development of Collocations as Constructions in L2 Writing** .................................................................................................................. 20

Graedler, Anne-Line - Inland Norway University of Applied Sciences - **L2 students’ choice of cohesive devices in translation** .................................................................................................................. 21

Grieve, Jack and Clarke, Isobelle - University of Birmingham - **Tracking Stylistic Change on the Donald Trump Twitter Account** .................................................................................................................. 21

Grieve, Jack, University of Birmingham, Clarke, Isobelle, University of Birmingham, Popoola, Olumide, University of Birmingham and Waibel, Emily, Aston University - **Native Language Identification of the Guccifer 2.0 Corpus** .................................................................................................................. 22

Hardy, Jack, Oxford College of Emory University and Payant, Caroline, Université du Québec à Montréal - **Conference and Research Article Abstracts: A Corpus-Based Comparison** .................................................................................................................. 22

Howe, Chad - University of Georgia - **A corpus analysis of category change: Spanish temporal constructions as syntactic hybrids** .................................................................................................................. 23

Imao, Yasuhiro - Osaka University - **Going beyond simple word-list creation using CasualConc** ................. 23

Ivaska, Ilmari, Bernardini, Silvia and Ferraresi, Adriano - University of Bologna - **Detecting Traces of Constrained Communication: A Corpus-Driven Comparison of Learner and Translated English** .......... 24

Jin, Bixi - The University of Hong Kong - **Using Multidimensional Analysis to Investigate the Discussion Sections of Research Articles in Chemical Engineering** .................................................................................................................. 24

Kang, Sanghee, Gwangmyeongbuk High School, Shin, Yu Kyoung, Georgia State University and Yoo, Isaiah WonHo, Sogang University - **Formulaic-language-based preposition instruction: An instructional experiment with Korean high school EFL students** .................................................................................................................. 25

Kauffmann, Carlos - São Paulo Catholic University - **Exploring stylistic dimensions of Brazilian nineteenth century fiction** .................................................................................................................. 25

Keller, Daniel - Northern Arizona University - **Do registers have psychological reality?: Evidence for register based priming in lexical access** .................................................................................................................. 25
Kia, Elnaz - Northern Arizona University - **Directive Use across Disciplines in University Classroom Discourse** .................................................................................................................................................. 26

Kim, Dongsung - Ewha Womans University - **Gutenberg Project Frequency & KeyWord-In-Context (KWIC) Viewer** ..................................................................................................................................................... 26

Kim, Jong-Bok, Kyung Hee University and Davies, Mark, BYU - **Transparent Free Relatives in English: A Corpus-based Perspective** .......................................................................................................................................................... 27

Kim, Minkyung and Crossley, Scott - Georgia State University - **Predicting second language writing and speaking scores using a testing corpus: The roles of linguistic features and individual differences** .......... 28

Kim, Susie - Michigan State University - **Predicting CEFR proficiency level using grammatical criterial features** ........................................................................................................................................................................... 28

Kim, Youjin, Georgia State University, Kang, Sanghee, Gwangmyeongbuk High School, Korea and Tywoniw, Rurik, Georgia State University - **Examining Dynamic Development in Speaking versus Writing among EFL Learners with Different Proficiency Levels: A Longitudinal Multi-Case Study** ................................................................................................................. 29

Kunath, Stephen, Georgetown University, Weinberger, Steven, George Mason University and Nelson, Jill, George Mason University - **Crowdsourcing Phonetic Transcriptions: Experimental Results from a New System for Transcription Management** ......................................................... 29

Kyle, Kristopher, Laflair, Geoffrey and Ziegler, Nicole - University of Hawai'i at Manoa - **The TMLE-Corpus project: Academic language in technology-mediated contexts** .................................................................................................................................... 29

Lake, William and Cortes, Viviana - Georgia State University - **Using lexical bundles to reassess disciplinary norms in Spanish & English humanities and social science writing** ........................................................................................................... 30

Lamance, Rachel - University of Arizona - **Visualizing the Rhetorical Moves of Source Incorporation: A Corpus Analysis** ................................................................................................................................................................. 30

Larsson, Tove, Université catholique de Louvain, Callies, Marcus, University of Bremen, Hasselgård, Hilde, University of Oslo, Judith Laso, Natalia, University of Barcelona, Paquot, Magali, Université catholique de Louvain, van Vuuren, Sanne, Radboud University and Verdaguer, Isabel, University of Barcelona - **Adverb placement in EFL academic writing: Going beyond syntactic transfer** .................................................................................................................................................. 31

Lindley, Jori - Purdue Fort Wayne - **How do features of constructions’ grammatical subjects (1st/2nd/3rd person and human/non-human status) affect usage patterns: Two case studies** ................................................................. 31

Liu, Jing - Wuhan University - **Conclusion Sections of Research Articles: Multi-dimensional Analysis** ...... 32

Liu, Yingying and Lu, Xiaofei - The Pennsylvania State University - **The Count/Mass Properties of Nouns and Chinese EFL Learners’ Article Use: A Corpus-Based Investigation** ............................................................................................................................ 32

Lu, Xiaofei, Casal, J. Elliott and Liu, Yingying - The Pennsylvania State University - **A corpus-based study of the rhetorical functions of syntactically complex sentences in research article introductions** ................. 33

Lu, Xiaofei, Yoon, Jungwan and Kisselev, Olesya - The Pennsylvania State University - **Matching phrase-frames to rhetorical moves in social science research article introductions** ......................................................... 33

Lutzky, Ursula - Vienna University of Economics and Business - “I sent my tweet because I couldn’t find an answer on that page.” British airlines’ use of Twitter for customer communication purposes.................. 34
<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mayer, Cristina</td>
<td><strong>Canonical dimensions of Web registers: correlating functional and lexical dimensions of register variation</strong></td>
<td>34</td>
</tr>
<tr>
<td>Mbodj, Ndye Bineta</td>
<td><strong>Exploring the discipline and writing experience interface in learners’ use of lexical bundles</strong></td>
<td>35</td>
</tr>
<tr>
<td>Monteiro, Kátia and Crossley, Scott</td>
<td><strong>Monolingual and bilingual processing norms as explanatory variables of L2 writing quality</strong></td>
<td>35</td>
</tr>
<tr>
<td>Mostafa, Tamanna</td>
<td><strong>Exploring complexity measures as indices of L2 oral development: A corpus-based study</strong></td>
<td>35</td>
</tr>
<tr>
<td>Motschenbacher, Heiko</td>
<td><strong>Language use in media coverage before and after coming out: A corpus-based study of texts on Ricky Martin</strong></td>
<td>36</td>
</tr>
<tr>
<td>Nam, Yunjung</td>
<td><strong>Korean EFL Learners’ Use of ‘give’ Ditransitive Constructions: Evidence from Corpus Research</strong></td>
<td>36</td>
</tr>
<tr>
<td>Nekrasova-Beker, Tatjana and Becker, Anthony</td>
<td><strong>Discipline-specific use of language patterns: The case of Engineering</strong></td>
<td>37</td>
</tr>
<tr>
<td>Nelson, Kyra and Hashimoto, Brett</td>
<td><strong>Using Corpora for Improving the Discourse Completion Task</strong></td>
<td>37</td>
</tr>
<tr>
<td>Nelson, Kyra</td>
<td><strong>Gendered Insults in YouTube Comments</strong></td>
<td>38</td>
</tr>
<tr>
<td>Novikov, Aleksey</td>
<td><strong>L1 Typology Effects on Complexity Features in Academic Writing</strong></td>
<td>38</td>
</tr>
<tr>
<td>Pacheco, Aline</td>
<td><strong>Exploring the discourse of international aviation from the Corpus of Pilot and ATC Communications (CORPAC)</strong></td>
<td>38</td>
</tr>
<tr>
<td>Paquot, Magali</td>
<td><strong>Relational co-occurrences as an index of oral proficiency in the Trinity Lancaster Learner Corpus</strong></td>
<td>39</td>
</tr>
<tr>
<td>Peng, Siyao and Zeldes, Amir</td>
<td><strong>Validating and Merging a Growing Multilayer Corpus – The Case of GUM</strong></td>
<td>39</td>
</tr>
<tr>
<td>Picoral, Adriana</td>
<td><strong>Compilation and automated annotation of a Portuguese learner corpus: Challenges and lessons learned</strong></td>
<td>41</td>
</tr>
<tr>
<td>Puga, Karin</td>
<td><strong>Using prosodically annotated corpora for language teaching and learning</strong></td>
<td>41</td>
</tr>
<tr>
<td>Quinley, Jason</td>
<td><strong>Modal Verbs in Requests and Proposals: Game-Theoretic Modeling Meets Corpus Data</strong></td>
<td>42</td>
</tr>
<tr>
<td>Ren, Haoshan and Crossley</td>
<td><strong>Psycholinguistic Features and Lexical Sophistication in TOEFL Writings: The Effect of Proficiency and L1 Distance</strong></td>
<td>43</td>
</tr>
</tbody>
</table>
Reppen, Randi, Northern Arizona University and Olson, Shannon, Laureate Education, Inc - **A look at lexical bundles across disciplines: Consistency and variability** ................................................................. 43

Roberson, Audrey - Hobart & William Smith Colleges - **Perpetrators or victims: A corpus-based critical discourse analysis of newspaper reportage on suicide** ................................................................. 44

Satake, Yoshiho - Surugadai University - **The effects of corpus use on error identification in L2 writing** .. 44

Schanding, Brian - Bluegrass Community and Technical College - **Cataphoric Shell Noun Use in the Construction of Modal Expression in Advanced Social Science Argumentative Writing** ................................................................. 45

Schnur, Erin, Hacking, Jane and Rubio, Fernando - The University of Utah - **MuSSEL: Designing and building a corpus of multilingual second language speech** ................................................................. 45

Serigos, Jacqueline - George Mason University - **¿Español or English? An automated tool for identifying Anglicisms in Spanish** ........................................................................................................ 45

Silva, Barbara - São Paulo Catholic University - **Cultural representations of age groups in Google Books.** 46

Skalicky, Stephen and Crossley, Scott - Georgia State University - **Examining creativity in the Japanese Learners of English corpus** ........................................................................................................ 46

Smith, Jordan - Iowa State University - **Building a Representative Corpus of Blogs: Methods, Challenges, and Constraints** ........................................................................................................ 46

Soon, Suet Ching and Chung, Siaw-Fong - National Chengchi University - **EFFECT-related Nouns in Thesis Abstracts: A Contrastive Study of Mandarin and English** ........................................................................................................ 47

Staples, Shelley, University of Arizona, Reppen, Randi, Northern Arizona University, Holcomb, Chris, University of South Carolina, Liu, Qiandi, University of South Carolina, Novikov, Aleksey, University of Arizona, Picoral, Adriana University of Arizona and Lan, Ge, Purdue University - **Lexico-grammatical Patterns in First Year Writing across L1 Backgrounds** ........................................................................................................ 49

Stewart, Ian, Pinter, Yuval and Eisenstein, Jacob - Georgia Institute of Technology - **Catalan independence and linguistic identity on social media** ........................................................................................................ 49

Stratton, James – Purdue - **A Diachronic Analysis of the Adjective Intensifier well from Early Modern English to Present Day English: A Corpus-Based Study** ........................................................................................................ 49

Suethanapornkul, Sakol, Georgetown University and Supasirapraka, Sarut, National Institute of Development Administration - **Second Language Users’ Knowledge of the Introductory-IT Construction: Corpus and Psycholinguistic Evidence** ........................................................................................................ 50

Tasker, David - Northern Arizona University - **Exploring Dimensions of Complexity in English through Multidimensional Analysis** ........................................................................................................ 51

Twyoniw, Rurik - Georgia State University - **Investigating Tense-Aspect Constructions and the Aspect Hypothesis using Collostructional Analysis** ........................................................................................................ 51

Veirano Pinto, Marcia, Federal University of Sao Paulo (UNIFESP) and Cortes, Viviana, Georgia State University - **Lexical bundles across North American movie genres: Discursive functions and levels of routinization** ........................................................................................................ 51
Whitty, Lauren - Victoria University of Wellington - A corpus-based analysis of ‘can’ and its dependency on categorical criteria and expanded context .................................................................................. 52

Wright, Heidi - Carnegie Mellon University - A Multidimensional Analysis of Stand-Alone Literature Review: Six Dimensions, Old and New .......................................................................................... 52

Wuiff, Stefanie, University of Florida and Gries, Stefan Th., UCSB - Particle placement in learner English: Measuring effects of context, first language, and individual variation ................................................................. 52

Yilmaz, Selahattin and Römer, Ute - Georgia State University - Phraseological Variability in English as an Academic Lingua Franca Writing .................................................................................. 53

Yoon, Jungwan and Casal, J. Elliott - The Pennsylvania State University - P-frames and Rhetorical Moves in Applied Linguistics Conference Proposals ................................................................................. 54

Zeldes, Amir - Georgetown University - A Multi-Dimensional Analysis of RST Discourse Relations in Eight Genres ............................................................................................................... 55

Zuppari, Maria Carolina - Catholic University of Sao Paulo - Lexical Variation in Written Academic English ................................................................................................................................. 56

POSTER PRESENTATIONS ........................................................................................................ 57

Becker, Amelia - Georgetown University - Corpus investigation of a selected finger markedness hierarchy in American Sign Language .................................................................................................. 57

Bullock, Barbara, The University of Texas at Austin, Serigos, Jacqueline, George Mason University and Toribio, Almeida, The University of Texas at Austin - Predicting the degree of language mixing in an open bilingual corpus ........................................................................................................ 59

Bychkovska, Tetyana - George Mason University - The Influence of Language and Context of Publication on the Use of Self-Mention in Single-Authored Ukrainian Research Articles .................................................................. 59

Diniz, Luciana, Portland Community College and Concordia University, Bunting, John, Georgia State University and Botelho, Sergio, Software engineer consultant - If I were to teach if-clauses: A corpus analysis of actual uses with ‘if’ .................................................................................. 60

Sylviane, Granger and Paquot, Magali - Université catholique de Louvain - The Louvain English for Academic Purposes Dictionary: From design to evaluation ........................................................................ 60

Kuhn, Tanara Zingano and Ferreira, José Pedro - CELGA-ILTEC, University of Coimbra - Introducing CoPEP, the Corpus de Português Escrito em Periódicos (Corpus of Portuguese from Academic Journals) ............. 61

Lavrentovich, Alexandra University of Florida - Using classifier features to detect linguistic transfer on morphemes .................................................................................................................. 61

Liu, Dilin, University of Alabama and Lei, Lei, Huazhong University of Science and Technology - Synonym under-differentiation in L2 vocabulary use: A corpus analysis of Chinese EFL learners’ use of “circumstance,” “demand,” and “significant” ...................................................................................... 61

Liu, Yang - Georgetown University - The Contributions of the Verb Complements "到(dào)" and "见 (jiàn)" to the Main Verb in Mandarin Chinese .............................................................................. 62
Liu, Shuyuan - The Pennsylvania State University - Inter-Group Variations of Discourse Markers Use Among Non-Native English Speakers: A Comparison between L1 Spanish and L1 Mandarin Chinese Groups in Advanced Academic Setting ................................................................................................................. 62

Olsen, Michael - University of Georgia - Establishing and investigating American sociopolitical keywords: the Bush-Obama Corpus ............................................................................................................................................. 63

Riccobono, Philip - Aichi Bunkyo University - Towards English for Baseball Purposes (EBP): A Corpus-based Investigation of Contemporary Vocabulary for Non-native English Speaking Specialists and Learners of Baseball English ........................................................................................................................................... 64

Starcevic, Katarina, maslansky + partners, Arthur, Paige, maslansky + partners and Butler, Anne, Ascension Health - Applications of corpus-based methodologies for market research: Three case studies ............... 64

Terrill, Kristin - Iowa State University - A corpus-based study of language ideologies in charter school performance audits ........................................................................................................................................ 64

Tywoniw, Rurik and Crossley, Scott - Georgia State University - Using Cohesive Features to Automatically Classify Integrated and Independent L2 Writing Assessment and Predict Human Judgments of Quality 65

Zhang, Genggeng - Georgia State University - A Corpus Approach of Teaching If-Conditionals to Chinese Aviation Maintenance Students .................................................................................................................. 65
WORKSHOP 1: An Introduction to Natural Language Processing Tools: The Curious Case of the Lexicon

Facilitators
Scott Crossley, Georgia State University, scrossley@gsu.edu
Kristopher Kyle, University of Hawai‘i, Manoa, kkyle@hawaii.edu

Time
1:00 pm – 5:00 pm, with breaks

Technical Requirements for the Workshop:
All participants will need to bring their own laptops with the following capabilities:
– TAALES 2.8 (http://www.linguisticanalysistools.com)
– MS Excel or similar
– JAMOVI (https://www.jamovi.org)

Data Requirements
The workshop organizers will provide participants with access to different set of learner corpora for natural language processing (NLP) analyses. However, it is strongly recommended that participants bring their own learner corpus for analysis (BYOC). Participants that bring their own corpus should ensure that:
1. The corpus is separated into individual files wherein each file represents output from a single learner at a single time point.
2. Each file is saved as a simple plain text format (.txt).
3. Each file has at least 50 words of learner production (the more the better though). For the purposes of this workshop, the learner samples should be spell corrected if possible.
4. The corpus is large enough to be representative (at minimum 100 samples).
5. Each file has a dependent variable of interest (i.e., something to predict such as proficiency/grade level, length of study, demographic information, individual differences, test performance)

Workshop Description
This workshop will focus on introducing participants to the basic notions underlying natural language processing tools and how they can be used in second language research. While the workshop will provide an overview of NLP tools and language features commonly assessed using NLP tools, this workshop will focus on using NLP tools to examine lexical sophistication. No computer science background is needed to join the workshop, but participants should be familiar with data analytics, data formatting, and basic statistical analyses.

The workshop will be divided into the following sections:
1. Overview of NLP and NLP research in second language settings
2. Available NLP tools (focusing on those that are open-source)
3. Common methods and pitfalls in NLP analyses
4. An introduction to the Tool for the Automatic Analysis of Lexical Sophistication (TAALES)
5. Data analysis
WORKSHOP 2: Lexical Multi-Dimensional Analysis

Facilitator
Tony Berber Sardinha, Catholic University of Sao Paulo, tonycopuslg@gmail.com

Time
10am to 12 noon

Workshop Description
Multi-dimensional analysis is an approach to the study of register variation introduced by Douglas Biber in the 1980s (Biber 1988; Berber Sardinha and Veirano Pinto, 2014; in press). Biber designed the MD analysis framework in order to capture the parameters underlying register variation, which are called dimensions. Originally, a dimension is a set of correlated linguistic features, captured through a factor analysis, that is interpreted in functional terms. In such functional MD analyses, the linguistic features in question are typically structural features, such as parts of speech (nouns, adjectives, personal pronouns, etc.), clause types, and stance constructions. In contrast, in a lexical MD analysis, the features are entirely lexical, such as the actual tokens, lemmata, n-grams, or collocations present in the texts. The resulting lexical dimensions can reflect a range of linguistic phenomena enacted by the lexis, such as the prevailing discourses, the cultural representations, and the major topics discussed in the texts, among others. In this workshop, we will cover such topics as how to sample and organize data for a lexical MD analysis, how to conduct factor analyses based on lexical data, and how to interpret lexis-based factors into lexical dimensions.

Technical Requirements for the Workshop
In preparation for the workshop, please have a working copy of SPSS installed on your laptop. A trial version is available for free at https://www.ibm.com/account/reg/us-en/signup?formid=urx-19774. Please note that this trial version will work for 14 days only and therefore attendants should time the installation so that it will be working during the workshop. In addition, please download data files and other support materials at https://bit.ly/2sByFsA.

References
WORKSHOP 3: Complexity in Writing Development: Untangling Two Approaches to Measuring Grammatical Complexity

Facilitators
Bethany Gray, Iowa State University, begray@iastate.edu
Shelley Staples, University of Arizona, slistaples@email.arizona.edu
Jesse Egbert, Northern Arizona University, jesse.egbert@nau.edu

Time
10:00am – 5:00 pm, with a lunch break and afternoon coffee break (on your own)

Technical Requirements for the Workshop
All participants will need to bring their own laptops with the following capabilities:
– AntConc 3.5.7 or newer
– MS Excel or similar
– Text Editor (e.g., Notepad, Wordpad, etc.)

Workshop Description
Linguistic complexity is often investigated with respect to language development, based on the assumption that more advanced L1/L2 writers use more complex language and produce more complex texts. But what constitutes complex language? How do we operationalize complexity to measure it in language production? How are approaches to complexity similar and different? How is complexity mediated by proficiency or level, register or genre, and other contextual factors?

This workshop focuses on these questions for one type of complexity – grammatical complexity – as it relates to L1 and L2 writing development. The goal of the workshop is to explore how fundamentally distinct measures approach the same underlying construct, to gain practice analyzing grammatical complexity in written texts, and to see a selection of complexity variables applied in research on writing development. The workshop begins with a brief comparison of two major approaches to grammatical complexity: the holistic (T-Unit) approach and the register/functional approach. Then, the workshop is divided into three parts:

Part 1: Hands-On, Practice-Oriented Session in Coding and Compiling Complexity Variables

Part 2: Research Synthesis on the Development of Complexity in Academic Writing

Part 3: Roundtable Discussion

Part 1 is a hands-on, practice-based session in which participants will use manual and automatic corpus tools to analyze authentic texts for a range of complexity measures from both the holistic (T-Unit) and register/functional traditions. Participants will code and annotate complexity features and then quantify their coding using automatic procedures with AntConc 3.4.4 (Anthony, 2014). Issues such as reliability (precision, recall) and interrater reliability will be addressed.

Part 2 is a research synthesis of recent work by the workshop organizers and colleagues on the development of grammatical complexity in academic writing, focusing on studies within the register/functional tradition.

Part 3 provides an opportunity for extended discussion between workshop participants on the issues practiced and discussed in Parts 1 and 2.
Berber Sardinha, Tony - Catholic University of Sao Paulo, Brazil - Corpus Linguistics and Historiography

Tony Berber Sardinha is a professor with the Graduate Program in Applied Linguistics and the Linguistics Department, the Catholic University of Sao Paulo, Brazil. His recent publications include Multidimensional Analysis: 25 Years on (2014, John Benjamins), Working with Portuguese Corpora (2015, Bloomsbury), Metaphor in Specialist Discourse (2015, John Benjamins), and Multidimensional analysis (forthcoming, Routledge). His main research interests are multi-dimensional analysis, the use of corpora for historiography and historical discourse analysis, the development of corpus methods for metaphor retrieval, the application of corpus techniques in forensic linguistics, corpus-based analysis of translated texts, and the interface between corpus linguistics and language teaching. He is on the board of several journals and book series such as the International Journal of Corpus Linguistics, Corpora, Metaphor and the Social World, Studies in Corpus Linguistics, Register Studies, and Metaphor in Language, Cognition, and Communication.

Abstract

Historiography can mean both the development of principled approaches for the study of history and the study of history based on such approaches. The goal of this talk is two-fold: first, to make the case that corpus linguistics is well-suited to engage with historiography and, second, to present a study that demonstrates corpus linguistic historiography. The historical accounts derived from a corpus perspective are entirely text based—that is, they represent the discourses that actually emerge from the texts, mined directly from primary sources. The principal research questions asked in this kind of investigation include: What are the different discourses that exist in the field? How are they constituted? How do these discourses vary over time and across the publications in the field? To answer such questions, corpora comprising as long a time frame as possible need to be collected. The corpus used in the study covers a 70-year span of publications, containing the full collection of papers published in major journals. The corpus was tagged for part of speech and analyzed by means of a lexical multi-dimensional analysis. Three multivariate statistical procedures were employed in the analysis of the corpus. First, a factor analysis was used to determine the groupings of lexical items cooccurring across the texts. These factors were interpreted in terms of such constructs as lexical fields, topic selection, and semantic preference, thereby giving rise to the lexical dimensions underlying the variation among the texts over time. The second procedure was a discriminant function analysis, which was employed to test the robustness of the dimensions as predictors of the time periods in which the texts were published. The third procedure was cluster analysis, which was run on the dimension scores in order to group the texts around the prevailing discourses in the field. Finally, a timeline was derived from the cluster analysis, reflecting the major time periods in the discourse of applied linguistics over the course of its history.
Gray, Bethany - Iowa State University, Ames, Iowa - The Role of Register Analysis in Academic Writing Research

Bethany Gray is Associate Professor of English at Iowa State University (program in Applied Linguistics and Technology). Her research uses corpus linguistics methods to explore register variation in English, with a particular focus on the phraseological and grammatical characteristics of texts, English for Academic Purposes and disciplinary variation, and second language writing development. Her work has appeared in journals such as Applied Linguistics, International Journal of Corpus Linguistics, Corpora, TESOL Quarterly, Journal of English for Academic Purposes, English for Specific Purposes, Journal of English Linguistics, and English Language & Linguistics, and has been published in book form through Cambridge University Press and John Benjamins. She is a co-founding editor of the journal Register Studies (John Benjamins).

Abstract

Register analysis provides a framework for the systematic consideration of situational factors that underlie functional linguistic variation across texts. This presentation explores the importance of register analysis as a key component of corpus-based research on the linguistic characteristics of academic writing, arguing that such research can benefit from greater register awareness and specificity. Analyses of grammatical variation in expert and L2 production demonstrate how an explicit consideration of registers and sub-registers within academic writing has implications for corpus design and representativeness, corpus-based research design, investigations of disciplinary variation, and studies of writing development.
Gries, Stefan Th. - University of California, Santa Barbara and Justus Liebig University Giessen - Towards a Unified Tupleization of Corpus Linguistics

Abstract

In a sense, this is the final talk in a series of three plenary talks this year on how corpus linguistics should change its quantitative analyses of corpus data. In two previous talks, I argued that corpus linguistics is held back by our (often implicit) decision(s) to not utilize all the different dimensions that corpus data have to offer and to use measures that sometimes even mask information that would be relevant to our analyses. In those previous talks, I then proceeded by discussing dimensions of information that we need to use more (e.g., variability and/or dispersion) and measures that are better suited to providing us with the relevant information (e.g., directional association measures and/or entropy). In this talk, I will try and, to say it with a great amount of hubris, exemplify how I think we should be doing corpus statistics. More specifically, the field is teeming with many different measures for different purposes (there dozens of association measures, more than a dozen dispersion measures, many association measures are also used for key word statistics, …) that suffer from some of the problems alluded to above, that are diverse and often not particularly comparable, and that defy a unified perspective on the kinds of distributional data we are always working with. I will therefore propose and of course exemplify an alternative statistic that can be applied to most standard uses in corpus-linguistics statistics; avoids many shortcomings existing measures exhibit; has a solid theoretical grounding utilized across a huge number of other disciplines (including cognitive science); allows us to keep different pieces of information that existing approaches routinely and unhelpfully conflate separate; even offers new avenues of research.
A considerable number of research studies investigated the relation between lexical complexity and raters’ judgment of writing quality. Three main aspects of lexical complexity have been generally posited to affect the human judgment of writing quality: diversity, sophistication and lexical cohesion (Crossley & McNamara, 2012; Guo, Crossley & McNamara, 2013; Knoch, Macqueen and O’Hagan, 2014). Despite the general agreement on the positive correlation between lexical complexity and writing proficiency, little is known on how aspects of lexical complexity can be influenced by task characteristics (e.g. narrative vs argumentative, integrated vs independent). Such knowledge is of a paramount importance especially for developing scoring rubrics of integrated writing tasks. Hence, there is a need for more studies that take task characteristics into consideration when selecting appropriate measures of lexical complexity. Bearing this in mind and in light of the importance of lexical complexity in reflecting the development of ESL writing, the current study attempts to examine how ESL learners of various placement levels differ in their use of diverse, sophisticated and cohesive vocabulary. A corpus of 582 texts representing three ESL placement levels was analyzed using Indices from TALLES (Kyle & Crossley, 2015) and Lexical Complexity analyzer (Lu, 2012) (free available tools online). Six indices of lexical complexity were entered in MANOVA as dependent variables with the placement level as the independent variable. The results revealed significant differences among the placement levels in their scores on the three aspects of lexical complexity. The subsequent univariate tests revealed a statistically significant effect of the placement level on all indices -with the exception of hypernym nouns and verbs. Differences between adjacent lower levels were however not consistent across the three complexity aspects. The findings highlight the need of interpreting the results of lexical complexity indices in terms of written tasks’ characteristics and requirement. Implications on the assessment of integrated written tasks and pedagogical practices in ESL writing courses are discussed accordingly.

Elaborated topical development in which each sentence has typically a topic and a comment has been posited as an indication of textual coherence (Connor, 1987, Lautamatti,1987). Analysis of organizational patterns requires identification of three main organizational elements: topical subjects, mood subjects and initial sentence elements. Topical subjects in authentic texts are usually preceded by initial sentence elements (ISEs) and or mood (grammatical) subjects. Organizational patterns of topics and the linguistic realizations of mood subjects and ISEs provide not only insights on how proficient writers introduce their arguments to the reader (Goutsos,1997; North, 2005) but also how these discourse organizational patterns are influenced by contextual factors of discourse: purpose, targeted audience and context (Nwogu & Bloor,1991; Wang,2007). Unfortunately, ESL research on writing gave little attention to these aspects despite the fact that organizational patterns constitute a significant criterion in writing assessment. Little is known on linguistic realizations of ISEs and mood subjects and how these structural elements are used in contextual framing of topics in ESL writing. Such line of research is needed to highlight the connection between the use of certain grammatical structures with their rhetorical functions in discourse. With this in mind, this study examined 59 texts from three placement levels in terms of organizational patterns, syntactic realizations of mood subjects and ISEs along with the functions they serve in discourse. Texts were first divided into t-units to facilitate the identification of ISEs, topical subject and the mood subject of each T-unit. The structures in question were coded manually and a simple Python code was used to calculate the number of t-units, proportion of progression types, non-topical mood subject and ISEs per t-unit. The analysis revealed significant variation among the three levels in topical development patterns, frequency use and the linguistic realizations of mood subjects and ISEs in the three placement levels.

In this paper, we present an advanced n-gram generation and analysis tool called AntGram, which is designed to facilitate p-frame research and allow for more rigorous studies that go beyond an analysis of repeatedly used fixed word sequences. AntGram can generate and sort traditional n-grams of any length, displaying the results within the interface. The tool can also generate p-frames with any number of internal or global open slot positions. All results generated by the tool can be filtered by frequency or range (dispersion) and restricted to only those that contain or exclude certain lexical items. Various export options are included to facilitate further analysis.

AntGram has been applied in a range of different research tasks and data-driven learning (DDL) environments. In this paper, we will give an overview of the tool, explain its main features, and then apply it in the analysis of phraseological patterns in proficient student writing as captured in the 2.6-million word Michigan Corpus of Upper-level Student Papers (MICUSP). We will demonstrate how the tool allows us to gain a better understanding of the dominant patterns and connected functions expressed in successful student papers across a range of academic disciplines. We will also highlight the importance of extending...
Appel, Randy, Waseda University and Szeib, Andrzej, Trent University - Linking adverbials in L2 English academic writing: L1 related differences
Linguistic analyses of corpora
Appropriate and effective use of linking adverbials (e.g., furthermore, in addition, on the other hand) plays an important role in discourse cohesion. In existing studies on this subject, researchers have primarily relied on native/non-native writer contrasts to highlight production patterns related to specific groups of L2 English users (e.g., Carrio-Pastor, 2013; Chen, 2006. However, while this body of research has pointed to several tendencies associated with writers of particular L1 backgrounds, methodological limitations have prevented a distinction between L1-specific and universal features of L2 English writing from being made.

The goal of the current study was to move beyond a focus on native/non-native comparisons and use more closely comparable corpora to uncover a more accurate understanding of how L2 English writers make use of linking adverbials. To achieve this goal, this study used a specially designed corpus of 150 L2 English academic essays produced in an English as a Second Language setting where differences in target language proficiency, writing conditions, and essay type had all been controlled. Applying a contrastive interlanguage analysis targeting writers from three different L1 backgrounds (Arabic, Chinese, and French), the assembled corpus was used to document differences in how each L1 group made use of linking adverbials to help structure their academic English essays. Findings include overuse of additive linking adverbials (e.g., in addition, also) by L1 Arabic writers, contrastive linking adverbials (e.g., however, on the other hand) by L1 Chinese writers, and appositional linking adverbials (e.g., in fact, indeed) by L1 French writers. Methodological and pedagogical implications of these findings are discussed.

References


Araujo, Rafael Fonseca de, Berber Sardinha, Tony and Delfino, Maria Claudia - Sao Paulo Catholic University - Representations of American identity: a multidimensional analysis of Google Books trigrams around North, Central, South and Latin America
Linguistic analyses of corpora
According to Stubbs (1996 p. 185), national stereotypes can be constructed by the use of recurrent words and phrases. Naming and labelling are powerful ideological tools. Corpus analysis allows us to study directly the labels which are repeatedly applied to groups of people. The goal of this study is the identification of representations of the Americas through the analysis of the most common trigrams found in the Google Books Ngrams from 1800 to 2008. We will report on a lexical multidimensional analysis that takes as a basis the trigrams formed with the phrases North American, Central American, South American and Latin American. For each of these, a different lexical MD analysis was carried out. A lexical MD analysis uses as input solely lexical words, as opposed to a mainstream functional MD analysis, which utilizes mostly grammatical linguistic features. A factor analysis identified the major sets of correlated words for each subcontinent. These factors were interpreted with respect to the representations indexed by the lexis. As a result, 23 lexical dimensions were identified which encapsulate some of the most common Pan-American social representations, such as ‘American Society and Culture’ and ‘Economic Power’ for North American, ‘Political Revolution and Economic Development’ and ‘Immigrants, Culture and Women’ for Latin American, ‘Political Crisis’ and ‘Civilizations and Colonies’ for Central American and ‘Neighboring Nations and Jungle’ and ‘Business’ for South American, the one with the most dimensions. The paper will present the individual dimensions and compare them, in order to determine both unique and shared representations.

Balasubramanian, Chandrika - Sultan Qaboos University - Now we are having lot to think about, isn’t it? Indian English: What happens next?
Linguistic analyses of corpora, Applications
The rising status of English as a world language has resulted in the emergence of several new varieties of English that have been legitimized by such expressions as New Englishes and New Varieties of English. Accepting the idea of English as a World Language has allowed a much-needed movement away from the previously accepted notions of nativeness and non-nativeness, and today, they are seen as systems unto themselves, as opposed to deviant forms of traditional native varieties. Corpus linguistics methodology has aided tremendously in the investigation of New Englishes, and scholars such as Lange (2011) are now calling for New Englishes to serve as pedagogical models, particularly in ESL and EFL contexts.

Indian English, one new variety, has been extensively studied. While initial research focused on identifying characteristic ‘features’ of the variety, more current research has focused on describing the variation within the variety (Balasubramanian, 2009; 2016).
The aims of the current paper are three-fold: to provide a) a description of six grammatical features in a small sub-corpus of spoken and written English of young users of English in India to illustrate that that Indian English has entered Schneider’s Endonormative Stabilization and Differentiation Phases (Schneider, 2003; 2007) of development, b) a discussion of whether there is agreement among language teaching practitioners in India about the acceptability of Indian English in academic English in classrooms in India based on a study conducted in 2017; and c) to open a discussion of the broader pedagogical and social implications of using Indian English as a norm-providing model, even in ESL contexts like India.

Becker, Kimberly and Feng, Hui Hsien - Iowa State University - Stance in Unpublished Student Writing: An Exploration of MICUSP
Linguistic analyses of corpora

Applied linguists have long sought to uncover variation in the stance features of academic language -- the ways that academic writers employ objective versus subjective discourse and the degree to which they hedge or boost claims. Many examinations of stance have an underlying assumption that findings about expert writing can be a model for novices learning to navigate the requirements of university language (Biber, 2006; Hyland, 1999). However, most studies whose results purport such models draw information from a few specific academic registers, such as research articles or book reviews. At a student level, there is much research on dissertations/theses (Charles, 2006; Flowerdew, 2015; Geng & Wharton, 2016), but less on other student-oriented registers (e.g., reports, proposals, essays, etc.). It is because of this gap that the present study seeks to examine stance in student writing within other registers, specifically, those encountered by university-level students.

This presentation reports on an investigation of the stance features within texts from the Michigan Corpus of Upper Level Student Papers (see O'Donnell & Römer, 2012 and/or Römer & O'Donnell, 2011) (MICUSP). The analysis employs an adaptation of Biber's (2006) stance framework to examine lexi-co-grammatical realizations of stance (and their meanings/functional types) as operationalized by adverbs, modal verbs, and stance complement clauses controlled by verbs, adjectives, and nouns. The study examines per-text observations via automated and manual analysis of the 826 student papers in MICUSP. The goal of the presentation is to describe the frequency of linguistic stance features and explore how they vary across both registers and disciplines. The results expand the literature about the nature of university language by providing a view of the written work done by students. This description could be a move toward building a more robust model of successful student writing, thus providing more access to the norms of a discourse community. The results of this study contribute to the larger conversation about stance in academic writing -- especially as it applies to successful student writing.

References

Belcher, Diane, Friginal, Eric and Yilmaz, Selahattin - Georgia State University - Exploring the Academic Blogging Practices of English as A Lingua Franca (ELF) Scholars
Linguistic analyses of corpora

Blogging is an emerging academic genre that can be viewed as an open and less formal platform where scientific dialogues and professional identities are freely shared (Kirkup, 2010, Shenma, Bar-Ilan, & Thelwall, 2012). While research on academic blogging is becoming increasingly common, little is known about EAL (English as an additional language) academics’ blogging practices. Although limited, English as a lingua franca (ELF) writing research shows that EAL academics tend to produce discursively hybrid digital texts that are being widely read in academia (Carey, 2013; Tribble, 2017). Furthermore, exploring ELF writers' blogging could help us gain deeper insights into the characteristics of this relatively new academic genre and the hybridity of ELF writing in it.

This presentation focuses on the analysis of the blog entries by 40 ELF academic writers in a sub-corpus of the corpus of Written English as a Lingua Franca in Academic Settings (WiELFA, 2015). The multidimensional analysis tagger (MAT) by Nini (2015) was used for the multidimensional linguistic analysis (MDA) of the texts (Biber, 1988). Biber's (1988) six dimensions of textual features such as narrative, persuasive and scientific style were calculated and compared to the different genres such as
Over the years, television programs have been classified in a large number of different ways, however, to date no classification scheme has been developed that relied primarily on the linguistic features of the shows as a basis for the taxonomy. The goal of the current study is to develop such a linguistic taxonomy of television programs shown in the United States. Our taxonomy has been developed that relied primarily on the linguistic features of the shows as a basis for the taxonomy. The goal of this typology of American television programs differs from previous taxonomies with respect to both the number of types and the underlying parameters of variation among texts, and are based on groupings of correlated linguistic characteristics. The corpus employed for this analysis was the (name omitted) corpus, consisting of 31 registers (programs), totaling 5.3 million words. The corpus was carefully designed so as to represent the multitude of programs presented on contemporary American television (terrestrial and cable). In addition, the size of each corpus section was calibrated so as to reflect the inherent linguistic variation among the texts, following Biber’s (1993) proposal for corpus representativeness (cf. Berber Sardinha, 2014). The linguistic typology was based on a cluster analysis of the dimension scores of each text, following Biber’s (1989) proposal for text type identification. In an MD text typology, texts types are [g]roupings of text that are similar in their linguistic form’ (Biber, 1989: 13). Six clusters were identified, which upon interpretation correspond to the major text types for American television programming. Overall, this typology of American television programs differs from previous taxonomies with respect to both the number of types identified and the categories determined. In the paper presentation, examples of each cluster will be provided, in addition to interpretable labels of the clusters and a detailed discussion of the results.

References


Berber Sardinha, Tony, Sao Paulo Catholic University and Veirano, Pinto Marcia, Sao Paulo Federal University - A multi-dimensional text typology of American television programs

Linguistic analyses of corpora

Over the years, television programs have been classified in a large number of different ways, however, to date no classification scheme has been developed that relied primarily on the linguistic features of the shows as a basis for the taxonomy. The goal of the current study is to develop such a linguistic taxonomy of television programs shown in the United States. Our taxonomy has been developed from a multi-dimensional (MD) corpus-based perspective (Biber, 1988), using the dimensions of variation across American television registers uncovered by Berber Sardinha and Veirano Pinto (in print). The dimensions are the underlying parameters of variation among texts, and are based on groupings of correlated linguistic characteristics. The corpus employed for this analysis was the (name omitted) corpus, consisting of 31 registers (programs), totaling 5.3 million words. The corpus was carefully designed so as to represent the multitude of programs presented on contemporary American television (terrestrial and cable). In addition, the size of each corpus section was calibrated so as to reflect the inherent linguistic variation among the texts, following Biber’s (1993) proposal for corpus representativeness (cf. Berber Sardinha, 2014). The linguistic typology was based on a cluster analysis of the dimension scores of each text, following Biber’s (1989) proposal for text type identification. In an MD text typology, texts types are [g]roupings of text that are similar in their linguistic form’ (Biber, 1989: 13). Six clusters were identified, which upon interpretation correspond to the major text types for American television programming. Overall, this typology of American television programs differs from previous taxonomies with respect to both the number of types identified and the categories determined. In the paper presentation, examples of each cluster will be provided, in addition to interpretable labels of the clusters and a detailed discussion of the results.

References


**Bertoli, Patricia and Shepherd, Tania - University of the State of Rio de Janeiro - Corpus-informed analyses of research papers in English as a lingua franca: the case of concessive structures**

Linguistic analyses of corpora

Academic knowledge as constructed by a research paper has been described as a socially accomplished, albeit tacit “agreement between a writer and a potentially skeptical audience” (Hyland, 1999: 342). As a result, the research paper may be seen as the product of two needs, namely, the researchers’ need to make their research public, but, more importantly, their need to write persuasively. This latter task is significantly harder if the research is to be shared in English in international journals by academics whose L1 is not English. The question often asked is whether these research papers present any homogeneity in terms of persuasive features regardless of the writer’s L1, in order for them to be accepted for publication. Drawing from Charles’ (2011) investigation into persuasive patterns within academic writing produced in British universities, this paper focuses on a specific linguistic feature which contributes towards the construction of arguments in the presentation of research: signals of contrast/concession and accompanying ‘semantic sequences’ (Hunston, 2008 and 2011), i.e., patterns of “meaning elements”. The present study uses the SciEFL (Science as a Lingua Franca) corpus as data. This was compiled by Mauranen and collaborators (2015), and consists of 150 research papers (789,925 tokens) written by a variety of L2 users of English. The SciEFL papers have all been accepted for publication internationally but have not undergone proofreading by native speakers of English. Given that English as a Lingua Franca (or ELF), as posited by Seidhoffer (2012: 137) is “a linguistic resource which is drawn on as a common means of communication” in the globalized world, this study hypothesizes that the phenomenon of concessive structures may display similar quantitative and qualitative features in the SciEFL corpus as those identified within the data studied by Charles (2011).

**Biber, Doug, Egbert, Jesse and Keller, Daniel - Northern Arizona University - Reconceptualizing register as a continuous construct**

Linguistic analyses of corpora

Registers have traditionally been treated as discrete, non-overlapping categories, and most corpora are organized in terms of such categories (e.g., fiction, academic prose, press reportage). But the register of a text is not always obvious. A recent analysis of register variation on the web employed raters to identify discrete situational characteristics of documents (e.g., interactivens, narrative purpose), leading to identification of a register category (see Egbert et al. 2015). However, raters often disagreed in their coding. In that analysis, these disagreements were interpreted to reflect the fact that documents are often instances of ‘hybrid’ registers (see references below).

The prevalence of these hybrid registers suggests that documents and registers could be investigated from the outset in a quantitative, continuous situational space (see Sharoff, to appear). That is, situational parameters could be operationalized as continuous variables, allowing coders to evaluate the extent to which a document is, for example, interactive or opinionated.

We explore this possibility in the present talk, based on a reanalysis of 1,000 documents randomly selected from the CORE corpus. Each document is coded by two raters for 23 situational parameters (e.g., ‘the writing is based on personal experience’, or ‘the purpose of this text is to persuade the reader’). Each parameter is expressed as a 6-point likert scale, allowing us to treat the parameters as continuous variables. Using multivariate statistics, we analyze the underlying dimensions of situational variation. The resulting groupings of documents thus represent registers that are well-defined in a continuous space of situational variation. Further, any individual document can be described for how core or peripheral it is within its category. We discuss what our findings reveal about register variation on the web, in addition to the theoretical implications of our study as they relate to a general reconceptualization of register as a continuous construct.

References


A corpus-based approach

Linguistic analyses of corpora, Applications

The present study examines how variable slots within a fixed formulaic construction are accepted by second language learners and native speakers of English via an acceptability survey. Of particular interest is whether frequency of occurrence of a variable construction plays a determining role in acceptance. In addition, proficiency is considered. A corpus is used to determine frequency of the selected variable constructions and a comparison of the acceptability judgments is made (Hunston, 2002; Hyland, 2008). Results indicate that overall, acceptability of variability within formulaic sequences closely matches the frequency of use as demonstrated in a corpus independent of status as a native or second language speaker of English. However, proficiency seemed to impact results as acceptability was more closely related to the frequency of use in a corpus for participants of higher proficiency levels. These results support hypotheses made by Usage-based theorists as they indicate frequency is the driving force behind acceptance of a variable construction (Ellis & Wulf, 2015). This has several pedagogical implications as it demonstrates how frequent exposure to a construct may help develop a learner’s familiarity with that construct. Additionally, it demonstrates how we can use knowledge of natural language patterns to inform our instruction.

Classroom Discourse (CD) and Teacher Talk (TT) have received much attention over the years across a range of research perspectives. Much of recent CD work draws on Conversation Analysis (CA) and, more recently, there has been growing synergy between CA and Corpus Linguistics (CL). In the area of Language Teacher Education (e.g. CELTA), trainee teachers are introduced to the concept of TT as something to be minimised and very little mention is made of the types of TT used in the classroom. In other words, the focus is on reducing the quantity of TT rather than on the nature of the talk itself.

This paper will report on a quasi-longitudinal mixed-methods study of TT which uses data from classroom interactions over 9 months and also draws on interviews with the teachers. This 150,000 word corpus comprises 15 Native (NS) and Non-Native English (NNS) speaking teachers, at varying career stages, ranging in age from 23 to 36. Specifically, this paper will compare and discuss findings on teachers’ spoken feedback in the classroom in relation to the Native English-Speaking teachers and their Non-Native English Speaking counterparts. The results indicate that, regardless of career stage, the NNS teachers are more direct in their approach to corrective feedback, with more examples of direct repair and form-focused feedback than the NS teachers. Additionally, while the majority of teachers interviewed stated that they were robust in their positive feedback, the corpus data showed a disparity between NS and NNS feedback whereby the NNS teachers seemed to show more positive reinforcement than their NS counterparts. This brings to light the usefulness of using corpus data in teacher training, in a mixed methods format, not least of all to highlight where there is disparity between teacher beliefs about TT and how they actually use it in the classroom context.

In this software demonstration, we briefly introduce two brand new software tools #LancsBox and BNC lab, both developed at Lancaster University. Following the recent debate in the field (e.g. McEnery & Hardie 2011; Kilgarriff 2012; Gries 2013; Lijffijt et al. 2014; Brezina & Meyerhoff 2014; Brezina et al. 2015; Gablasova et al. 2017) and responding to the challenges identified in the debate, we have developed a software tools that incorporates a number of existing analytical techniques and add new innovative methods that enable more efficient and sophisticated exploration of the data. #LancsBox and BNC lab can be used by linguists, language teachers, translators, historians, sociologists, educators and anyone interested in quantitative language analysis. They are free to use for non-commercial purposes.

This software demonstration highlights innovative features of the new tools with the focus on data visualization. In addition, we also discuss more general principles of statistical data analysis and visual data display that can be used for effective presentation of quantitative findings based on language corpora.
University administrative or institutional writing aims often to provide faculty, staff, students, alumni, and donors understanding and guidance on institutional aims and goals. Administrative texts, or institutional writing (Biber, 2006), includes, but is not limited to, mission statements (Kreber & Mhina, 2007), school prospectuses (Bano & Shakir, 2015) and university strategic plans (Chan, 2009). Overall linguists have typically analyzed these texts a reflection of institutional power imbalances and a financial focus through genre analysis (Chan, 2009; Comut, Giroux, & Langley, 2012) and corpus guided approaches (Berger, Friginal, and Roberts, 2017; Nasti, Venuti, & Zollo, 2017) or multidimensional analysis (Bano & Shakir, 2015; Alafan, 2016). This interest in university administrative texts has begun to include a newer text type, International Strategic Plans (Engel and Siczek, 2017). Lacking from the observation that universities are adopting more business like discourse is a systematic comparison into the linguistic variation of academic institutional writing against non-academic, business writing. This paper aims to add International Strategic Plans (ISPs) to the register of university administrative documents and explore its register variation through comparing three corpora: U.S. based university International Strategic Plans (N = 41 texts), City and Government Strategic Plans (N = 41 texts) and Corporate Annual Reports (N = 40 texts), using Multidimensional Analysis. Each corpus was analyzed using the Multidimensional Tagger (Nini, 2005), which assigns dimensions as designed by Biber (1988, 1995) to texts. Once the dimension scores were identified, an ANOVA was ran to detect significant differences between dimension scores. Texts with extremely high or low dimension scores within the significant findings were then analyzed for their linguistic patterns and trends to find examples of their situationally-defined varieties (Biber, 1995; Conrad & Biber, 2001, Biber 2006). Results from this exploratory study indicate International Strategic Plans demonstrate no significant difference in terms of their use present tense, 2nd person pronouns, and private verbs use when compared to the other corpora. However, one significant finding demonstrated that the Annual Reports corpus had a higher amount of abstract or technical language when compared to ISPs. An analysis of ISPs' and Annual Reports' technical writing reveals a clear distinction between expected readership and different approaches to financial or legal prose. These ideas along with examples will be discussed to further our understanding of the similarities and differences between university administrative writing and corporate writing.

Automated lexico-grammatical annotations can facilitate and enrich corpus analyses, including second language (L2) analysis (Van Rooy, 2015). However, since most parsers and Part of Speech (POS) taggers have been trained on native speaker (L1) data, high accuracy rates and thus the utility of these annotations is not guaranteed on L2 data (Granger, 2008). In this study we explore the affordances and limitations of three such annotators: 1) a Probabilistic ContextFree Grammar parser developed by The Stanford Natural Language Processing Group (Klein & Manning, 2003); 2) a statistical dependency parser, namely MaltParser (Nivre & Hall, 2005); 3) and a rule and frequency-based lexico-grammatical tagger (Biber, 1988 & 2006). We use POS tags from CLAWS (Garside & Smith, 1997) with the first two parsers. The MaltParser was trained with a combination of features, including Universal and Biber tags, and both L1 and L2 data. We focus on lexico-grammatical patterns, especially complexity and stance aspects, including both phrasal (pre-modifying nouns, attributive adjectives) and clausal (relative clauses, verb and noun complement clauses, finite adverbial clauses) features. Our initial corpus consists of first year writing from four institutions across the US (over two million words in total), which was then carefully sampled for argumentative essays to evenly represent five L1s (Spanish, Korean, Arabic, Chinese, English). These data have been tagged and parsed using the tools mentioned above, and then examined for the lexico-grammatical patterns that can be derived from the different output. We will discuss whether the annotation tools allow us to extract the target language features, detailing performance across the L1s in our corpus. A comparison and analysis of these tools, including statistical results using random sampling with replacement (Berg-Kirkpatrick, Burkett & Klein, 2012), will be presented. We will also report error rates both per feature and per tool, discussing the ability to reliably compare statistics from different packages.

REFERENCES


**Charles, Maggie - Language Centre, Oxford University - Corpus- assisted editing for doctoral students: More than just concordancing**

**Applications**

This paper presents and evaluates a course which teaches doctoral students how to use corpus tools for editing their theses. For this purpose, students built two self-compiled corpora consisting of: 1) research articles in their field; 2) their thesis chapters. Although much has been written about the benefits of data-driven learning, particularly for academic writing (Boulton and Cobb, 2017; Yoon, 2011), that research focused primarily on concordancing. Little attention has been paid to the potential of other corpus tools; nor has their application to the task of editing been much explored. This paper argues that each corpus tool has specific affordances that are relevant to the task of editing a thesis and that using the full suite of tools (Concordance, Clusters, Collocates, N-grams, Concordance Plot, Word List, Keyword List) allows students to address problems not just of lexicogrammar, but also of organisation and content. For example, Concordance Plot provides a graphic representation of the distribution of a search term throughout the corpus files. By applying this tool to their thesis corpus and using a term that is central to their argument, students can see how their text is organised and the content developed over the course of the whole thesis. Similarly, Keyword List can be used to make a keyword list of one thesis chapter when compared to a reference corpus of the whole thesis. By identifying the words that occur more (or less) frequently in one chapter than expected, this tool can provide information on the content of the chapter which is difficult to obtain by reading linearly. To date, the course has run thirteen times and evaluation data are available for 90 students. This paper discusses further the course and the affordances of corpus tools for editing, illustrating the findings with examples of student searches.

**References**


**Cotos, Elena, Gray, Bethany and Smith, Jordan - Iowa State University - Combining Multi-Dimensional Analysis with Rhetorical Move Analysis to Research Writing Across Disciplines**

**Linguistic analyses of corpora**

Research into the characteristics of academic writing has often focused on identifying differences across disciplines, locating patterns of variation in (a) the linguistic characteristics of texts through analysis of lexical, phraseological, grammatical, and lexicogrammatical features, or (b) the rhetorical structuring of texts, often through Swales’s (2004) move analysis. Comparatively less research, however, has focused on combining the two perspectives to study underlying linguistic patterns that realize rhetorical moves. Such research has often focused on a particular feature (e.g., lexical bundles in Cortes, 2013, anticipatory it in in Ädel, 2014) in a range of disciplines or a wider range of grammatical features in a restricted set of disciplines (e.g., Kanoksilapatham, 2007). The purpose of the present study is to extend this line of research by investigating a wide range of both disciplines and linguistic features through multi-dimensional (MD) analysis (Biber, 1988).

This study uses a corpus of experimental research articles from 30 disciplines (900 articles, 5.2 million words). The corpus was divided into Introduction-Method-Results-Discussion sections, and each section was annotated for moves and steps, as reported in Author 2 (2015). The annotated corpus was tagged using the Biber tagger, and rates of occurrence for more than 100 linguistic features were obtained with the associated Biber Tagcount program. Each move was considered an ‘observation’ in the MD analysis. Exploratory factor analysis was then carried out to identify ‘dimensions’ of variation, or co-occurrence patterns among a wide range of linguistic features. The resulting dimensions were interpreted according to the shared functions of co-occurring features. Each ‘move’ in the corpus was then characterized according to its use of these dimensions in order to identify patterns of linguistic variation associated with the use of rhetorical moves across disciplines. Implications for future research and pedagogical application will be discussed.

**REFERENCES**


Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. Journal of English for Academic Purposes, 12, 33-43.
"Modifiers embedded in noun phrases are the major type of structural complexity found in academic writing." (Biber & Gray, 2010). Present-day academic writing in the form of journal articles and college textbooks, to which our students are most exposed through their coursework readings, have been distinguished from oral language such as classroom lectures, by a preponderance of noun phrase structures involving nouns and their cohorts: articles, prepositions, and attributive adjectives (Biber et al. 1999). Parkinson & Musgrave (2013) confirm that "written academic prose favors heavy nominal groups." Taguchi et al. (2013) found greater use of phrasal (rather than clausal) modification by way of attributive adjectives, nouns as nominal pre-modifiers, and post-nominal prepositional phrases in more highly rated freshman essays. Lu (2011) demonstrates the greater use of complex nominals by L2 writers at higher proficiency levels. Brown & Aull (2017) provide a detailed analysis of "what is rewarded (or not) in high-stakes writing assessments" in the way of grammatical features and their effects.

DO WE REWARD SUCH COMPACTNESS IN NOVICE COLLEGE WRITING?

Do we reward such compactness in novice college writing? Do students emulate the 'compressed' style of writing that they are most certainly being exposed to in their academic reading? Writing teachers value substance, clarity, and conciseness, and for students, the grade is of ultimate concern.

As observed, quality writing of substance can be partially achieved through the use of semantically dense nominal phrases and the prepositional phrases that modify them. This study compares two corpora: one of approximately 50K words by L1 Chinese writers (n = 16) receiving higher grades and the other of approximately 40K words by those (n = 8) receiving lower grades in first-year writing courses taught by instructors across the curriculum at a small, liberal arts college in the southeastern U.S. Part of speech frequencies are determined using the CLAWS5 Tagset. Noun frequencies are compared along with their cohorts: articles, prepositions, and some attributive adjectives. How students most often modify those nouns is then determined through the use of Antconc concordancer software, and a comparison of pre- and post-nominal modifying structures is made.

RESULTS INDICATE ELLS AT BOTH PROFICIENCY LEVELS ARE USING NOUNS AT RATES IN LINE WITH ACADEMIC WRITING IN GENERAL, BUT MODIFIER TYPES DIFFER. Musgrave & Parkinson (2014) suggest certain "complex noun phrase tasks...to increase learners' understanding and use of [especially] noun-noun phrases" in their academic writing. The ongoing challenges of corpus-building and tagging and their utilization to inform ESL academic writing instruction are discussed as well.

COLLABORATION IN L2 RESEARCH HAS GENERALLY BEEN DEFINED AS TWO OR MORE LEARNERS WORKING TOGETHER TO HELP ACCOMPLISH A TASK (Davin & Donato, 2013; Ohta, 2001). However, there is considerable variation in terms of how collaboration is assessed. Researchers have pursued two approaches to evaluate collaboration during peer interaction: rater judgments of collaboration (e.g., Ahmad & Sedeghi, 2016; Winke, 2013) and qualitative coding of interactional patterns (e.g., Galaczi, 2008; Storch, 2002). Largely absent, however, has been any attempt to describe the linguistic features of collaborative interactions.

To gain further insight into the nature of collaboration, the present study uses corpus linguistic techniques to identify the markers of collaborative and non-collaborative peer interactions. ESL students (N = 42) enrolled in an intensive English program carried out a paired oral test as part of the program's formative assessment procedures. Their interactions were audio-recorded and rated using an analytic rubric with three categories (collaboration, task completion, and style). The interactions were also analyzed in terms of 146 linguistic features using the Biber Tagger (1988, 2006) along with length features (words, turns, words per turn).

The results indicated that high collaboration interactions were longer, had more turns, and had fewer words per turn than the low collaboration interactions. Linguistic features associated with high collaboration included first and second person pronouns, Wh-questions, that deletion and subordinate conjunctions. In contrast, the low collaboration interactions were characterized by nominal forms, emphatics, definite articles, and attributive adjectives. The findings are interpreted in terms of the collaborative and non-collaborative functions served by these linguistic features.

Crawford, Bill, Northern Arizona University, McDonough, Kim, Concordia University and Brun-Mercer, Nicole, California State University-Fullerton - Identifying linguistic markers of L2 collaboration: A lexicogrammatical approach

COLLABORATION IN L2 RESEARCH HAS GENERALLY BEEN DEFINED AS TWO OR MORE LEARNERS WORKING TOGETHER TO HELP ACCOMPLISH A TASK (Davin & Donato, 2013; Ohta, 2001). However, there is considerable variation in terms of how collaboration is assessed. Researchers have pursued two approaches to evaluate collaboration during peer interaction: rater judgments of collaboration (e.g., Ahmad & Sedeghi, 2016; Winke, 2013) and qualitative coding of interactional patterns (e.g., Galaczi, 2008; Storch, 2002). Largely absent, however, has been any attempt to describe the linguistic features of collaborative interactions.

To gain further insight into the nature of collaboration, the present study uses corpus linguistic techniques to identify the linguistic markers of collaborative and non-collaborative peer interactions. ESL students (N = 42) enrolled in an intensive English program carried out a paired oral test as part of the program's formative assessment procedures. Their interactions were audio-recorded and rated using an analytic rubric with three categories (collaboration, task completion, and style). The interactions were also analyzed in terms of 146 linguistic features using the Biber Tagger (1988, 2006) along with length features (words, turns, words per turn).

The results indicated that high collaboration interactions were longer, had more turns, and had fewer words per turn than the low collaboration interactions. Linguistic features associated with high collaboration included first and second person pronouns, Wh-questions, that deletion and subordinate conjunctions. In contrast, the low collaboration interactions were characterized by nominal forms, emphatics, definite articles, and attributive adjectives. The findings are interpreted in terms of the collaborative and non-collaborative functions served by these linguistic features.

Crawford, Bill, Northern Arizona University, McDonough, Kim, Concordia University and Brun-Mercer, Nicole, California State University-Fullerton - Identifying linguistic markers of L2 collaboration: A lexicogrammatical approach

COLLABORATION IN L2 RESEARCH HAS GENERALLY BEEN DEFINED AS TWO OR MORE LEARNERS WORKING TOGETHER TO HELP ACCOMPLISH A TASK (Davin & Donato, 2013; Ohta, 2001). However, there is considerable variation in terms of how collaboration is assessed. Researchers have pursued two approaches to evaluate collaboration during peer interaction: rater judgments of collaboration (e.g., Ahmad & Sedeghi, 2016; Winke, 2013) and qualitative coding of interactional patterns (e.g., Galaczi, 2008; Storch, 2002). Largely absent, however, has been any attempt to describe the linguistic features of collaborative interactions.

To gain further insight into the nature of collaboration, the present study uses corpus linguistic techniques to identify the linguistic markers of collaborative and non-collaborative peer interactions. ESL students (N = 42) enrolled in an intensive English program carried out a paired oral test as part of the program's formative assessment procedures. Their interactions were audio-recorded and rated using an analytic rubric with three categories (collaboration, task completion, and style). The interactions were also analyzed in terms of 146 linguistic features using the Biber Tagger (1988, 2006) along with length features (words, turns, words per turn).

The results indicated that high collaboration interactions were longer, had more turns, and had fewer words per turn than the low collaboration interactions. Linguistic features associated with high collaboration included first and second person pronouns, Wh-questions, that deletion and subordinate conjunctions. In contrast, the low collaboration interactions were characterized by nominal forms, emphatics, definite articles, and attributive adjectives. The findings are interpreted in terms of the collaborative and non-collaborative functions served by these linguistic features.
Grammar and mechanics skills are important aspects of both implicit and explicit language knowledge for second language (L2) learners. These skills are also important indicators of proficiency level and can strongly affect human judgements of L2 learners’ language ability. While many grammar and mechanics error checkers exist for the purposes of providing feedback to writers, to our knowledge, no tool that quantifies sentence-level errors for research purposes exists in the field. In this presentation, we introduce the Grammar And Mechanics Error Tool (GAMET), which can automatically measure the incidence of various grammar, spelling and style errors in texts. GAMET is freely available, runs through a simple Python script, allows for batch-processing, and provides incidence counts for a variety of errors at the micro- and macro-level. Grammar errors reported by the tool include agreement errors, fragments, and comma errors. Style errors include redundancy, verbosity, and word choice.

Initial analyses examining associations between the incidence scores for error counts as reported by GAMET in a corpus of independent essays and human rating of essay quality support strong predictive validity for the tool. Correlations indicated strong effect sizes for all errors ($r = -.528$) and medium effect sizes for spelling errors ($r = -.477$) and grammar errors ($r = -.400$). Small effect sizes were reported for white space errors ($r = -.282$), verb phrase errors ($r = -.281$), typographical errors ($r = -.254$) noun phrase errors ($r = -.222$), adjective errors ($r = -.182$), and duplications ($r = -.146$). All correlations were negative indicating that a greater number of errors was strongly associated with lower ratings of writing quality. Implications for second language research and automatic essay scoring will be discuss as will data on tool accuracy and reliability.

Researchers have successfully described the linguistic characteristics of academic language use from a variety of perspectives (Biber et al., 2002; Biber, 2006; Csomay 2005), including the use of academic vocabulary (Coxhead 2016). While the majority of the studies focus on academic vocabulary looking at written texts that students are exposed to, or discuss how to teach it, relatively few studies explore how students actually use these kinds of words in their scholarly papers. Examples of such descriptive studies are recent as they investigate both native and non-native student writing, and with the goal to explore developmental patterns, for example, in the use lexical bundles (Yan and Staples 2017) or in the use of academic words (Csomay and Prades 2018).

The present study adds to the previous descriptive accounts by investigating the rate at which non-native English speakers (EFL students) use academic vocabulary in their writing assignments. A total of about 3000 naturally occurring classroom-assignment papers were collected in lower- and upper-level composition classes as well as in other General Education areas such as History and Political Science, from students attending an American STEM degree program offered in Tbilisi, Georgia (Eurasia). In the composition courses, each of the paper assignments is designed to have at least two drafts while in the other classes, only the final draft is available. Using Gardner and Davis’ (2014) Academic Vocabulary List (AVL) generated from the Corpus of Contemporary American English (COCA), the percentage of academic vocabulary was counted in each student paper with the assumption that, as students progress in their studies gaining higher proficiency levels in English, they would use more academic words in their papers. Further investigations include the difference in AVL use for their humanities classes versus their social science classes.

Preliminary results indicate a) that students use incrementally more academic vocabulary as they progress in their coursework; b) that academic vocabulary is used differently across the text-types students produced; and c) that students use words differently in the three levels of instruction. No difference has been found between drafts. Pedagogical as well as programmatic implications are discussed.

There has been an explosion of interest among law professors, judges and lawyers in using corpus linguistics to gather information about the “original meaning” of provisions of the US Constitution. See, e.g., Corpus Linguistics & Original Public Meaning: A New Tool To Make Originalism More Empirical, 126 Yale Law Journal Forum 21 (2016). The author, a law professor, is currently teaching a research seminar in which students are learning (with assistance from a trained linguist) how to use a new BYU corpus, Corpus of Founders Era English, to explore the meaning of important words and phrases of the Constitution and Bill of Rights in the era those provisions were enacted. This paper will describe the procedures used to design the course and summarize the findings from the students’ papers. The paper will also discuss possible ways for linguists to work in an organized way to make corpus linguistics research on legal texts available for possible use in court cases.
Cupery, Maria - Purdue University - Price-Based Lexical Variation in Turkish Online Restaurant Reviews

Tools and methods, Linguistic analyses of corpora

Previous research has demonstrated price-based lexical variation in American online restaurant reviews and advertising (Freedman and Jurafsky, 2011; Jurafsky, Chahneau, Routledge & Smith, 2014). The current study provides comparative data by examining a corpus of 996 Turkish-language reviews on Yelp of Istanbul restaurants. The purpose is to provide not only a cultural comparison of food reviews, but to test whether language variation based on perceived socioeconomic information is evident in price-based reviews. By using the price of a restaurant experience as a proxy for socioeconomic level, and by analyzing differences in lexical, grammatical and etymological choices across four price rating levels, this study fulfills two goals. It extends the research on price-based lexical variation to a dissimilar language and culture, and provides a novel methodology for identifying language variation in the context of perceived social class, a construct that proved difficult to measure in linguistic research (Block, 2013).

The results, triangulated between keyword, semantic frequency and etymological analyses, reveal price-based variation in the Turkish reviews that are unique to the Turkish corpus. For example, relative to American reviews, the lexical and semantic frequency analyses reveal that when reviewing inexpensive restaurants, Turkish reviewers stress health, homey authenticity, and price, and that these concerns decrease significantly for restaurants that are more expensive. As prices increase, the reviewers use more specific, European terms for serving staff. The grammatical analysis shows no significant differences, but the etymological analysis displays a growing preference for words from Germanic and Romance languages as prices increase. Thus, price-based lexical variation is measurably present, though the patterns differ widely from the patterns in the American corpora. This supports the claim that reviewers use differing criteria and vocabulary in evaluating experiences that are perceived to be class-based.

References


Davies, Mark - Brigham Young University - Using the new 14 billion word iWeb corpus for teaching and research

Tools and methods, Linguistic analyses of corpora, Applications

The iWeb corpus was released in April 2018. It has many features that make it unique, even compared to the other BYU corpora such as COCA, COHA, NOW, GloWbE, and BYU-BNC.

First, the corpus is huge. It contains 14 billion words of text, which makes it about 25 times as large as COCA. The corpus is based on nearly 100,000 websites (about 23 million web pages), which were selected based on criteria from Amazon’s Alexa website (https://www.alexa.com/). On average, there are 250 web pages for each of these 93,000 websites, containing about 150,000 words each. Users can easily search by and compare websites, and in just 3-4 seconds they can create “virtual corpora” (with thousands of web pages and millions of words) for a particular topic. (This makes it similar to BootCat, but arguably much faster and easier to use.) In spite of the size, the corpus is 10-20 times as fast as the handful of other 10-20 billion word corpora (just as Corpora from the Web (COW) and Sketch Engine).

The corpus also allows many types of searches that are not found in the other BYU corpora. The KWIC displays and collocates are greatly enhanced, such as automatic grouping and display by part of speech. Researchers can also now see related “topics” (perhaps even more powerful than collocates) and “clusters” (advanced n-grams). Users can also more easily use synonyms, WordNet entries, and word families to find related words, and they can seamlessly move between all of the different displays for information on any of the top 60,000 words of English. They can also browse the 60,000 most frequent words in the corpus (as with https://www.wordandphrase.info/, but in even more powerful ways), and see the full range of information on each lemma – frequency, collocates, related topics, synonyms, clusters, and concordance lines – including a “summary” page for each lemma, which links to all of this information. This makes the new corpus potentially very useful for language learners.

In this presentation, I will discuss how the new features of the corpus might best be used in the classroom. In addition, I will present many searches that show how the 14 billion word corpus can be used to find a wealth of information on English, which is not available from any other corpus.
Deshors, Sandra C. - Michigan State University - Does the passé composé influence L2 learners’ use of English past tenses? A semantic exploration of the Present Perfect in French-English interlanguage

Linguistic analyses of corpora

This study focuses on the uses of the present perfect (PP) and simple past (SP) by French learners of English and assesses how those uses differ from those in native English and from those of the passé composé (PC) in native French which, semantically, overlaps with PP and SP. While typological differences between French and English in past tense formation are an obstacle for French learners of English, so far studies on French English learners’ uses of PP/SP are yet (i) to include cross-linguistic data and (ii) use methodologies that allow for a large-scale assessment of the tenses’ semantic contexts of use. The present study therefore addresses this gap by investigating over 3,000 contextualized occurrences of the three investigated tenses from the International Corpus of Learner English, the Louvain Corpus of Native Essays and the Corpus de Dissertations Françaises. Methodologically, this cross-linguistic study consists of a cluster analysis (to identify general patterns of (dis)similarity between language varieties) and a collostructional analysis (to assess degrees of attraction between PP, SP and PC and their respective semantic contexts of use). Overall, relatively native-like form-function mappings in learner English emerge from the analyses, suggesting that, semantically, despite the formal resemblance between the English PP and the French PC, advanced French learners of English have integrated the uses of past tenses including fine-grained contextual information characteristic of their individual uses. As a result, the influence of the PC over learners’ uses of the PP/SP emerges as relatively weak. While more traditional form-based methodological approaches have often reported non-native-like distributions of PP in learner English, the present work shows how methodological designs geared to capture, in context, the subtle, yet systematic, semantic usage patterns of PP, SP and PC can lead to varying conclusions on the nativelike or non-nativelike usage patterns of PP in interlanguage.

Doolean, Stephen - Texas A&M University - Corpus Christi - A corpus-based analysis of undergraduate, source-based L1 student writing

Linguistic analyses of corpora

Source-based writing is a complex task that requires writers to comprehend, select, and appropriate source material into their own writing, often in the forms of summaries, paraphrases, and quotations (Spivey, 1990). Additionally, source-based writing is a frequently occurring task type in post-secondary education (Rosenfeld, Leung, & Oltman, 2001). Recent overviews of source-based writing research have emphasized the difficulty students have with this type of writing (Cumming, Lai, & Cho, 2016) and the importance of direct instruction to improve the integration of source material into student writing (Grabe & Zhang, 2016). Yet, while research on source-based has often investigated assessment practices (Plakans, 2009a; Gebri & Plakans, 2014), plagiarism (Bloch, 2012; Keck, 2014), and theoretical frameworks (Kirkland & Saunders, 1991; Kucer, 1985), few studies have used corpus-based methods to investigate student performance on reading-to-write tasks.

This presentation will report on a study of English as a first language (L1) students’ in-class essays (N = 200), using stratified random sampling from a larger population of 415 essays. For data collection, students were randomly assigned one of four possible informative essay prompts with each prompt containing two source readings. With data in hand, the analysis was conducted in three stages. First, essays were cleaned and tagged for lexical and morpho-syntactic features. Next, frequency scores were obtained for 30 features predicted (based on a literature review) to correspond with more and less effective source-based writing (e.g., communication verbs, that complement clauses, passives, third-person pronouns). In the second phase of this analysis, the 200 essays were assigned holistic scores to measure the efficacy of source use in student writing. In the final stage of analysis, frequency counts from the 30 linguistic features were used to predict the quality of source text use in student writing. Results are discussed as they relate to existing corpus-based studies of student writing, as well as existing research on source-based writing more generally.

Durrant, Philip - University of Exeter - Lexical development in English school children's writing from six to sixteen

Linguistic analyses of corpora

This paper describes a newly-completed project focusing on the development of written language in school children in England. The need for research of this kind has become pressing in England in recent years due to an increased curricular emphasis on explicit teaching of the linguistic features of writing. The current National Curriculum states that students should be taught to ‘draft and write by: selecting appropriate grammar and vocabulary, understanding how such choices can change and enhance meaning’ (DfE, 2013a) and specifies the ages at which children are expected master specific features of written grammar and vocabulary (DfE, 2013b). A convincing linguistic research base against which such policies can be evaluated does not yet, however, exist.

The project described in this presentation was developed in response to this need. It uses corpus methods to understand the lexical development of English children’s language throughout the course of their compulsory education. Our team has collected a corpus of approximately 3,000 educationally-authentic texts from children in schools across England from ages six to sixteen with the aims of understanding what distinguishes texts written at different ages in terms of their use of key features of syntax and vocabulary.

This presentation will focus in particular on findings with regard to lexical development. It will discuss the methodological and theoretical issues involved in measuring lexical sophistication in children, outline our key findings, and discuss their implications for models of writing development and for educational practice.
Linguistic analyses of corpora

The need to become a member of a scientific international community of any specific area is widespread for university graduate students in many countries. CorAE, Corpus of Specialized Articles, composed of Chemistry, Applied Linguistics and Biology article corpora, has been compiled for linguistic analysis and pedagogical purposes. Studies on the use of phrasal modifiers across written registers have pointed out that nouns and attributive adjectives are the main grammatical classes that function as noun phrase modifiers (Biber, Gray, 2016). The students we work with face difficulties with such structures as they come in pre-modifying position. In this paper we present an analysis of attributive adjectives as nominal premodifiers in two corpora: CorAQuim (Corpus of Chemistry Journal Articles) and CorALA (Corpus of Applied Linguistics Journal Articles), so as to shed light on the understanding of how different disciplines build informational sequences of adjective and noun, investigating if classifier adjectives are more frequent than the descriptive ones as Biber and Gray (2016) pointed out. Each corpora has 500,000 million words. For this study a subset of each corpus was selected: CorAQuim (92,368 tokens) and CorALA (383,084 tokens). First, the subsets were grammatically and semantically tagged with CLAWS 7 (Garside, 1987) and USAS (Archer et al., 2002). Second, concordance lines were generated and the adjectives of each subsets could be manually categorized into two semantic groups: descriptors and classifiers (Biber et al., 1999). The results show that classifiers are more frequent in both corpora. Yet, CorAQuim presented 56% of classifiers and 44% of descriptors. CorALA reveals a higher use of classifiers (79%) rather than descriptive adjectives (21%). In conclusion, the use of adjectives as nominal premodifiers in CorAQuim tend to be more descriptive than in CorALA, which is prone to favor adjectives that delimit or restrict the noun.

Edalatishams, Ide - Iowa State University - Development of a Corpus of TA Classroom Speech
Tools and methods

Teaching Assistants (TAs) are responsible for teaching a large number of undergraduate students in a variety of subjects and disciplines. According to the National Center for Education Statistics, International Teaching Assistants (ITAs) comprise 30% of all graduate assistants in the US higher education institutions. ITAs’ effective classroom communication is essential not only for their students’ successful learning experience, but also for the ITAs’ successful teaching and further professional identity development. Previous research has addressed the issue of ITAs’ speaking problems through qualitative analyses of ITAs’ and their students’ perceptions of such problems (Arshavskaya, 2015; Chiang, 2006; Rubin & Smith, 1990), resulting in the development of different guidelines for ITAs (Gorsuch, 2006, 2011; Sarkisian, 2006). However, literature is lacking studies on systematic identification of ITAs’ speaking problems using corpus linguistics. This presentation reports on the ongoing development of a corpus of classroom speech by 50 TAs and ITAs, aimed at specifying ITAs’ most common speaking problems compared to their native-speaking counterparts. Details on the compilation, annotation, and mark-up of this spoken corpus will be discussed, followed by the regression analyses run to determine the differences between TAs’ and ITAs’ speech at the level of segmental (consonants and vowels) and suprasegmental (intonation, pitch, stress) features. Qualitative evaluations provided by linguistics and ESL experts will complement the quantitative findings from the corpus analysis.

References:


Egbert, Jesse and Biber, Douglas - Northern Arizona University - Incorporating text dispersion into keyword analyses
Tools and methods

Keyword analysis is one of the most widely used methods in corpus linguistics and corpus-assisted discourse studies (CADS), being applied to identify the words that are especially characteristic of the texts in a target discourse domain (Scott, 1997; Scott & Tribble, 2006). But surprisingly, the statistical computation of keyness usually makes no reference to those texts. Rather, once a corpus has been constructed, it is treated as a homogeneous whole for the computation of keyness. This ‘corpus frequency’ approach often results in lists of keywords that are relatively frequent in the target corpus, but not necessarily widely dispersed across the texts of that corpus, and thus not truly representative of the target discourse domain (see Baker, 2004).
The purpose of this study is to propose a new method for keyword analysis—text dispersion keyness—that is based on text dispersion rather than corpus frequency. Text dispersion keyness identifies words that occur in significantly more texts in the target corpus than in the reference corpus. In this study, we compare the effectiveness of text dispersion keyness to four widely used methods for computing keyness, carrying out a series of case studies to identify the keywords that are typical of online travel blogs. A variety of quantitative and qualitative analyses are carried out to compare these methods based on their content-generalizability and content-distinctiveness. Our results demonstrate that text dispersion keyness is a superior measure for generating content-distinctive and content-generalizable keyword lists.

References


Fahy, Matthew, Northern Arizona University, Egbert, Jesse, Northern Arizona University, Biber, Douglas, Northern Arizona University and Szmacsanyi, Benedikt, Katholieke Universiteit Leuven - Comparing logistic regression, conditional inference trees and random forests applied to ternary variables

Tools and methods

Most research carried out within the variationist tradition focuses on binary linguistic variables. Examples of this include the dative (to variant: ‘Jason gave the paper to Claire’ vs. double object variant: ‘Jason gave Claire the paper’) and the genitive (‘s: ‘FBI’s Director’ vs. of: ‘Director of the FBI’). However, variation in linguistic choice is not restricted to two-way alternations. For example, genitive meaning can also be expressed through noun+noun sequences (‘FBI Director’). Whereas two-way variation has traditionally been modeled using binary logistic regression, there are no standard methods for modeling ternary variation. Some studies have carried out multiple logistic regressions in order to account for every possible pair of variants (see, e.g., Szmacsanyi, Biber, Egbert & Franco, 2016; Hinrichs, Szmacsanyi & Bohmann, 2015). However, this approach is not ideal because results are not based on a simultaneous analysis of all relevant variants, and they are difficult to interpret. We apply conditional inference trees (also known as classification trees) and random forests to the problem of modeling choice between ternary variables and compare these methods to logistic regression. We explore the use and interpretation of these methods using a large dataset of interchangeable English genitives (‘s, of, and noun+noun) from the ARCHER corpus.

REFERENCES


Garner, James - University of Florida - The Longitudinal Development of Collocations as Constructions in L2 Writing

Linguistic analyses of corpora

Longitudinal learner corpus studies utilizing different approaches (native speaker judgements, raw frequency, association strength) to phraseological unit identification have demonstrated how second language (L2) learners develop their ability to produce phraseological units in their speech and writing over time (Bestgen & Granger, 2014; Crossley & Salsbury, 2011; Garner & Crossley, in press; Li & Schmitt, 2009, 2010). However, these studies are often limited in that they utilize only one or two measures of production and neglect part of speech information when calculating association strength. The current study addresses these limitations by defining lexical collocations as constructions (“collocations”) within a Construction Grammar framework and analyzing their longitudinal development using multiple indices of collocation production.

The data in this study comes from a corpus of freewrites by ESL learners collected over the course of one year (Salsbury, 2000). Adjective-noun, verb-noun, and adverb-adjective collocations were extracted from the learner texts and analyzed using indices of frequency, diversity, and strength of association. Longitudinal develop was assessed at the group level using linear mixed effects models and at the individual level using min-max graphs, and moving correlation graphs, two methods common to Dynamic Systems Theory (DST; Verspoor, 2015).

The linear-mixed effects models found no significant effect of time for any of the indices, suggesting that, for the entire group, there was no significant development in collocation production over the study period. However, the min-max and moving correlation graphs revealed that individual learners did develop their ability to use collocations in their writing. One learner was found to develop her ability to produce more strongly associated adjective-noun collocations, while another learner was able to increase how frequently she produced verb-noun collocations as well as how strongly associated those collocations were. These results will be discussed in terms of their implications for L2 phraseology research and learner corpus research.

AAACL 2018
REFERENCES


Graedler, Anne-Line - Inland Norway University of Applied Sciences - L2 students’ choice of cohesive devices in translation

Linguistic analyses of corpora

Translation is being reestablished as an effective tool of L2 learning, based on arguments connected to the notion of the awareness of language in general, and of contrastive aspects in particular, both in relation to vocabulary, grammar and text coherence, and also in relation to stylistic, pragmatic and cultural differences. Hence, L2 translation can serve documentation for general assessment of the language students’ textual competence.

This paper will present advanced Norwegian language learners’ competence and awareness of text coherence from a cross-linguistic perspective, with a focus on so-called grammatical cohesion based on structural content (reference, substitution and ellipsis, and conjunction; cf. Halliday & Hasan 1976), with the emphasis on what types of grammatical cohesive devices are displayed in the Norwegian source texts (L1), and whether those are transferred or changed into English target texts, i.e. in what way the L2 target texts display equivalent (or non-equivalent) cohesive relations.

The research data is the Norwegian-English Student Translation corpus (NEST), which consists of 32 source texts and 348 target texts, submitted by 141 advanced Norwegian learners of English from different tertiary level institutions (where the translation courses were part of regular English studies, not specific translation programs). The NEST corpus also contains metadata about the students’ language and educational background.

Since both Norwegian and English are Germanic languages, there are many similarities related to vocabulary, syntactic and thematic structures, conjunctions, etc., but there are also several distinctions. The research includes some basic statistical analysis of contrastive discourse, but will also contribute to explore and discuss the advanced English language learners’ general competence and awareness of text coherence.

Grieve, Jack and Clarke, Isobelle - University of Birmingham - Tracking Stylistic Change on the Donald Trump Twitter Account

Tools and methods, Linguistic analyses of corpora, Applications

The @realDonaldTrump Twitter account has been an important platform for Donald Trump and his team, identified by many observers as a key part of the success of his presidential campaign. At the same time, the account has been the subject of constant critique, not only from the opposition and the media, but from members of his own party. Regardless of one’s opinion of its tone and content, @realDonaldTrump is undoubtedly one of the most influential social media accounts in the world.

The goal of this study is to use methods from quantitative corpus linguistics to track how the style of language used on the Trump Twitter account has changed over time so as to better understand the underlying communication strategy employed by Trump and his team. Specifically, we investigate three research questions:

1. What are the main dimensions of stylistic variation on the Donald Trump Twitter account?
2. How does the style of posts on the account change over time?
3. What do these results tell us about the communication strategy of Trump and his team?
To answer these questions, we conduct a multidimensional analysis (MDA) (Biber 1988) of all posts on Trump’s Twitter account from 2009-2017 (21,320 Tweets), extracting the most important dimensions of stylistic variation in this corpus and plotting change in these dimensions over time.

Because Tweets are short, we apply a new form of short-text MDA, which only considers whether or not a range of lexical and grammatical features occurs in the texts, as opposed to their relative frequencies, as in standard MDA. Consequently, rather than identifying common patterns of stylistic variation in the corpus using Factor Analysis, we use Multiple Correspondence Analysis, which is suitable for the multivariate analysis of categorical data. We then interpret the dimensions functionally following standard MDA procedure.

We identify five main dimensions of stylistic variation on the Trump Twitter account, which we interpret as being related to the degree of opinion, prediction, advice, promotion, and critique expressed by a Tweet. We then plot stylistic variation over time in the corpus based on these five dimensions, demonstrating that shifts in tweeting style generally align with shifts in Trump’s public persona, campaign, and presidency. Finally, we argue that these results provide new insights on Trump’s changing and at times highly successful strategies for communicating with the public using social media.

Grieve, Jack, University of Birmingham, Clarke, Isabelle, University of Birmingham, Popoola, Olumide, University of Birmingham and Waibel, Emily, Aston University - Native Language Identification of the Guccifer 2.0 Corpus Tools and methods, Applications

The hacking of the Democratic National Committee (DNC) was a major event in the 2016 US Presidential Election. The release of thousands of private DNC emails on WikiLeaks revealed that the DNC had undermined the campaign of Bernie Sanders. As a result, the DNC issued an official letter of apology its chair, Debbie Wasserman Schultz, resigned. Online, a hacker known as Guccifer 2.0 took responsibility for the breach of the DNC servers. Although Guccifer 2.0 claimed to be Romanian and to have no connection to Russia, the hack has been attributed to Russia by many intelligence agencies and firms, leading to highly publicised allegations that Russia interfered in the election.

The goal of this paper is to conduct a linguistic analysis of the corpus of Guccifer 2.0 communiqués to explore the possible first language of the author or authors responsible for the Guccifer 2.0 persona. We apply a novel semi-automated approach to native language identification through back translation.

First, we identify apparent lexical errors in the corpus. For example, Guccifer 2.0 writes that “together we’ll be able to throw off the political elite”, which appears to contain the lexical error “throw off” for “overthrow”. Second, we translate the error into 34 major world languages using Google Translate (GT), recording the list of possible translations and their possible English back-translations. Third, we judge whether these possible English back-translations have the meaning that Guccifer 2.0 intended to express in the original sentence. For example, GT provides “свергать” as a possible Russian translation of “throw off”, which can be translated in English as “depose” and “overthrow”, suggesting that a Russian author writing in English might make this type of error. Finally, we compare the results for the complete set of lexical errors across the complete set of languages, to identify languages and language families that have the largest number of back-translation matches.

Based on a preliminary analysis of 8 apparent lexical errors in the initial posts, we have identified a Slavic signal in the Guccifer 2.0 corpus. Such research is especially helpful as proposals for conference presentations is an occluded genre, hidden from novices who may want to examine and model their own proposals off of them.

Hardy, Jack, Oxford College of Emory University and Payant, Caroline, Université du Québec à Montréal - Conference and Research Article Abstracts: A Corpus-Based Comparison Linguistic analyses of corpora

Conference abstracts (CAs) play an important role in the professionalization in the field of Teaching English to Speakers of Other Languages (TESOL). This genre is usually learned through the drafting of successful (and unsuccessful) proposals. As a result, many novice researchers lack knowledge of the typical features present in CAs. Previous research has identified important discourse features of successful CAs (e.g., Cutting, 2012; Egbert & Plonsky, 2015) and prominent sequencing of rhetorical moves (Payant & Hardy, 2016). Such research is especially helpful as proposals for conference presentations is an occluded genre, hidden from novices who may want to examine and model their own proposals off of them.

Unlike CAs, a similar genre, the research article abstract (RAA), is easily available to readers. To date, studies have not systematically compared these two similar genres. Given the fact that RAs are available to the public but CAs are not, the goal of this study is to explore the extent to which TESOL CAs and RAs compare in linguistically and functionally.

Our corpus includes 3,183 proposals from the TESOL 2017 International Convention, 250 RAs from TESOL Journal published between 2013-2017, and 250 from TESOL Quarterly (2010-2017). To automatically count important lexical and grammatical features, we used the Multidimensional Analysis Tagger (Nini, 2014). The features from Egbert and Plonsky (2015) along with other communicative functions associated with the multidimensional analysis by Biber (1988) are used to explore variation in the data according to the type of proposal (CA versus RA). The close examination of these genres is imperative given the availability of both types to community members who currently do not have access to the occluded genre of CAs.
Croft (2001:46) observes that “syntactic categories are derivative of — in fact epiphenomenal to — the representation of grammatical knowledge”. This paper considers Croft's claim from a 'constructional' perspective, focusing on a set of structures in Spanish that employ the verb hacer ‘to do/make’ and a temporal complement, as shown in examples (1) and (2). These structures have largely eluded a clear categorization given that they display properties of both (i) their verbal origins (from the fully productive verb hacer) and (ii) the largely adjunct-like and largely de-verbal behavior displayed in natural language data (see Howe 2011). Following Michaelis (2010), this analysis argues that this construction should be characterized in terms of “multiple inheritance hierarchies”, whereby a given construction may belong to more than one type (e.g., verb, noun, etc.), in this case a construction that is both verbal and adjunctival. This, I argue, is at the core of the behavior of these constructions, given that these constructions represent a type of syntactic blend where formal and functional properties of different lineages converge in a single construction (Croft 2000).

(1) Hace mucho tiempo que vi el tráiler de esta película y no me llamó la atención

‘A long time ago I saw the trailer for that movie and it didn’t get my attention’ (CdE)

(2) Un día me pondré y la veré! Hace mucho tiempo la vi en el colegio

‘One day I’ll set down and watch it! A long time ago I saw it in high school.’ (CdE)

Using data extracted from both the oral and written portions of the Corpus del Español (Davies 2002), this paper focuses specifically on the distribution of this construction with respect to the use of the complementizer que, which is expressed in example (1) and unexpressed in example (2). Crucially, this analysis does not assume that (1) and (2) can be distinguished a priori (Herce 2017). Rather, the analysis of the corpus data reveals a gradient distribution of the structural properties of this construction that is incompatible with a discrete syntactic categorization.

REFERENCES


Imao, Yasuhiro - Osaka University - Going beyond simple word-list creation using CasualConc

Tools and methods

The use of web-based corpus tools with a large volume of text data has becoming a norm of corpus analysis of modern-day languages. However, those are mostly available only for English. Also, there still is need for small-scale specialized corpora to explore more detailed analysis of language use. Some easy-to-use GUI-based application for Mac and Windows have been developed, but many of them have limited functionality for simple KWIC and word-listing with some basic statistics.

CasualConc, a concordancer for macOS was introduced in 2008 as a more Mac-app-like alternative to AntConc. It originally had some basic functionality, but has since evolved to utilize some outside application for tagging and statistical analysis as well as basic macOS functionality. Also developed are companion programs for corpus building.

As other similar application, CasualConc can handle basic KWIC, word/n-gram list, and collocation analysis. But these are only a fraction of what CasualConc is capable of. With a help of Stanford CoreNLP/TreeTagger and MeCab (for Japanese), CasualConc can handle batch pos-tagging of text file easily. Also with the help of built-in macOS functionality, it can tokenize some non-western languages.

In addition, CasualConc, with its functionality to easily create word lists for multiple files individually (which is essentially a word frequency table), opens up a possibility for multivariate analysis of corpus data with a few mouseclicks. This is enabled by calling a statistics application R within CasualConc.

Other applications are of great help when building a specialized corpus. CasualTextractor and CasualTranscriber can help these extracting and tidy up text from web or documents. Then comes CasualTagger for not only for tagging text, but also for editing with KWIC functionality.

All in all, a suite of macOS application will help text analysis research move one step further and with ease!
Corpus studies of learner language (L2) and translated first language (L1T) share an interest in contrasting these varieties to “native” production (L1O), as seen in research on typical features of L2 and L1T (vs. L1O), that rely on similar methods and assumptions (Granger et al. 2009). We derived our L2 English texts from the ICLE corpus (Granger et al. 2009) and the LOCNESS corpus, supplementing them with translations into English collected ad hoc, for a total of 1224 texts. To control for cross-linguistic effects, two typologically distant first/source languages are represented in the corpus, Finnish and Italian. We annotated the data using the UDPipe parser (Straka/Straková 2017), and extracted frequencies of parts-of-speech, morphological features, and syntactic dependencies, normalized over 1,000 words per text. These were used as features in key structure analysis (Ivaska/Siitonen 2017), a method relying on random forests to distinguish linguistic differences between language varieties.

We identified multiple linguistic features whose frequency distinguishes both L2 and L1T from L1O at all levels of annotation. Some differences are language pair-specific, suggesting that direct CLI plays a role, while others hint at more general constrained communication effects. Taken together, these results shed light on the nature of the common ground between translated and L2 English.

REFERENCES


Jin, Bixi - The University of Hong Kong - Using Multidimensional Analysis to Investigate the Discussion Sections of Research Articles in Chemical Engineering

Linguistic analyses of corpora

In the process of writing up research articles, previous studies have shown that writing the discussion of results section can be a particular challenging undertaking for EAL novice scientists (e.g. Bitchener and Basturkmen, 2006). This study investigates the linguistic dimensions of RA discussion sections in chemical engineering discipline, and also the variation in patterns of language of successful and less successful written samples.

A total of 213 Discussion sections in chemical engineering RAs were collected. Multi-dimensional Analysis Tagger (Nini, 2014) was adopted for automatic tagging. The study made use of Biber’s multi-dimensional (MD) analysis to identify the linguistic dimensions. They were: Dimension 1: Involvement and interactivity; Dimension 2: Non-narration vs. narration; Dimension 3: Further explanations and elaborations on evaluative statements; Dimension 4: Informational focus; Dimension 5: Stating results/claims; Dimension 6: Expression of denial relations towards statement or experimental findings. The follow-up t-test analysis compared and contrasted the differences of discussion sections in 110 successful and 103 less-successful discussion sections in the established corpus along these six dimensions. Dimensions 1, 3, and 5 were found to be significantly different in two sub-corpora and may suggest that successful discussion sections made use of more stance-taking expressions (i.e., hedges, boosters, attitude markers and first-person pronouns).

Pedagogical implications can be drawn from the MD results. EAP writing instructors can make use of the findings to develop corpus-informed teaching materials, helping novice scientists construct discipline-specific language patterns for the discussion of findings. In addition, a series of consciousness-raising tasks can be devised to help novice writers notice the stance features underlying successful RA discussions to produce convincing and compelling claims.
This presentation explores the potential of teaching prepositions embedded in frequently occurring sequences of four words, or lexical bundles (LBs; Biber et al., 1999). Research suggests that English learners focus on key content words of sequences because function words, which lack semantic content and are unstressed, are less salient; thus, when learners produce the sequences, they often edit out the function words (Schmitt et al., 2004; Wray, 2004). Considering that all LBs previously identified in academic genres (Biber et al., 1999, 2004) include prepositions, their potential as a pedagogical tool to teach prepositions seems worth exploring.

The presentation has two goals. The first is to identify LBs with embedded prepositions frequently misused by English learners. The bundles identified by Biber et al. (1999, 2004) were searched for in a learner corpus of argumentative essays by 5,600 graduating Korean high school students. The learners frequently misused or left out prepositions, producing, e.g., in the end of and the same time for the LBs at the end of and at the same time.

The second goal is to report the results of an instructional experiment exploring the possibility of teaching prepositions using 10 LBs frequently but incorrectly used in the learner corpus. Two instructional treatments will center on the LBs' discourse functions and employ a set of pictures describing each bundle. One treatment explicitly teaches the prepositions in LBs while the other treats the LBs as single units. The data will be collected over eight weeks in pretests/posttests/delayed posttests, in which participants (90 English learners from a Korean high school: one control group, two treatment groups, n = 30 each) will write sentences based on the pictures. The findings will shed light on the effectiveness of LBs as a tool to teach prepositions and on the most beneficial type of LB instruction.

Kaufmann, Carlos - São Paulo Catholic University - Exploring stylistic dimensions of Brazilian nineteenth century fiction

Linguistic analyses of corpora

One of the basic traits of literary fiction is style, that is, the characteristics that separate one particular author from the others. Few studies have looked at style in literary fiction from a corpus perspective, such as Biber & Finegan (1994), Semino & Short (2004), Fischer-Starke (2010), and Mahlberg (2013). The previous literature has focused mostly on English literature, and therefore more research is needed in corpus-based analysis of literary style in languages other than English. The goal of this study is to provide a multidimensional (MD) account of style across the major authors of Brazilian fiction, writing in Portuguese between 1850-1910. The study takes as its initial reference the work of Egbert (2012), which analyzed ten American and British nineteenth-century authors in terms of their stylistic similarities and differences. In Egbert (2012), three stylistic dimensions were identified from a full MD analysis generated from a broad reference corpus that covered a range of more than three centuries of English-fiction samples. In the current study, a balanced corpus of novels and short stories was collected from 20 different Brazilian authors related to the focused period, totaling 3.2 million words. The corpus was tagged with the PALAVRAS parser (Bick, 2014), the major tagger for Portuguese, for more than 100 features. An MD analysis was conducted on the corpus, and five different dimensions were identified. The paper will present the dimensions, and discuss how they correspond to the particular styles of the different authors, especially Machado de Assis, the leading author of Brazilian fiction in the 19th century. The dimensions were shown to be good predictors of Machado’s texts (precision of 0.839 and recall of 0.860).

Keller, Daniel - Northern Arizona University - Do registers have psychological reality?: Evidence for register based priming in lexical access

Linguistic analyses of corpora

No understanding of linguistic variation is complete without accounting for register differences. Register to register variation is thought to be driven by corresponding variation in the non-linguistic context of communication. Registers are thus frequently conceptualized as pairings of situations and text types (Biber & Conrad, 2009). While this way of thinking about registers is both intuitive and useful, it obfuscates the fact that texts are not created by situations directly. Rather, texts are created by language users who shape their linguistic behavior to match mental representations of the situation. Despite this, little research has explored the psycholinguistic processes involved in register representation (though see Berber Sardinha’s work on lexical priming and register variation, 2017). The current study addressed this gap by investigating priming effects of register expectations on psycholinguistic processes of lexical access. Participants were shown register primes, short video clips evocative of either news or recipe registers, and instructed to quickly decide whether a following string of characters was an English word or not (a lexical decision task; Neely, 1977). Preliminary results suggest that when the string of characters formed a keyword for an immediately primed register (a facilitating condition) or a more distantly primed register (a suppressing condition), participants completed the lexical decision faster than when the characters formed a control word matched to the target on morphological and psycholinguistic characteristics $X^2(2)=54.03, p<.001$. Pairwise comparisons indicate significant differences in mean response time (in milliseconds) between the control condition ($M = 635.15, SD = 16.97$) and facilitating condition ($M = 596.91, SD = 16.96$) and between the control condition and suppressing condition ($M = 576.31, SD = 16.96$). These results suggest that a productive avenue of triangulation in register research may involve psycholinguistic investigations of links between situational cues and linguistic forms.
Kia, Elnaz - Northern Arizona University - Directive Use across Disciplines in University Classroom Discourse

**Linguistic analyses of corpora**

Directives—"attempts by the speaker to get the hearer to do something" (Searle, 1976, p. 11)—are one of the recurrent speech acts in university classrooms (Barbieri, 2008; Garcia, 2004). Directives are used by the instructors to perform important functions, such as assigning homework and guiding class work. Previous research has revealed a relationship between linguistic variation and situational features of the classroom discourse (Barbieri, 2008; Biber, 2006; Csomay, 2005). However, there is little research examining the linguistic variation of directives in university lectures across situational contexts using a large corpus (Barbieri, 2008). The current study attempts to fill this gap by examining structural and situational variation (i.e., level of interaction, level of instruction, and discipline) in directive use in a large corpus of university lectures.

The data in this study comes from a 1.2 million-word corpus of lectures sampled from the TOEFL 2000 Spoken and Written Academic Language corpus (T2K-SWAL, see Biber et al., 2002). Possible linguistic patterns of directives were identified by manually analyzing sample lectures from various situational contexts. Subsequently, five major structural types of directives (i.e., imperatives, obligation, intention, permission verbs, and directive vocabulary) were selected for the automatic analysis, according to their frequency and lexico-grammatical explicitness. Python scripts were used to automatically identify directive utterances with the five structural types. Structural variation of directives was analyzed in relation to three situational variables: (1) level of interaction (low, medium, high), (2) level of instruction (freshman/sophomore, junior/senior, graduate), and (3) discipline (business, engineering, humanities, natural sciences, social sciences, and education). Preliminary results reveal that these situational variables affect the use of directives in important ways. This study has implications for the field of teacher education by describing the structural variation of directives with respect to disciplines, interactivity, and level of instruction of the university lectures.

Kim, Dongsung - Ewha Womans University - Gutenberg Project Frequency & KeyWord-In-Context (KWIC) Viewer

**Tools and methods**

Recently Google Ngram Viewer (Michel et al. 2013) is the publicly available tool to visualize a word(s) or phrase(s) based on digitization of tons of millions of books. We can detect a particular trend or historical events (or even our grammaticality). This process enables cultromics which reads human behaviour and cultural trends, having a temporal insight into cultural dynamics (Aiden & Michel 2013). Inspired by Google Ngram Viewer, we have been building the text visualization and KWIC viewer, using the Gutenberg Project. We have been using IMS Corpus Workbench (Stefan & Hardie 2011) in order to extract real frequencies and KWIC of words-phrases. Since Gutenberg Project offers copy-right free ebooks, the majority of data is concentrated in the 1920s’. Our system based on the Gutenberg Project cannot show chronological cultural trends as expected. Therefore, we need to apply a new statistical model to uniformize each different time period. We have tried to adjust statistics of different chronological texts using relative frequency (Gries & Hilpert 2008, Gries 2010), logarithmic ratio (Gulordava & Baroni 2011), logarithmic spaced bins (Bochkarev et al. 2016), and dispersion measures (Liberman et al. 2007, Liberman et al. 2011). We think that Google Books is the entire corpora and Gutenberg Project is sub-corpus. As McEnery & Wilson (1996) noted, sub-corpus has similarity with the entire corpora. We assume that Gutenberg Project has similar behavior in showing cultural trends as Google Books has. We compared adjusted statistical results with those of Google Ngram Viewer in showing chronological transition of 1920s’ slang-phrase patterns. For instance, Google Ngram Viewer (Figure 1) shows that ‘cake-eater’ is popular in 1920s’, diminishing after 1940s’. We found out that logarithmic ratio approach produced the best result in showing similar patterns as shown in Google Ngram Viewer.

**References**


Michel J. B., Shen Y. K., Aiden A. P., Veres A., Gray M. K., Pickett J. P., Hoiberg D.,

Kim, Jong-Bok, Kyung Hee University and Davies, Mark, BYU - Transparent Free Relatives in English: A Corpus-based Perspective

The so-called ‘transparent’ free relatives (TFR, e.g., I bought [what seemed to be a guitar]) display quite striking properties in that its content nucleus a guitar is within the bracketed clause, as if the underlined expressions are transparent. The transparent properties can be attested from corpus data (COCA: Corpus of Contemporary American English):

(1) a. [What appears to be a pale blue painting] turns/*turn into something else entirely. (COCA 1994 NEWS)
   b. [What we call coincidences] are/*is limited to the ones we happen to notice. (COCA 2008 NEWS)

The number value of SFRs (standard) is singular or determined by the wh-phrase, but that of TFRs is dependent upon the boldfaced nucleus. There have been three main approaches to the TFRs: parenthetical placement with backward deletion (Wilder 1999), shared structures (Van Riemsdijk 2001, 2006), and configurational structures with movement operations (Grosu 2003). Authentic data obtained from our corpus search, however, show that none of these previous analyses are satisfactory enough to account for the various uses of the construction.

Previous literature focuses on TFR data whose nucleus category is an NP as in (1) and shares the view that the transparent expressions are rather parenthetical. However, corpus data indicate that they are core syntactic expressions: unlike SFRs which externally act like nominal clauses in their distribution, TFRs have much wider distributional possibilities including AP, AdvP, PP, and others as well as NP:

(2) a. You’re definitely not [what anyone would describe as ecstatic]. (COCA 2007 MAG)
   b. In that process I begin to work [what I would call creatively]. (COCA 1992 MAG)
   c. She definitely wasn’t [what she’d call in love with Sam Butler]. (COCA 1990 FIC)
   d. You make him responsible for [what I call trumping the center]. (COCA 2007 NEWS)

Attested coordination corpus data also show us the headedness of the nucleus in the TFRs and a transparent effect (see Kajita 1977, Van Riemsdijk 2001, 2006 also):
Given such flexible data and nonconstituent properties of the transparent expressions, it seems hard to claim that they are
parenthetical or deleted by certain movement operations. The most viable way to account for such vibrant properties of the
TFRs as attested by corpus search is to allow tight interactions between the lexicon and constructional constraints. The key
property starts from the fact that what is 'lexically' underspecified for categorical and semantic information. In terms of
constructional constraints, we assume that there is a TFR construction as a primitive grammatical element. This construction is
a special type of filler-gap unbounded construction with its own constructional properties. It is peculiar in that the head of the
construction is what, whose category value is determined by the semantic nucleus. An additional lexical property of the
construction is that it is only raising verbs (e.g., call, consider, take, assume, describe, seem, appear) that can introduce the
TFR. This means that raising verbs can change its predicative argument into any nonverbal argument (NP, PP, Adv, AP) whose
categorical information (including POS and number values) is shared with that of the head what. These lexical specifications can
immediately explain two important constraints in the TFR: what is the only possible element in the TFR and its possibility of
referring to a human being, and only raising verbs can introduce the TFR. There are other welcoming consequences of the
present analysis. For example, since the predicative expression eventually determines the syntactic category of the whole
clause, we can explain why the distributional possibilities of TFRs are determined by the predicative expression. It further
accounts for the preposition restriction (He speaks in ‘at what linguists call a Northern dialect COCA 2001 ACAD) as well as
coordination facts in (3). This is possible since the property of the whole TFR in question is determined by the nucleus
expression whose syntactic and semantic features are identified with the expression what functioning as the head of the clause.

Selected References:
  Theory 21: 247–331;
- Van Riemsdijk, Henk. 2006. Free Relatives. In M. Everaert and H. Riemsdijk (Eds.), The Blackwell Companion to Syntax Vol. II:
  338–382. Blackwell;

Kim, Minkyung and Crossley, Scott - Georgia State University - Predicting second language writing and speaking
scores using a testing corpus: The roles of linguistic features and individual differences
Linguistic analyses of corpora

While previous second language (L2) studies have examined relationships between linguistic features and speaking/writing
scores (e.g., Biber, Gray, & Staples, 2016; Crossley & McNamara, 2012), and between individual differences and
speaking/writing scores (e.g., de Jong et al., 2012; Staher, 2008), few, if any, studies have simultaneously investigated links
among individual differences, linguistic features, and speaking/writing scores. This study using a testing corpus fills that gap by
examining how speaking/writing scores can be respectively predicted by individual differences and linguistic features.

A corpus of 295 test-taker performance data on the Examination for the Certificate of Competency in English (ECCE) was
examined. Linguistic features were measured using the tool for the analysis of lexical sophistication (TAALES; Kyle & Crossley,
2015) and the tool for the automatic analysis of text cohesion (TAACO; Crossley, Kyle, & McNamara, 2016). Individual
difference variables included language scores (reading, listening, and vocabulary) assessed by the ECCE and demographic
information (age, gender, and first language).

Linear mixed-effects models explained 49.9% of the variance in speaking scores and 44.5% of the variance in writing scores,
respectively. In both models, higher-rated speaking and writing samples included n-grams whose associations were stronger (as
measured by mutual information and Delta P). Differences were also found such that higher-rated speaking samples contained
more function word types, and greater function-word and adverb overlap, while higher-rated writing samples contained more
lemma types, more academic words, and greater verb repetitions. The speaking model also indicated that higher listening
scores were predictive of higher speaking scores, whereas the writing model indicated that higher reading and vocabulary
scores were predictive of higher writing scores and that female learners showed better writing performance than male learners.
Overall, these findings provide evidence to support distinctions between oral language proficiency and written language
proficiency.

This presentation reports on research funded through CaMLA’s Spaan Research Grant Program, 2017.

Kim, Susie - Michigan State University - Predicting CEFR proficiency level using grammatical criterial features
Linguistic analyses of corpora

In a previous study using a Korean EFL corpus, I investigated the use of the grammatical criterial features that have been
identified as characteristic of the Common European Framework of Reference (CEFR) levels by Hawkins and Filipović (2012).
Research on the use of such level-specific lexico-grammatical features has largely been descriptive. This study contributes to
better understanding the relationship between learner language and CEFR proficiency levels by empirically examining how well
the use of the set of grammatical features predict learners’ proficiency. The current analysis included 6,042 narrative and
argumentative essays written by 3,021 learners. Each learner’s performance had been holistically evaluated on the CEFR scale,
ranging from A2 to B2. With the data from the previous study where I examined the occurrences of ten A2-, fifteen B1-, and ten
B2-level grammatical criterial features, I tabulated how many different types of the level-specific features each learner used and
the average normalized frequencies of the occurrences. In other words, the analysis included six predictor variables: the number of total A2-, B1-, and B2-level features used and the frequencies of these level-specific features. An ordinal logistic regression analysis was performed to predict the proficiency level based on the six predictors. The final model showed a good discrimination among the proficiency groups (R-sq = .31), with 67% correct classification rate. All six predictors were statistically significant in predicting the proficiency level. The odds of a learner being placed at a higher level than the A2 level were approximately 1.24-1.87 times greater with one standard deviation increase in each predictor. The findings indicate that the frequency as well as the diversity in the use of the grammatical features serve as important predictors in determining the proficiency level. The results have implications for CEFR level descriptors and assessment using the CEFR scale.

Kim, Youjin, KAIST, Kang, Sanghee, Georgia State University and Tywoniw, Rurik, Georgia State University - Examining Dynamic Development in Speaking versus Writing among EFL Learners with Different Proficiency Levels: A Longitudinal Multi-Case Study

Linguistic analyses of corpora

Over the last decades, second language acquisition (SLA) researchers have demonstrated that language learners exhibit dynamic patterns of language development, particularly in the areas of complexity and accuracy (Larsen-Freeman, 2006; Verspoor, de Bot, & Lowie, 2011). However, the development patterns of speaking and writing by learners with different proficiency levels have not been examined. From dynamic complexity theory perspectives, the current study analyzed a spoken and written corpus of 13 Korean English as a Foreign Language (EFL) learners in terms of complexity, accuracy and fluency. Complexity was operationalized as both syntactic complexity (e.g., # of clauses per t-unit/AS-unit, # of words per clauses) and lexical sophistication (e.g., lexical frequency indices using TAALES, Kyle & Crossley, 2014). Accuracy was operationalized as the number of error free T-units/AS-units, and fluency was operationalized as the number of syllables per minute (speaking) and the number of words per minutes (writing).

A total of thirteen high school learners completed multiple independent writing and speaking tasks independently using sample TOEFL prompts over one year (10 writing samples and 14 speaking samples per learner). Approximately 60,000 words in the spoken corpus and 130, 000 words in the written corpus were transcribed and analyzed in the study. To control for topic effects, topics were all related to academic life and were counter-balanced. Their proficiency level was measured using the Test of English Proficiency. Following methodological suggestions from the dynamic complexity theory, the longitudinal data were analyzed using the Monte Carlo technique and moving min-max graphs. Findings showed different patterns of variability among learners depending on their proficiency levels and the target linguistic areas (complexity, accuracy, and fluency). The presenters will shed light on the relationship between the development of speaking and writing among instructed EFL learners from different proficiency levels.

Kunath, Stephen, Georgetown University, Weinberger, Steven, George Mason University and Nelson, Jill, George Mason University - Crowdsourcing Phonetic Transcriptions: Experimental Results from a New System for Transcription Tools and methods, Applications

A key challenge for the construction of phonological corpora relates to the phonetic transcription of speech samples. Phonetic transcription requires transcribers to be appropriately trained to produce accurate transcripts and familiarity with an appropriate method of adjudicating discrepancies between differing transcriptions. A potential solution to this quandary is using crowdsourced transcriptions where individuals each attempt to offer transcriptions that somehow are integrated together into a single representative transcription. Crowdsourcing brings its own challenges where transcribers of different skill levels transcribe various samples and could produce some variation amongst their different returned transcriptions. Therefore in order to harness the possibilities of crowdsourced phonetic transcriptions a new process of managing the many returned phonetic transcriptions must be developed. This paper presents results of using crowdsourced phonetic transcription on samples from the Speech Accent Archive hosted at George Mason University. The Archive contains over 2,500 speech samples but only around half have associated phonetic transcriptions. Finding a means to rapidly construct phonetic transcriptions would prove tremendously useful to both this corpus and be applicable to a number of other corpora. In our experimental system over 500 different phonetic transcriptions were provided by over 100 different volunteer transcribers from across the world. Each transcription was aligned with other transcription and then the differences we analyzed to identify variability of transcriptions. The results of this study show that while variability exists among provided transcriptions that variability is limited to differences in a discrete set of phonetic features. Using multiple transcribers in a phonological corpus offers the possibility of rapidly transcribing data that might otherwise never be evaluated. The results of our study suggest the crowdsourcing is a reasonable technique for phonetic transcription and our system provides an initial capability for unifying the crowdsourced transcription into a single reference transcription.

Kyle, Kristopher, Latflair, Geoffrey and Ziegler, Nicole - University of Hawaii at Manoa - The TMLE-Corpus project: Academic language in technology-mediated contexts

Tools and methods

An important aspect of a validity argument for a language assessment tool such as the TOEFL is a demonstrated alignment between the linguistic demands of the target language use domain and the assessment tasks (Chapelle, Enright, & Jamieson, 2008). Corpus analyses are well suited to generate such evidence (Biber et al., 2004), but researchers are constrained by whether appropriate corpora exist for the target domains. Currently, there are a number of corpora that represent various types
of language that university students encounter and/or produce in traditional academic settings, such as BAWE (Alsop & Nesi, 2009), MiCUSP (Römer & O’Donnell, 2011), and T2KSWAL (Biber et al., 2004). Increasingly, however, a typical university experience may be supported with technology mediated learning environments (TMLEs) (Jacoby, 2014; Means et al., 2013), which are not yet represented in extant academic corpora.

The TMLE-Corp project seeks to address this gap in three stages. First, a large-scale survey will be conducted to determine the types of texts that are typically encountered and produced in TMLE in universities across the USA. Based on the survey results a corpus of the typical texts that are representative of TMLEs will be collected. Finally, linguistic analyses will be conducted to explore the linguistic features of TMLEs, examine any linguistic variation across disciplines, and compare the linguistic features of TMLE to other registers represented by existing academic corpora. The main outcomes of the project include the addition of a large, representative corpus of TMLE texts to current resources and an in-depth report of the linguistic features of typical TMLEs, including how these features compare with the linguistic features of more traditional university learning environments.

This presentation will outline the TMLE-Corp project, including the results of the TMLE survey. Protocols for the next steps of the project (corpus collection and linguistic analysis) will also be discussed.

Lake, William and Cortes, Viviana - Georgia State University - Using lexical bundles to reassess disciplinary norms in Spanish & English humanities and social science writing

Linguistic analyses of corpora

Quantitative corpus studies on research article genre conventions have bolstered the design of pedagogical materials in scientific research writing (Robinson et al., 2008; Kanoksilapatham, 2015). However, humanities and social sciences writing pedagogy for languages other than English is currently less grounded in formulaic language research (Nance, 2015).

As a point of departure for the design of Spanish for academic purposes curricula, a partial replication of MacDonald’s (1994) comparison of sentence-level differences in scientific and humanities writing was conducted. First, six corpora of literary criticism, history and psychology articles in English and Spanish (six million words total) were compiled. Following MacDonald’s taxonomy, nouns in 600 randomly sampled sentences from each corpus were classified as either phenomenal (e.g. persons, places, objects or attributes) or epistemic (abstract reasoning nouns, analytical techniques or prior research). Next, using frequency and range criteria, lexical bundle lists were generated from each corpus. These bundles were then classified functionally and structurally. A functional comparison of the resulting bundles revealed that referential bundles in English & Spanish humanities writing contained more phenomenal noun phrases, while bundles containing epistemic noun phrases were significantly more prevalent in Spanish psychology articles.

These results lend support to MacDonald’s assertions about recurring differences between writing in the three disciplines. That is, that psychology’s tendencies toward epistemic means of persuasion are visible at the both the individual sentence and formulaic language levels. Conversely, bundles in the humanities contained more phenomenal noun phrases. As such, much of their rhetoric is based on describing historical and textual events, rather than data analyses, in support of a thesis. Second, since a high volume of psychology-specific lexical bundles was discovered, these findings may prove useful for the design of Spanish scientific writing teaching materials. The session will conclude with recommendations for courses in Spanish for academic purposes.

Lamance, Rachel - University of Arizona - Visualizing the Rhetorical Moves of Source Incorporation: A Corpus Analysis

Linguistic analyses of corpora

Analysis of rhetorical moves (Swales, 1990) has been helpful in identifying organizational patterns and functions within texts. Although previous research has combined the study of moves with positioning (Chang & Schleppegrell, 2011), move organization was not the focus. This project analyzed the use and organization of rhetorical moves to examine how students positioned themselves in relation to source texts. The moves included here were based on a pedagogical framework often used in writing courses: Point (introducing topics or ideas), Illustration (providing support from outside sources), and Explanation (interpreting or extending support from sources). Variations in linguistic features among these moves were also examined to determine how students linguistically signalled functions of their writing.

Three corpora, each consisting of 30 argumentative essays, were analyzed: English as a first language (L1) students enrolled in first-year writing (FYW), English L2 FYW students, and English L1 final-year undergraduates. All FYW essays were randomly selected from the corpus described in Staples and Reppen (2016), and all final-year undergraduate essays were randomly selected from the Michigan Corpus of Upper-Level Student Papers (MiCUSP). Each text file was manually tagged for PIE moves, then AntConc and other computational tools were used to determine the order, frequency, and length of moves. To assess linguistic variation, passages belonging to each move were compiled to form corresponding sub-corpora. Word frequencies generated by AntConc highlighted the linguistic cues student writers employed to signal shifts among Point, Illustration, and Explanation. A Python script was created to translate tagged text files into color-coded html visualizations. Results, which will be presented as visualizations, showed that final-year undergraduates tended to use a greater number of shorter moves throughout their writing. Findings from this project can help writing instructors better understand how student writers from different linguistic backgrounds and at different stages incorporate sources in terms of structure.
Larsson, Tove, Université catholique de Louvain, Callies, Marcus, University of Bremen, Hasselgård, Hilde, University of Oslo, Judith Laso, Natalia, University of Barcelona, Paquot, Magali, Université catholique de Louvain, van Vuuren, Sanne, Radboud University and Verdagne, Isabel, University of Barcelona - Adverb placement in EFL academic writing: Going beyond syntactic transfer

Linguistic analyses of corpora

Adverb placement has received considerable attention in second language research. However, previous studies suffer from four main limitations: (1) they often focus primarily on misplaced adverbs, (2) they tend not to distinguish between different types of adverbs (cf. Rankin, 2010), (3) very few studies use comparable native and/or expert language data, and (4) they rarely consider factors such as positional variability and register, despite the fact that these have been found to influence adverb placement (cf. Biber et al., 1999).

The main objective of this study is to investigate adverb placement in six components of the VESPA learner corpus (EFL learners whose L1 is either French, Spanish, Norwegian, Swedish, German or Dutch) as compared with speech (LINDSEI), British students’ academic writing (BAWE) and experts’ academic writing (LOCRA) to explore the effect of L1 background, register and the positional variability of adverbs. We focus on adverbs denoting epistemic modality (e.g. clearly, maybe, possibly). As these adverbs are at the intersection of word order, information structure and stance marking (cf. Osborne, 2008; Callies, 2009; Larsson, 2017), they can be expected to present varying difficulties for the different L1 populations under scrutiny. Methodologically, our study uses Hasselgård’s (2010) syntactic classification for adverb placement and builds on Granger’s (2015) Contrastive Interlanguage Analysis for corpus comparison.

Preliminary results confirm that students whose L1 has V2 word order (e.g. Swedish) more readily place adverbs such as possibly clause-initially than those whose L1 has SVO word order (e.g. French). However, the results also suggest that register and the adverbs’ positional variability play a more important part in explaining the differences found between the learners, the native-speaker students and the experts. The study nuances the, at times, somewhat categorical view of correct vs. incorrect adverb placement found in the literature, thereby enabling new pedagogical recommendations.

References


Lindley, Jori - Purdue Fort Wayne - How do features of constructions’ grammatical subjects (1st/2nd/3rd person and human/non-human status) affect usage patterns: Two case studies

Linguistic analyses of corpora

Features of a construction’s grammatical subject—particularly 1st/2nd/3rd person or human/non-human status—correlate with certain discourse functions or usage patterns. Yet example sentences in grammars and ESL/EFL textbooks over-rely on 3rd person human subjects (e.g., “John is always laughing”), which can be misleading. I show, through two analyses, that constructions are better understood when we consider their grammatical subjects.

The first study is on “always” + progressive, e.g., “you’re always calling.” Grammars (Sinclair 1990, Biber et al. 1999, etc.) characterize these as complaints, but in my study of the Corpus of Contemporary American English over 70 percent of the 752
tokens were used for neutral descriptions. However, their likelihood of being complaints fluctuated depending on subject type. We complain least often about ourselves and non-humans, more often directly to others (2nd person), and most often about absent humans (3rd person). A configurational frequency analysis showed the relation between person and function to be significant at the p < 0.01 level. In particular, the negative correlation between 1st person and complaining was very strong.

The second study is on “broke a/my finger” type utterances, involving inalienable possessions. This has been treated as a matter of “a” being able, sometimes, to implicate closeness rather than distance (Grice 1975, Horn 1984). Analyzing data from the internet, I found that this is, instead, best viewed as a lexical decision (“a” versus “my”) driven by salience, where “a” is associated with lower salience. This choice is additionally affected by person: Speakers use “a” more often with “you” than “I”, and most of all with “he/she/they”, a finding I relate to it being easier to empathize with ourselves than others (Kuno 1987). My work on “broke a/my finger” and “always” shows that considering their grammatical subjects enriches our understandings of linguistic expressions.

References


Liu, Jing - Wuhan University - Conclusion Sections of Research Articles: Multi-dimensional Analysis

Linguistic analyses of corpora

Multi-dimensional analysis has been applied to study linguistic features of articles from different registers, especially concerning the investigation of research articles in recent years. However, little research has been done to explore what linguistics variations exist in sub-section of the research article. The purpose of this paper is to investigate the linguistic features of the Conclusion section of the research article (RA) from six disciplines (Applied Linguistics, Biology, Engineering, Physics, Philosophy, and Sociology) using multi-dimensional analysis. The study adopts Gray’s (2015) four dimensions (Dimension 1: Academic Involvement & Elaboration vs. Information Density; Dimension 2: Contextualized Narration vs. Procedural Description; Dimension 3: Human Focus vs. Non-human Focus; Dimension 4: ‘Academese’) to analyze the whole corpus. The Conclusion corpus includes Conclusion sections from 4,459 journal articles. The corpus was tagged for part-of-speech and then put into SPSS for calculating Z-score. The results show that great differences exist between natural science and social science along four dimensions. Social science articles have the higher level of academic involvement and elaboration than that in natural science; While there are more procedure descriptions in natural science than in social science. It is also interesting to note that social science and natural science get the similar mean scores along Dimension 4: ‘Academese’. After examining variations between natural science and social science in general, we probed deep into the linguistic variations of six disciplines along the four dimensions. Along Dimension 1 (Academic Involvement & Elaboration), Philosophy has the highest score. Along Dimension 2 (Contextualized Narration vs. Procedural Description), it is interesting to point out that Physics has less procedural description than that in Gray’s (2015) study. Meanwhile, it is strange to note that Sociology which supposes to involve more human activities gets the least score along Dimension 3 (Human Focus vs. Non-Human Focus). It is also beyond our expectation that Engineering gets the highest score while Philosophy has the lowest along Dimension 4 (‘Academese’), which shows that sometimes natural science articles might be more academic than social science articles. This paper fills in the gap that few studies investigated the linguistics variations of sub-sections of research articles using multi-dimensional analysis. The study has both an empirical implication of further conducting MD analysis in different sub-section of research articles as well as a pedagogical implication which presents novice writer the characteristics of the conclusion sections of research articles across various disciplines.

Liu, Yingying and Lu, Xiaofei - The Pennsylvania State University - The Count/Mass Properties of Nouns and Chinese EFL Learners’ Article Use: A Corpus-Based Investigation

Linguistic analyses of corpora

Articles present a major challenge to English as a Foreign Language (EFL) learners, particularly those whose L1 is article-less, such as Mandarin Chinese. Previous research has shed useful light on the complexity of English article use (Biber et al., 1999) and provided robust empirical evidence for learner difficulties with articles (McEnery & Xiao, 2010; Dewey, 2009). However, much such research has focused on learners’ holistic performance in article use at the sentence and discourse levels (e.g., context-dependent specific/non-specific reference), while scarce attention has been paid to learners’ word-level, item-specific knowledge of noun countability.

This paper investigates Chinese EFL learners’ use of English articles, with the specific goal of revealing their knowledge of the count/mass properties of nouns demonstrated in their use. To this end, we compared the use of English articles in two corpora of comparable size, i.e., the written component of the Corpus of Contemporary American English (COCA; Davies, 2008) and the
Written English Corpus of Chinese Learners (WECL; Wen, Wang, & Liang, 2005). Using a self-designed analytical toolkit, we identified 34 article-noun combinations unique to WECL (Chinese EFL learner data) and 144 unique to COCA (L1 English speaker data). The learner-unique combinations were classified into four categories, and each category was qualitatively analyzed to closely examine the count/mass properties of the nouns involved and the articles used by L1 English speakers with those nouns. The results suggest that a large proportion of Chinese EFL learners’ issues with English article use were related to their confusion of the count/mass properties of nouns. We conclude with a discussion of the pedagogical implications of our findings, particularly regarding helping learners develop depth of their vocabulary and idiomatic knowledge.

Lu, Xiaofei, Casal, J. Elliott and Liu, Yingying - The Pennsylvania State University - A corpus-based study of the rhetorical functions of syntactically complex sentences in research article introductions

Linguistic analyses of corpora

This paper investigates the rhetorical functions of syntactically complex sentences in research article (RA) introductions. There has been a growing interest in the “integration of genre analysis and corpus-based investigations” (Flowerdew, 2005, p. 5) in English for Academic Purposes research, which allows researchers to explore the linguistic realization of rhetorical moves (e.g., Cortes, 2013; Durrant & Mathews-Aydinli, 2011), rather than profile the moves themselves. Meanwhile, while many studies have quantitatively assessed the relationship of syntactic complexity to writing quality (e.g., Beers & Nagy, 2009) and language development (e.g., Lu, 2009, 2011), studies that examine the rhetorical functions of syntactically complex sentences in academic writing are scarce (e.g., Ryshina-Pankova, 2015). The current function-based analysis of syntactic complexity aims to advance our understanding of how expert writers accomplish their rhetorical goals in RA writing.

Our data consists of RA introductions from the Corpus of Social Science Research Articles (COSSRA), which we compiled from 600 RAs published in 2012-2016 in six social science disciplines. From each discipline, 100 RAs from five top journals were sampled according to impact factor and expert judgment. Four widely used operationalizations of syntactic complexity are explored: mean length of sentence, left embeddedness (i.e., number of words before the main verb), subordination (i.e., number of subordinate clauses per sentence), and nominalizations. The L2 Syntactic Complexity Analyzer (Lu, 2010), the D-Level Analyzer (Lu, 2009), and other tools are used to measure syntactic complexity and identify sentences which meet thresholds for each measure. The data is also manually tagged for rhetorical moves according to a modified version of the Create a Research Space model (Swales, 1990, 2004). The rhetorical functions of the syntactically complex sentences identified are systematically described. The implications of our findings on the relationship between syntactic complexity and rhetorical functions for writing research and pedagogy will be discussed.

Lu, Xiaofei, Yoon, Jungwan and Kisselev, Olesya - The Pennsylvania State University - Matching phrase-frames to rhetorical moves in social science research article introductions

Linguistic analyses of corpora, Applications

In response to the growing interest in the study of formulaic sequences in English for Academic Purposes (EAP) (e.g., Ackerman & Chen, 2013; Simpson-Vlach & Ellis, 2010) and building upon emerging research on matching formulaic sequences to rhetorical functions in EAP writing (e.g., Cortes, 2013; Durrant & Mathews-Aydinli, 2011), this study aims to derive a pedagogically applicable list of phrase-frames (p-frames), a type of multi-word sequence with a variable slot, from the Introductions subcorpus of the Corpus of Social Science Research Articles (COSSRA) and to systematically match the p-frames derived to the rhetorical moves and steps they are used to materialize. Academic written discourse has been found to rely heavily on p-frames (Biber, 2009) and to also feature highly variable p-frames (Gray & Biber, 2013). As such, novice EAP writers stand to benefit from expanding their linguistic repertoires to include academic p-frames. From a genre approach, they will also benefit from knowledge of the rhetorical functions the p-frames are associated with.

COSSRA consists of 600 research articles (RAs) published in 2012-2016 in 30 leading academic journals in six social science disciplines. We first extracted a list of p-frame candidates from the Introductions sub-corpus of COSSRA using kNgram. We then filtered the candidates using a series of procedures to ensure their pedagogical applicability, including imposing thresholds for the frequency and number of variants of the p-frames as well as the number of RAs and disciplines they occur in, disregarding linguistically incomplete candidates, and eliminating candidates that crossed clausal boundaries. The resulting list contained 413 and 102 p-frames of length 5 and 6, respectively. The subcorpus was fully annotated for rhetorical moves and steps using a modified version of Swales’ CARS (Create a Research Space) model (Swales, 1990, 2004). This allowed us to systematically describe the p-frames commonly associated with different rhetorical moves and steps. For example, the frame in the present study we’ve often used to “announce present research descriptively and/or purposively” (Move 3 Step 1) with such variants as focus, examine(d), and investigate.

The p-frame list compiled and the systematic description of the associated rhetorical functions of the p-frames provide useful resources to facilitate novice EAP writers’ acquisition of academic p-frames and their rhetorical functions in the RA genre. We conclude by illustrating how matching p-frames and formulaic sequences in general to their rhetorical functions provides a productive way for enhancing the pedagogical value of formulaic sequence research.

References

Lutzky, Ursula - Vienna University of Economics and Business - “I sent my tweet because I couldn't find an answer on that page.” British airlines' use of Twitter for customer communication purposes

Linguistic analyses of corpora

Anyone who has ever had to find specific information on an airline's website or tried contacting them to discuss a specific issue, such as lost luggage, flight delays or cancellations, will know that not only does the maze of subpages and hyperlinks need advanced navigation skills, but especially the time spent waiting on hold requires high levels of patience. It is thus not surprising that passengers increasingly accept airlines' offer of getting in touch with them on Twitter (see e.g. the Contacts page for British Airways).

This study addresses the application of the social media site Twitter in the world of business and focuses in particular on its interactive uses in customer communication, which have not been studied extensively from a linguistic perspective to date (but see Page 2014). It is based on the British Planes Twitter Corpus, which comprises 6.7 million words and includes tweets that customers directed at 11 British airlines over a period of four months as well as the airlines' replies.

The aim of this talk is to show how corpus linguistic analyses can uncover patterns of language use and their communicative functions in the context of customer service exchanges online. This includes the study of keywords to highlight areas of customers' concern and needs (e.g. cancelled, stuck), of Twitter specific features such as hashtags (e.g. #greatservice, #poor; see also Zappavigna 2018), and of frequent clusters that shed light on customers' (dis)satisfaction or companies' attempts at restoring face (e.g. no info, very sorry). Thus, this talk provides an example of a practical application of corpus linguistics in a business context and offers further insights into the features of customer communication on social media.

References


Mayer, Cristina - PUC-SP - Canonical dimensions of Web registers: correlating functional and lexical dimensions of register variation

Linguistic analyses of corpora

This paper presents the results of a Canonical correlation analysis (Tabachnick & Fidell, 2007), revealing the relationships between two multidimensional (MD) analyses of register variation, namely a functional MD analysis (Biber, 1988) and a lexical MD analysis (Berber Sardinha, forthcoming), across Web text varieties, specifically social networks and consumer generated content. Web registers have been the object of several MD investigations (Biber et al., 2011; Berber Sardinha, 2014; Biber & Egbert, 2015, 2016); however, these studies have not looked at content such as comments, reviews, and complaints. Thus, to date, no known studies have examined the correlation of both functional and lexical analyses in the same corpus, which this paper aimed at. For this purpose, a corpus of 15 registers was designed and compiled, the COP – Corpus of Online Posts, and two MD analyses were conducted. The first one was a functional MD analysis, in which the texts were tagged for part of speech using the Biber Tagger, the features were counted and normed and entered in a factor analysis, in which 4 dimensions were interpreted, representing the functional parameters underlying the variation across the registers. In the second analysis, the data were groups of words, identified by programs designed for this study (Berber Sardinha, forthcoming). The data were also entered in a factor analysis and 5 dimensions were interpreted, reflecting semantic groupings rather than functional considerations. A canonical correlation analysis was then run in order to explain the relationship between the functional and lexical characteristics. In this analysis, 4 canonical dimensions were interpreted, showing the correlations between the 4 functional and 5 lexical dimensions interpreted previously. All analyses led to the study of the use of the language by Web users, as well as a new perspective for the lexicogrammar paradigm through the Canonical correlation analysis.
Mbodj, Ndaye Bineta - Georgia State University - Exploring the discipline and writing experience interface in learners’ use of lexical bundles.

Linguistic analyses of corpora, Applications

Lexical Bundles (LBs), “statistically the most frequent recurring sequences of words in any collection of texts” (Hyland, 2012; p. 150), have been shown to be pervasive in academic prose (Biber, Johansson, Leech, Conrad, & Finegan: 1999; Wray & Perkins, 2000). As such, they are seen as markers of proficiency (Cortes, 2004; Martinez & Schmitt, 2012). Studies that looked at learners’ use of LBs in academic writing have found that novice writers rarely use the same bundles as expert writers (Cortes, 2004, Hyland, 2008a; 2008b) and have therefore advocated explicit teaching of LBs to help learners develop preferred rhetorical practices in their fields of study. However, the handful of studies that trialed the explicit teaching of LBs reported only marginal (Jones and Haywood, 2004; Li & Schmitt, 2009) or no significant gains (Cortes, 2006). This suggests a necessity to explore how learners acquire LBs in order to better help them develop their use of such formulaic expressions.

The present study explored the potential interaction between discipline and writing experience, two factors that have been identified in the literature as affecting novice writers’ use of LBs in some way or other (Hyland 2008a; 2008b; Cortes, 2004; Howarth, 1998). The study compared the use of 4-word LBs (previously identified in a corpus of medical research articles) by novice and expert writers in the medical field and in five other academic fields. The findings indicated that a combination of both writing experience and field-specific knowledge are necessary for increased use of target LBs. The results also indicated field-specific knowledge as a pre-requisite with novice medical writers using more target LBs than expert writers in other fields. Novice medical writers’ use of LBs was further analyzed and discussed with regard to the reported reliance on and repetition of a limited number of LBs at this level (Staples et al., 2013).

Monteiro, Kátia and Crossley, Scott - Georgia State University - Monolingual and bilingual processing norms as explanatory variables of L2 writing quality

Linguistic analyses of corpora

Eye-tracking technology provides rich moment-to-moment source of lexical processing information (Conklin & Pellicer-Sánchez, 2016). However, few studies have used eye-tracking data to investigate corpora and L2 learning. The exploratory study reported here used word reading information from monolingual and bilingual speakers collected from a large corpus of eye-tracking data to predict L2 writing quality.

The Ghent Eye-Tracking Corpus (GECO; Cop, Keuleers, Drieghe, & Duyck, 2015), which includes eye-tracking data from a novel read by English monolinguals and Dutch-English bilinguals, was used to derive monolingual and bilingual norms for lexical processing at the word level. These included lexical norms such as gaze duration and go-past time. Processing norms for individual words were computed from 427 timed argumentative and 427 integrated essays written by English L2 learners that had been scored by expert raters for writing quality. The bilingual and monolingual norms were used as fixed effects in linear mixed effect (LME) models with participant as random effect to predict essay scores. Additional fixed effects included age, first language distance from English, and task (independent or integrated writing). An LME using only GECO indices indicated that three monolingual eye-tracking features related to average word reading time, go past time, and third fixation duration were significant predictors in an LME model that reported a marginal R2 of .02 and a conditional R2 of .61. A full model using all fixed effects reported a marginal R2 of .13 and a conditional R2 of .65 and included the three GECO indices in the initial model along with language distance and task. The findings suggest that eye-tracking data for individual words explains a small amount of variance in L2 writing and that this variation remains when a number of other fixed effects are co-varied. Implications for L2 writing assessment, pedagogy, and theory are discussed.

Mostafa, Tamanna - Georgia State University - Exploring complexity measures as indices of L2 oral development: A corpus-based study

Applications

Most studies on language complexity and proficiency either focus on how different measures and indices contribute to oral or written proficiency scores (e.g., Iwashita, Brown, McNamara, & O’Hagan, 2008; Bulté & Housen, 2015) or how complexity measures are used as indices of second language (L2) proficiency (e.g., Lahmann, Steinkrauss, & Schmid, 2016). L2 development is also a distinct construct, but there is a lack of research on how L2 complexity is related to L2 development. In response to this gap in research, the present study investigates whether complexity measures can be used as indicators of L2 learners’ interlanguage development in an instructional context and whether or not learners of different proficiency levels produce developmentally distinct forms of structures.

A specialized learner corpus was created consisting of monologic oral production data collected from beginner, intermediate, and advanced English as second language (ESL) classes. Data was collected at two points in a 16-week long semester: beginning (week 2-5) and end (week 11-14). The individual oral activities that the participants (total number 100) performed as part of their regular classes were audio-recorded and transcribed (56,550 words in total). The developmental sequences for the acquisition of morphemes, negation, and questions as outlined in Ortega (2009) was used as the basis for developing three different complexity measures. Results of one-way ANOVA showed that there was no significant difference in the production of developmentally higher-level morphemes between the learners of three different class levels. However, results from five paired-sample t-tests showed that beginner level participants, as opposed to intermediate and advanced learners, significantly developed their morpheme complexity scores from the beginning to the end of the semester. Additionally, advanced level learners significantly developed their complexity scores in negation and question from the beginning to the end of the semester. Pedagogical implications will be presented and discussed.
**Motstchenbacher, Heiko - Western Norway University of Applied Sciences - Language use in media coverage before and after coming out: A corpus-based study of texts on Ricky Martin**

Linguistic analyses of corpora

This paper presents research carried out within the project Linguistic Dimensions of Sexual Normativity. It seeks to contribute to a strand of recent research that has focused on the discursive construction of celebrities’ public coming out in the media (King 2017, Magrath, Cleland & Anderson 2017) and conducts a corpus-based critical discourse analysis of the linguistic representation of Latino pop star Ricky Martin (Rivera Santana et al. 2010) in media texts. Drawing on a corpus of English-language newspaper and magazine articles, the study seeks to investigate whether and how the language used to write and talk about Ricky Martin has changed after his public coming out as a gay man in 2010. For this purpose, two sub-corpora (pre-2010, post-2010) are compared and various types of quantitative and qualitative analysis are carried out, using the corpus tool AntConc. These include a comparison of the frequency lists of the two corpora (Archer 2009), concordance and collocation analyses of lexical items that prove to be particularly common in the two sub-corpora (Baker 2016), and a keyword analysis, which highlights word-forms that occur unusually (in)frequently in one of the corpora when compared to the other (Baker 2004). The findings are discussed in relation to the theorisation of coming out (Chirrey 2003) and normativity (Motstchenbacher 2014) in language and sexuality studies.

References


**Nam, Yunjung - Georgia State University - Korean EFL Learners’ Use of ‘give’ Ditransitive Constructions: Evidence from Corpus Research**

Linguistic analyses of corpora

A usage-based constructional approach posits that knowledge about constructions emerge or develop through exposure to language in use (Ellis, O’Donnell, & Römer, 2013; Goldberg, 2006) and prototypical exemplars with high frequency are considered to play a crucial role in usage-based model (Bybee, 2013; Goldberg, 2006). Within the frame work of usage-based constructional approach, this study aims to investigate English ditransitive constructions of the verb ‘give’ used by Korean EFL learners. The ‘give’ ditransitive construction was chosen since the verb ‘give’ is prototypical of the ditransitive construction, which plays an important role in construction acquisition. Another reason is that investigating the use of ‘give’ helps understand cross-linguistic interferences since Korean, as a case-marking language, does not have linguistic cues comparable to the English ditransitive construction. Data will be collected from the Korean learner corpus - Yonsei English Learner Corpus (YELC), consisting of argumentative essays written by pre-college students. After extracting all the instances containing the verb ‘give’, the data will be analyzed in terms of different types of construction patterns and their frequencies. Then, only the ditransitive constructions of ‘give’ will be analyzed in terms of the three variables as in Xu (2016): pronominality (whether IO or DO is a pronoun), weight (lengths of IO or DO), and semantic classes (transfer, communication, enablement, permission, causation, feeling, schematic interaction, or idiom) suggested by Newman (1996). The results will be compared with the findings of Xu (2016), which conducted a contrastive analysis of ‘give’ ditransitive constructions between a Chinese EFL learner corpus and a native English corpus – The Louvain Corpus of Native English Essays. The study will conclude with a discussion of pedagogical implications for EFL teachers and learners.
Discourse Completion Tasks (DCTs) are widely used to collect pragmatic data. However, a number of studies have brought into question their validity and their ability to elicit authentic-like data (Beebe & Cummings, 1998; Yuan, 2001; Golato, 2003). Critics of DCTs argue that responses are contrived, that participants respond with what they think they would say in a real situation rather than what they would actually say, and that participants say what they think the researcher wants to hear. Other concerns include using a written task to collect spoken data and the potential for participants to be given tasks they are unfamiliar with.

This study examines how corpora can be used to improve DCT instruments by helping to correct some of these issues. In this study, a DCT was created using data drawn from the Q+A Corpus (Baker & Egbert, 2015), which consists of questions and answers on the Yahoo! Answers web forum. The DCT was administered via a Qualtrics survey, and items were formatted to model those on the Answers website. Prior to completing the DCT, participants were required to spend time becoming familiar with the Answers website. After completing the DCT, participants were interviewed about their experience and their perception of politeness strategies in online forums.

After collecting DCT data, responses were coded for both form-based and content-based strategies. An ANOVA was used to compare directive speech act strategy use in the DCT responses to corpus data using normalized counts. Preliminary results suggest no significant difference between the authentic corpus data and the elicited DCT data with either form-based strategies, $F(1,4) = 0.002, p = 0.96, d = 0.02$ nor content-based strategies, $F(1,8) = 0.0005, p = 0.98, d = 0.01$.

These findings suggest that corpora can be useful in improving the DCT instrument. Corpora can aid researchers in finding authentic prompts and situations to create more valid items and that participants are likely to be familiar with. Corpora can also be used to expose participants to the context of the task.

Furthermore, the qualitative analyses used in developing and validating this DCT highlight how a mixed-methods approach using techniques such as DCTs and interviews are useful in developing coding schemes and aiding pragmatics researchers in deciding which features to look at in a corpus.
Insults and gender have both been widely studied in linguistics; however, few studies have attempted to analyze both together. A number of studies have noted differences in the language used to address men and women (see Kjellmer, 1986; Baker, 2010; Ashwell, 2016). The present study examines whether there are differences between insults directed at men and women in the comments section of YouTube music videos.

A corpus of YouTube comments on music videos of popular solo artists was created using a web-scraping code in Python. The corpus used for analysis contained 25,912,009 words. There are comments from ten total videos in the corpus. Five are from the female artists Adele, Beyoncé, Lady Gaga, Katy Perry, and Rihanna. The other five videos are from male artists Drake, Ed Sheeran, Eminem, Kanye West, and The Weeknd. Artists to be included in the study were drawn from a list of Google’s “Most mentioned on the web.”

Frequency lists were created for the male and female subsets of the data to determine which insults were most common in the data for each group. Based on the keyword analysis, three words (fat, bitch, and fake) were chosen for closer examination. A random sample of comments containing each of the target words were coded by hard to determine whether the insult was directed at the artist, another commenter, a backup dancer or video cast member, or somebody else.

The words fat and bitch were found to be common insults used against females but were rarely used to attack males. When fat and bitch did appear in music videos featuring male artists, the insults were most frequently directed at female dancers in the video or other commenters. A qualitative analysis revealed that insults directed at men were often intended to emasculate the man or attack his sexual partners. The results of the study suggest insults are prevalent in the comments section of YouTube and that different types of insults are used against men and women.

Novikov, Aleksey - University of Arizona - L1 Typology Effects on Complexity Features in Academic Writing

This study analyzed effects of L1 transfer in L2 writing in English. To this end, two typologically different L1s, Russian and Mandarin Chinese, were selected for the study with two complexity features being distinctly different in the respective L1s, specifically premodifying nouns (NNs) and relative clauses (RCs).

According to Bongartz (2002), transfer can be illustrated by two ways of noun combinations: “phrasal noun combinations” (e.g., the song about love) and “incorporation structure” (e.g., the love song). The former is typologically characteristic of Russian, while the latter is a common feature of Mandarin Chinese. The study by Staples & Reppen (2016) showed that incorporation structure abounds for L1 Chinese. In terms of RCs, Chinese, unlike English and Russian, allows only prenominal RCs. Schachter (1974) demonstrated that L1 Chinese learners avoid using RCs for that reason.

Drawing from these typological L1 distinctions, this study asked two research questions: 1) Will L1 typology affect the use of NNs in an L2? 2) Will L1 typology affect the use of RCs in an L2?

The analysis was performed on argumentative essays by high intermediate L1 Chinese (160 texts) and L1 Russian writers (274 texts) from the International Corpus of Learner English (ICLE). The hypothesis in this study was that L1 Chinese typology, namely the “incorporation structure” would result in a greater number of NNs compared to L1 Russian, while an increased number of RCs would be seen in the L1 Russian compared to L1 Chinese.

The results showed that L1 Chinese writers have a significantly quantitative preference for NNs compared to L1 Russian writers. In contrast, L1 Russian writers used significantly more RCs compared to L1 Chinese. These findings are in accordance with other studies on typology effects and suggest that L1 typology should be taken into consideration when analyzing L2 complexity.

Pacheco, Aline, The Pontifical Catholic University of Rio Grande do Sul, Friginal, Eric, Georgia State University, Roberts, Jennifer, Embry-Riddle Aeronautical University and Cavallet, Joao, The Pontifical Catholic University of Rio Grande do Sul - Exploring the discourse of international aviation from the Corpus of Pilot and ATC Communications (CORPAC)

This paper examines the linguistic characteristics and patterns of cross-cultural miscommunication (i.e., cross-talk) of a specialized corpus of International English Pilot and ATC Communications (CORPAC). We utilize corpus-based [critical] discourse analysis and multidimensional analytical (MDA) (Biber, 1988, 2006) framework in exploring the discursive practices across the cultural structures and task dimensions of this language domain, focusing especially upon pilot and controllers’ understanding of identities, role-relationships, and power dynamics. Over the years into the present, international aviation has exponentially employed airline pilots who are second language speakers (or non-native speakers) of English, tasked to communicate in English with native and non-native English speaking air-traffic controllers on the ground. In some extreme cases, there have been fatal aviation accidents attributed to miscommunication as well as cross-cultural pragmatic factors from these interlocutors. The specialized corpus analyzed in this study (120 total texts, with 234,563 number of words) was collected from actual flight logs and recordings from multiple data sources, in English, with air-traffic controllers who are predominantly
native English speakers and international pilots. An iterative cycle which combines critical approaches to data extraction and a progression of stages involving quantitative and functional analyses (Baker et al., 2008; Gentil, 2013) appears to show how cross-talk can be further described and defined using evidence from corpora. At the same time, Biber’s (1988) multi-feature, multidimensional analytical framework allows for a relevant point of comparison on how speakers utilize linguistic co-occurrence features of spoken professional discourse from various sub-registers and tasks. Results show interesting similarities and differences in how pilots and air-traffic controllers use linguistic features across co-occurring dimensions of spoken discourse. We will discuss implications for macro/micro language policies and how corpus-based data could be used for materials design, language assessment, and in developing training programs in this register of professional discourse.

Paquot, Magali, Université catholique de Louvain, Brezina, V., Vaclav, Lancaster University, Gablasova, Dana, Lancaster University and Naets, Hubert, Université catholique de Louvain - Relational co-occurrences as an index of oral proficiency in the Trinity Lancaster Learner Corpus

Linguistic analyses of corpora, Applications

Recent studies have shown that statistical co-occurrences, i.e. co-occurrences extracted and ranked with the help of association measures such as the mutual information (MI) score, can be used to describe EFL learner performance across proficiency levels (e.g. Durrant & Schmitt, 2009; Granger & Bestgen, 2014). Paquot (2017), for example, focused on relational co-occurrences (i.e. where the co-occurring words appear in a specific structural relation) in French EFL learner academic writing and showed that phraseological indices based on the MI are better able to gauge language proficiency than traditional measures of syntactic and lexical complexity.

Most studies so far, however, have explored EFL learner use of statistical co-occurrences in upper-intermediate to advanced writing and focused on learner groups representing a limited number of first languages or language families. The main objective of this study is therefore to explore learners’ use of co-occurrences in the Trinity Lancaster Learner Corpus (Gablasova et al. 2017), i.e. roughly 4 million transcribed words from the Trinity College London spoken language exams, and answer the following main research questions:

- To what extent can relational co-occurrences be used to describe L2 oral performance at different proficiency levels (from B1 to C1/C2)?
- Does phraseological competence develop in the same way across different learner groups (i.e. Chinese vs. Hindi vs. Spanish speakers of English)?

We focus on verb + object co-occurrences as these structures have repeatedly been shown to be a major hurdle for English L2 learners (e.g. Nesselhauf, 2005). Co-occurrences were extracted from the learner spoken corpus with the help of regular expressions and evaluated on the basis of MI scores computed from the ENCOW16AX corpus (see Paquot, 2017 for more information on the methodology).

Preliminary results suggest that phraseological competence develops (slowly) from B1 to C1/C2 in spoken language but with variability within and across proficiency levels, L1 groups and task types.

References


Peng, Siyao and Zeldes, Amir - Georgetown University - Validating and Merging a Growing Multilayer Corpus – the Case of GUM

Tools and methods

This paper reports on expanding a class-sourced, richly-annotated and freely available corpus of English Web genres called GUM (Georgetown University Multilayer corpus, Zeldes 2017). Expanding the existing corpus of news, interviews, how-to guides and travel guides, we add four new genres: academic writing, biographies, fiction, and reddit discussions. These are annotated by students in the classroom using multiple online annotation tools to add TEI-XML structural markup, rough speech act information, POS tagging, dependencies, entity and coreference annotations, and discourse parses in Rhetorical Structure Theory (Mann & Thompson 1988).
In the talk, we discuss challenges in maintaining high quality annotations in new genres across a wide range of annotation types. We present error-highlighting strategies in annotation tools (Figure 1) and validating merging tools (Figure 2), which collate and compare dependency annotations, POS tags, sentence types, discourse parses and more (cf. Dickinson & Meurers 2003 on catching errors in individual layers).

For example, our tools:
- Verify closed class configurations, e.g. ‘mwe’ annotations linking unlisted multiword expressions
- Check entity-type identity across coreference links
- Compare sentence and phrase borders in discourse parses, dependency parses and sentence type annotation
- Rule out implausible combinations, e.g. imperatives cannot dominate a subject function

We evaluate using an older corpus version created without these facilities. Results show that despite careful manual curation, per 10,000 tokens the merging tools catch an additional:
- 20 tagging errors
- 8 lemmatization errors
- 34 dependency errors
- 6 sentence type/border errors
- 11 coreference/entity errors
- 3 discourse parsing errors
These include errors preventing valid merging of multilayer data from different annotation tools and, are vital to maintaining high corpus quality.

References


Picoral, Adriana - University of Arizona - Compilation and automated annotation of a Portuguese learner corpus: Challenges and lessons learned

Tools and methods

Learner produced language data can be used to investigate second language acquisition (SLA) and develop language classroom materials that target the specific characteristics of different teaching settings (Mäkinen & Hiltunen, 2016). Although the total number of learner corpora available to teachers and researchers around the world has increased greatly in the past 30 years (López-Couso et al., 2016), “the number of learner corpora for Brazilian Portuguese is still limited and this number is even lower if we consider a corpus which is specific for Spanish speakers” (Torres, Rodrigues & Aluíso, 2014, p. 100). There is also few resources for automated annotation developed specifically for Portuguese.

This study describes the compilation of a learner corpus (about 500,000 tokens) based on classroom assignments from a Portuguese language program collected longitudinally at a university with a large Spanish-English bilingual population (L1 Spanish/L2 English, L1 English/L2 Spanish, and L1 Spanish/L1 English). The corpus has been annotated for both part of speech (POS) and dependency labels. For POS tagging, the HunPos tagger (Halácsy, Kornai & Oravecz, 2007) was used with two different training data sets: 1) the Floresta Sintática corpus (Freitas, Rocha & Bick, 2008), a publicly available Treebank for Portuguese; and 2) the Brazilian Portuguese UD, a dataset converted from the Google Universal Dependency treebanks v2.0 (McDonald, Nivre & Zeman, 2013). Accuracy rates for universal POS are then compared with two vanilla bigram implementations of Markov Model (MM). For the dependency parsing, MaltParser (Nivre & Hall, 2005) and the dependency parser developed by the Computational Language Understanding (CLU) Lab at University of Arizona were used. Accuracy rates range between 84.47% and 96.43%, and statistical significance for accuracy comparisons was calculated using random sampling with replacement (Berg-Kirkpatrick, Burket & Klein, 2012). A linguistic analysis of the most common tagging and parsing errors follows.

Puga, Karin - Justus Liebig University Giessen - Using prosodically annotated corpora for language teaching and learning

Applications

L2 learners of English have been described to differ significantly from native speakers in their perception (cf. Mok et al. 2016; Puga et al. 2017) as well as production of English prosody (cf. Gut 2009). This matter is complicated by a general neglect of prosody in English Language Teaching (ELT) (see e.g. Derwing et al. 2013; Gass & Selinker 2001; Hirschfeld & Trouvain 2007). Teachers (especially nonnative teachers) often feel insecure about how to detect and correct deviant L2 prosody (Hirschfeld & Trouvain 2007). Intonation training has been shown to exert an even greater influence on the acquisition of tonal patterns than stays abroad in English speaking countries (e.g. Grosser 1997; Gut 2009) and should therefore be considered in ELT.

The present paper will show how prosodically annotated corpora can be incorporated in ELT. It will do so by first showing the results of corpus-based analyses (e.g. Gut 2009; Mennen 2004) in which common deviations of advanced learners were identified, including peak alignment and the replacement of rises with falls and vice versa. It will then evaluative existing prosodic corpora for their appropriateness for use in ELT. Then, based on this background knowledge, it will show the potential for using self-collected local learner corpora and suggest different types of paper- and computer-based exercises for intonational phrasing and pitch range. These exercises are based on existing and freely available prosodically annotated corpora (e.g. LeaP corpus; Gut 2010) and self-annotated corpora (e.g. LINDESI; Gilquin et al. 2010). Thus, the present study demonstrates how prosodically annotated corpora can be used together with speech analysis software like Praat (Boersma & Weenink 2017) to teach intonational phrasing and the use of intonation in different speaking styles, languages (German & English), and varieties of English (native vs. learner language) to advanced learners of English in order to raise awareness of the differences and deviations of L2 prosody.

References


Quinley, Jason - University of Tuebingen - Modal Verbs in Requests and Proposals: Game-Theoretic Modeling Meets Corpus Data

Linguistic analyses of corpora

We explore the rational basis for differentiating requests from proposals with modal verbs, using data from the Stanford Politeness Corpus[Danescu-Niculescu-Mizil et al., 2013] and the online Politeness Classifier tool. For instance, requests seen in the corpus include:

- Yeah, it’s just search, sadly. Any chance I could get an accept for that? :)
- Could you suggest me a different way to reach the result? Another rom, or a market app?

Requests like these employ modals invoking possible-world semantics [Brown and Levinson, 1987], yet a proposal like Would you marry me? seems infelicitous compared to Will you marry me?. Insights from game-theoretic pragmatics Franke [2009], Burnett [2016], the economics of signaling, and social contracts over public goods [Skyrms, 2004, Huttegger and Zollman, 2011] give us several rational bases for this choice.

The first is that requests encode a cooperative dilemma, much like the Prisoner’s Dilemma, whereas proposals involve switching between stable arrangements within a social contract, much like the Stag Hunt, Rousseau’s original metaphor for risky but rewarding communal efforts. As requests do not benefit the hearer in the short-term, there is a face-threat involved in asking for one’s preference or commitment to fulfilling the request. As proposals can benefit the hearer, one can ask for preference or commitment, assuming common knowledge of preferences, a second reason for would creating a non-binding proposal, or weak commitment (Table 1).

Sample Game Tables

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>a,a</td>
<td>0.4</td>
</tr>
<tr>
<td>D</td>
<td>4.0</td>
<td>2.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>5.5</td>
<td>0.4</td>
</tr>
<tr>
<td>D</td>
<td>4.0</td>
<td>2.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>III</td>
<td>3.3</td>
<td>0.4</td>
</tr>
<tr>
<td>D</td>
<td>4.0</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Table 1: A trio of public goods games Stag Hunt and Prisoner’s Dilemma depending on a parameter a. We have a Stag Hunt if a > 4(II), and we have a Prisoner’s Dilemma if a < 4(III). A proposal with common knowledge of preferences revealed can reveal which game the agents are playing.

Cognitive, Evolutionary, and Epistemic Approaches

A third thread of inquiry examines mechanisms altering the standard decision functions used in game theory. This includes sympathetic payoffs[Sally, 2003] and novel relational heuristics based on Kahneman et al. [1991], Fiske [1992]. This also includes some evolutionary stability and probabilistic modeling Figure 1. This line

AACL Atlanta 2018

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>W</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>5.5</td>
<td>4.5</td>
<td>0.4</td>
</tr>
<tr>
<td>W</td>
<td>5.4</td>
<td>4.4</td>
<td>2.4</td>
</tr>
<tr>
<td>H</td>
<td>4.0</td>
<td>4.2</td>
<td>2.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>W</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stag(s)</td>
<td>A</td>
<td>W</td>
<td>R</td>
</tr>
<tr>
<td>A</td>
<td>5.5</td>
<td>4.5</td>
<td>2.2</td>
</tr>
<tr>
<td>W</td>
<td>4.5</td>
<td>4.5</td>
<td>3.3</td>
</tr>
<tr>
<td>R</td>
<td>2.2</td>
<td>3.3</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Table 2: Weak commitment adds a strategy that takes advantage of the other player’s cooperation like the Prisoner’s Dilemma but minimizes risk. Adding sympathy s > 0 to the Stag Hunt with weak commitment removes non-strict Nash equilibria (boxed). Here s = .5 for readability.
allows us to trace the history of the English auxiliary will as a proxy for want in parallel to French and German constructions of proposals. I.e. knowing the preferences of our partner can create common knowledge of their equilibrium strategy from a game-theoretic perspective, just as knowing what someone wants can provide pragmatic inference into what they will commit to do. We invite insights from those working with historical corpora in this formative work.

Figure 1: Uncertainty over Stag Hunt or Prisoner’s Dilemma.

References


Ren, Haoshan and Crossley, Scott - Georgia State University - Psycholinguistic Features and Lexical Sophistication in TOEFL Writings: The Effect of Proficiency and L1 Distance

Linguistic analyses of corpora, Applications

The Test of English as a Foreign Language (TOEFL) is widely used to measure English language ability of non-native speakers in various linguistic areas. Among many predictors that are strongly associated with TOEFL writing scores, lexical features generally explain the greatest amount of variance (Guo, Crossley, & McNamara., 2013; Kyle, & Crossley, 2016). While previous research has shown that language background (i.e., first language) and language proficiency level can interact with lexical features, the connection between the linguistic distance between a writer’s L1 and English and psycholinguistic lexical features of L2 writing across proficiency levels in the TOEFL has yet to be investigated.

The current study utilizes the publicly available corpus of non-native English writing TOEFL11 to investigate how psycholinguistic properties of L2 learners’ word use vary across proficiency levels and across language backgrounds (linguistic distance). The TOEFL 11 comprise 12,000 TOEFL essays written on 8 number of prompts. First language and language proficiency levels for writers are available in the corpus. For this study, five psycholinguistic features were examined: concreteness, contextual distinctiveness, word processing effects, neighborhood effects, and word frequency. Linguistic distance was operationalized using a scalar measurement introduced by Chiswick and Miller (2008). Linear mixed effect (LME) models were constructed to measure the effects of language distance and writing proficiency on the five lexical features. Prompt was entered as a random effect. Significant effects for all five lexical features were found for proficiency level and language distance, indicating that both language distance and writing proficiency were predictive of lexical sophistication. Results indicated that higher proficiency writers and writers with an L1 closer to English produced more sophisticated language. No interactions were reported. Implications for lexical acquisition, testing and classroom pedagogy will be discussed.

Reppen, Randi, Northern Arizona University and Olson, Shannon, Laureate Education, Inc - A look at lexical bundles across disciplines: Consistency and variability

Applications

Although known by many different labels (e.g., n-grams, clusters, formulaic expressions, lexical bundles), these ‘pre-packaged’ language units are considered to be useful building blocks of language. This presentation examines over 700 four word lexical bundles (Biber et al 1999) in a corpus of over 25 million words from 900 readings in introductory courses across nine disciplines (Architecture, Business, Culinary Science, Digital Arts, Fashion Design, Film, Hospitality Industry, Interior Design and Studio
The procedure is as follows: English (COCA). They were given 20 Japanese intermediate (CEFR B1) EFL learners. The participants employed Corpus of Contemporary American English (COCA). Participants’ corpus use was compared with their dictionary use along with their non-use of both. The participants included 55 Japanese intermediate (CEFR B1) EFL learners. The participants employed Corpus of Contemporary American English (COCA). They were given 20-minute instruction to search for a target word and interpret the concordance lines. They used English–Japanese dictionaries of their choice.

The procedure is as follows:

**Roberson, Audrey - Hobart & William Smith Colleges - Perpetrators or victims: A corpus-based critical discourse analysis of newspaper reportage on suicide**

**Linguistic analyses of corpora**

Suicide is the tenth cause of death nationally, claiming the lives of 44,000 Americans each year, and public health officials suggest that societal stigma might prevent those at risk from seeking help (Center for Disease Control). The theory of critical discourse analysis (Fairclough, 2001) argues that mass media can disseminate and reinforce the dominant ideology around power-laden issues; in practice, nonprofits have disseminated journalistic guidelines advising against sensational or simplistic reporting of suicide (American Foundation for Suicide Prevention). It appears, however, that no corpus analysis of the reporting of suicide has been published, although suicide notes have been considered as genre (Samraj and Gawron, 2015) and suicide prevention chats have been analyzed (Cooren et al., 2017).

This study is a corpus-based critical discourse analysis of an approximately 220,000 word specialized corpus, comprised of 300 recent news articles about suicide from three national papers representing a range of political orientations. Using WordSmith Tools, a keyword search for “suicide” and related lemmas was conducted. Next, 10 words on either side of the keyword occurrences were analyzed for frequent collocations, from bigrams to 5-grams. Semantic prosody analysis (Sinclair, 1991) of the representation of suicide victims was also conducted for the most frequent collocations.

Results suggest that the recommended shift toward more responsible and nuanced reporting of suicide has not yet become journalistic practice. For example, “commit”, which is semantically associated with acts of crime or sin, was the most frequent verb collocate of “suicide”, despite recommendations to use the phrase “death by suicide”. In addition, suicide victims were often described as troubled or impulsive, suggesting they acted from moral failure or in response to a single event. Results could help suicide prevention advocates shift media coverage of suicide from a discourse of criminality to one of public health.

**References**


**Satake, Yoshiho - Surugadai University - The effects of corpus use on error identification in L2 writing**

**Applications**

The strengths of corpora in language learning have been reported (Flowerdew, 2010); however, error identification in data-driven learning (DDL) settings has only been explored in a few studies despite that appropriate error identification is necessary for error correction. This study examined the effects of corpus use on error identification in L2 writing.

Participants’ corpus use was compared with their dictionary use along with their non-use of both. The participants included 55 Japanese intermediate (CEFR B1) EFL learners. The participants employed Corpus of Contemporary American English (COCA). They were given 20-minute instruction to search for a target word and interpret the concordance lines. They used English–Japanese dictionaries of their choice.

The procedure is as follows:
1. Timed essay task (25 minutes). The participants wrote an essay on a topic given by the author without consulting COCA and/or dictionaries.

2. Peer feedback session (10 minutes). The participants exchanged essays and highlighted their classmates’ errors. They were instructed to consult COCA and dictionaries at least once each. The above procedure was repeated every two weeks on four and five occasions in 2014 and 2015, respectively.

3. The author collected participants’ essays, made error-annotated corpora and analyzed the effects of both types of references and the lack thereof.

The results revealed that the participants considered COCA more useful for error identification than dictionaries because its easy access to the target phrases and frequency information helped. The corpus was particularly effective for identifying collocational errors, such as errors of omission and addition, and form-related errors, such as those of number and agreement; inductive inference based on example sentences of target phrases worked well. The findings suggest that effective corpus use for error identification requires teachers to consider error types. Fine adjustments of DDL for accurate and appropriate error identification would improve learners’ L2 writing.

References


Schnur, Erin, Hacking, Jane and Rubio, Fernando - The University of Utah - MuSSeL: Designing and building a corpus of multilingual second language speech

Tools and methods

In this presentation, we will describe the ongoing process of designing and compiling the Multilingual Spoken Second Language (MuSSeL) learner corpus. When completed, the corpus will be freely-available on-line, and contain approximately 18,000 spoken texts produced by 1,800 second language (L2) learners of six languages, spanning multiple proficiency levels, ages, and contexts of learning. While extant learner corpora cover a wide range of languages, learner characteristics, and text types, few large-scale, freely-available multilingual corpora of learner language currently exist. Additionally, the majority of existing learner corpora contain written rather than spoken texts and represent the language production of learners at higher proficiency levels (Gablasova, Brezina, & McEnery, 2017; Gilquin & Granger, 2015; Granger, 2002). The MuSSeL project addresses these gaps. The corpus draws from professionally-rated oral proficiency exam data and comprises samples from 300 L2 speakers of each of six languages: Mandarin, French, German, Portuguese, Russian and Spanish. Texts come from learners in three contexts of learning: 3rd and 5th grade students enrolled in Utah’s Dual Language Immersion Program, adult classroom learners, and adults who have acquired their L2 through in-country immersion. Files are transcribed into both a simple (.txt) version and a version using the .cha annotation conventions (developed for users of the Child Language Exchange Data System, MacWhinney & Snow, 1984). We will discuss the specific challenges of corpus design, data collection, and transcription that arose in creating a pilot version of the corpus, and how these challenges have impacted the on-going corpus development process. In addition, we will present the results of a study using the pilot corpus to investigate lexical diversity and complexity in novice and intermediate Mandarin, French, and Spanish learners to demonstrate how the corpus can be used to investigate learner language using multiple variables.


Serigos, Jacqueline - George Mason University - ¿Español or English? An automated tool for identifying Anglicisms in Spanish

Tools and methods

The spread of global English has increased presence of Anglicisms in numerous languages around the globe. While Anglicisms are seemingly ubiquitous in that countless languages have adopted them into their lexicons, Anglicisms often appear quite infrequently within a text or dialogue (rates as low as one token per 100,000 words are common). Thus contact linguists studying Anglicisms often require corpora with word counts in tens of millions or more in order to get a sizeable sample. With such large datasets, manual methods frequently employed within contact linguistics are rendered insufficient and automated methods are needed. This paper presents one such method, an algorithm that detects Anglicisms within a Spanish text. Building on the previous language classifiers (Alex, 2008; Andersen, 2005; Leidig, Schlippe, & Schultz, 2014; Mansikkaniemi & Kurimo, 2012), the algorithm uses a combination of n-gram and look-up methods to distinguish English from Spanish tokens and adds layers of annotation to address the role of Named Entities (NEs), lemmas, and loan phrases. Several parameters of the classifier were tested to optimize the performance, resulting in an F-score of 79.41% on the training data and 76.25% on unseen test data. It outperforms several recent models in the literature (Andersen, 2005; Leidig et al., 2014). While it is outperformed by Alex (2008), the model presented here is less computationally taxing and doesn’t suffer from internet search limits, which can present challenges when processing large datasets. This model addresses the important role of NEs and loan phrases, which have been so far ignored by previous anglicism identifiers. NEs can be particularly problematic for newspaper corpora, often used in anglicism studies, in that they can lead to an over-identification of loanwords.

References


Silva, Barbara - São Paulo Catholic University - Cultural representations of age groups in Google Books.

Linguistic analyses of corpora

The goal of the current study is to identify cultural representations of two main distinct age groups, namely teenagers and the elderly, through the examination of the patterns of usage of the following words: ‘teen’, ‘teenager’, ‘elder’ and ‘elderly’. In order to achieve that, the data set from Google books covering the period from 1800 through 2008 was queried via the BYU online interface to Google Books. The method employed in this project is based on the Multidimensional (MD) framework for register variation introduced by Biber (1988). Following this analytical framework, the adjective collocates of each word were identified and analyzed through the application of multivariate statistics, which grouped the words for each age category according to their patterns of use. The groups were interpreted as the underlying parameters of cultural representation that helps frame the way each age group is represented historically in Google Books since the 1800s. The detection of such representations from linguistic units requires very large quantity of texts in order to draw evidence of significant indications of their weight on the language. The goal of the current study is to identify cultural representations of two main distinct age groups, namely teenagers and the elderly, through the examination of the patterns of usage of the following words: ‘teen’, ‘teenager’, ‘elder’ and ‘elderly’. In order to achieve that, the data set from Google books covering the period from 1800 through 2008 was queried via the BYU online interface to Google Books. The method employed in this project is based on the Multidimensional (MD) framework for register variation introduced by Biber (1988). Following this analytical framework, the adjective collocates of each word were identified and analyzed through the application of multivariate statistics, which grouped the words for each age category according to their patterns of use. The groups were interpreted as the underlying parameters of cultural representation that helps frame the way each age group is represented historically in Google Books since the 1800s. The detection of such representations from linguistic units requires very large quantity of texts in order to draw evidence of significant indications of their weight on the language. The detection of such representations from linguistic units requires very large quantity of texts in order to draw evidence of significant indications of their weight on the language.

Skalicky, Stephen and Crossley, Scott - Georgia State University - Examining creativity in the Japanese Learners of English corpus

Linguistic analyses of corpora

The use of learner corpora has allowed researchers to better understand second language (L2) development from a variety of perspectives, such as learners’ use of constructions, lexical development, and employment of linguistic creativity (e.g., humor and teasing) over time (Bell et al., 2014; Crossley et al., 2010; Römer et al., 2014). However, no learner corpus research has investigated potential connections between creative ability (i.e., ability to provide effective and original solutions to problems; Runco & Jaeger, 2012) and L2 proficiency.
The purpose of this study is to investigate perceptions of creativity in a learner corpus in order to determine if creative ability in L2 output is associated with increased L2 proficiency. We analyzed picture description narration tasks from the Japanese Learners of English (JLE) corpus (Izumi et al., 2003, 2004), which is comprised of transcribed oral English proficiency interviews and includes proficiency ratings for the learners based on their oral performance. We selected a random sample of 466 picture description tasks of at least 50 words in length for analysis, with proficiency levels ranging from 2 (the second lowest) to 9 (the highest). The picture description tasks were rated for creativity by human raters using an analytic rubric designed to measure creativity as a function of idea generation, elaboration, originality, and appropriateness.

The results reported a significant, positive correlation between English proficiency and perceptions of creativity ($\rho = .450$), suggesting that increased English proficiency (as operationalized by JLE oral proficiency scores) is associated with increased perceptions of creativity. Results also indicated that a greater number of words resulted in higher creativity scores for learners at lower proficiency levels but not for those at higher proficiency levels, suggesting an interaction between quantity and quality of learner output and perceptions of creativity. Our findings further highlight connections between creativity and L2 development.

Smith, Jordan - Iowa State University - Building a Representative Corpus of Blogs: Methods, Challenges, and Constraints

Tools and methods

In their pioneering work identifying internet registers, Biber and Egbert (2016) and Biber, Egbert, and Davies (2015) note different communicative functions that blogs perform (e.g., narrate, describe, or express opinion). Because of this variation, blogs in general do not constitute a formal register but instead are simply “a means of publishing writing online” (Egbert, personal communication, January 19, 2018). As a result, additional work needs to be done in order to better understand how the situational variation among blogs influences the linguistic variation found in blogs.

Many of the tools that facilitated previous research of blogs such as directories (e.g., globeofblogs.com), ranking tools (e.g., Technorati) and blog-specific search engines (e.g., Google’s Blog Search) have either radically changed their functionality or are now defunct, further complicating the already-challenging task of compiling a representative corpus of blog texts. In this session, I discuss the criteria I used for building a corpus of blogs that represents informal written American English and the constraints and limitations I faced in building the corpus. Specific topics I discuss include

- Determining which search engine (or combination of search engines) is best to use in order to collect blog URLs to include in the corpus
- Making use of advanced search features such as site-restricted searches (Grieve, Asnaghi & Ruette, 2015) to help ensure the content returned by the search engines is usable
- Using Amazon’s Mechanical Turk crowd-sourcing network (see Biber, Egbert & Davies, 2015) to identify situational characteristics of a sample of the blog posts included in the corpus

I conclude my presentation by discussing the ways I plan to make use of this corpus for future research, particularly in investigating the relationship between prescriptive language ideologies and informal online writing.

References


Linguistic analyses of corpora

This paper investigates EFFECT-related nouns in self-constructed database comprising English and Chinese academic abstracts. The Chinese and English abstracts are mostly parallel for both languages are required in all theses written in English. Three high frequency nouns, i.e. effect, result and impact, used to shell the concept of an outcome were selected. According to Schmid (2000), these nouns share the semantic features of [FACTUAL] and [CAUSAL], and are categorized under the ‘Result’ family that highlights “the EFFECT component of causal relations” (pp. 106).
Instances containing effect, result and impact were retrieved as well as the equivalent terms in the Chinese abstracts. Text alignment was carried out manually, see Figure 1.

<table>
<thead>
<tr>
<th>Items</th>
<th>English</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCCU9013</td>
<td>(3) the <strong>impact</strong> of long-term <strong>effects</strong>; the discovery of human capital, wage income, government spending on these three variables the number of small and medium-sized start-up companies have a negative <strong>impact</strong>, in which to pay income in the three most affected</td>
<td>(3) 在長期的 <strong>影響</strong> 下，發現人力資本、薪資所得、政府支出這三個變數對中小企業創業公司家數均有負面的 <strong>影響</strong>，而在此三項中以薪資所得的影響最大</td>
</tr>
<tr>
<td>TKU9001</td>
<td>The <strong>results</strong> show the relationship of Sulfur dioxide (SO2) and per capita income in inverse-U type, while Waste gas in not.</td>
<td>実際結果顯示，二氧化硫(SO2)和人均所得之間存在倒“U”形的關係，在工業廢氣方面，則不存在EKC的現象。</td>
</tr>
<tr>
<td>NCKU5007</td>
<td>In the estimation of spillover <strong>effect</strong>, the marginal externality in demand-following model is far greater than the supply-leading model.</td>
<td>在估算模型之外溢效果時，需求引領模型的邊際外部性遠大於供給領導模型的邊際外部性。</td>
</tr>
<tr>
<td>NCCU9009</td>
<td>The empirical <strong>results</strong> showed the two futures markets both have long memory.</td>
<td>實證結果顯示，這兩個期貨市場長期的波動皆具有長期記憶。</td>
</tr>
<tr>
<td>NCNU9025</td>
<td>It is expected to enforced the consistency of <strong>result</strong> through the two different databases.</td>
<td>期望能通過兩種不同資料庫，加強實證結果的一致性。</td>
</tr>
<tr>
<td>FIU9002</td>
<td>Main <strong>results</strong> are as follows: (1) the impulse response function pattern using GARCH-SVAR model can be explained by interest rate parity theory, Purchase parity theory and portfolio approach.</td>
<td>由實證結果發現，SVAR-GARCH模型下的衝擊反應函數，可以解釋利率平價理論、購買力平價理論與匯率資產分析法的成立。</td>
</tr>
<tr>
<td>NTU9002</td>
<td>3. Government monetary policy <strong>effect</strong> and crowding out <strong>effect</strong>.</td>
<td>3. 貨幣政策 <strong>影響</strong> 與金融排擠 <strong>影響</strong>。</td>
</tr>
</tbody>
</table>

Figure 1: Examples of the aligned data

A total of 77 instances taken from 28 abstracts, constituted by effect 44.1%, result 42.9%, and impact 13%. With reference to collocations, result expressed neutral outcomes; effects conveyed preferences when highlighting the changing of event; while impact referred more to negative events. Compounds elaborating influences such as ‘knock-on effect’, ‘spill-over effect’ were commonly found. Frequent collocations of result included ‘empirical results’, ‘surprising results’, etc. Polarity modifiers such as ‘positive/negative’ and ‘direct/indirect’ are mostly followed by impact and effect.

From the corresponding Chinese instances, 55.9% of effect were translated consistently as ‘效果’xiàoguǒ while the remaining varied to a great extent; 66.7% of result were translated as ‘結果’jiéguǒ, and 40% of impact were translated as ‘影響’yǐngxiǎng.

We then used the Chinese equivalents to make new queries to see if these terms also match other English terms. About 24% of the total instances used different English terms. We also found 37% of ‘影響’yǐngxiǎng in verb form, translated as ‘to affect’, ‘to influence’. In addition to the above, the result showed a consistency usage referring to ‘outcome’.

References


Cambridge Dictionary https://dictionary.cambridge.org
Staples, Shelley, University of Arizona, Reppen, Randi, Northern Arizona University, Holcomb, Chris, University of South Carolina, Liu, Qiandi, University of South Carolina, Novikov, Aleksey, University of Arizona, Picoral, Adriana University of Arizona and Lan, Ge, Purdue University - Lexico-grammatical Patterns in First Year Writing across L1 Backgrounds  

Linguistic analyses of corpora

While an increasing number of studies examine L1 and L2 writing in English at the university level, fewer studies have examined writing within the context of writing classes, particularly undergraduate first year writing (FYW) courses (Leki, Cumming, & Silva, 2008). Notable exceptions include Lee et al. (2017) and Staples and Reppen (2016). Adding to this existing research, this study investigates one of the most common genres in the First Year Writing classroom, research-based argumentative essays, to determine differences in the use of lexico-grammatical patterns across L1 backgrounds. Utilizing a corpus of FYW across four institutions (over two million words in total), we created a balanced sample of argumentative essays carefully selected to evenly represent writing across five L1s, English, Chinese, Arabic, Korean, and Spanish, that is similar in situational characteristics (such as use of sources and the expression of an overt argument). We quantitatively and qualitatively investigate a selection of complexity and stance features, including both phrasal (pre-modifying nouns, attributive adjectives) and clausal (relative clauses, verb and noun complement clauses, finite adverbial clauses) features. We also identify lexico-grammatical patterns associated with these grammatical forms, providing insight into the variability of the writers’ phraseology. In addition, our qualitative analysis examines how students’ use these lexico-grammatical patterns for rhetorical purposes such as creating arguments, packaging information, expressing stance, and providing cohesion within their essays. Finally, we briefly investigate a subset of the texts for variation across institutions to explore the impact of assignment goals within different writing programs. Preliminary findings indicate that L1 English writers used significantly more relative clauses, that relative clauses, and finite adverbial clauses, when compared with L2 English writers. This study has implications for integrating language awareness into both L1 English and L2 English FYW classrooms as well as for EAP instruction in other contexts.

Stewart, Ian, Pinter, Yuval and Eisenstein, Jacob - Georgia Institute of Technology - Catalanian independence and linguistic identity on social media  

Linguistic analyses of corpora

Political identity is often manifested in language variation, such as the choice to use a local language over a supraregional language (Woolard 1985). However, the relationship between the two is still relatively unexplored from a quantitative perspective. This study examines the use of Catalan, a language local to the semi-autonomous region of Catalonia in Spain, versus the use of Spanish on Twitter in discourse related to the 2017 independence referendum. Following a prior study which focused on the Scottish referendum in 2014 (Shoemark et al. 2017), we mine tweets posted one year prior to the Catalanon referendum and identify Twitter users who support or oppose independence based on the hashtags they use in their tweets. The language used in the tweets is identified automatically using the langid algorithm (Lui & Baldwin 2012).

With respect to political stance, we find first that Twitter users who support independence are more likely to express themselves in Catalan than anti-independence users. This corroborates the prior study in the Scottish referendum that shows a similar correlation. Contrary to that study, however, we find that Twitter users are more likely to use Catalan in referendum-related discourse than in other contexts. We also find that Twitter users are less likely to use Catalan in messages aimed at a smaller audience, which seems counterintuitive. This suggests a strong role for the Catalan language in the expression of Catalan political identity, with different mechanisms of audience design acting than those in the Scottish scenario. This study also calls for further investigation into the relationship between political and linguistic identity, particularly in societies such as Spain where local languages can serve as a symbol for political movements (Moreno, Arriba & Serrano 1998).

References


Stratton, James – Purdue - A Diachronic Analysis of the Adjective Intensifier well from Early Modern English to Present Day English: A Corpus-Based Study  

Linguistic analyses of corpora

The adverbial well can modify participial adjectives and a limited number of regular adjectives such as well aware and well capable in most Standard varieties of English (Stenström 2000: 177). However, in British English well can also modify regular adjectives such as it’s well good and that was well boring. This use as an adjective intensifier has been attested throughout the history of the English language but previous scholarship suggests that its frequency declined after the mid-14th century.
Linguistics, 7, 140

Röm Corpus Linguistics, 16(1), 72


Suethanapornkul, Sakol, Georgetown University and Supasiraprapa, Sarut, National Institute of Development Administration - Second Language Users’ Knowledge of the Introductory-IT Construction: Corpus and Psycholinguistic Evidence

Linguistic analyses of corpora

The introductory-IT construction is prevalent in written academic discourse (Biber et al., 1999; Peacock, 2011); however, it is problematic for L2 writers, who seem to underuse it in both frequency and functions (Römer, 2009). One difficulty could stem from selective appearances of the adjectives in two nearly-synonymous variants: the adj-that frame (e.g., ‘it is [clear/obvious/etc] that’) and the adj-to frame (‘(it is [difficult/impossible/etc] to’). Drawing both on corpus and psycholinguistic evidence, we report on the specific lexical items selected by each variant and advanced L2 writers’ usage of adjectival collexemes.

In our corpus analysis, we employed a distinctive collexeme analysis (Gries & Stefanowitsch, 2004) to test the strength of associations between the variants and adjectival collexemes as attested in COCA. For example, clear and difficult were the most distinctive collexemes of adj-that and adj-to variants (collostructional strength = 514.501 and 218.675). Each adjective appeared only in one frame, providing a strong cue (e.g., ΔP (construction|word) = 0.702 for clear). Both frames exhibited a Zipfian distribution with the three most frequent adjectives accounting for ≈50% of the data (Hrel = 0.629 and 0.612).

In our psycholinguistic experiment, we utilized a timed free-association task (Ellis et al., 2014) to probe production of adjectives in both variants in 23 international graduate students at US universities. Participants’ production demonstrated variable knowledge of the occupancy of adjectives in the two frames of the construction. For example, clear was extended to the adj-to frame (e.g., ΔP (construction|word) of ‘clear’ = 0.207). Participants’ amount of English moderately correlated with target-like use of adjectival collexemes.

The study provides new perspectives on factors influencing the usage of adjectival collexemes and combines corpus and production evidence to reveal difficulties inherent in L2 writers’ usage of this highly productive construction.

References


Tasker, David - Northern Arizona University - Exploring Dimensions of Complexity in English through Multidimensional Analysis

Linguistic analyses of corpora

While linguistic complexity is a multicomponential construct (Bulté & Housen, 2014), few studies have comprehensively explored how sets of complexity features relate to each other, and what latent variables might underlie patterns of co-occurrence. This study examined the functional dimensions underlying linguistic complexity features from the Biber, Gray, and Poonnpon (2011) developmental framework across spoken and written registers of English. 18 lexico-syntactic features from the framework representing several clause- and phrase-based structures were investigated in a balanced corpus consisting of 179 texts each from face-to-face conversation and academic journal articles (approximately two million words). Complexity features were retrieved programatically from Biber-tagged texts. Multidimensional analysis (Biber, 1988) revealed two dimensions accounting for approximately half of the total variance in feature appearance. The first dimension revealed both positively and negatively loading features interpreted as follows: Highly Developed Noun Phrase Compression features used to convey technical information, and negatively loading Less Developed Clausal Elaboration features used to convey and situate stance expression. The second dimension consisted of a single set of features interpreted as Intermediate-High Development, Noun- and Adjective-based Elaboration structures used to construct abstract arguments. Factor loadings of face-to-face conversation texts exhibited a relatively narrow degree of variation in the appearance of these complexity features, while academic journal articles exhibited regular variation by discipline: Physics texts scored highest on the first dimension and lowest on the second; Philosophy texts scored highest on the second dimension and lowest on the first; and Applied Linguistics and Political Science texts were more balanced. Findings suggest that functional dimensions underlie co-occurrence patterns of complexity features, that the nature of linguistic complexity may vary by register, and that face-to-face conversation is not characterized by the same variation in the use of these linguistic complexity features as academic texts.

References


Tywnwar, Rurik - Georgia State University - Investigating Tense-Aspect Constructions and the Aspect Hypothesis using Collostructional Analysis

Tools and methods, Linguistic analyses of corpora, Applications

The tense-aspect (TA) system of English is multidimensional and nuanced, making it complicated for language learners to approach. Functional approaches to SLA and constructionist theories of language agree that grammatical form, lexical information, and discursive functions are tightly interwoven. For TA-pairings (e.g. “present perfect”), one hypothesis explaining how TA is tied to lexis and theoretically distributed verb usage is the Aspect Hypothesis, which states that the inherent lexical aspects of verb arguments (e.g. “accomplishments”, “actions”, etc.) influence the choice of TA use beyond structural rules.

Previous studies analyzed association between verb-TA constructions and lexical aspect in narratives, finding support for the aspect hypothesis. The current research extends these findings, examining verb-TA constructions in a corpus of highly-rated academic writing and verifying statistically, through Collostructional Analysis (CA), verb preferences for certain TA-constructions based on lexical aspect. This study specifically uses Multiple Distinctive Collexeme Analysis, whereby frequency data is gathered for a linguist. The first hypothesis of analysis was that TA constructions and lexical aspect are tightly interwoven. For TA-pairings (e.g. “present perfect”), one hypothesis explaining how TA is tied to lexis and theoretically distributed verb usage is the Aspect Hypothesis, which states that the inherent lexical aspects of verb arguments (e.g. “accomplishments”, “actions”, etc.) influence the choice of TA use beyond structural rules. As such, previous studies showed that verb-TA constructions are calculated using Fisher-Yates Exact Test (FYE) based on a contingency table of occurrences of each particular lexicogrammatical pair. We then compared the resulting statistical preferences of verbs to appear in collostructions with tense-aspect pairs to our understanding of lexical-aspectual information within the verbs.

We found that verb and TA-pairings exhibit prototypically, with verbs either attracting or repulsing collostructional combinations based on the verbs lexical aspect. Additionally, we found association strengths of collostructions to vary across and within verbs, suggesting that lexical-aspectual categories, such as “achievement” and “accomplishment,” are ranges of collostructional strengths as opposed to discrete categories. This research ultimately informs pedagogical practices related to TA instruction, English reading and writing instruction, and theoretical understanding of tense-aspect. Numerous limitations and avenues for future research will be described.

Veirano Pinto, Marcia, Federal University of Sao Paulo (UNIFESP) and Cortes, Viviana, Georgia State University - Lexical bundles across North American movie genres: Discursive functions and levels of routinization

Linguistic analyses of corpora

North American movie genres have been extensively studied from a semiotic and cultural studies perspective (Altman 2009; Julier and Marie 2009; Stam 2006), but few studies have analyzed how lexico-grammatical patterns shape the verbal language of North American cinema. Veirano Pinto (2013, 2014) used Multidimensional Analysis to locate movie genres along the five major dimensions of register variation found by Biber (1988) and to identify the dimensions of variation across North American movies. In order to complement the information provided by these studies and to provide some insight into the lexical profile of
movie language, the present study focuses on the identification and analysis of lexical bundles, frequently occurring corpus-driven multi-word expressions (Biber et al., 1999).

This study tries to profile the formulaic language of the long-established cinematic genres action/adventure, comedy, drama, and horror/suspense/mystery by identifying the most frequent 4-word bundles in each genre, classifying these bundles according to their structures and discourse functions, and determining the extent of bundle variation across movie genres. The North America Movie Corpus (NAMC), a diachronic corpus covering 80 decades of movies (640 transcripts, over five million words) was used in the study. The results show that in terms of lexical bundle use, some genres seem to be more formulaic than others. Comedies, for example, used many more bundles than action/adventure movies. In addition, while some bundles performing epistemic stance (what do you think, I don’t know what) were present in all the study sub-corpora, some specific genres made more frequent use of a particular type of bundle. For example, horror/suspense/mystery movies used more imperative attitudinal stance bundles (get out of here) than the other genres. The presentation will also include a comparison of the bundles identified in these movie genres and those previously identified in other registers such as everyday conversation.

**Whitty, Lauren - Victoria University of Wellington - A corpus-based analysis of ‘can’ and its dependency on categorical criteria and expanded context**

**Tools and methods**

Through investigation of the central modal auxiliary ‘can’ in the British National Corpus (BNC), this paper examines the roles of criteria and context when performing corpus-based analyses. I first review the semantic categories utilized in previous modal auxiliary studies of ‘can’ and then present the categories established in my study, as well as the criteria for the category ‘ability’. In comparison to prior studies, this study relies on more categories than previously identified, and recognizes the function of prototypicality in assigning usage categories. Next, I consider the role of expanded context, which is twofold. Expanded context is required for understanding ‘can’ in use, as well as for reporting these uses. I demonstrate how in order to identify the categories of use for ‘can’, many instances require expanded context, which, to anyone studying ‘can’, the need for context is not a new finding. However, the extent of context required in some cases may come as a surprise. With regard to reporting analyses, too often, linguists present one sentence (or shorter) extracts as examples. For instance, in Biber et al. (1999), “He goes, I can’t swim” (p. 492) is given as an example of ‘ability’ but when searching “I can’t swim” in the BNC, there are additional readings, such as ‘possibility’. This paper recognizes the importance of: (a) making usage criteria as explicit as possible; (b) expanded context in understanding ‘can’ in use; and (c) expanded context in reporting analyses of instances. I show how a strong corpus-based analysis of ‘can’ is dependent upon these three factors, factors which are also applicable to other lexical items under examination using corpus-based methods.

**Wright, Heidi - Carnegie Mellon University - A Multidimensional Analysis of Stand-Alone Literature Reviews: Six Dimensions, Old and New**

**Linguistic analyses of corpora**

Since its inception (Biber, 1986, 1988), multidimensional (factor) analysis has offered a way to examine how sets of grammatical features in texts function together to help define different genres and registers. Political speeches and fiction, for example, differ in the quantity and exact set of narrative features they include (Biber, 1988). To date, multidimensional studies have been conducted on many academic genres including student writing (Nesi & Gardner, 2012), research articles (Gray, 2011), and textbooks (Egbert, 2014). However, other respected academic genres such as the stand-alone literature review (Swales, 2004) have not yet been analyzed, and the differences between reviews across time, discipline, and review type (i.e., qualitative, quantitative, peer reviewed, or invited) have not been catalogued. To address this gap, I conducted a multidimensional analysis of stand-alone literature reviews from three time periods and three disciplines. A corpus of 417 stand-alone literature reviews (e.g., Annual Review of Medicine) in medicine, psychology, and education was compiled from the years 1950, 1980, and 2010. Reviews were categorized as invited or peer-reviewed, qualitative or quantitative, having methods present or absent. The corpus was grammatically tagged, counted, and normed (Biber, 1993). A factor analysis of 72 of the grammatical features was then conducted. The analysis found a six factor (i.e., dimension) solution consisting of a technical/ academic vs. human focus, explicit vs. situation dependent reference, expression of persuasion, author vs. process focus, abstract vs. concrete focus, and methodological concerns. Z-scores for each text were computed. Mean scores for groups of texts (e.g., qualitative) were mapped onto each dimension. Findings revealed that some review type sub-groups (e.g., invited vs. peer reviewed) had similar scores. However, other sub-type divisions (e.g., qualitative vs. quantitative) showed clear distinctions between mean scores on more than one dimension as suggested by Gray (2011).

**Wulff, Stefanie, University of Florida and Gries, Stefan Th., UCSB - Particle placement in learner English: Measuring effects of context, first language, and individual variation**

**Linguistic analyses of corpora**

This study applies a multi-step regression method (Gries & Wulff 2013; Lester to appear; Wulff & Gries 2015; Wulff, Gries & Lester 2018) to examine the factors governing the learners’ placement of the particle in transitive particle verbs (The squirrel picked up the nut | The squirrel picked the nut up). This alternation has been extensively studied in native speaker production, yet has received comparatively little attention in learner production.

We retrieved an exhaustive sample of over 4.9K attestations from the ICLE, LINDSEI, LOCNEC, and LOCNESS corpora and annotated them for a number of contextual predictors known to govern this alternation: direct object length, complexity,
concreteness, and definiteness; rhythmic and segmental characteristics; and idiomaticity. We also included learner-specific variables such as their first language (L1) background and a (dispersion-based) proxy to each learner’s L2 proficiency. First, we trained a maximally predictive regression classifier on native speaker data to predict learners’ particle placement choices given the contextual configurations exhibited in the learner data. This classifier allowed for non-linear effects as well as random effects for verbs, particles, and speakers and was also cross-validated. In a second step, we built a second classifier that checked whether and to what extent the learners had made native-like choices. We show effects plots resulting from the second classifier that visualize in which contexts learners make more and less native-like choices, and that help identify individual learners that experience more difficulty with target-like particle placement than others.

The results indicate that this alternation, like others, is largely governed by processing-related factors; measurably modulated by the learner’s L1 experience; and characterized by substantial individual variation. We close our presentation with a discussion of the implications and applications of complex statistical modeling as outlined here for understanding learner production, assessment of learner production, and custom-tailored teaching materials.

References

Yilmaz, Selahattin and Römer, Ute - Georgia State University - Phraseological Variability in English as an Academic Lingua Franca Writing
Linguistic analyses of corpora

Academia relies on international networks of research and education, which has always created a need for a common language of communication, that is, a lingua franca (Mauranen, 2010). The majority of English speakers in the international academic community today are non-native speakers of English who use the language for professional communication (Mauranen, Hynninen & Ranta, 2010). The need to publish in English for academic success also strengthens the role of English in non-English speaking countries remarkably (Flowerdew, 2008). Therefore, it is of great value to understand how English is used in these international contexts. While most ELF research has focused on spoken data, little is known about English as a Lingua Franca writing. Initial research indicates that, compared to both language learners and native speakers, spoken and written ELF discourse uses less conventional but still acceptable patterns (Carey, 2013; Keckes, 2007; Seidhofer, 2009; Tribble, 2017).

This study aims to further explore the phraseological variability in a corpus of written English as a Lingua Franca. This study uses the Scientific English as a Lingua Franca Corpus (SciELF, 2015), a sub-corpus WrELFA (2015) , the corpus of Written English as a Lingua Franca. SciELF consists of 150 unedited research articles collected from authors from 10 different first language backgrounds. A combination of n-gram and phrase-frame analysis was used to find the most commonly used continuous and discontinuous patterns using the software tools kFNgram 1.3.1 (Fletcher, 2012) and AntGram 0.0.3 (Anthony, 2017). This presentation will discuss the variability and predictability levels and discourse functions of these phrases. It will also talk about implications for future research and pedagogy.

References

AACl 2018
Academic conference proposals remain under-researched, despite their centrality in disciplinary knowledge construction and dissemination. Given that conference proposals represent a high-stakes, yet “occluded” (Swales, 1996) genre and provide novice scholars with valuable access to academic socialization (Simon-Maeda, 2016), analysis of these genre practices can provide insights for English for Academic Purposes (EAP) pedagogy. Furthermore, despite increased attention to the relationships between language and rhetoric more broadly in EAP (e.g., Cortes, 2013), existing research on conference proposals explores either broad rhetorical move structures (e.g., Halleck & Connor, 2006) or characteristic linguistic features (e.g., Simon-Maeda, 2016). Continuing a trend towards the “integration of genre analysis and corpus-based investigations” (Flowerdew, 2005, p. 5), this study explores the formulaic phrase frames (p-frames, Fletcher, 2007) that successful academic writers employ to linguistically realize their rhetorical goals in conference proposals.

Our data consists of 625 proposals accepted by the 2017 AAAL Conference. Proposals were freely available on the organization website without login. The corpus was manually tagged according to a rhetorical move framework, based on past research (Cotos, Huffman, & Link, 2017; Swales, 1990) and modified to fit our data. P-frames of 5 and 6 words (threshold of 17 per million) were extracted using KfNgram (Fletcher, 2007) and correlated with rhetorical moves. Results show that some p-frames occur across rhetorical moves/steps, (e.g., “attention has been paid to”) serving multiple functions, while others are more specialized. Often it is the variant that determines the function of the frame itself. For example, when the frame “attention has been paid to” is used with variants particular and significant, it establishes “research territory”, while variants little and scant are used to establish “niche”. Moreover, some rhetorical moves/steps (e.g., Occupying the niche) seem to be highly formulaic. The presenters will discuss implications for teaching proposal writing.

References

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. Journal of English for Academic Purposes, 12, 33-43


This paper examines genre variation from the viewpoint of discourse relations in the framework of Rhetorical Structure Theory (RST, Mann & Thompson 1988). Discourse relations, such as ‘cause’ or ‘condition’ in Figure 1, are signaled using a range of linguistic means (Das & Taboada 2017), including discourse markers, syntactic devices (e.g. relatives), semantics (e.g. coreferentiality) and genre conventions (e.g. use of headings).

While Multi-Dimensional analyses (Biber 2009) have shown the importance of lexical, morphological and syntactic cues across genres, genre studies of discourse relations remain few (see Taboada & Lavid 2003, Gruber & Huemer 2008). Using the Georgetown University Multilayer corpus (GUM, Zeldes 2017), we compare data covering 85,000 tokens/7,400 instances of 20 relations from eight text types taken from the Web: news, interviews, how-to guides, travel guides, reddit forums, and academic, biographical, and fiction writing.

We discuss and integrate two kinds of results into a Multi-Dimensional analysis: on one hand, genres vary widely in proportions and arrangements of discourse relations. On the other, ways in which the same relations are signaled depend strongly on genre. For example, we can consider the contrastive ‘contrast’, ‘antithesis’ and ‘concession’ in Figure 2.

While neutral ‘contrast’ is balanced, ‘antithesis’ (presenting an alternative to favor the first option), is used most in interviews (often polemically), and in how-to guides (‘whow’ in Figure 2) to warn against not following instructions, but less in news and travel guides. A concession, which concedes value in an opposing idea, is common in travel guides (‘voyage”), but less in news, as viable alternatives are emphasized in instructional texts, but not in news narrative. Starting from these distributions, we use machine learning methods to explore recurring words and annotations signaling the same relations in different genres.

References

Das, Debopam & Maite Taboada (2017), Signalling of Coherence Relations in Discourse, Beyond Discourse Markers. Discourse Processes, online first, 1-29.


Zuppari, Maria Carolina - Catholic University of Sao Paulo - Lexical Variation in Written Academic English

Linguistic analyses of corpora

English has become the lingua franca of academia worldwide. Content and language integrated learning and the teaching of English for specific purposes are becoming commonplace in higher education classrooms around the world. In Brazil, this movement has gained momentum with the entry of international networks in the higher education market and the offer of publicly subsidized exchange and foreign language programs. Though English has been increasingly used as a medium of instruction worldwide in different fields of knowledge, the focus of English as a Foreign Language teaching in universities has been mostly General English rather than other areas, such as health sciences, engineering, information technology and humanities. There is a wide range of studies on English for Academic Purposes, however, few have focused on collocations and their semantic implications across different fields of knowledge. This study sets out to describe, through a multidimensional analysis perspective, lexical variation of different fields of knowledge, namely health sciences, engineering and IT, business, and social sciences, by identifying the main collocations of each area in addition to their lexical parameters and the semantic associations underlying the variation present in these areas of study. The corpus being used is the CACT (Corpus of American College-Level Texts), comprised of textbooks and research articles from 5 discipline areas, according to the National Science Foundation (NSF): Biological Sciences; Computer and Information Science and Engineering; Engineering; Mathematical and Physical Sciences; and Social, Behavioral and Economic Sciences; totaling 60.4 million words across 1,150 texts. The presentation will focus on the dimensions of collocation identified for the Social, Behavioral and Economic Sciences discipline area and will briefly discuss their implications to teaching both English for Specific Purposes and English for Academic Purposes.
Becker, Amelia - Georgetown University - Corpus investigation of a selected finger markedness hierarchy in American Sign Language

Linguistic analyses of corpora

Type frequency, token frequency, and child acquisition of selected finger combinations (SFCs), a handshape feature (Mandel 1981; Brentari 1998; Sandler 1986; Caselli et al. 2016) were investigated using data from Sign Language Acquisition, Annotation, Archiving and Sharing (SLAAASH; Lillo-Martin and Pichler 2008), a corpus of adult and child American Sign Language (ASL) use. SLAAASH annotation is ongoing. The present study is a small-scale analysis to be expanded once more data are annotated. The language analyzed was captured in 11 video recorded sessions (9 hours, 10 minutes total) of conversational ASL produced by a deaf child ABY (ages 1;04.22 - 3;04.07), ABY’s parents, deaf and hearing researchers, and several other adults. All signs were coded for SFC following Caselli et al. 2016. Acquisition was operationalized as the session in which ABY first produced a lexical item which has a given SFC in citation form.

Adult SFC type and token frequency (Figure 1) and order of SFC acquisition (Figure 2) converge to suggest an SFC markedness hierarchy which is at least partially attributable to the musculoskeletal (Figure 3) and innervation (Figure 4) structures of the hands. Fingers with independent muscle attachments showed higher frequency and earlier acquisition as the only selected finger in a lexical item. This suggests a role of articulatory ease in the frequency and acquisition of SFCs. Similarly, the SFC of index and middle finger was both highly frequent and acquired early. These fingers, along with the thumb, share innervation by the median nerve. An SFC of ring finger alone was the least frequent and was acquired late. Innervation of this finger is split between the median and ulnar nerves. It is argued that shared innervation of fingers in an SFC aids in acquisition by facilitating sensory feedback for handshapes with these SFCs.

Acknowledgment

The research reported here was supported in part by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R01DC013578 and award number R01DC000183. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


Figures

Figure 1. Adult SFC type and token frequencies, categorized as high, mid, and low frequency. Selected fingers are highlighted in blue.

Figure 2. SFCs acquired by ABY, categorized as early and late acquisition.
Figure 3. Musculoskeletal structure of the hand, showing fingers with independent muscle attachments.

Figure 4. Innervation structure of the hand. Blue indicates innervation by the ulnar nerve, red by the median nerve, and green by the radial nerve.

Bullock, Barbara, The University of Texas at Austin, Serigos, Jacqueline, George Mason University and Toribio, Almeida, The University of Texas at Austin - Predicting the degree of language mixing in an open bilingual corpus

Tools and methods, Linguistic analyses of corpora

We present the Spanish in Texas Corpus [spanishintexas.org], an open resource of over 500,000 words spoken by 97 Spanish-English bilinguals living in Texas. A central goal of the corpus is to allow users to explore local varieties of Spanish, which are often disparaged due to purported English interference. Researchers can freely access the corpus materials, which include audio-video files, time-aligned transcripts, and language and POS annotations, as well as rich metadata on the speakers: region, gender, age and social network along with self-reported level of Spanish speaking/writing/reading ability, language of education, and language usage with various interlocutors. This type of detailed metadata is particularly relevant to research on language switching, as an individual's language choice has been shown to be highly affected by extra-linguistic factors. To explore the actual degree and diffusion of mixing in the corpus, we calculate the probability of switching languages between word n-grams as an Integration-Index (I-index) value. We find that the actual probability of switching, as measured by the I-index scaled from 0-1 is quite small (.06), and the appearance of switching through time, as measured by a Burstiness metric (Goh & Barabási 2008), is sporadic, a hallmark of insertional rather than alternational code-switching (Poplack, Sankoff & Miller 1989; Adamou 2016). To better understand how speakers' degree of language mixing varies, we model speakers' I-index value in a linear regression as a function of the various factors provide by the metadata listed above. The contribution of this presentation are as follows: we introduce a fully annotated and open resource, The Spanish in Texas Corpus; we introduce the metrics that we employ to quantify language mixing at the word level in any corpus; we use these metrics and the corpus metadata to explore the nature of insertional code-switching in Spanish in Texas.

Bychkovska, Tetyana - George Mason University - The Influence of Language and Context of Publication on the Use of Self-Mention in Single-Authored Ukrainian Research Articles

Applications

The analysis of self-mention, an important expression of stance in academic writing, can provide insights into how discourse is constructed in particular settings (Hyland, 2005). Cross-cultural studies have examined the use of self-mention in various languages (e.g. Flottum, 2005; Molino, 2010; Yakhontova, 2006); however, few studies (Lafuente-Millan, 2014) have investigated how writers’ use of this feature differs in first and second languages in different publication contexts. Thus, this poster presentation reports on an exploratory corpus study that compared first person pronouns (I, we) and possessive adjectives (my, our) in single-authored linguistics research articles written by Ukrainian scholars in two languages for local and international publication. Three corpora were compared, each consisting of 15 articles: papers in Ukrainian from local Ukrainian
journals (38,233 words), in English from local journals (33,713 words), and in English from international journals (32,117 words). For comparison, a corpus of articles written in English by native-speaking authors with American affiliation (124,725 words) was also compiled. After normalization per thousand words (ptw), the comparison indicated that papers published by Ukrainians in local journals in English contained more instances of self-mention (1.39 ptw) than articles in the same context written in Ukrainian (1.18 ptw). A greater quantitative difference appeared in the texts from international journals in English (2.43 ptw). All three Ukrainian corpora showed more frequent self-mention use in comparison to articles by American authors (0.67 ptw). Also, American-affiliated scholars relied almost exclusively on singular first-person pronouns (I, my), while Ukrainian linguists, irrespective of publication context and language, employed plural pronouns (exclusive we and our). The results suggest that linguistic means of expression, audience awareness, and academic cultural norms are among the factors influencing self-mention use in the examined research articles. These findings lead to implications for cross-cultural research regarding data selection and for teaching second language academic writing.

**Diniz, Luciana, Portland Community College and Concordia University, Bunting, John, Georgia State University and Botelho, Sergio, Software engineer consultant - If I were to teach if-clauses: A corpus analysis of actual uses with ‘if’**

**Applications**

Conditionals are usually divided into four main types (zero, first, second, and third). However, studies have shown that there are many more structures of conditionals commonly used in English (e.g., Jones and Waller, 2011; Phoocharoensil, 2014). Even though textbooks tend to divide them neatly into four groups, they are still a challenging grammar topic to teach to ESL students. One of the main reasons is because the form and the meaning of unreal conditionals do not always match. For example, the structure “if + subject + simple past, would” (also known as second conditional) has a present (unreal) meaning even though the verb in the if-clause is the past tense. This brings a great deal of confusion to students who until that point in their learning process were conditioned by teachers and textbooks to think of simple past as a tense used to refer to past actions. Also, in contrast to the second conditional, other types of if-clauses also have simple past in the if-clause, but the independent clause has simple present or simple past (for example, “if you did your homework, please leave it on my desk”). In this case, the meaning of the dependent clause is actually simple past (real), which contributes to students’ confusion. Perhaps this is the reason why it is often ignored or given minimal attention in ESL textbooks. Another type of if-clauses: “if + was/were + to (base form), would,” is also not given attention to in ESL textbooks, but it is commonly used in spoken English.

Anecdotally, our colleagues appear to rely heavily on their own intuition to determine the categories, relevance, and usage of if-clauses. In order to have a better understanding of how conditionals are used in regular conversations and academic writing/speaking, the authors will present a corpus study on how if-clauses are used in a variety of registers. Our main goal is to identify which constructions and “meaning sets” are more frequent in different genres, which ones should be emphasized in specific ESL/EFL courses and teacher training workshops, and how they should be presented. Findings from the corpus analysis on if-clauses are presented, along with strategies for incorporating the results into coursework for teacher training and ESL/EFL programs.

**Sylviane, Granger and Paquot, Magali - Université catholique de Louvain - The Louvain English for Academic Purposes Dictionary: From design to evaluation**

**Applications**

The Louvain English for Academic Purposes Dictionary (LEAD) is a web-based dictionary-cum-writing-aid tool for intermediate to advanced learners of English, which has been running on the Louvain intranet since 2009 and was released online in 2017 (https://leadlico.uclouvain.be/login). The tool can be used autonomously or as a complementary resource within an academic English course (Granger & Paquot 2015). Native and learner corpora of academic writing were used to identify the typical phraseological patterning (collocations and lexical bundles) of 1,150 cross-disciplinary academic words and to highlight attested problem areas.

One of the innovative features of the dictionary is its adaptable user interface, which automatically customises the content on the basis of the user’s discipline and mother tongue background. Before using the dictionary, users select a domain (currently literature, sociology, anthropology, education, psychology, economics, business, law, political science) and specify their L1 background (currently Chinese, French, Dutch, German and Spanish). Domain selection makes it possible to illustrate the search word with domain-specific example sentences and provides direct access to domain-specific corpora. L1 background selection customizes the feedback to errors and problems that a specific L1 population typically encounters and provides exercises targeting these L1-related difficulties.

Another distinguishing feature of the LEAD is the number of different types of access it offers: by word, by function and by translation. We are also currently testing a newly implemented dynamic interface where users paste their text and the tool identifies all words and word sequences that have an entry in the dictionary and highlights potential errors.

The main aims of this poster presentation are to describe how the LEAD has evolved since we gave a talk on the tool at AACL 2014 and report on a first evaluation of the LEAD by MA students and researchers at Louvain.

**References**

The objective of our contribution is to introduce CoPEP, the Corpus de Português Escrito em Periódicos (Corpus of Portuguese from Academic Journals). CoPEP was especially developed for the compilation of an on-line corpus-driven dictionary of Portuguese for university students (Kuhn, 2017). It contains around 10,000 texts totaling over 40 million tokens extracted from academic journals published in the Brazilian and Portuguese national collections of SciELO (a scientific electronic library), distributed among three schools of knowledge, and further divided into six great areas, following the CAPES classification of areas of knowledge (Brazil).

CoPEP is a synchronic corpus, with only 2% of its texts having been published before 2000. It is markedly multi-variety, with particular effort put in evenly balancing the subcorpora for each language variety, which ended up with almost exactly the same absolute size and consist of a similar number of tokens per both school of knowledge and great area. We kept the text metadata in order to allow for advanced search constraints such as year of publication or scientific area, as well as ISSN, also used for retro-indexing with SciELO. CoPEP is available for analysis on the TEITOK (Janssen, 2016) and Sketch Engine (Kilgarriff et al., 2004) web corpus tools.

We will start by reporting on the rationale for compiling this corpus, outlining our design choices and justifying our options. The methodology of data acquisition will then be described, with a special focus on the challenges we met and solutions we found. The corpus building final phase addresses the sorting out of subcorpora texts, as well as the methodology for variety and scientific domain balancing. We wrap up with some notes on how a corpus such as this can be key in developing non-contrastive multi-variety lexicographic works for pluricentric languages such as Portuguese or English.

Lavrentovich, Alexandra University of Florida - Using classifier features to detect linguistic transfer on morphemes

Linguistic analyses of corpora

We perform a Native Language Identification (NLI) task on a new corpus of English learner essays culled from a virtual online English school (Huang et al., 2017; Geertzen et al., 2013). The NLI task determines the native language background (L1) of a writer based solely on their written production in the second language. We focus on Chinese, Portuguese, Russian, and Spanish learners of English and their overuse, underuse, and erroneous use of English grammatical morphemes across four proficiency levels. We demonstrate that a Support Vector Machine classifier can use only English morphemes to predict the L1 background. The outcome of the computational task is connected to a position in second language acquisition which holds that all learners, regardless of native language background, acquire English grammatical morphemes in the same order (Ortega, 2009). We use the NLI task as a tool to uncover cross-linguistic influence on the developmental trajectory of English morphemes. We show within-group similarities based on L1 background on the use patterns of morphemes, and between group differences. To increase the reliability and generalizability of the NLI output, we perform a cross-corpus evaluation along similar L1 backgrounds and proficiency levels. The second corpus consists of English learner essays for the Test of English as a Foreign Language (Blanchard et al., 2013). We show evidence for cross-linguistic influence on English grammatical morphemes.

References


Liu, Dilin, University of Alabama and Lei, Lei, Huazhong University of Science and Technology - Synonym underdifferentiation in L2 vocabulary use: A corpus analysis of Chinese EFL learners’ use of “circumstance,” “demand,” and “significant”

Linguistic analyses of corpora

(Near)-synonyms are extremely common albeit challenging even for native language speakers. However, there have been only a few studies on L2 learning/use of synonyms. The most recent such study (Author 2016) shows that even advanced ESL/EFL learners have difficulty in this area as they often over/under/misuse some synonyms due to various factors including L1 interference. This corpus study extends the author’s (2016) work by examining Chinese ESL learners’ use of “circumstance,” “demand,” and “significant,” each being a member of a separate synonym set (“circumstance” synonymous with “context, instance, situation…”; “demand” with “request” and “require”; “significant” with “important”). The data used were the 2.65 million-word Spoken and Written English Corpus of Chinese Learners. The researcher first identified all the tokens of the three words (including their inflected forms when applicable) respectively by using Anthony’s (2014) AntConc program: 125 tokens for “circumstance,” 54 for “demand,” and 84 for “significant.” Then, three native-English-speaker graduate research assistants in Applied Linguistics and the author each read individually all the tokens in concordance lines to determine their appropriateness.
by using the following ranking for each item: appropriate (=2), questionable (=1), or inappropriate (=0). The appropriateness score for an item could thus range from 0 (0x4 raters=0) to 8 (2x4=8), with 0-2 being considered inappropriate, 3-4 being somewhat-inappropriate, 5-6 being somewhat-appropriate, and 7-8 being appropriate. A substantial number (51%) of the tokens of the three words were rated inappropriate or somewhat-inappropriate (e.g., “Pollution will destroy our beautiful circumstance” where “circumstance” was erroneously used for “environment”; and “We demanded our teacher to give us more time for our work” where “request” or “ask” would be a much more appropriate choice for “demand” in this context). Such inappropriate uses, which could result in serious semantic or pragmatic misunderstanding in communication, reveal a substantial under-differentiation among the three words and their respective (near)-synonyms by the Chinese learners of English. An analysis of the results against the use of the three words in the Chinese language as well as the teaching materials and practices related to these three English words in China suggests that this under-differentiation has resulted from interference from Chinese, which has only one word for each of the three English (near)-synonym sets, and/or from inadequate or incorrect descriptions and instructions concerning the synonyms in question. Pedagogical and research implications will also be discussed.

Liu, Yang - Georgetown University - The Contributions of the Verb Complements "到(dào)" and "見(jiàn)" to the Main Verb in Mandarin Chinese
Linguistic analyses of corpora

A distinct feature of Mandarin Chinese is that some verbs, in addition to the direct object, also require a resultative complement, either “到(dào)” or “见(jiàn),” immediately following the main verb. To be specific, the two structures are: verb + resultative complement “到” + direct object and verb + resultative complement “见” + direct object. The contributions of these two resultative complements to the main verb are semantically trivial for some verbs like “看(kàn)” to see whereas semantically non-trivial for other verbs like “撞(zhuàng)” to bump into/hit. The direct objects following these two resultative complements can fall into different semantic camps, but the difference is not well-aware of by the native speakers and is of great difficulty for language learners to distinguish one and the other as well.

In general, some verbs can use the two structures interchangeably without changing the meaning too much or without causing ungrammaticality. However, the underlying differences still exist. Specifically, according to 萧红(2016), the resultative complement “到” emphasizes the expectation/consequence aspect whereas the resultative complement “见” addresses the objectivity of the action. In other words, there is no extra entailment introduced by “见”.

This paper is going to further substantiate the claim in 萧红 and 陈一(2016) with more verbs attested in Chinese Treebank 9.0 Corpus. In addition, this paper is also going to see whether the claim in 萧红(2011) regarding the limited generalization on verbs of the mixed use of the resultative complements “到” and “见” in the development of Mandarin Chinese holds true with the examples obtained from Chinese Treebank 9.0 in ANNIS Corpus Search.

References

Liu, Shuyuan - The Pennsylvania State University - Inter-Group Variations of Discourse Markers Use Among Non-Native English Speakers: A Comparison between L1 Spanish and L1 Mandarin Chinese Groups in Advanced Academic Setting
Linguistic analyses of corpora, Applications

Most studies on ESL/EFL learners’ use of discourse markers to date concentrated on a categorical contrast between native English speakers’ (NES) use and that of non-native English speakers (NNE). Usually the learner corpora analyzed comprise only one L1 group representing the category of NNE, which might homogenize variations across speakers with different L1 backgrounds. The current study examines and compares the production of six discourse markers (DM) by NNE with L1 in Mandarin Chinese or Spanish. The six DMs “well, you know, actually, cuz, sort of, now” are constantly reported as being restrictively used by NNS compared with NES (Trillo, 2002; Fung & Carter, 2007; Aijmer, 2011; Polat, 2011; Li, 2015). Based on two parallel sub-corpora of MICASE (Michigan Corpus of Academic Spoken English), quantitative and qualitative analysis are combined to compare the frequency, range of contexts and functional roles of these DMs used by both groups. Additionally, a sub-corpora of native speakers of American English from MICASE is used as a reference corpus to be respectively compared with the learner corpora. The multi-categorical framework (Maschler, 1994; Fung & Carter, 2007) is adopted to qualitatively analyze the functional roles of DMs used by the two focal groups. Results reveal statistically significant variations between two L1 groups. The Spanish group displays more frequent use of the six DMs (especially “well, actually, now”) across a wide range of contexts. Both groups employ the six DMs to perform various communicative functions, with the Spanish group exhibiting more cases in fulfilling the interpersonal function. Possible reasons for these inter-group variations are discussed. This study demonstrates the practicality of utilizing a widely accessible academic corpus to generate findings that inform pedagogical and
professional practices. As the student and faculty body in higher education institutions become more linguistically diverse, the teaching and using of discourse markers should consider the diverse needs across various L1 groups.

References:


Olsen, Michael - University of Georgia - Establishing and investigating American sociopolitical keywords: the Bush-Obama Corpus

Linguistic analyses of corpora

In this study, I investigate keywords across a time of cultural change in American society: the Bush and Obama presidencies. My motivating assumption shares Williams’ (1983) and Stubbs’ (1996) views that cultural keywords reflect social values and that these words are prone to changes in meaning and usage. I first constructed a small pilot corpus (hereafter, the Bush-Obama Corpus, or BOC) of political news articles from three sources representing a left-moderate-right ideological spectrum (Jeffries & Walker 2012). The three-million-word BOC contains full articles from the NY Times, USA Today, and the NY Post; the Bush and Obama subcorpora each contain approximately 1.5mil tokens.

Keyword lists were generated in AntConc (Anthony 2014) based on frequency comparisons between the two subcorpora (Baker 2004). The initial lists were then reduced using a minimum Log-likelihood (Dunning 1993) of 15.13, frequency cutoff of 100 tokens, and intuitional judgments of cultural keywords (Jeffries & Walker 2012). The 36 BOC keywords reflect the changing political climate over the corpus. Bush keywords tended to relate to war (LL=1381), security (LL=367.299), and faith (LL=113.401), while Obama keywords included values such as class (LL=119.94), climate (LL=83.557), and progressive (LL=33.54).

A case study of the Bush keyword freedom (LL=48.04, 224 tokens; 95 tokens in the Obama subcorpus), reveals how this “American” ideal was shaped in the media. During the Bush years, freedom was predominantly used to justify the Iraq War in collocates such as “spreading” and “defending” freedom. By contrast, freedom in the Obama years collocated with “personal”, “individual”, and “religious” freedom and was described as being under threat by Democratic policies such as the Affordable Care Act. These findings reveal the extent to which cultural keywords are susceptible to change depending on sociopolitical climate (Williams 1983). Further study will expand the BOC and analyze additional keywords from the Bush-Obama years.

References


Riccobono, Philip - Aichi Bunkyo University - Towards English for Baseball Purposes (EBP): A Corpus-based Approach

Linguistic analyses of corpora, Applications

Traditionally known as the “All-American” sport, baseball in recent decades has become increasingly internationalized. Indeed, non-native English-speaking players comprised 29.8% of the 2017 opening-day Major League Baseball team rosters, representing a record 19 nations and territories (MLB, 2017). Moreover, in 2012 international players at the minor league level filled 3,382 spots on team rosters (MLB, 2012). Additionally, English serves as the lingua franca for an array of international baseball events (e.g., the 2020 Olympics, the World Baseball Classic, sanctioned international baseball tournaments). To address the increasing globalization of baseball, a new sub-type of English for Occupational Purposes (EOP), English for Baseball Purposes (EBP) is needed, one containing technical vocabulary essential for communicating within this niche (Coxhead, 2013; Nation, 2008). Hitherto, no published research on this specific topic has emerged. Thus, the present study employs mixed methodological approaches to investigate technical and key baseball lexis across various genres of Baseball English (BE). Ethnographic interviews with N = 6 stakeholders of BE are combined with corpus-based and intuition-based approaches to formulate BE lexis beneficial to non-native English-speaking baseball specialists. This study deploys a Baseball English Corpus (BECO) compiled by the researcher (for his doctoral dissertation) from representative English texts using advanced corpus analysis applications and praxis. In keyword and phrase comparisons across BECO genres, findings indicate baseball core user spoken lexis as BECO’s most unique lexis. This study provides learners and educators of BE with insight into characteristics of BECO as well as its specific technical vocabulary and an exploration of pedagogical implications and praxis in the emerging field of EBP.

Starcevic, Katarina, maslansky + partners, Arthur, Paige, maslansky + partners and Butler, Anne - Ascension Health - Applications of corpus-based methodologies for market research: Three case studies

Tools and methods, Applications

Corpus frameworks have been used to understand the linguistic patterns of corporate communication aimed toward a consumer audience (Rutherford, 2005; Pérez-Sobrino, 2016; Poole, 2017) and to analyze consumer language across social media and online review platforms (Vasquez, 2014; Rahmani et al., 2017). While studies like these investigate consumer and consumer-facing language, few have applied corpus methodologies in an explicit market research context. We argue that using corpus-based methods in business contexts—specifically in market research—offers opportunities for both academic fields and the business world. The goal of this study is to complete a feedback loop: applying corpus tools in a market research setting, while also using that application to inform and improve those same corpus tools. To accomplish this, in this study we explore three case studies that involved compiling and analyzing corpora for market research purposes. In the first case study, we reflect on research conducted for a healthcare company, where we constructed and analyzed a corpus of user reviews for telemedicine mobile device applications. In the second and third case studies, we consider the analysis and coding of open-ended survey responses for a healthcare and an insurance company, respectively. In all three cases, the results of the analysis drove decisions around business strategy and investment, and involved demonstrating the business value of corpus analysis to stakeholders with no prior exposure to such methods. Throughout, we identify best practices for adapting corpus methodologies to a business context, and address potential limitations. Moreover, we explore how collaboration between scholarship and business applications can inform one another, and produce a more robust understanding of the applications of corpus-based methodologies.

References:


Terrill, Kristin - Iowa State University - A corpus-based study of language ideologies in charter school performance audits

Applications

Although the complexities of language learning in classroom contexts resist reductive formulas for success, second language acquisition scholarship suggests that some emerging approaches to second language teaching are more effective than others. Standards for teaching language in public schools may restrict the degree to which more effective approaches can be implemented. While “school choice” is politically controversial, an exciting affordance that this movement may offer to educators
of bilingual students is the freedom to apply emerging and possibly more effective approaches to bilingual education. The purpose of this study is to begin to assess the degree to which charter schools have autonomy to implement unconventional pedagogical approaches to bilingual education. The study focuses on Arizona, a state which ranks among the highest in the U.S. for both multilingual students and charter schools. A triangulated method that incorporates keyword analysis and critical discourse analysis is used to analyze documents collected from a division of the Arizona Board of Education tasked with oversight of charter schools. A list of charter schools whose websites promote bi/multilingual education was compiled, and a corpus of performance audits was collected for these schools, as well as a comparison corpus for other schools. In the analysis, particular attention was paid to ideologies surrounding linguistic intellectual development and ideologies related to the role of language in school communities. Results suggest that, although some charter schools promote bi/multilingual education on their websites, their performance audits do not contain strong evidence that multilingualism is a prioritized goal. Qualitative analysis of keywords in this documentation suggests that accountability reinforces the monolingual outcomes as well as a deficit-view of bilingualism and bilingual learners. This research uses a corpus of publicly available documents to demonstrate how power imbalances and entrenched ideologies are reinforced between larger (state-level) and smaller (school-level) communities.

Tywonix, Rurik and Crossley, Scott - Georgia State University - Using Cohesive Features to Automatically Classify Integrated and Independent L2 Writing Assessment and Predict Human Judgments of Quality

Connectedness and organization are important aspects of L2 writing development, L2 writing instruction, and assessment of L2 writing quality. Key to a text’s connectedness are the use of explicit lexical, syntactic and textual features to connect ideas throughout a text (i.e., cohesion). In this study, cohesion features were calculated for a corpus of 960 essays from the Test of English as a Foreign Language (TOEFL) in order to examine differences in the use of cohesion devices between integrated (source-based) writing and independent writing samples. The essays were written by 480 test-takers at two administrations of the TOEFL, with each of the test-takers writing an integrated and an independent essay. Local and global indices of cohesion were measured using two different automated textual analysis tools: the Tool for the Automatic Assessment of Cohesion (TAACO) and ReaderBench.

Multivariate Analysis of Variance found significant differences in the use of cohesive features for the two writing tasks. A discriminant function analysis correctly classified essays as either integrated or independent in 92.3 per cent of cases. The analyses suggest integrated writing is marked by higher use of specific connectives and greater lexical overlap of content words between textual units, whereas independent writing is marked by greater lexical overlap of function words, especially pronouns. Linear regression analyses found that cohesive indices which distinguish test tasks predicted human judgments of writing quality more strongly in independent writing. However, the strongest predictor (lexical overlap of function words) was the same for both tasks.

The findings demonstrate that text cohesion is a multidimensional construct shaped by the writing task, yet the measures of cohesion which affect human judgments of writing quality are not entirely different across tasks. These analyses allow us to better understand the role of cohesive features writing tasks and has implications for automated writing assessment.

Zhang, Genggeng - Georgia State University - A Corpus Approach of Teaching If-Conditionals to Chinese Aviation Maintenance Students

Although a great number of miscommunication incidents in aviation maintenance are caused by verbal and documentation communication failures (Ma, et al., 2010), very few research studies investigate the communication problems resulting from the mismatch of language use between English training and the manuals (Alshabe et al., 2017; Borowska, 2017). Aircraft maintenance, repair, and overhaul (MRO) manuals written in Simplified Technical English (STE) represents a controlled language that is used in technical documents to improve its comprehensibility to non-native speakers (Van der Eijk, 1997; Gabrielatos, 2006; Pendić et al., 2016). Following the ASD-STE100 standards (2017), manuals are restricted to use simple verb tenses (present, past, and future) and limit the sentence length within 20 words, which is also reflected on the simplified use of conditionals. Among all conditional clauses, if-conditional clauses are one of the most difficult English grammatical structures to Chinese learners (Lester, 2008; Wu, 2012). Based on evidence from AntConc analysis, if-conditional clauses are also one of the most frequently used conditionals in MRO manuals, which indicates the necessity of teaching if-conditional clauses to Chinese maintenance personnel to avoid miscommunication. However, does the knowledge of if-conditional clauses acquired from the aviation English textbook fully prepare the Chinese aviation maintenance students to understand MRO manuals and generate comprehensible communication at workplace (Pazyura, 2015; Alshabeb et al., 2017; Borowska, 2017)? This study collects and compares a corpus of Chinese aviation training materials in use and a corpus with Boeing 737 maintenance manuals. The type, frequency, and distribution of if-conditional clauses are illustrated with AntConc. Findings demonstrate how if-conditional clauses are used in training textbooks and the aviation manuals. Training applications, especially the corpus driven approaches for teaching conditionals are discussed.