FROM DATA TO EVIDENCE
BIG DATA
RICH DATA
UNCHARTED DATA
ORGANISED BY THE RESEARCH UNIT FOR VARIATION, CONTACTS AND CHANGE IN ENGLISH (VARIENG)
The d2e conference gratefully acknowledges the support and cooperation of the following institutions:
Dear Colleagues,

Welcome to Helsinki and the conference “From data to evidence in English language research: Big data, rich data, uncharted data” – or “d2e”, as we organizers have become accustomed to calling the event over the last eighteen months of preparations.

The VARIENG Research Unit has several reasons for organizing this conference in 2015. Now that Digital Humanities has become something of a buzz word in many disciplines, we would like to engage in and diversify the current discussion in our field by focusing on data explosion and the use of new sources, historical and modern, in English language research. We were delighted to discover that, in response to our call, the papers to be given at d2e address topics that range from big data to small but richly annotated data, both including hitherto uncharted materials, and all focused on the kinds of evidence they can provide for future research in the field.

Our decision to hold this conference in 2015 was also motivated by two external reasons. First, this year marks the 20th anniversary of VARIENG as a research unit. Although no official foundation document has survived, the decisive step taken to form the unit is documented in the 1995 Annual Report on Research and Teaching at the English Department, University of Helsinki.

The year 2015 also marks the 375th anniversary celebration of our alma mater, the first university to be established in Finland, founded in Turku in 1640 and moved to Helsinki in 1828. The Power of Thought exhibition tells the whole story in the University of Helsinki Museum, located opposite our conference rooms on the third floor of the Main Building. Please feel free to explore it.

On behalf of the Organizing Committee and VARIENG as a whole, I would like to wish you a most rewarding conference and an open exchange of ideas over the next few days.

Terttu Nevalainen
Chair of the d2e Organizing Committee
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.15–9.30</td>
<td>Opening words by Terttu Nevalainen</td>
</tr>
</tbody>
</table>
| 9.30–10.30   | **PLENARY: Tony McEnery**  
The Criminalized Poor - exploring a marginalized group in public discourse in the 17th century  
Consistorium (Chair: Irma Taavitsainen) |
| 10.30–11.00  | Coffee break                                                          |
| 11.00–12.30  | Session 1a  
Room 6 (Chair: Lieselotte Anderwald)  
**James McCracken**  
Integrating historical frequency data with the Oxford English Dictionary: What can it do, and what are the limitations?  
Session 1b  
Room 7 (Chair: Radoslaw Dylewski)  
**Lucia Siebers**  
Uncharted data: African American letters from the 18th and 19th centuries |
|              | **Tyler Kendall**  
Making old data sources into new data sources: On the aggregation of sociolinguistic datasets and the future of real-time and cross-study analysis |
|              | **Jack Grieve et al.**  
Big data for the analysis of language variation and change |
| 12.30–14.00  | Lunch break                                                          |
| 14.00–15.30  | Session 2a  
Room 6 (Chair: Justyna Robinson)  
**Javier Martín Arista & Ana Elvira Ojanguren López**  
Making sense of heterogeneous data with a relational database of Old English  
Session 2b  
Room 7 (Chair: Elizabeth Peterson)  
**Marianna Hintikka & Minna Nevala**  
Charting the social margins: Fallen women in late 19th-century English newspapers |
|              | **Stefanie Degaetano-Ortlieb et al.**  
The taming of the data: Using text mining in building a corpus for diachronic analysis |
|              | **Turo Vartiainen**  
Covert expressions of attitude in criminal records: Rape cases in the Old Bailey Corpus, 1720–1749 |
| 15.30–16.00  | Coffee break                                                          |
| 16.00–16.30  | **DEMO: Jonathan Hope**  
Ubiquity: a web-based text tagger  
Room 6 (Chair: Antoinette Renouf) |
|              | **Mikko Tolonen**  
DEMO: History in the ESTC catalogue |
|              | **Eetu Mäkelä, Terttu Nevalainen & Tanja Säily**  
DEMO: Developing an interface for historical sociolinguistics |
|              | **Anni Sairio**  
Uncharted data from the social margins: a case study of 18th-century L2 English |
|              | **Christer Geisler**  
Analyzing non-native English letter-writing by Swedes in the 17th and 18th centuries: How ‘small’ and ‘big’ data meet |
| 18.00–20.00  | University of Helsinki Reception  
(University Main Building, Teachers’ Newspaper Lounge) |
# TUESDAY — 20 October

## 9.00–9.30

| Session 4a | Session 4b |
| Room 6 (Chair: Marc Alexander) | Room 7 (Chair: Stefan Diemer) |
| DEMO: Gerold Schneider, Mennatallah El-Assady & Hans Martin Lehmann<br>Tools and Methods for Processing and Visualizing Large Corpora |

## 9.30–10.30

| Antoinette Renouf<br>Big Data: opportunities and challenges for English corpus linguistics | Beke Hansen<br>The ICE metadata: A window to the past? An exploratory study of Hong Kong English |
| Tanja Säily & Jukka Suomela<br>DEMO: types2: Exploring word-frequency differences in corpora | Mirka Honkanen<br>Multilingual resources and authentication in the computer-mediated communication of U.S.-Nigerians: The role of African American Vernacular English in the repertoire |

## 10.30–11.00

Coffee break

## 11.00–12.30

| Session 5a | Session 5b |
| Room 6 (Chair: Joanna Kopaczyk) | Room 7 (Chair: Magnus Levin) |
| Hendrik De Smet<br>How gradual change progresses | Sanna Franssila<br>The expressions of betrayal in American political discourse |
| Karolina Rudnicka<br>How to investigate grammatical obsolescence? A case study from Late Modern English grammar | Marije van Hattum<br>‘I am roary on the hill’: identity-construction in 19th-century Irish English threatening notices |
| Claudia Claridge & Merja Kyö
Intensifiers on their way out: full, right, and real in the Old Bailey Corpus | Joe McVeigh<br>Too Much to Say in One Subject Line...: Exploring trends in Big Data marketing communication |

## 12.30–14.00

Lunch break

## 13.30–15.30

| Session 6a | Session 6b |
| Room 6 (Chair: Mikko Laitinen) | Room 7 (Chair: Jukka Tyrkkö) |
| Svetlana Vetchinnikova<br>What can a cognitive corpus tell us about grammar as an emergent property of language at the communal level? | Olli Silvennoinen<br>Not only apples and apples but also apples and oranges: a register perspective on contrastive negation in English |
| Marie-Louise Brunner, Stefan Diemer & Selina Schmidt<br>“... okay so good luck with that ((laughing))” - Managing rich data in a corpus of Skype conversations | Marjut Salokannel<br>How copyright law could support text and data mining |

## 18.30–19.30

City of Helsinki Reception<br>(Old City Hall)
### Session 7a
Room 6 (Chair: Tanja Rütten)

**Turo Hiltunen & Jukka Tyrkkö**  
Academic vocabulary in Wikipedia articles: frequency and dispersion in uneven datasets

**Jefrey Lijffijt & Tanja Säily**  
Adjusting p-values for heterogeneity in collocation analysis

**Ljubica Leone, Rita Calabrese & Bruno Bisceglia**  

**Rik Vosters, Mike Kestemont & Folgert Karsdorp**  
Open-set authorship verification and language history from below: Determining authorship in a corpus of 18th-century correspondence

**David Brett & Antonio Pinna**  
Words (don't come easy): the automatic retrieval and analysis of popular song lyrics

**Joseph Flanagan**  
Big Data and Reproducible Research: Strategies, Tools and Workflows

### Session 7b
Room 7 (Chair: Hendrik De Smet)

**Lieselotte Anderwald**  
Empirically charting the success of prescriptivism

**Morana Lukač**  
Charting out the discourse of linguistic prescriptivism in the Complaint Tradition Corpus

**Nuria Yáñez-Bouza**  
Paratext, title-pages and eighteenth-century grammar books

**Gerold Schneider, Mennatallah El-Assady & Hans Martin Lehmann**  
Changes in language, topics and society according to the COHA corpus

### Coffee break & Poster exhibition

### Session 8a
Room 6 (Chair: Christer Geisler)

**Lieselotte Anderwald**  
Empirically charting the success of prescriptivism

**Michael Pace-Sigge**  
Mono-collocates: How fixed Multi-Word Units (MWUs) with of or to indicate diversity of use in different corpora

**Morana Lukač**  
Charting out the discourse of linguistic prescriptivism in the Complaint Tradition Corpus

**Maggie Leung**  
Studying phrasal verbs in a corpus of engineering English

**Nuria Yáñez-Bouza**  
Paratext, title-pages and eighteenth-century grammar books

**Gerold Schneider, Mennatallah El-Assady & Hans Martin Lehmann**  
Changes in language, topics and society according to the COHA corpus

### Session 8b
Room 7 (Chair: Svetlana Vetchinnikova)

**Robert Daugs**  
Modal and quasi-modal development in 19th and 20th century American English: Evidence from COHA and COCA

**Sara Norja**  
From manuscript to digital edition: Middle English alchemical texts as a resource

**Benedikt Heller & Melanie Röthlisberger**  
Big data on trial: Researching syntactic alternations in GloWbE and ICE

**Thomas Kohnen**  
Commonplace books: Charting and enriching complex data

**Mark Kaunisto & Juhani Rudanko**  
Exceptions to Bach’s Generalization in Regional Varieties of English: a Case Study with Data from the GloWbE Corpus

**Radosław Dylewski**  
Confederate textbooks – an unexplored source of historical dialectal data

### Lunch break

### Session 9a
Room 6 (Chair: Merja Kytö)

**Robert Daugs**  
Modal and quasi-modal development in 19th and 20th century American English: Evidence from COHA and COCA

**Sara Norja**  
From manuscript to digital edition: Middle English alchemical texts as a resource

**Benedikt Heller & Melanie Röthlisberger**  
Big data on trial: Researching syntactic alternations in GloWbE and ICE

**Thomas Kohnen**  
Commonplace books: Charting and enriching complex data

**Mark Kaunisto & Juhani Rudanko**  
Exceptions to Bach’s Generalization in Regional Varieties of English: a Case Study with Data from the GloWbE Corpus

**Radosław Dylewski**  
Confederate textbooks – an unexplored source of historical dialectal data

### Session 9b
Room 7 (Chair: Martti Mäkinen)

**Robert Daugs**  
Modal and quasi-modal development in 19th and 20th century American English: Evidence from COHA and COCA

**Sara Norja**  
From manuscript to digital edition: Middle English alchemical texts as a resource

**Benedikt Heller & Melanie Röthlisberger**  
Big data on trial: Researching syntactic alternations in GloWbE and ICE

**Thomas Kohnen**  
Commonplace books: Charting and enriching complex data

**Mark Kaunisto & Juhani Rudanko**  
Exceptions to Bach’s Generalization in Regional Varieties of English: a Case Study with Data from the GloWbE Corpus

**Radosław Dylewski**  
Confederate textbooks – an unexplored source of historical dialectal data

### Conference Banquet
(Unioninkatu Banquet Hall)
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Chair/Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.30–10.30</td>
<td><strong>PLENARY: Päivi Pahta</strong> Multilingual practices in Late Modern England: Evidence of multilingual repertoires in historical data</td>
<td>Consistorium (Chair: Thomas Kohnen)</td>
</tr>
<tr>
<td>10:30–11.00</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>11.00–12.30</td>
<td><strong>Session 10a</strong> Room 6 (Chair: Michael Pace-Sigge)</td>
<td><strong>Session 10b</strong> Room 7 (Chair: Nuria Yáñez-Bouza)</td>
</tr>
<tr>
<td></td>
<td><strong>Mikko Laitinen, Magnus Levin &amp; Alexander Lakaw</strong> Advanced non-native English on a continuum of Englishes: Charting new data sources</td>
<td><strong>Tanja Rütten</strong> Behind the scenes of diachronic corpus studies - uncharted discourse types in historical corpora</td>
</tr>
<tr>
<td></td>
<td><strong>Elizabeth Peterson, Turo Hiltunen &amp; Johanna Vaattovaara</strong> Tracking the meanings of ‘pliis’: combining evidence from multiple sources</td>
<td><strong>Alpo Honkapohja</strong> Manuscript abbreviations in Latin and Middle English medical treatises: a corpus-linguistic approach</td>
</tr>
<tr>
<td></td>
<td><strong>Steven Coats</strong> Gender and lexical type frequencies in Finland Twitter English</td>
<td></td>
</tr>
<tr>
<td>12.30–13.00</td>
<td><strong>Marc Alexander</strong> DEMO: How Can We Stage 470,000 Words?</td>
<td>Consistorium (Chair: Susan Fitzmaurice)</td>
</tr>
<tr>
<td>13.00–13.15</td>
<td>Closing words (Room 6)</td>
<td></td>
</tr>
</tbody>
</table>
PLENARIES
**Mark Davies (Brigham Young University)**

**Why size alone is not enough: the importance of historical, genre-based, and dialectal variation in language**

Anyone attending this d2e conference already understands the importance of big data, and so any additional evidence for this would simply be “preaching to the choir”. Nevertheless, to begin the presentation I will briefly review two types of phenomena in which corpus size is indeed a crucial consideration. The first phenomenon is syntactic change involving lexical subcategorization, which has typical been often carried out with “ad-hoc”, proprietary corpora (newspapers, magazines, etc) – but much less often with large, publicly-available, balanced corpora. The second phenomenon is lexical change and variation, which is often omitted altogether in corpus-based studies of historical and dialectal variation, due to researcher’s continuing reliance on corpora that are too small to model such variation.

Recognizing the limitations of smaller corpora, several groups (e.g. Sketch Engine, Corpus from the Web) have created very large corpora that are based primarily (or exclusively) on web pages, which are quite easy to obtain. For certain types of investigations, such corpora are invaluable. The question, however, is how data from these huge, undifferentiated corpora of web pages correlate with data from more carefully constructed, balanced corpora. I will provide several examples that show how – depending on the particular phenomena under investigation – these large web-based corpora are “all over the map” in terms of their relationship to the genres in corpora like the BNC or COCA – sometimes agreeing most with newspapers, sometimes with magazines, sometimes with academic, etc. We must therefore be quite careful in claiming that data from these web-based corpora in any way model “standard English” (or Spanish, or Finnish, or Chinese...)

An additional limitation of these web-based corpora is that (in addition to their problematic relationship to standard genres) they are often oblivious to historical and dialectal variation in language. I will provide several examples (mostly from the domain of syntax) where there is important dialectal and (ongoing) historical variation, which would never really be accounted for in these large web-based corpora. These examples will be taken primarily from the Corpus of Historical American English (COHA; 400 million words, 1810s-2000s) and the Corpus of Global Web-based English (GloWbE; 1.9 billion words from 20 different English-speaking countries).

Variation matters. Many small corpora are limited primarily to looking at just high-frequency phenomena like modals or other auxiliaries. On the other hand, many people would argue that newer “mega-corpora” claim to model languages well, but they fail to account for linguistic variation that is crucial to understanding these languages. As we will suggest, the best solution is to create corpora that are very large, but whose composition and architecture are also explicitly designed to provide data on genre-based, historical, and dialectal variation.

**References**

Biber, Douglas; Jesse Egbert; and Mark Davies. 2015. Exploring the Composition of the Web: A Corpus-based Taxonomy of Web Registers. *Corpora*.


Davies, Mark. 2014. The importance of robust corpora in providing more realistic descriptions of variation in English grammar. *Linguistic Vanguard* (online)


Davies, Mark. 2012. Expanding Horizons in Historical Linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7: 121-57.

Tony McEnery (Lancaster University)

The Criminalized Poor - exploring a marginalized group in public discourse in the 17th century

Who were the criminalized poor in the 17th century? What did they do? Where did they live? Who did they associate with? What was associated with them? Did the way they were talked about change over time? In this talk I will explore these questions by looking at modern lexicographical resources, 17th century lexicographical resources and the EEBO corpus. In doing so I will cast light on these questions while also exploring the potential and shortcomings of the resources that are used in the study. In particular I will focus upon how the corpus can help us to come to a fuller view of these questions than dictionary resources currently permit. I will also reflect upon and explore ways of dealing with the volatility of collocates over time.
Multilingual practices in Late Modern England: Evidence of multilingual repertoires in historical data

The last few years have witnessed an increase in research on multilingualism in historical data, including functional and structural aspects of the phenomenon. This research has established that multilingual practices are present in various types of writings in the earlier history of the English language (see e.g. Schendl & Wright 2011, Pahta 2012). While most studies have focused on the medieval period, which was a time of extensive societal multilingualism in Britain, data from later periods have also proved valuable for studying the use of multilingual resources in a historical perspective.

This paper examines language use of the eighteenth and nineteenth centuries, based on research conducted in an on-going project on Multilingual Practices in the History of Written English at the University of Tampere. The material originates from the 34-million-word multi-genre Corpus of Late Modern English Texts 3.0. These texts contain multilingual passages that have been identified using complementary automated and semi-automated techniques, including a new corpus tool, Multilingualiser, developed specifically for processing multilingual data. This paper demonstrates that English writers of the late modern period made use of their multilingual resources in many different types of communicative events and for various communicative purposes. The statistical analysis of multilingual passages in relation to enhanced text-typological information and sociolinguistic information on the writers, which have been collected in the project, indicates that some sociolinguistic and textual factors, such as education and genre, clearly had an impact on multilingual practices. However, the patterns in the large dataset are complex, and there is great variation in the multilingual repertoires of individuals evidenced in the texts. To complement the statistical analysis, this paper also examines sample repertoires to provide a closer view of how the writers used their multilingual resources in their communicative tasks, without observing the boundaries of individual languages in expressing themselves.

References

Corpus of Late Modern English Texts 3.0 (CLMET3). Compiled by Hendrik De Smet, Hans-Jürgen Diller and Jukka Tyrkkö. See https://perswww.kuleuven.be/~u0044428/.


Multilingualiser. Software developed by Jukka Tyrkkö at the University of Tampere.


Tackling complexity in humanities big data: from parliamentary proceedings to the archived web

One of the key characteristics of big data for the humanities is its complexity, whether we are dealing with the text of digitised nineteenth-century newspapers or the vast quantities of born-digital data generated by social media platforms such as Twitter. It has been produced over different periods of time, using different and often undocumented methods; at best it may be only partially structured; and on occasion we may not even know precisely what it looks like. How can humanities researchers develop theoretical and methodological frameworks for dealing with such complex material? What tools and skills do we need to help us to work effectively with big data? How can we analyse data at scale, while retaining an understanding of the people and stories which are woven into its fabric? What, ultimately, can the humanities in general, and history in particular, bring to big data research? This paper will address some of these questions, focusing on two very different types of data: parliamentary proceedings in the UK, the Netherlands and Canada (from c. 1800 to the present day); and the archive of UK web space from 1996 to 2013. Both offer fascinating insights into language, politics, culture and society, but both also present challenges, some of which we are only just beginning to identify, let alone to solve. It is vital that we work together to tackle these questions, or contemporary decisions about how we describe, publish and preserve big data may hamper the humanities researchers of the future.
Empirically charting the success of prescriptivism

One source of "uncharted" data that have not really been used in research so far is, surprisingly enough, prescriptive grammar writing of the nineteenth century – the century still known first and foremost for its prescriptivism – even though materials abound (e.g. as listed in Görlach 1998). While a number of scholars have opened up the field of non-scholarly historical grammars for research, these explorations have mostly concentrated on the eighteenth century (cf. the contributions in Tieken-Boon van Ostade 2008; Beal 2010; Auer 2009; or Yáñez-Bouza 2008); the following century has seen little systematic interest so far. Contrary to received opinion, my investigations of a whole range of linguistic features undergoing change at the time show that prescriptivism quite generally does not seem to have had much influence. Even though many new phenomena were criticized harshly, this does not seem to have affected their introduction; even though some old-fashioned features were supported, this does not seem to have prevented their demise. Against this overall ineffectiveness of prescriptive grammar writing, I want to concentrate in this talk on those intermittent periods where prescriptivism can be argued to have been successful. As we will see, this success is typically short-lived (it extends to periods of a decade or two), small-scale (the statistical effect ranges over few percentage points), and temporary (overall trends are at best halted briefly, but the following dynamic more than makes up for this halt). As my discussion will show, there are general methodological problems involved in identifying potentially successful cases of prescriptivism. However, with the help of "uncharted" data available through many unconventional sources (only one of which is Google Books), corpus linguistics (or rather: corpus linguistics plus) can now at least begin to undertake these worthwhile investigations.

References:


Words (don’t come easy): the automatic retrieval and analysis of popular song lyrics

A text type that has been by and large ignored by mainstream corpus linguistics research until recently is that of the lyrics of popular songs. Three recent works, by Kreyer and Mukherjee (2007), Werner (2012) and Bértoli-Dutra (2014), are groundbreaking studies, however they are based on relatively small samples.

The current work will describe the compilation of a large corpus of popular song lyrics in English (>10 M tokens) divided into sub-genres. The texts were gathered by web crawling the index pages of an online song repository. It will then analyse the keywords of each genre and “key keywords” (Scott and Tribble, 2006) highlighting similarities and differences between genres.

The first part of this paper will discuss the procedures adopted to retrieve the song lyrics, along with metadata such as date, author, album and genre. The repository proved somewhat unreliable regarding the attribution of artists to musical genres, therefore alternative semi-automatic processes had to be developed. Several other reliability issues will be discussed, for example, songs in foreign languages, covers, variation in song titles and artist names are all factors that had to be filtered out or normalised.

The second part will present preliminary results concerning the analysis of keywords. While each sub-genre (Alternative Rock, Country, Hip Hop, Heavy Metal, Pop, R’n’B and Rock) had a considerable number of keywords, we noticed that those of some genres, such as Hip-Hip and Heavy Metal, were highly characteristic lexical items, those of others, such as Pop and R’n’B were mainly grammatical items with very high frequencies. The latter two genres share so many keywords that it could be argued that, at least on a textual basis, they are essentially the same genre.

References


SCOTT, MIKE, and CHRISTOPHER TRIBBLE, 2006, Textual Patterns: keyword and corpus analysis in language education, Amsterdam: Benjamins.

Managing rich data in a corpus of Skype conversations

Spoken computer-mediated communication (CMC) presents an increasingly complex challenge for corpus creation. While big-data approaches work well with written data, as Google Books and GloWbE (Davies 2011-, 2013) show, rich conversation corpora pose major problems at the recording, transcribing, annotation and querying stages (Diemer, Brunner, and Schmidt 2015). Many elements of conversation corpora move beyond current transcription standards, are a matter of debate (e.g. gestures and gaze), or raise organizational issues.

This paper discusses rich data in spoken CMC corpora, using examples from CASE, the Corpus of Academic Spoken English (Diemer et al. 2015), compiled at Saarland University, Germany. CASE consists of Skype conversations between English as a Lingua Franca (ELF) speakers. CASE data allows research on a wide range of linguistic features of informal spoken academic CMC discourse as well as pragmatic, intercultural, and didactic aspects (Brunner 2015). It illustrates the benefits and the challenges rich data poses, particularly during transcription and annotation.

Further developing Thompson’s (2005) guidelines for spoken corpora, Diemer, Brunner and Schmidt (2015) as well as Sauer and Lüdeling (2015) propose a layered corpus organization for this purpose. Following this approach, CASE comprises multiple layers, including orthographic, part-of-speech-tagged, and pragmatic layers (cf. Schmidt, Brunner, & Diemer 2014). The paper discusses the limitations of transcription and annotation in view of pragmatic (paralinguistic, e.g. laughter, and non-verbal cues, e.g. gestures), conversational (overlap, pauses, etc.), and organizational (turn-taking, topic choice) features, as well as the feasibility of adding further layers reflecting ELF, English as a Foreign Language, and CMC features (echo, lag, etc.). The paper endeavors to provide an extensive overview of possible areas of research that can be included in the creation of rich data corpora.

References


Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries. Online: http://corpus.byu.edu/glowbe/.

Davies, Mark. 2011-. Google Books Corpus. (Based on Google Books n-grams). Online: http://googlebooks.byu.edu/.

Diemer, Stefan et al. 2015. CASE: Corpus of Academic Spoken English. Saarbrücken: Saarland University; Sofia: St Kliment Ohridski University; Forlì: University of Bologna-Forlì; Santiago: University of Santiago de Compostela; Helsinki: University of Helsinki, Hanken School of Business; Birmingham: Birmingham City University. To appear.


Intensifiers on their way out: *full*, *right*, and *real* in the Old Bailey Corpus

The study of waning forms leading to fossilized expressions or even loss of forms has been a somewhat neglected area in the study of the history of English. Intensifiers such as *full*, *right* and *real* – as in *full late*, *right true*, and *real good* – provide an intriguing test case in this respect. These uses were typical of Middle and Early Modern English (Méndez-Naya 2008, Nevalainen 2008), but are no longer common in Present-day standard British English. Previous research points to the Late Modern English period as the crucial time when these items were losing ground heavily (Méndez-Naya 2007, 2008).

Our research questions include the following:

- What are the latest attestations of intensifying uses?
- In which lexi-co-syntactic contexts do late uses occur? Are there signs of (phraseological) fossilisation?
- Which are the most conservative types of users in sociolinguistic respects?
- Can one detect any overarching regularities in the loss-behaviour of these three types?

As for the early stages of an innovative form, it is necessary to have a large corpus base for the (final) phases of outgoing forms. In our study we turn to the *Old Bailey Corpus* (OBC) for data. The OBC is a balanced subset of the *Old Bailey Proceedings* (134 million words), and its unique basis in speech, which is more authentic than e.g. drama, makes the corpus a perfect tool for the study of intensifiers, a feature that is often more frequent in conversation. The 14-million-word online version of the OBC has particular sociolinguistic and sociopragmatic relevance to our study, with its full socio-biographical speaker annotation (gender, age, occupation, social class), pragmatic information (speaker role in the courtroom: defendant, judge, victim, witness etc.) and textual information (scribe, printer and publisher of the Proceedings). The three forms are sufficiently common already in the online OBC to yield intensifying uses, i.e. *full*: 1760, *right*: 6540, and *real*: 219. If more data is necessary for specific questions, we will be able to tap the larger versions of the OBC and also the complete Proceedings materials.

References


Gender and lexical type frequencies in Finland Twitter English

English is playing an expanding role as a language of informal online communication in many communities where it has hitherto not been widely used as a language of local communication. This is particularly evident on global social media platforms such as Twitter. Some research has found small but significant differences by gender for the use of grammatical and lexical features in spoken and written language, including online varieties such as chat, instant messaging, and Twitter (Baron 2004; Herring & Paolillo 2006; Squires 2012; Bamman et al. 2014). In this study, the frequency distributions of selected standard and non-standard lexical types and their correlation with gender are considered in a corpus of English-language Twitter messages originating from Finland.

In a first step, a corpus of geo-located English-language Twitter user messages from Finland was created by accessing the Twitter Streaming API and using an automated language detection tool to remove non-English user messages. After disambiguating author gender by automated methods, the frequencies and distributional profiles of selected lexical types were examined and compared with those derived from a corpus of English-language Twitter messages worldwide subject to the same processing procedures.

The analysis supports some previous findings pertaining to gendered language use, but also suggests that the manifestation of gender in lexical type frequencies in Finland Twitter English reflects sociolinguistic considerations, particularly for those lexical features most strongly associated with the discourse of the Twitter platform itself. The analysis sheds light on the dynamics of a geographically specified online English variety and considers how sociolinguistic factors interact with technological considerations to contribute to the differentiation of online Englishes.

References

BAMMAN, DAVID; JACOB EISENSTEIN; and TYLER SCHNOEBELEN. 2014. Gender Identity and Lexical Variation in Social Media. Journal of Sociolinguistics 18.135–60.


Modal and quasi-modal development in 19th and 20th century American English: Evidence from COHA and COCA

It is generally accepted that the modal verbs (e.g. shall, must) have been significantly decreasing in their frequency of use, while, at the same time, the so-called quasi-modals (e.g. BE going to, HAVE to) undergo a remarkable overall increase (cf. e.g. Krug 2000, Leech 2003, Biber 2004, Leech 2011). However, evidence from COHA and COCA can shed new light on these developments and relativize earlier results.

My study focusses on three aspects concerning the diachrony of modal expressions in AmE. First, I will provide data from COHA on a relatively uncharted research field, i.e. modal/quasi-modal variation and change in 19th century AmE. Secondly, while my data confirm a general decline in the frequency of the modal verbs in AmE over the 20th century, a closer look at their individual developments suggests that some of Leech’s near-canonical findings, particularly Leech’s (2003) and (2011) subdivision of the modals into frequent and infrequent ones and the observed ‘bottom weighting of the frequency loss’ (Leech et al. 2009: 73), need revision. The evidence from COHA clearly indicates that the high-frequency modals (will, would, can, could) are historically not a homogenous group and analyzing them as such obscures the facts that (i) up to the 1950s can and could cannot be clearly affiliated with either the frequent or the infrequent modals, and (ii) that the decline in the frequency of will over 20th century is responsible for almost one-fourth of the overall modal decrease. Interestingly, the demise of will as the main future marker in PDE is not counterbalanced by the significant increase in the use of BE going to, thus raising questions regarding a possible overall change in referring to future time. To complete the overall picture, my final part addresses register-specific variation of modals and quasi-modals in both COHA and COCA.

References


Leech, Geoffrey; Marianne Hundt; Christian Mair; and Nicholas Smith. 2009. Change in Contemporary English. Cambridge: Cambridge University Press.
How gradual change progresses

Grammatical changes are typically gradual. That is, a change affects different grammatical contexts at a different time or pace (see e.g. Harris & Campbell 1995). One way of explaining this gradualness is by what may be called a ‘cascade model’ of change. The central idea in such a model is that when one grammatical context is affected by a change, this facilitates extension of the change to other related contexts. In gradual changes, then, each step triggers the next (De Smet 2012).

This paper provides evidence to support the cascade model of gradual change. The change under study is the development of the noun key into an adjective (Denison 2001). By hypothesis, the change is gradual, starting with increased productivity of compounds with key as specifying element, leading later to ‘debonded’ (Norde 2009; Van Goethem 2012) uses, as in (1b), and clearly adjectival uses, as in (1c-d).

(1)

a. I turn briefly to the key question of licensing. (1977, HD)
b. This figure is a key one in the argument. (1978, HD)
c. I believe that it is an absolutely key industry in many respects [...]. (1971, HD)
d. [...] discretionary awards, which are so key to what we are considering [...] (1996, HD)

Evidence that the steps of change trigger one another is drawn from four different sources, using data from Parliamentary debates in the British Houses of Parliament. First, qualitative analysis of the corpus data shows that early debonded uses are typically built on well-entrenched compound uses. Second, patterns of cross-individual variation in the adoption of the change are consistent with the cascade model, in that individuals whose compounding uses are low in productivity or who did not adopt debonded uses of key are unlikely to have (respectively) debonded and clearly adjectival uses. Third, changes over language users’ lifetime are also consistent with the cascade model, in that individuals are likely to adopt debonded and adjectival uses only if their compounding uses also increase in productivity. Fourth, analysis of Parliamentary exchanges shows that if speakers are primed by a compounding use of key in a previous turn they are more likely to produce a debonded or adjectival use then if they are not primed.

All of the evidence indicates that increased availability of early-stage constructions (be it through entrenchment or through priming) facilitates the adoption of later-stage constructions. This is in line with the cascade model of gradual change.

References


The taming of the data: Using text mining in building a corpus for diachronic analysis

Social and historical linguistic studies benefit from corpora encoding contextual metadata (e.g. time, register, genre) and relevant structural information (e.g. document structure). While small, hand-crafted corpora control over selected contextual variables (e.g. the Brown/LOB corpora encoding variety, register, and time) and are readily usable for analysis, big data (e.g. Google or Microsoft n-grams) are typically poorly contextualized and considered of limited value for linguistic analysis (see, however, Lieberman et al. 2007). Similarly, when we compile new corpora, sources may not contain all relevant metadata and structural data (e.g. the Old Bailey sources vs. the richly annotated corpus in Huber 2007).

For corpora with rich metadata and structural data, we can draw on well-established methods of analysis, from descriptive statistics to machine learning (see e.g. Kilgarriff 2004 for an overview). For the analysis of corpora with few/no metadata or structural information, we first need to learn more about our data. Relevant methods are found in data mining (Witten et al. 2011), which is concerned with detecting patterns in complex and potentially noisy datasets. This is what we have when building a corpus from uncharted material.

We are currently building a corpus from the Philosophical Transactions and Proceedings of the Royal Society of London (see e.g. Atkinson 1998; Taavitsainen et al. 2010), covering the first two centuries (1665–1869) of publication (Khamis et al. 2015). The sources (obtained from JSTOR) contain some but not all relevant metadata (year of publication and author, but not discipline) and no structural data. We apply a combination of pattern-based techniques and text-mining methods (e.g. clustering, classification, topic modeling) to explore the data. Apart from understanding our data better and (semi-)automatically enriching it with relevant contextual and structural information, we obtain positive effects regarding data quality (detection of artifacts like OCR errors, text duplicates, and running headers/footers).

Acknowledgements: This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grants SFB 1102: Information Density and Linguistic Encoding (http://www.sfb1102.uni-saarland.de/) and EXC-MMCI: Multimodal Computing and Interaction (http://www.mmci.uni-saarland.de/). We are also indebted to Peter Funkhauser (IdS Mannheim) for his continuous support in questions of data analysis.

References


Khamis, Ashraf; Stefania Degaetano-Ortlieb; Hannah Kermes; Noam Ordan; and Elke Teich. 2015. A resource for the diachronic study of scientific English: Introducing the Royal Society Corpus. The 8th International Corpus Linguistics Conference (CL), Jul 21–24, 2015. Lancaster University, Lancaster, UK.


Taavitsainen, Irma; Peter M. Jones; Päivi Pahtaa; Turo Hiltunen; Ville Marttila; Maura Raita; Carla Suhr; and Jukka Tykkö. 2011. Medical texts in 1500–1700 and the corpus of Early Modern English Medical Texts. Medical writing in Early Modern English, ed. by Irma Taavitsainen and Päivi Pahtaa, 9–29. Cambridge: Cambridge University Press.

Confederate textbooks – an unexplored source of historical dialectal data

According to Kopp (2009), an exceptionally large number of English textbooks were published in the territory of the Confederate States of America in the years 1861-1865. The number of these publications, excluding reprints, amounts to at least 136 textbooks; they include prescriptive grammars, spellers, readers or textbooks on Southern geography, which, apart from instruction, contain texts aimed at preservation of Southern way of life and strengthening Southern patriotism. This patriotism, or rather Southern nationalism, has been discussed at length in existing works on textbooks published in the Southern States during the times of Civil War (for instance, Bernath 2010, Kennerly 1956, or Kopp 2009). These works, however, do not analyze language norms that can be found in the aforementioned primary sources and their authors merely make scant comments on certain linguistic features considered erroneous at that time. Traits that were deemed not to appear in the language of “cultured speakers” must have enjoyed popularity in coeval spoken language. In other words, such grammatical, stylistic, or pronunciation “errors”, so much frowned upon by the 19th-century Southern purists are, in fact, dialectal features typical of (historical) Southern or African American English (*pen-pin* merger, counterfactual *liketa*, *a*-prefixing with *-ing* forms).

To my knowledge, up to date the only scholar who culls linguistic data from the Confederate textbooks is Montgomery (2004). The present essay builds upon Montgomery’s keynote address, focuses on the hitherto neglected sources of dialectal knowledge, and aims at a comprehensive analysis of the prescriptive norms and guidelines found in Confederate publications (mainly accessible at www.archive.org), which might contribute to the ongoing discussion on the incipience of Southern American English(es).

References


Documenting current language use in Zimbabwe: evaluating literary sources as linguistic evidence

In the last thirty or so years, postcolonial Zimbabwe has witnessed demographic, social and economic changes amidst political upheaval, changes which have been documented in writing by Zimbabweans of all colours and races, a great deal of it in English. These depictions of places, people and events prompt the consideration of how far fictional constructions of the language use of Zimbabweans might be regarded as data on the language variation that ranges across race and class. In this paper, I evaluate literary material as evidence for the linguistic construction of hitherto undocumented varieties of English in Zimbabwe. To do so, I draw upon a range of contemporary literary texts and then evaluate their utility by inspecting the spoken narratives that I have collected from Zimbabweans.

Big Data and Reproducible Research: Strategies, Tools and Workflows

One of the hallmarks of scientific research is that its results are reproducible. Traditionally, the responsibility for ensuring that the results of an analysis are reproducible has been borne by the materials and methods sections of a scientific article. Within the last few years, the failure to replicate high-profile studies in a number of different fields has made it painfully obvious that the traditional means of ensuring reproducibility are not up to task. This is especially the case with research involving “big” -- or even relatively big -- data, where the traditional methods section is simply too short to allow a full discussion about what data was selected for inclusion in the analysis, how it was selected, and what computational procedures were used to obtain the published results. In this paper, I will present five strategies designed to make analyses involving big, medium, or small data reproducible and show how they can be implemented in R Studio, an IDE (integrated development environment) for the statistical programming language R. While these tools and practices have become increasingly common in certain fields and disciplines, they are not yet widely known - let alone practiced - within the humanities. I will conclude the paper with some thoughts about the obstacles preventing the strategies I outline from becoming standard practice within the humanities and will offer some suggestions for how we might overcome those obstacles.
The expressions of betrayal in American political discourse

Moral outrage and expressing it is a significant feature in political persuasion. Lakoff (1996) has claimed that in the American political arena, moral inference is strong in discourse and that conservatives and liberals differ from each other regarding morality. The lexis of moral expressions, however, has not been studied extensively and quantitatively (see however Ahrens 2011).

In this study, I shall explore the use of three verbs: betray, abuse and sell out. I will investigate if conservative and liberal opinion media or the Presidents of The United States use these terms differently:

(1) Are there any quantitative differences in the conservative or the liberal usage?
(2) What is betrayed or abused, is there a context difference?

I will evaluate the value of big data in the study of political language. Can large corpora help in discovering unseen aspects of political discourse? As data I will use corpora compiled of American conservative and liberal opinion articles in the years 1993-2013 and a corpus of Presidential speeches. They are as follows:

(1) Opinion pieces on the Presidents’ politics (Bill Clinton, George W. Bush, Barack Obama) (2.5M words)
(2) Opinion pieces on social issues (education, immigration and equality) (0.6M words)
(3) Opinion pieces on conservative or liberal ideology (0.8M words)
(4) Presidential Radio Addresses by B.Clinton, G.W. Bush and Obama (0.4M words)

There are 684 expressions of betrayal in the data. Preliminary results indicate that liberal media and Democratic Presidents use more of the lexis of betrayal than conservatives. Topics that liberals refer to with betrayal most are civil rights, abuse in prisons (torture) and rights of privacy. It also seems that there is more self-criticism applied with expressions of betrayal in liberal media than in conservative media. The results seem to confirm the old common sense wisdom of practical, optimistic conservatives and ideological, pessimistic liberals.

References


The political opinion articles. Unpublished corpus. Compiled by Sanna Franssila.

“How now, sir John?”: Locating social class in Early Modern drama

In this paper, I argue that social class can be identified and understood through the use of vocatives. Jonathan Culpeper and Dawn Archer (2003) claim that using nouns in the vocative mode is an unwise measure for identifying gentry-class figures in spoken discourse; however, as I demonstrate in this paper, new resources in the form of the Historical Thesaurus of the Oxford English Dictionary (HTOED; 2010) and the public-domain release of the first 25,000 hand-keyed and machine-readable texts from the Early English Books Online Text Creation Partnership (EEBO-TCP, phase I, 2015) make it possible to challenge this claim.

In the second part of this paper I ask if vocatives marking for class status correlate to character construction in early modern dramatic storyline. Character construction is a highly relevant process in dramatic storylines, in which individuals are introduced and information is gathered about them, which would then presumably drop off as characters become more established. Dramatic structure and narrative arcs are driven by storylines; Freytag’s (1863) pyramid is based on patterns he found in Shakespearean and Greek tragedy; many scholars would argue that issues of social mobility underlie the storylines of many early modern plays.

I use the HTOED to identify vocatives marking for social class and apply these to a sample corpus of Shakespearean drama. I then use a corpus of 332 Early Modern London plays curated from the EEBO-TCP initiative to explore issues of dramatic structure. Through a combination of concordance plots and close textual analysis, I examine whether vocatives can be used to identify characters of different social classes and test whether they have a role in character construction.

References


NEVALAINEN, TERTTU, and HELENA RAUMOLIN-BRUNBERG. 1996. Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence. Amsterdam: Rodopi.
Analyzing non-native English letter-writing by Swedes in the 17th and 18th centuries: How ‘small’ and ‘big’ data meet

The study of early non-native English writing in Sweden is only in its infancy. This paper introduces a project intended to move this field forward by cataloguing, transcribing, and analyzing texts written in English by Swedes in the 17th and 18th centuries. Part of the paper investigates the use of English in a series of letters from 1718 by a Swedish administrator, Johan H. Frisenheim. The purpose of the project is to explore the English language produced by non-native language users at a time when the English language was still in early stages of its subsequent global expansion. In particular, the research project focuses on the morphology and syntax of the 18th-century material, as well as postulated interference from 18th-century Swedish.

Currently, transcribed English manuscript material includes letters from 1653–4 by Count Christer Bonde, Sweden’s ambassador to Britain, a travelogue and personal letters by a young nobleman, Edvard Gyldenstolpe, from 1698–9, a series of letters from a young student Johan Schult to the Royal Society from 1703, and a travelogue by Henrik Gahn, a young surgeon, from 1772. Although the materials total only 30,000 words for the time being, I argue that exploring ‘small data’ is an essential first step, launching work on this variety of early English when large-scale materials are not yet available.

More specifically, the paper analyzes the use of English in Frisenheim’s letters to Governor Hugo Hamilton. The example below points to the frequent use of Swedish words and expressions, such as the noun feltwebel ‘sergeant major’ and the verb berättar ‘tells’.

My Lord, Last night at 12 a Klock came on here a feltwebel with 22 Men, and so many horses [men] from Drontheim Comanded at Capt: Longstrom to Come hither, but upon the ways, the Enemy lying in Embusquades, fired upon this party, and Schot Capt: Longstrom, what the fältwäbel berättar I Send herewith, he and none Knows what the Capt: his orders has beene, only he tould to these men that they Schould Safe themselves as good as they [... ] doe (Letter from Johan Henrik Frisenheim to Governor Hugo Hamilton dated December 4, 1718)

References


Big data for the analysis of language variation and change

In this presentation, we argue that compiling and analyzing multi-billion word corpora of natural language data harvested from social media sites online allows for a variety of new and important research questions relating to language variation and change to be pursued for the first time. To demonstrate how adopting a big data approach to data collection opens up new avenues for linguistic research, we briefly present the results of a series of linguistic analyses based on an 8.9 billion word corpus of American Twitter data, which was collected between October 2013 and November 2014 using the Twitter API. This corpus contains nearly 1 billion tweets written by over 7 million different users and each message is time-stamped and geo-coded with the precise time and longitude and latitude of the user when they posted that message. In particular, we look at how big data can be used for conducting research related to the identification and analysis of the spread newly emerging words, the analysis and mapping of highly uncommon linguistic forms (e.g. double modals), the large-scale systematic mapping of common words, and the development of automated methods for geographical linguistic profiling. Based on these studies we show how big data can be used for conducting exiting new types of linguistic research and how this approach to analyzing language variation and change can overcome several limitations with traditional corpus-based research, as well as research based on data elicited through linguistic surveys and interviews.
The ICE metadata: A window to the past? An exploratory study of Hong Kong English

Some ICE corpora provide rich metadata with biographical information on the speakers (including information on age, gender, mother tongue, etc.). The potential of the ICE metadata for research has not yet been fully realised. Hence there are only a few studies which make systematic use of the ICE metadata to study variation and change in New Englishes (Fuchs & Gut 2012, Höhn 2012, Lange 2012, Schweinberger 2012, Fuchs & Gut 2015). The ICE-Hong Kong (ICE-HK) metadata have not been included in studies so far, although information on speaker age may offer a window to the past by studying changes in apparent time. The use of the metadata may thus partly compensate for the lack of diachronic corpus data for Hong Kong English. The aim of my study is to explore the potential of the ICE metadata by studying changes in apparent time in the use of the modal and semi-modal verbs of obligation and necessity in the spoken component of ICE-HK.

My findings show consistent variation in the use of must, have to and (have) got to according to age. The use of the core modal must is gradually decreasing from the oldest to the youngest age group, while the semi-modal have to is on the increase in ICE-HK; in fact, the youngest age group strongly prefers have to over must. My study also reveals that the semi-modal (have) got to is on the decline in apparent time in ICE-HK. Apart from age-specific variation, there is also gender-related variation in the use of must, have to and (have) got to. By exploiting the ICE metadata, my study complements large-scale inter-varietal studies by explaining variation from within rather than against the yardstick of the 'parent' variety.

References


Big data on trial: Researching syntactic alternations in GloWbE and ICE

This study sets out to evaluate the use of big corpora in large-scale variationist studies. Since the arrival of mega corpora (e.g. COHA), researchers have hotly debated their usefulness for linguistic research (see Davies & Fuchs 2015). While supporters of traditional corpora have pointed out the advantages of careful sampling, accurate tagging, and whole-text access of smaller corpora (see Hundt & Leech 2012), more data is needed to investigate low-frequency phenomena. Davies and Fuchs (2015) have argued for the use of GloWbE in addition to traditional corpora to research cross-varietal phenomena.

To corroborate this, we looked at two very well-studied syntactic alternations – the dative alternation (1) and the genitive alternation (2) across four varieties of English (British English, Canadian English, Indian English, and Singaporean English) and compared results from GloWbE – which samples data from the world-wide-web – to findings from ICE (1-million-word corpora).

(1) Dative alternation
   a. That will give [the panel]_recipient [a chance]_theme to expand. [ICE-GB:S1B-036] (ditransitive dative)
   b. [T]hat gives [a chance]_theme [to Bhupathy]_recipient to equalise the points. [ICE-IND:S2A-019] (prepositional dative)

(2) Genitive alternation
   a. [Singapore]_possessor’s [small size]_possessum meant it could be quick to respond to changes. [ICE-SIN:W2C-011] (s-genitive)
   b. the [size]_possessum of [the eyes]_possessor is to help them at night. [ICE-GB:W2B-021] (o-genitive)

Research has shown that speakers of different varieties of English choose a grammatical variant based on underlying probabilistic constraints (e.g. Bresnan & Hay 2008, Bresnan & Ford 2010). While the influence of variety on the ordering of constituents is known (e.g. Bresnan & Hay 2008), a preliminary study (Szmrecsanyi et al. 2016) has illustrated that genre differences play only a minor role. Thus, we hypothesize that the data from GloWbE and ICE should yield the same probabilistic distributions. The cross-varietal differences are analyzed using mixed-effect logistic regression analysis and conditional inference trees.

References


Davies, Mark, and Robert Fuchs. 2015. Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE). English World-Wide 36.1–28.


In the little over a decade since its launch in 2001, Wikipedia has become the largest and most widely used encyclopaedia in the world. Currently, the English-language Wikipedia comprises more than 4.75 million articles. Organised on the principles of shared authorship and crowdsourcing, Wikipedia benefits from contributions by more than 75,000 registered volunteers and countless occasional editors.

Despite being widely used, the status of Wikipedia in higher education (HE) settings is somewhat problematic. Along with concerns about their factual inaccuracies and lack of stability, Wikipedia articles also show considerable variation when it comes to register and style, due to their varied topics and the heterogeneous background of editors and contributors. This being the case, the Wikipedia article – which Kuteeva (2011) calls a “new academic genre” – is likely to be less suitable as a model of academic style, which HE students are required to master, than published expository texts, including traditional encyclopedia texts written and edited by experts. That said, Wikipedia does provide extensive writing guidelines in its Manual of Style, and new and edited content is monitored in a variety of ways.

This exploratory paper takes a data-driven approach to assessing academic style in Wikipedia articles. Using the Westbury Lab Wikipedia Corpus (see Shaoul and Westbury 2010) of slightly less than 30 million articles, and the well-known Academic Word List compiled by Coxhead (2000), we employ methods of statistical data analysis to classify samples from the corpus according to the frequencies of academic words. The unsupervised classification procedure will group the articles according to academic content regardless of topic, which allows us to measure register-specific similarities. We shall also address some methodological issues with this type of data, including the classification of Wikipedia articles and the treatment of texts of different length.

References


Charting the social margins: Fallen women in late 19th-century English newspapers

Historical newspapers provide the contemporary linguist with data in large quantities, yet the challenge lies in the way the material can be mined to map the previously unmapped. Our paper stands at the crossroads between big data and uncharted data: out of a vast quantity of already charted material we aim to find the untold pragma-semantic story of the socially marginalised.

One way of defining the social margins is to look at how the members of the lowest classes were linguistically evaluated. We will, firstly, chart the lexico-semantic field of prostitution and the conceptual relations between terms for prostitutes and metaphorically related concepts in England during the late 19th century (for conceptual metaphor, see Cameron & Deignan 2006 and Rakova 2003). One of our aims is to look at a set of terms used for prostitutes and to determine how and to what extent they map conceptual parallelisms within the larger social tapestry of value relations, and whether these conceptual relations display a discernible regularity.

Secondly, our aim is to study the terms from a socio-pragmatic perspective by looking at how the social status and character of these “fallen women” are expressed in public writings of the time. Moreover, we will focus on how the terms in newspapers could be used to oppose prostitutes and keep them as a group of their own outside respectable society (for newspaper language, see Conboy 2010).

We will look at the terminology relating to prostitutes in English newspapers published between 1850 and 1900 (19th Century British Library Newspapers Database). The study is a continuation to our previous research on the pragma-semantic variation and change in the terms used of common prostitutes in early and late modern England (Nevala & Hintikka 2009).

References


Mirka Honkanen (University of Freiburg)

Multilingual resources and authentication in the computer-mediated communication of U.S.-Nigerians: The role of African American Vernacular English in the repertoire

In the past 30 years, new forms of mass migration and the rise of the Internet have cut the ties between particular geographical locations, communities, and linguistic varieties taken for granted in earlier dialectological and variationist sociolinguistic studies. Inspired by the sociolinguistics of globalization advocated by Blommaert (2010), my PhD project explores one context emblematic of this mobility of people and linguistic resources, examining the computer-mediated communication of Nigerians in the United States. The aim is to describe the role of different varieties of English and other relevant languages in the digital ethnolinguistic repertoires (Benor 2010, Heyd & Mair 2014) of U.S.-Nigerians, and how they employ their multilingual language resources in negotiating their race, ethnicity, identity, and authenticity in written online interactions.

My dataset is in the form of a very large corpus (>840 million words) downloaded from the Nigerian online discussion forum nairaland.com, spanning 2005–2014. Due to the pervasive multilingualism and nonstandard orthography present on the forum, parsing or tagging the corpus was not considered viable. However, the data can be accessed via an online search interface NCAT (Net Corpora Administration Tool), which allows for searching for strings or substrings, and restricting the search to certain users or time periods. My analysis zooms in on the most active members of the community of practice who are located in the U.S., describing their multilingual repertoires mainly qualitatively but also quantitatively when the ‘messy’ data allow for it. The second strand of investigation, focusing on the negotiation of authenticity and processes of “authentication” (Bucholtz 2003), relies on detailed qualitative analyses of chosen threads of discussion. In this presentation of a work in progress, I will present my first findings, zooming in particularly on the role of African American Vernacular English in these U.S.-Nigerians’ repertoires.

References


Manuscript abbreviations in Latin and Middle English medical treatises: a corpus-linguistic approach

Medieval manuscripts used an abbreviation and suspension system, in which words or morphemes were frequently substituted with logographic elements. While these abbreviations and their expansions constitute useful data, their corpus linguistic potential is mostly unexplored, as most historical corpora inherit the practice used in printed editions of expanding them silently.

In the study, I examine differences in abbreviation for lexical words (nouns and verbs) and function words (articles, prepositions, conjunctions, adverbs and pronouns) in Latin and Middle English. The data consists of five manuscripts which are part of the Voigts-Sloane Group of medical manuscripts, characterised by a standard bilingual anthology of twelve texts. The Voigts-Sloane manuscripts date from ca. 1450–90, and were connected to commercial book trade in London shortly before and after printing. The corpus is parallel, which makes it possible to contrast how different scribes treated lexical and function words in the same text, as well as differences between Latin and Middle English. I use markup based on TEI P5 XML, in which both the abbreviation and its expansion are encoded, which makes it possible to examine both via quantitative corpus linguistic approach.

Results of a pilot study suggest that the abbreviations used for function words in the Voigts-Sloane Group tend to vary from manuscript to manuscript, revealing that each scribe had an individual repertoire, whereas the abbreviations used for lexical words are more uniform, suggesting the scribes reproduced the forms in their exemplar. This can be explained by the known origin and provenance of the manuscripts: the scribes, who were not medical practitioners themselves, followed the exemplar closely with medical terminology, but each scribe had his own personal inventory of abbreviations for function words.
Exceptions to Bach’s Generalization in Regional Varieties of English: a Case Study with Data from the GloWbE Corpus

Consider (1a-b), from the corpus of Global Web-Based English, GloWbE:

(1)

a. She strongly warned me against eating any of the foods listed above [...] (US)

b. But Abe warns against upgrading too soon because not all plugins are supported. (US)

(1a) involves object control with an overt controller (me) but (1b) lacks an overt controller, in violation of Bach’s Generalization. This generalization is that “in object control structures the object NP must be structurally represented” (Rizzi 1986: 503). The pattern of (1a) may be termed the overt object control pattern, and that of (1b), the covert object control pattern.

It has been noted in work on prepositional gerunds selected by verbs such as warn in recent American English that with such gerundial complements, violations of Bach’s Generalization have become more frequent than “regular” constructions with expressed objects, which observe the Generalization. In the present study both overt and covert constructions selected by warn are investigated in four subsections of the GloWbE corpus (the sections from web sites located in the USA, the UK, the Philippines, and Pakistan). The purpose is to find out whether violations of Bach’s Generalization are as widespread in other regional varieties as they are in recent American English and whether generalizations can be established among different regional varieties of English, relating to the inner and outer core. The results here shed light on an aspect of the Great Complement Shift in different regional varieties. A second major objective is to inquire into the nature of understood objects in the covert construction in different regional varieties. Here the central issue is whether understood objects in violations of Bach’s Generalization should be assigned general or specific interpretations, and whether regional differences exist with respect to the interpretation of such understood objects.

References
Tyler Kendall (University of Oregon)

Making old data sources into new data sources: On the aggregation of sociolinguistic datasets and the future of real-time and cross-study analysis

More and more data have become available in recent decades for the study of the English language, and for sociolinguistic work more generally. One important source for “new” data come from the fact that sociolinguistic data themselves are rapidly accumulating as sociolinguistic data acquisition (e.g. research based on sociolinguistic interview recordings; Labov 1966) enters its sixth decade and the sociolinguistic research community continues to grow. The aggregation of sociolinguistic datasets represents a key opportunity for advancing real-time investigation of language change on English and other languages. It also facilitates “cross-study” studies, the direct empirical comparison of datasets collected for different original projects. The aggregation and comparison of disparate datasets raise challenges, in addition to opportunities, for empirical research (see contributions to Yeager-Dror & Cieri, eds. 2014).

In this presentation, I survey a series of projects connected to the Sociolinguistic Archive and Analysis Project (SLAAP), a web-based sociolinguistic data repository and analytic toolkit housed at North Carolina State University in the U.S., with a range of affiliated research groups in North America. SLAAP houses over 4,000 recordings (as well as metadata, transcripts and annotations, publications, and other materials) from over 50 separate projects, from foundational projects of the 1960s to smaller Masters-level student projects of just a few recordings. In addition to its archival and data sharing features, SLAAP provides a suite of analytic tools to aid in common sociolinguistic (and more general audio-based) research tasks. As I focus on in this talk, by aggregating a wide-range of data collections and building specialize tools and interfaces, SLAAP allows for large-scale corpus-based inquiry into a range of old and new questions. The presentation surveys several recent investigations made possible through the SLAAP archive and software and introduces some of the new projects being undertaken in connection to the archive.

References


Commonplace books: Charting and enriching complex data

Medieval English commonplace books are collections of texts which were compiled for future reference or further use and contain a large variety of genres from different spheres of life (such as medical recipes, letters, proverbs, obituaries, chronicle entries, legal texts, prognostications, poems etc.). Texts in commonplace books reveal fairly complex communicative conditions. Apart from the arrangement for the sake of documentation, the texts are supposed to be re-used in several ways (both in terms of text reception and text production), but also reflect genre conventions from quite disparate domains (religious, administrative, medical, private etc.). Seen from a corpus-linguistic perspective, commonplace books are largely uncharted texts and ideal candidates for ‘communicatively enriched’ corpus data.

Commonplace books play only a peripheral role in current diachronic corpora (see, for example, the short extracts from the Commonplace Book of Robert Reynes of Acle (CMREYNES) in the Helsinki Corpus). Many of them are not even edited and those that are emphasise the ‘literary’ passages they contain (see, for example, Dybosky 1908, Robbins 1955, Rigg 1968; but see also Louis 1980, which covers all entries of Robert Reynes’s commonplace book).

The texts of commonplace books can be ‘enriched’ in several ways. First of all, metalinguistic information can be given on the context of the compilation process and the documentation / arrangement of the different texts in the collection. Secondly, the texts may be seen in the context(s) of their potential re-use(s), for example, not only as documentation but in further acts of text reception and text production. This is relevant, for example, because text sections may change their functions and pro-forms their reference, according to the different uses (see Kohnen 2011). Thirdly, the individual texts can be seen against the background of the different contemporary genres and genre conventions they reflect.

In my presentation I will first give an overview of the uncharted area of medieval English commonplace books and survey the field of edited (and unedited) texts that might be eligible for a diachronic corpus. I will then illustrate the different ways in which the texts may be enriched when edited for corpus-based analyses, following the three levels pointed out above. Thus, this paper seeks to contribute to a systematic and comprehensive account of the factors that are relevant for a ‘pragmatic enrichment’ of corpus texts.

References


Mikko Laitinen, Magnus Levin and Alexander Lakaw (Linnaeus University)

Advanced non-native English on a continuum of Englishes: Charting new data sources

This presentation discusses research in which the objective is to find new empirical and theoretical ways of approaching the ongoing globalization of English. A particular angle is to test the usability of corpus-based diachronic methods for studying advanced non-native Englishes and to investigate present-day non-native use as one stage in the long continuum of Englishes. Such an approach is motivated by calls from the English as a lingua franca (ELF) domain to provide diachronically-informed evidence of English in multilingual settings (Seidlhofer 2011) and by recent attempts in the study of indigenized World Englishes to take into account diachronic processes in shaping the outer circle Englishes (Noël, van Rooy & van der Auwera 2014). In particular, we investigate how ongoing grammatical variability, which is widely documented in many native varieties, is adapted in advanced non-native use. A key question is to investigate to what extent multilingual settings contribute to ongoing variability.

The presentation discusses requirements for sources of material and evidence, and its starting point is the fact that the ELF research has so far focused on meaning making in interaction, which is also reflected in the scope of corpus resources. We zoom into ongoing corpus compilation work in which the aim is to collect a representative multi-genre sample of English texts in multilingual settings. The objective is that the sampling frame should enable diachronic and diatopic analyses of advanced non-native use and make possible quantitative comparisons between our evidence and some of the existing English corpora, both native and non-native. The presentation discusses the diverse nature of our data and presents how we turn the data into evidence. We will introduce the set of grammatical structures, stemming from the corpus material, which have so far been investigated, and discuss a set of broader research questions to which this type of multi-genre corpus material of English texts in multilingual settings could shed more light.

References


The present paper aims at investigating the discourse strategies characterizing the documents which were released by the Vatican Council II in the years 1962-65. The main objective of the analysis is to investigate how language relates with the structures of power, authority and ideology in religious prose. Starting from the assumption that the study of transitivity and passive forms can give insights into the key stylistic features of a text (Simpson 2004), a corpus of the second Vatican Council documents (VC-II corpus) was compiled and automatically parsed by using the Visual Interactive Syntax Learning (VISL) language analysis tools (http://beta.visl.sdu.dk/) which can provide both syntactic and semantic information on a given text. The corpus consists of three different subcorpora each containing complete texts in the form of decrees, declarations and constitution. All the documents released during the second Vatican Council were included and the overall corpus brings up 169,000 words. A smaller control corpus of documents issued by the first Vatican Council in 1869-70 was also created and annotated to carry out a comparative analysis. Following a procedure adopted in a recent study on the automatic detection and extraction of syntactic structures expressing modality in religious prose (Bisceglia, Calabrese & Leone 2014), we have matched corpus-based evidence and linguistic diagnostics in order to: 1. Identify the main text functions in terms of exhortation, exposition, narration and argumentation; 2. Determine the degree of recursion and innovation, authority and liberality underlying/characterizing our data; 3. Compare the main principles and methodological procedures of Critical Discourse Analysis (CDA) with the corpus data.

The aim is to provide insights into the interpretation of the documents of the second Vatican council and to underline their innovative value for both theologians and non-specialized readers.

References


Studying phrasal verbs in a corpus of engineering English

Phrasal verbs have attracted a considerable amount of attention over the past thirty years. They have been described as a common feature of English frequently found in spoken or informal contexts. While much of the discussion in the literature is based on phrasal verbs in general English, the pattern of the use of phrasal verbs in more specialised contexts is under-researched.

This study uses a 9.2-million-word corpus, the Hong Kong Engineering Corpus to examine the use of phrasal verbs in engineering English in terms of their frequencies, forms and functions, and examines the extent of genre specificity of phrasal verbs. This profession-specific corpus is formed by 31 genres used in the field, such as agreements, code of practice, consultation papers, reports, speeches, etc. Phrasal verbs were extracted from the sub-corpora using Wmatrix (Rayson 2009) and WordSmith Tools (Scott 2012). ConcGram (Greaves 2009) is used to perform concordancing as it is able to show all possible configurations in the concordances of a single search. As different from traditional research on phrasal verbs, this study treats different inflectional forms as distinct phrasal verbs and investigates whether they have the same or different co-selections.

The methodology and results of the study have research and pedagogical implications for examining the use of words and phrases with big data. The study, in particular, provides insights into the authentic patterns of language use in the engineering industry.

Acknowledgements: The work described in this paper is substantially supported by a research studentship funded by the Research Grants Council of the Hong Kong Special Administrative Region (Project No.: 4-RTY2).

References


Adjusting $p$-values for heterogeneity in collocation analysis

Although heavily criticised, the use of $p$-values is pervasive across the sciences. A crucial problem associated with the use of $p$-values is that equal $p$-values from the same test may have different degrees of support from the data, especially in the case of heterogeneous data. For instance, if one value is supported by several samples and the other only by a single sample, the latter is much more uncertain. Corpus data, both big and rich, is often heterogeneous: texts have different authors, topics, audiences, etc.

For keyword analysis, there are statistical tests that account for the variance—also called uncertainty, burstiness, or poor dispersion—caused by heterogeneity (Bestgen 2014; Lijffijt et al. 2014). For identifying collocations, no such tests currently exist. Interestingly, the statistical problem is equivalent to that of META-ANALYSIS, as done in medicine, where Cochrane Reviews require researchers to compute a measure to assess heterogeneity (Higgins et al. 2003). However, the actual evidence level is computed using Fisher’s Exact Test, which is similar to the log-likelihood ratio test commonly used in collocation extraction (Evert 2009).

We present an algorithm that takes as input a statistical test and a data set, and gives as output an adjusted estimate of the $p$-value for that test. This estimate is more robust than the original if the assumptions of the test are not met by the data, e.g., when variance is too high. Comparisons of the log-likelihood ratio test and our method show that heterogeneity matters when assessing the strength of association between words in corpora. Our algorithm could thus be used to complement significance-based association measures for more reliable results.

References


LIJFFIJT, JEFREY; TERTTU NEVALAINEN; TANJA SÄILY; PANAGIOTIS PAPAPETROU; KAI PUOLAMÄKI; and HEIKKI MANNILA. 2014. Significance testing of word frequencies in corpora. Digital Scholarship in the Humanities 29(4). doi:10.1093/llc/fqu064
The perception of changes in the English language among members of the general public differs considerably from the observations of linguists analysing large-scale diachronic corpora (see Mair 2006; Hilpert 2007). Members of the general public tend to focus on single usage features in their anecdotal observations of language change, yet they often fail to notice general linguistic developments. The prescriptive tradition is one of the principal causes for a discrepancy between lay and expert observations. Due to the practices of reproducing prescriptive language rules in usage guides, in school settings and in public discussions, minor changes in usage, such as the spread of the ‘split infinitive’ construction, keep being commented on, as in the following example taken from a letter to the editor of The New York Times.

Split infinitives are like putting an insurmountable obstacle in the way of a speeding train that must stop to clear the tracks before picking up speed again. We lose the thrust and impact when we separate preposition and verb, for an infinitive by its very nature needs to be taken as one unit.

(16 June, 2009)

The practice of complaining about ‘bad usage’, often referred to as the ‘complaint tradition’ (Milroy & Milroy 2002:24–46), in what are in effect bottom-up standardisation efforts has been enabled through media outlets which offer forums for the general public. In this paper I analyse linguistic complaints in a self-compiled Complaint Tradition Corpus consisting of linguistic complaints taken from three different registers: letters to the editor, comments on news websites, and blog comments. This paper explores the range of linguistic features identified by the authors of complaints: which linguistic features are identified as problematic and why? Furthermore I analyse the register differences in the Complaint Tradition Corpus (letters—blog comments—news comments) by drawing on the findings from the multidimensional analysis approach (e.g. Biber & Conrad 2009). The register differences in the language discussions reflect general differences in public discussions in traditional and new media.

References


HILPERT, MARTIN. 2007. Just because it’s new doesn’t mean people will notice it. English Today 23.29—33.


Making sense of heterogeneous data with a relational database of Old English

In spite of the wealth of data on the Old English language collected for, at least the last one hundred and fifty years, the different traditions, the various sources as well as the divergent readings, have resulted in a vast amount of information that is sometimes inconsistent and very often hard to interpret. Against this background, one of the aims of the Nerthus Project (www.nerthusproject.com) is to turn the heterogenous data on Old English into significant information and to disseminate it on the Internet by means of the resources and procedures characteristic of Digital Humanities. The data are processed by means of a relational database called The Grid, which comprises six blocks of related information (layouts), including an unlemmatized part, a lemmatized part, a fragment concordance, a word concordance, a reverse index and an index of secondary sources. The Grid relies on several lexicological and lexicographical sources. On the lexicological side, it draws on various secondary sources for the study of Old English morphology, syntax, semantics and dialectology, as well as a number of seminal works in the field of the old Germanic languages. On the lexicographical side, The Grid is based on the standard dictionaries of Old English, An Anglo-Saxon Dictionary, A Concise Anglo-Saxon Dictionary and The student’s Dictionary of Anglo-Saxon. The attested forms have been retrieved from The Dictionary of Old English Corpus. After a presentation of the building blocks of the database and a discussion of the relations holding across the layouts, the conclusion is reached that it is necessary to refine the system used for relating alternative spellings and inflections to the relevant headword so that the visibility across the layouts can be maximized.

References
James McCracken (Oxford University Press)

Integrating historical frequency data with the Oxford English Dictionary: What can it do, and what are the limitations?

The Oxford English Dictionary (OED) may be regarded as a forerunner of ‘big data’ in English language research, anticipating large-scale data collection, analysis, synthesis, and crowd-sourcing, albeit conducted in a much more manual way than is possible now. As well as being a human-readable dictionary, the OED can be mined as a database of various features of the English lexicon over time, including etymology, orthography, meaning, topic, and usage. In addition, the recent integration of the Historical Thesaurus (historicalthesaurus.arts.gla.ac.uk) gives a powerful concept-oriented dimension to the OED.

But the utility of this data has always been severely limited by the failure of the OED to provide anything more than the most impressionistic idea of changing frequencies over time, whether for lemmas or for senses. The release of the Google Books Ngram data sets in 2009 and 2012 provided a chance to begin to remedy this (at least back to the late 1700s). In this presentation I discuss attempts to map the Google Books n-grams to OED lexemes, the benefits of this marriage, and the necessary limitations and approximations involved. I demonstrate some of the uses to which statistically-informed OED data can now be applied. These include exploration of the changing ‘shape’ and composition of the English language over time; and quantitative and stylistic analysis of historical texts.

Finally, I look ahead to opportunities to improve the integration of OED with corpus statistics, to cover more of the history of English (i.e. before the late 1700s), and to re-analyse diachronic frequency data in terms of individual senses rather than lemmas.

Joe McVeigh (University of Jyväskylä and University of Helsinki)

Too Much to Say in One Subject Line...: Exploring trends in Big Data marketing communication

Big Data collection and analysis is exploding throughout every field of research. Now that many forms of communication have gone electronic, the possibility to research them is greater than it has ever been. With Big Data there is also the possibility to correlate the linguistic data with metadata to uncover interesting patterns. And in certain cases, linguistic analyses of Big Data sets can have benefits outside the field of linguistics, such as in marketing, where there is a substantial economic value placed on language. Linguistic descriptions of marketing data therefore have a commercial appeal since they can be applied directly to the creation of future texts. This paper researches a corpus of 33,000 marketing email subject lines, which were together sent over five billion times. The subject lines are coupled in the analysis with metadata, such as how many emails were opened and the sales figure for each email. The texts exhibit interesting linguistic features, especially in terms of non-standard variations in spelling (bbbbbBBBLACK is bbbbbBBBACK) and even spacing (S A L E !). These variations are investigated to show which of them correlate to the relative economic success of each subject line. The challenges of dealing with non-standard features in Big Data are also discussed. The results suggest another layer of linguistic analysis for research on Big Data sets.
‘To be sold, a parcel of valuable Negroes’ and ‘Ten Dollars reward for TOM, the property of Mrs. T O Elliott, who absconded yesterday’: Slaves in the advertising sections of 18th and 19th century American newspapers

Starting with the first weekly newspapers in colonial Boston in the early 1700s and continuing up to the end of the Civil War in the case of the Southern states, North American newspapers offered slave owners a convenient way to inform the public of matters concerning their human property. Potential buyers and sellers regularly advertised in the papers, as did slave owners placing notices for fugitive slaves and sheriffs notifying the public of captured runaways. The online database America’s Historical Newspapers allows access to digital facsimiles of thousands of early American newspapers, thus also providing researchers with an abundance of advertisements in their original context. From this database, I have collected the slavery-related notices from a sample of approximately 250 newspaper issues, representing papers from Massachusetts, New York and South Carolina and spanning from early 18th to mid 19th century. The aim of my paper is to investigate the language used in these advertisements in reference to the slaves, focusing particularly on the ways it can be used to reinforce the idea of slaves as merchandise and chattel, instead of individual human beings. Under scrutiny are for instance some of the most commonly used adjectives for describing the slaves in these various types of notices, as well as the presence and frequency of words and constructions that explicitly demote these people to the status of property.
Alchemical texts can provide new material for exploring e.g. the process of the vernacularisation of science in the late Middle Ages, the study of which has focused on medical texts (cf. Taavitsainen & Pahta 2004). In addition to issues of vernacularisation, the Bacon texts also provide material for the study of multilingualism.

In this presentation I will demonstrate some aspects of ME which can be studied through alchemical texts. This previously under-explored material, which mostly exists only in manuscript form, can increase our understanding of the development of English scientific prose and medieval multilingualism, as well as provide new material for historical corpus linguistics.

References


TAAVITSAINEN, IRMA, and PAIVI PAHTA (eds.) 2004, Medical and Scientific Writing in Late Medieval English. Cambridge: Cambridge University Press.
Mono-collocates: How fixed Multi-Word Units (MWUs) with *of* or *to* indicate diversity of use in different corpora

There is evidence that particular words in the English language do not freely collocate; similarly not all words of a single word class fit in the same colligational structure. Thus, John Sinclair and others pointed out that commonly found structures often rely on a specific node word. Corpus-based investigations provide evidence for this: “You can talk about 'an atomic explosion', but you do not say, 'The explosion was atomic'. You can talk about 'outdoor pursuits', but you do not say 'Their pursuits are outdoor'” (Sinclair 1990: 80). Francis, (1993: 140f.) also points out constructions which are “accounting for over 98% of all the citations of the structure in the corpus.”

In a detailed investigation of the uses of the items *of* and *to* (Pace-Sigge 2015) it has become apparent that there are, in fact, a number of node words that are close collocates with these items. An example is *able* which only appears in the construction *able to* in almost all texts. It can be said that such two-word units are, being found far more often than other possible collocates, *mono-collocational*. To name an item mono-collocate, it should appear bound to either to or of eight (or more) out of ten times. Alternatively, a single item can be deemed mono-collocational if other collocates are of low relative frequencies.

In this paper, *of* and *to* usage has been investigated in six corpora: two are casual spoken British English (spontaneous spoken), a further two look at public speeches (prepared spoken), and a final two look at British fiction of the 19th and 20th century. Such mono-collocates appear to create MWUs which are essential building blocks of communication. (see Hoey 2005). The strength in their bonds can either be genre- or even corpus-specific; on the other hand, a number of word units appear regardless of genre.

References


Tracking the meanings of ‘pliis’: combining evidence from multiple sources

In a plenary address at the conference New Ways of Analyzing Variation in 2012, the linguist Sali Tagliamonte used the parable of the elephant to demonstrate how multiple means of looking at just one linguistic variable may offer a focused interpretation of the meaning of the linguistic form in question, as well as any variation and change relating to it. Our work on the borrowed politeness marker pliis ‘please’ in Finnish, like Tagliamonte suggests, has made use of various types of data, including attitude data (see Peterson & Vaattovaara 2014), Twitter data, uncharted “opportunistic data” deriving from an online forum hosted by the Helsinki Transit Authority (Peterson, Hiltunen & Pyykkö 2014), and “traditional” corpus data from FinnishWaC, a large web-based corpus. With this paper, we present the main goals of using these types of data, and the challenges and successes associated with the respective analyses. We propose that, in the absence of “ideal” spoken or naturalistic data, a collection of data sources work together to contribute to the overall understanding of the social, grammatical, and pragmatic meaning of the form in question.

References


Big Data: opportunities and challenges for English corpus linguistics

This paper draws on our experience of working with a large diachronic corpus, namely 1.3 billion words of Guardian and Independent news text, from 1984-2013 and ongoing. Big data is thus, for us, both quantitative and temporal. The data exist as raw text and as analysed databases, created by AVIATOR (1990-3), APRIL (1997-0), WebCorpLSE (2000-) and other tools. We also refer to the COCA corpus (Davies, 2008).

Our research focus is on lexis, and such big data is thus desirable (Sinclair 1991; Linquist 2009). The lexicon comprises a few high frequency words, but many more medium-low frequency words, and a majority of hapax legomena. Big data increases scope and enhances granularity of study, allowing rare and intuitively unaccessible features to be glimpsed (Renouf 1983). Thirty-plus years of diachronic text brings the corpus linguist an evolving understanding of language innovation and change (Renouf 2013; Renouf & Kehoe 2013).

On the other hand, big data presents challenges for the corpus linguist. High and even medium-frequency words and affixes begin to retrieve too much data; hapax legomena, since they are primarily studied collectively at sub-word level, retrieve enormous numbers of tokens for analysis, supplemented by typographical and tagging errors in the corpus sump which, in earlier decades, could be ignored (Renouf, forthcoming). Moreover, whilst it undoubtedly allows microscopic analysis, a very large corpus reveals details of language use which complicate descriptions, and can entice the linguist down time-consuming paths of enquiry which prove fruitless or excessive. At this point in corpus linguistic history, large-scale language corpora are available in advance of the necessary tools for automated analysis.

The paper will try to illustrate some of the opportunities and challenges of big data experienced recently.

References


How to investigate grammatical obsolescence? A case study from Late Modern English grammar

Based on first results from my ongoing PhD project, the present paper discusses an approach to investigating the under-researched (Hundt 2014) topic of grammatical obsolescence. The first variable (subordination of purpose) and its set of variants are presented in the context of a potential change in progress in Late Modern English. The paper draws on evidence from traditional small corpora such as ARCHER and the Brown family and from some of the more recent mega-corpora, such as COHA. This makes it possible to assess advantages and disadvantages of using corpora of various size.

The paper raises challenging data analysis-related questions appearing in the context of declining frequencies of use in corpora. In particular, investigation of constructions which have become rarer in the last few decades may lead to difficulties like data sparsity and general data deficiency. To trace the development over time in such cases, we need more fine-grained data (Gries 2008). Elaborating on these issues, the paper shows to what extent the findings from big and small corpora may confirm or complement each other.

These are two charts showing the correlation of time and frequency per million words in the case of 8 variants of the first investigated variable (Data extracted from COHA (Davies 2010–)):

The hypothesis that the decline observed in the above figures is not a mere fluctuation is further confirmed by correlation testing. The tests show very high negative correlation between the time (measured in decades) and the frequency of occurrence per million words in the COHA, in the case of 2 constructions, high and intermediate negative correlation in the case of 3 further constructions.

These preliminary findings and their potential implications will be discussed in the framework of usage-based approach.

References


Behind the scenes of diachronic corpus studies - uncharted discourse types in historical corpora

Current corpus-based historical research can indulge in a wealth of digitally accessible databases that cover an impressive range of genres. And - as a consequence - it usually does. An issue that is rarely addressed is which discourse types we disregard when we make corpus-based claims about linguistic change that is related to genres. How robust are our findings about genre conventions, genre change and, ultimately, language change in the light of "uncharted" material?

In my contribution I will highlight data that, so far, has remained "behind the scenes" in a twofold sense. First, there is linguistic material that is digitally accessible but generally overlooked. Old English has quite a range of genres which have seldom been in the centre of linguistic interest, for example prognostications or charms. This type of data is behind the scenes in the sense of "simply ignored" for an analysis of linguistic/pragmatic change. Secondly, I will discuss material that is not included in historical corpora, for whatever reason. Here, all sorts of non-institutional discourse types may be named (with the notable exception of private correspondence), for instance, manorial accounts and household management books. This type of data is behind the scenes in the sense of "simply uncharted".

I will provide two case studies that show how such material will qualify or even change our understanding of established pragmatic change phenomena. First, I discuss non-canonical and previously overlooked instructional discourse from the Old English period, contained in the Dictionary of Old English Corpus. Here, I will point out how prognostications further qualify our conception of legalistic language and its functions. Secondly, I will comment on Early and Late Modern household management books. These show that we may have to modify our view on speech act conventions and address term systems when we find, for example, that maid servants are being addressed as sweet hearts (Hannah Woolley (1677) The Compleat Maid-Servant), a form of address more regularly projected to private (love) letters of the period (e.g. in the correspondence of Thomas Knyvett, see Helsinki Corpus CEPRIV2). Thus, I hope to identify linguistic material that may serve as testing ground for change phenomena already measured and thought to be received wisdom.

References

Dictionary of Old English Corpus, see http://www.helsinki.fi/varieng/CoRD/corpora/DOEC/

Helsinki Corpus, see http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/

Woolley, Hannah. 1677. The Compleat Maid-Servant or the young maidens tutor. London.
Anni Sairio (University of Helsinki)

Uncharted data from the social margins: a case study of 18th-century L2 English

Non-native varieties of historical English remain largely unmapped and unexplored (see Sharma 2012, Geisler 2013, and Tieken-Boon van Ostade 2012 for recent work), and in this paper I want to a) investigate the range of linguistic data available for the study of historical English as a second language, and b) explore this topic particularly through the concept of social margins. Second-language speakers who lacked access to linguistic resources, did not have a creditable identity in terms of language competence or cultural background, or occupied a marginal network position in relevant social groups can be characterised as operating in the social margins. I investigate variation and change in the language of a cultural outsider and social riser and the references of L1 letter-writers to this L2 speaker. As L2 material I use the personal letters and published autobiography of an 18th-century Armenian who migrated to England in 1751 to “learn the art of war” (Emin 1792: 140) and spent years doing menial labour before forming important relationships with figures of high society. This Late Modern migrant’s access to prestigious L1 language practices is evaluated through a comparative analysis of 18th-century English spelling variation.

References


TIEKEN-BOON VAN OSTADE, INGRID. 2012. Late Modern English in a Dutch context. English Language and Linguistics 16(2), 301–127.

Gerold Schneider (University of Constance and University of Zurich), Mennatallah El-Assady (University of Constance) and Hans Martin Lehmann (University of Zurich)

Changes in language, topics and society according to the COHA corpus

In this pilot study, we employ the Corpus of American English (COHA), the currently largest corpus of diachronic English in a largely data-driven fashion to describe changes in language, society and technology from 1800 to 2000.

Linguistic Changes in Late Modern English

We first investigate and interpret the linguistic changes in Late Modern English, as they have been described e.g. in Aarts, López-Couso, and Méndez-Naya (2012):

• increase of the progressive passive (see Figure 1)
• increase of the get-passive
• increased use of progressives
• decrease of be as perfect auxiliary
• decrease of periphrastic do in affirmative sentences
• tendency to replace finite complements by non-finite clauses
• replace to-infinities by -ing forms

Figure 1. Progressive passive forms over the COHA period

Noun Compounds as new Concepts and Keywords

One of the most striking changes in English over the last 50, 100, or 200 years, however,
has been the rapid increase of noun-compounds, as described e.g. in Leech et al. (2009), which can be used as a means to compress information (Biber and Conrad 2009). We trace the increase of noun-noun compounds and show that they are a major method to create new concepts, and thus show shifting trends and topics in society and technology. In particular, we show the lifespan of noun-noun compounds by treating them as alternations (Levin 1993), typically derived from noun-preposition-noun constructions. Figure 2 shows the frequency of alternating constructions in COHA.

**Interactive Topic Modeling**

We use the completely data-driven method of Topic Models (Blei, Ng, and Jordan 2003; El-Assady 2015) to visualize topics and changes over time. When using word features, we detect and describe shifting topics over time. There is a particular reason for noun-preposition-noun constructions to condense into noun-noun constructions: a new concept is formed, the relation between the two nouns in the noun-noun construction is no longer ambiguous. We show that noun-noun constructions, although much rarer and thus sparser, are better index terms for topic models, as they are more specific and more precise. They allow us to trace both single events in society, such as the landing on the moon, as well as recurrent crises like wars and economic recession over time. Figure 3 shows a small excerpt from a run with 100 hidden topics on the news section of COHA. Topic Models also allow us to detect changing associations in current debates. As an example, we show how the debate of the *climate change* has developed over time.

![Figure 2. Forms which alternate between noun1-preposition-noun2 and noun2- noun1 over the COHA period](image)

**References**


Uncharted data: African American letters from the 18th and 19th centuries

The aim of this paper is to report on uncharted data in the history of African American English: letters written by African Americans in the 18th and 19th centuries. Due to the high rate of illiteracy, it was often assumed that the few written documents that survived were either edited or written by amanuenses. Only in the 1970s and 1980s were several letter collections from African Americans published by social historians (e.g., Miller 1978, Wiley 1980). The letters presented in this paper are new unpublished material: they have all been recently unearthed in several U.S. archives. It is argued that this type of data is particularly valuable because, firstly, sources for the study of earlier forms of African American English are scarce, particularly for the antebellum period. Secondly, the letters present the language of common people, voices that are not normally heard.

The letters are part of a larger corpus aiming to provide a basis for the study of the origin and development of African American English. Since the corpus is mainly compiled with dialectological interests in mind, the letters were carefully selected to include only those which show vernacular traces, e.g., phonetic spelling and non-standard morphosyntactic features. The letter writers had little or no formal education, could barely write and have often been described as semi-literate. However, a great deal of heterogeneity obtains between these semi-literate writers. While some letters only contain very few occasional vernacular features, others are what Fairman has called 'nowhere-near-Standard' (Fairman 2007a: 172). The focus of this paper will be on a careful analysis of what Fairman labels the five 'literacy skills/subskills', orthography, lexis, grammar, syntax and handwriting (2007b: 268). A selection of letters will be examined to find characteristic traits of each skill and to shed more light on the question of who counts as semi-literate.

References


Not only apples and apples but also apples and oranges: a register perspective on contrastive negation in English

It is widely agreed that different registers draw on different grammatical resources even when expressing the same function; extreme examples are casual conversation and research articles (Biber 2012). Registers that are functionally divergent show robust differences (e.g., Biber et al. 1999) but also raise issues of comparability. Conversely, similar registers display subtle differences in the use of constructions but may lack the full range of variation.

My paper is based on my PhD project on contrastive negation in English. Negation is contrastive when it is juxtaposed with a parallel affirmative expression (Gates Jr. & Seright 1967, McCawley 1991). A basic distinction may be drawn between neg-first (as in example 1) and neg-second (as in example 2) variants of negative-contrastive constructions (Tottie 1991, examples from the BNC):

(1) When, I wonder, did it become fashionable for politicians to talk not about the world but about the planet?

(2) We have to accept this is showbiz now, not a sport.

Studies on contrastive negation have mainly been based on anecdotal or experimental data. In my previous corpus-based work, I have shown that newspaper editorials have a higher prevalence of negative-contrastive constructions than sports news of art reviews (Silvennoinen 2013), but much of the register variation remains unexplored. Negative-contrastive constructions are resistant to automatized queries and require manual searching. This poses obvious problems of workload. To solve these while also getting enough data for further quantitative findings, my research is based on two datasets drawn from the BNC: the entire broadsheet newspaper component and a subset of the conversation component. The newspaper data permit comparisons between closely related registers, such as editorials and art reviews – i.e., apples and apples. The conversation data permit comparisons between speech and writing – i.e., apples and oranges. An overarching concern is marrying quantity and quality in analyzing grammatical data.

References


‘I am roary on the hill’: identity-construction in 19th-century Irish English threatening notices

Throughout the nineteenth century, the rural midlands of Ireland suffered from agrarian violence by ‘secret societies’, whose membership according to contemporary sources consisted solely of the lower ranks of society. In the minds of the ruling Protestants, these so-called ‘Ribbon’ societies were the embodiment of a nationwide conspiracy of Irish Catholics to exterminate all Protestant in the country (Beames 1982). Though there were indeed outbreaks of violence, the Ribbon societies predominantly used fear and intimidation to achieve their goals. One of their means of intimidation was the sending of threatening notices. My paper will investigate a corpus of ca. 600 of these notices, as found in Gibbons (2004) and as collected by the author as part of the Irish component of the LALP corpus (see Auer & Fairman 2013). The threatening notices, which provide a source of lower-order writing from nineteenth-century Ireland hitherto unexplored for linguistic purposes, are useful for the analysis of the historical development of Irish English ‘from below’. However, the main aim of my paper is to explore the insights they offer into the negotiation of national, social and personal identity in nineteenth-century Ireland. The authors of these notices constructed an identity which was a) indexical of their Irish nationality to justify their actions to their peers; b) indexical of their social identity as members of ribbon societies to signal a collective rather than an individual threat; and c) purposefully not indexical of their personal identity in order to avoid detection by the authorities. Thus, it is hoped that my paper will shed light on the value of threatening notices for the analysis of identity-construction in nineteenth-century Ireland.

References


This paper focuses on evaluative language in reports of rape trials in the Old Bailey Corpus from 1720 to 1749. The OBC provides rich data for linguistic analysis as the corpus includes information of several sociopragmatic features, such as the speaker’s gender and the scribe and the printer of the reports (Huber 2007). In this paper, I focus on covert expressions of attitude in the trial reports by studying how child and adult plaintiffs were referred to in rape trials. The main result of this study relates to the way in which definite, lexical NPs are used as covert evaluative expressions. While an adult plaintiff is typically referred to in accordance with the “Accessibility hierarchy”, i.e. first with a lexical NP, then with pronouns (see e.g. Ariel 1990), a young victim is very often referred to with a lexical NP, typically “the child”, when a pronoun is expected. I will argue that the persistent use of gender-neutral lexical NPs and the avoidance of gendered pronouns functions as a subtle evaluation strategy that foregrounds the plaintiffs’ young age and backgrounds their sexuality. I propose that this is suggestive of the fact that the children were perceived as non-agentive and non-volitional participants in the events that had occurred; they were discreetly portrayed as being unable to give their consent to sexual relations even if they were not protected by the contemporary age of consent legislation (that is, if they were over ten years of age; Simpson 1987). I will also compare the spoken witness testimonies in the OBC to written narratives in the Proceedings of the Old Bailey. The results are similar in both data sets, suggesting that the attitudes expressed in the trial reports cannot be explained by scribal/editorial intervention; instead, they reflect the views of the trial participants.

References


What can a cognitive corpus tell us about grammar as an emergent property of language at the communal level?

In this paper we present a new corpus, the Diachronic Blog Community Corpus, which, by virtue of its design, enables observations not possible to date. We use it to test the hypothesis that grammar arises as a result of averaging across individuals who in fact exhibit wide inter-individual variation in the patterning of their language use. The corpus comprises comments posted to an exceptionally active blog by over 4,000 unique commenters over 7 years, amounting to 7.3 million words in more than 73,000 comments. This includes 1.75 million words of comments written by a single person, the author of the blog. Thus, the corpus allows sampling of language representation at three different levels: 1) the individual or the cognitive level – language use of the author of the blog and its most active commenters taken individually; 2) the communal level – all comments, excluding the heavily represented commenters; 3) and the inter-individual microsocial level (Mauranen 2012) – the interaction between the most active commenters, including the author of the blog. In this way, the corpus operationalises a complex systems perspective on language which distinguishes between several levels of language representation and postulates that properties of language at each of the levels are emergent and cannot be attributed to the lower-level properties (e.g. Beckner et al. 2009, Ellis 2011, de Bot and Larsen-Freeman 2013). By looking at several grammar patterns (Francis et al. 1996, 1998) as represented at different levels over time, we will test the hypothesis that Pattern Grammar ( Hunston and Francis 2000), as presumably any grammar, is an emergent property of language at the communal level. In contrast, patterns at the cognitive level are expected to be more fixed, creating wide inter-individual variation and narrow intra-individual variation.

References

BECKNER, CLAY; RICHARD BLYTHE; JOAN BYBEE; MORTEN H. CHRISTIANSEN; WILLIAM CROFT; NICK C. ELLIS; JOHN HOLLAND; JINYUN KE; DIANE LARSEN-FREEMAN; and TOM SCHOENEMANN. 2009. Language is a complex adaptive system: Position paper. Language learning 59(s1). 1–26.


Open-set authorship verification and language history from below: Determining authorship in a corpus of 18th-century correspondence

Language history has recently moved away from traditional elite sources – representing only the language of a small group of highly literate elite – to material from all social ranks. The underlying assumption is that less formal writings of ordinary language users more closely reflect the spoken vernacular. One of the reasons why such data ‘from below’ have remained uncharted until recently, however, is because such sources present researchers with significant problems concerning authorship: many people from the lower ranks of society sought out the help of family members, friends or even professional scribes to help them write a letter. A major challenge in this field is therefore to distinguish between the person responsible for the intellectual content of the letter (auctor intellectualis) and the scribe (auctor manualis) who drew up the letter.

To overcome this challenge, we present a case study of open-set authorship verification in a collection of Flemish soldiers’ correspondence from the Napoleonic era. Most contemporary approaches in computational authorship attributions have major limitations: (i) they tend to focus on the so-called “closed set-up”, assuming that the correct author is among the candidate authors available; (ii) they statically operate on a fixed set of stylistic features. We will report experiments with the recently developed imposter method for authorship verification (Koppel & Winter 2014). Using a background corpus of imposter authors, this method randomly resamples features to determine whether two texts are similar enough to be attributed to the same author. We aim to show how this method can help us to overcome the authorship problem in language history from below, turning ‘uncharted data’ into ‘rich data’.

References
Matylda Włodarczyk (Adam Mickiewicz University), Joanna Kopaczyk (Adam Mickiewicz University and University of Edinburgh), Elżbieta Adamczyk (Adam Mickiewicz University and University of Wuppertal)

Electronic Repository of Greater Poland Oaths 1386-1444 (ROThA): Lessons in text selection

Decisions concerning text selection criteria for a specialised corpus of “uncharted data” are a serious challenge, and require striking a balance between the specified research focus and potential usefulness of the tool to wider circles of users. This paper explores the ways in which English-language historical corpora, such as the multigenre Helsinki Corpus and ARCHER on the one hand, and the specialised CEEM corpus on the other, can inform new projects in terms of potential text selection criteria. The new corpus in question is the first Electronic Repository of Greater Poland Oaths (ROThA) covering the period 1386-1444 (National Science Centre grant, 2015-2018). The oldest Polish court oaths are the oldest extant collection of secular texts written largely in Polish, hence their central significance for the history of the language, in particular in its “national” dimension. However, the prevailing monolingual perspective on the data (e.g. Czachorowska 1988, Trawińska 2009) disregards the profoundly multilingual nature of scribal practices in Europe at the time (see Adamska 2013, Kopaczyk 2013). The main focus of the prospective corpus is to characterise multilingualism as a social and cultural phenomenon in late mediaeval Greater Poland and its representation in the court oaths, with specific focus on the Polish-Latin and Latin-Polish code-switching (see Pahta 2012 for a recent overview of CS in historical texts). The starting point for text selection is the philological edition of the material which covers 6350 oaths from six different locations (Kowalewicz & Kuraszkwicz 1959-1981). The following potential selection criteria are discussed in the paper: 1) temporal coverage, 2) regional representation (mutual influence in local books, scribal communities), 3) graphological criteria (scribes in groups based on graphemic similarities), 4) CS-related (information on the codes at the disposal of individual scribes) and 5) amount of CS data. The paper argues for a scheme incorporating some aspects of all the above-mentioned criteria in order to best represent issues of multilingualism in the Greater Poland region, in view of lessons learnt from previous corpus compilation projects.

References


Paratext, title-pages and eighteenth-century grammar books

This paper aims to contribute to recent cross-disciplinary research that combines historical corpus linguistics, book history and textual theory (Meurman-Solin & Tyrkkö 2013). It will present a pragma-linguistic analysis of title-pages, these being 'the single most important feature of a book' (Smith 2000) and an especially rich source of paratextual elements (Tyrkkö et al. 2013). Taking Genette’s (1997) theory of paratext and Suhr’s (2011) linguistic approach to book history as a dual framework, I will examine the title-pages of c.150 eighteenth-century grammars of the English language, with a view to identifying verbal items that fulfil the key functions of title-pages – identifying, describing and advertising the text. The analysis will address the following: the sequence of elements that constitute a typical title-page in an English grammar book; textual labels in titles and subtitles; lexicon about the target audience; authors’ credentials that signal authority; the text of the imprint. This study will thus shed light on the linguistic strategies used by publishers, printers and authors to persuade and engage the reader in a text. This is especially relevant in the context of eighteenth-century grammar writing, when grammar books and the English language were considered marketable commodities and instruments of advancement for those with social and political aspirations.

References


DEMONSTRATIONS
Marc Alexander (University of Glasgow)

**How Can We Stage 470,000 Words?**

This demonstration will focus on visual interpretations of the Historical Thesaurus of English and its daughter Mapping Metaphor project. The very size of the Thesaurus – 793,742 word forms from the last millennium of English all arranged into 225,131 semantic categories – means that it presents particular challenges with regards to visually displaying its size and scope of the Thesaurus; limiting its data to only those words we know to have been cited after the mid-nineteenth century still gives us almost half a million database entries to explore. Visualisation can also show unusual broad-scale patterns not otherwise easy to spot in the overall history of English, and the demonstration will discuss a patchwork-like effect growing throughout the language.

Jonathan Hope (University of Strathclyde)

**Ubiquity: a web-based text tagger**

Visualising English Print is a Mellon-funded project based at Strathclyde University, Wisconsin-Madison University, and the Folger Shakespeare Library. Our aim is to produce tools and methods to allow the exploration and analysis of large-scale text corpora, in particular the TCP corpus. Ubiquity is a web-based text tagger which allows users to upload their own text files for tagging. By default, Ubiquity tags texts with Docuscope, a linguistic analysis program developed at Carnegie Mellon University - additionally, however, users can write their own tag set and Ubiquity will provide interactive tagged texts, and a full statistical output in csv format.

Eetu Mäkelä (Aalto University), Terttu Nevalainen and Tanja Säily (University of Helsinki)

**Developing an interface for historical sociolinguistics**

Sociolinguists know that competition between linguistic variables is not unorganized and random, but guided by both language internal and external factors. Variation also reveals language ideologies that can lie behind any linguistic choice. By connecting linguistic data with language external factors such as place, time, social status, level of education, or ideological and political environment, we can observe how social meanings arise in context. Interfacing structured and unstructured data also enables new kinds of questions about linguistic variation and change, and offers opportunities to test and experiment on novel methods to reveal the logic behind ostensibly random variation.

This poses a challenge from a computer science perspective, as current tools are not able to fluidly cross-question linguistic phenomena and contextual information. The new open source tools to be discussed allow the user to define data subsets based on both linguistic features and the various extralinguistic criteria included in the corpus metadata. This constrained subset can be subjected to further linguistic analysis and visualization, as well as projected again through structured metadata into interactive visualizations combining the two, such as plotting the spread of linguistic phenomena through time and space. The exploration of visualization parameters allows us to detect interesting variations in the material, and to extract the relevant subset of the data for linguistic analysis.
Tools and Methods for Processing and Visualizing Large Corpora

We present several approaches and methods which we develop or use to create workflows from data to evidence. They start with looking for specific items in large corpora, exploring overuse of particular items, and using off-the-shelf visualisation such as GoogleViz. Second, we present the advanced visualisation tools and pipelines which the Visualisation Group at University of Constance is developing. After an overview, we apply TreeView, Lexical Episode Plots and Interactive Hierarchical Modeling to the vast historical linguistics data offered by the Corpus of Historical American English (COHA), which ranges from 1800 to 2000.

Tanja Säily (University of Helsinki) and Jukka Suomela (Aalto University)

types2: Exploring word-frequency differences in corpora

We will demonstrate the use of the types2 tool (Suomela 2014) to explore, visualise, and assess the significance of variation in word frequencies. Based on accumulation curves and the statistical technique of permutation testing, this freely available tool is especially well suited to the study of types and hapax legomena, which are common measures of morphological productivity and lexical diversity. We are currently developing a new version of the tool that will provide improved linking between the visualisations, metadata and corpus texts, which will facilitate the analysis of rich data.

The new version of our tool will be demonstrated using two data sets extracted from the Corpora of Early English Correspondence (CEEC) and the British National Corpus (BNC), both of which are rich in sociolinguistic metadata. Our software will be capable of processing metadata and search results in several different formats employed by corpora and concordancers, including CEEC, BNCweb and WordSmith Tools. The output is provided in three formats: web pages (see Figure 1), PDF images and raw statistics in a database. The web pages include interactive visualisations that are linked to the underlying data and metadata. Our demonstration will show how the linked data facilitate exploring and interpreting the results.

Figure 1. Interactive visualisation of change in the productivity of the suffix -ity in the CEEC, 1680–1800. The links above the graph are used to select what is shown on the graph, while the points on the graph are clickable links to the underlying data.

References

History in the ESTC catalogue

This demonstration concerns publication trends of history in early modern Britain and North-America, 1470-1800, based on the English Short-Title Catalogue (ESTC) data. Its most important contribution is a demonstration how digitized library catalogues can become a crucial tool for scholarship and part of reproducible research. The presentation shows the relevance of the analysis of paper consumption in early modern book production and it demonstrates in practice the importance of open science principles. The demonstration examines three main research questions: 1) how did publishing of history change over time in early modern Britain and North America 2) who wrote history, and 3) where was history published. It will be demonstrated that the average book size for history publications becomes smaller over time and that the octavo-sized book is the rising vehicle of history in the eighteenth century that tells us factually also about widening audiences. The presentation will also compare different aspects of most popular authors on history such as Edmund Burke and David Hume. While focusing on history publishing, these findings may reflect more widespread trends in publication in the early modern era, and the demonstration illustrates how some of the key questions in book history can be assessed by statistical analysis of large-scale bibliographic data collections.
POSTERS
Introducing the corpus of Late Modern English Medical Texts 1700-1800

Leena Kahlas-Tarkka and Matti Kilpiö (University of Helsinki)

20 years of EETACS collaboration

Susan Fitzmaurice, Iona Hine (University of Sheffield) and Justyna Robinson (University of Sussex)

Linguistic DNA: Modelling concepts and semantic change in English, 1500–1800

Daisuke Suzuki (Japan Society for the Promotion of Science)

The historical development of perhaps and related expressions

Stephanie Horch (University of Freiburg)

Charting conversion in Asian varieties with GloWbE

Terttu Nevalainen, Tanja Säily and Turo Vartiainen (University of Helsinki)

Language Change Database: a new online resource

Hendrik De Smet (KU Leuven), Susanne Flach (Freie Universität Berlin) and Jukka Tyrkkö (University of Tampere)

Introducing CLMET3.1 and the plans for CEM1.0