Table of contents

Welcome to ICAME 39 at Tampere! ................................................................. 1
Pre-conference workshops.............................................................................. 2
  WS1: Corpus approaches to conceptual metaphor and idioms in World Englishes ................................................................. 3
  WS2: Data visualization in corpus linguistics: Critical reflections and future directions ............................................................ 10
  WS3: Democratization in English(es): Synchronic and diachronic perspectives ................................................................. 18
  WS4: Discourse markers in the Outer and Expanding Circles: A corpus-linguistic perspective ............................................... 29
Plenaries ........................................................................................................... 40
Full papers ....................................................................................................... 45
Work-in-progress reports ............................................................................. 165
Posters ............................................................................................................ 187
Software demonstration .................................................................................. 204
List of delegates and participants ................................................................ 206
Welcome to ICAME 39 at Tampere!

On behalf of the organising committee we would like to welcome you to the 39th Conference of the International Computer Archive of Modern and Medieval English, held in Tampere from 30 May to 3 June. The theme of ICAME 39 is Corpus Linguistics and Changing Society. We hope you will enjoy the conference.

This booklet contains the conference abstracts and a list of the conference participants. If you have any further queries do not hesitate to approach the organising team.

Organizing Committee:

Mark Kaunisto
Juhani Klemola (chair)
Minna Nevala
Arja Nurmi
Päivi Pahta
Hanna Parviainen
Paula Rautionaho
Paul Rickman
Veera Saarimäki

University of Tampere Conference Services:

Johanna Lehto

Student helpers:

Dianhui Hu
Jouni Kivivuori
Joonas Lehto
Eetu Mäkelä
Riikka Savela
Iida Tanni
Satu Vallineva
Laura Viitanen
Pre-conference workshops
Using large electronic corpora to validate elicitation techniques in research on conceptual metaphor and idioms: The case of the “lexicon of corruption” in West African Englishes

Marcus Callies
University of Bremen

The various linguistic forms used to refer to the social phenomenon of corruption in West African Englishes have gained increasing attention in the last few years, with several studies aiming at identifying a “lexicon of corruption” used to express veiled bribes in these varieties of English (Skjerdal 2010; Bassey & Bassey 2014; Adegoju & Raheem 2015; Safotso 2015, 2017; Meutem Kamtchueng 2015, 2017). Polzenhagen & Wolf (2007) in their pioneering study of cultural conceptualisations of corruption in African Englishes discuss the conceptual mapping bribe is food as a special instantiation of the more general mapping money is food. In turn, a number of (lexicalized) expressions are motivated by this metaphor and mean ‘bribe’ in the appropriate contexts, for example kola ‘cola nut’ (a traditional food gift presented to guests), mimbo (Cameroon alcoholic beverage) or gombo ‘okra, okra sauce’, which additionally means ‘funds’, all of which can also be used in idiomatic expressions with the verbs give and take as in take kola / mimbo / gombo ‘to accept a bribe’ (Wolf & Polzenhagen 2007: 143). Further recent studies list various other idioms such as to settle a person ‘to bribe or to provide such gratification or take an action that defeats the course of justice or blinds the eye of the “settled” person to act justly’ (Adegbija 2003: 43; Ekundayo 2013: 22) and to wet the ground ‘to provide prior gratification or bribe that makes a later course of action or conduct smooth’ (Adegbija 2003: 43).

Most of the words and idioms listed in these previous studies were obtained either by means of elicitation through questionnaires or interviews, or sourced from participant observation and interaction with potential informants, online and printed materials (sometimes literary works). Thus, information about the actual frequency and extent of use of these words and idioms is severely limited. In this talk, I will present the findings of a case study to validate the (often anecdotal) evidence of the occurrence and use of the “lexicon of corruption” in West African Englishes (Nigeria, Ghana and Cameroon) as presented in published research against actual usage data by means of large electronic corpora of world Englishes such as the GlowbE and the NOW corpus (Davies 2013–, Davies 2016–).

References


Davies, M. (2013–). *Corpus of Global Web-Based English*: 1.9 billion words from speakers in 20 countries. [http://corpus.byu.edu/glowbe](http://corpus.byu.edu/glowbe)

Davies, M. (2016–). *News on the Web* corpus: 3 billion words of data from web-based newspapers and magazines from 2010 to the present time. [http://corpus.byu.edu/now](http://corpus.byu.edu/now)


---

**Metaphors of light and darkness in a corpus of memorable American political speeches**

*Marta Degani*

University of Verona

The paper aims at investigating the evocative power of metaphor in a number of remarkable American political speeches. The study will concentrate on the metaphorical framing of political issues in terms of light and darkness. As research in cognitive linguistics has demonstrated (e.g., Lakoff & Johnson, 1980; Lakoff & Turner, 1989) light and darkness play a significant role in people’s understanding of antithetical concepts such as knowledge and ignorance, life and death. In political discourse, this potential for antithesis is chiefly used to convey either positive (light) or
negative (darkness) evaluations, typically in compliance with more general rhetorical purposes of the speech.

The analysis will be carried out on a corpus of speeches which American rhetoricians ranked among the top 100 speeches of the twentieth century in a national survey (http://www.news.wisc.edu/misc/speeches). Considering the impact of these speeches, an investigation in terms of their metaphoricity is expected to provide additional insights into the reasons that made them memorable. Overall, the study intends to illustrate how the evocative concepts of light and darkness have facilitated the communication of central political ideas, values and beliefs in twentieth century American political rhetoric.

References


Conceptual metaphors of FAMILY in Hong Kong English

*Sarah Geers*

University of Bremen

This case study investigates metaphorical expressions as evidence of different conceptualisations of FAMILY in Hong Kong English. It is part of a larger dissertation project on “Cross-cultural Variation in Conceptual Metaphor. A corpus-based study” and is based on a corpus approach in order to examine conceptual metaphors in different English-speaking cultures. The objectives of this paper are two-fold. First, I will outline the methodological approach of the study and second, present findings from a first case study.

Corpus-based research on metaphor variation is increasing with the publications by Sharifian (2008, 2017), Callies (2017), Callies & Onysko (2017), Lederer (2016), Polzenhagen & Wolf (2007) and Stefanowitsch (2004) leading the way into this relatively young field of research. In order to examine cultural conceptualisations in figurative expressions two approaches can be identified: manual annotation of smaller samples (in keeping with the Pragglejaz Group’s (2007) Metaphor Identification Procedure) and keyword analysis in large corpora (similar to Stefanowitsch’s (2004) metaphorical patterns). The present project aims to combine these two different methodological approaches in order to achieve a better grasp needed to locate metaphorical expressions in corpora. Starting with manual annotation of 50 news articles from the Hong Kong-based newspapers *South China Morning Post* and *Hong Kong Free Press* a corpus of 33,000 words has been compiled and analysed. Half of the articles were chosen randomly, the other half was chosen with a view to coverage of the topic of family. The results point to a connection between MARRIAGE and MONEY as illustrated by expressions such as *bride price* and *naked marriage*. Social expectations connected to marriage seem to centre around economic stability and financial success. It seems to
be of paramount importance in an adult’s life to be financially successful while the individual is at the same time striving to fulfil the social expectations of marriage and children. The more detailed account of this conceptualisation which emerged after a first exploratory analysis is then checked with lexical keywords, similar to Stefanowitsch’s semantic triggers, in the Corpus of Global Web-based English (GloWbE, Davies & Fuchs 2015).

In the presentation I will first give a short overview of the methodology and then concentrate on the linguistic expressions as manifestations of conceptual metaphors. Finally, an evaluation of the results from a meta-level with regard to the coherence of the different approaches will be presented.

References


Methodological challenges for the retrieval and identification of conceptual metaphor in corpus data: Insights from the Study of Emotion Metaphors in New Englishes

Barbara Ann Güldenring
University of Marburg

In line with the ever increasing interest in the corpus-based study of World Englishes from a cognitive linguistic perspective, this contribution aims at presenting some important insights into the methodological challenges of such an undertaking. These insights have been informed by a doctoral study into emotion metaphors in New Englishes which aims at discovering metaphorical variation as well as commonalities in the metaphorical systems pertaining to the source domains
ANGER, FEAR and HAPPINESS in four institutionalized second-language varieties of English (Nigeria, Kenya, India and Singapore).

While developing a methodological framework which makes use of corpus data from the Corpus of Global Web-based English (GloWbE; Davies 2013), various decisions were contingent on finding appropriate and transparent ways of accomplishing the steps of identifying and retrieving metaphorical data from a corpus which has not been annotated for metaphor. This contribution will not only serve to outline these methodological steps, but will also zoom into three distinct challenges that a researcher interested in conceptual metaphor faces when working with corpus data. These will be discussed and illustrated on the basis of the following categorization: 1) authentification of variety-specific metaphors and corpus selection, 2) limitations in the method of retrieval of metaphorical data, and 3) identification of conceptual metaphors on the basis of metaphorical expressions.

First, the authentification of variety-specific metaphors is an important initial step for any research program aiming at discovering conceptual metaphors that reflect underlying conceptualizations attributable to varieties of distinct cultural backgrounds. The major challenge here concerns the extent of analysis that goes into determining if an individual instance (i.e., linguistic metaphor) can be truly considered as representative of a specific variety of English (and thus, in turn, can be considered an active mapping within the metaphorical system underlying the source domain under investigation). This challenge directly concerns the choice of corpus for metaphorical study, and the examples presented will directly compare results that can be gleaned from a reliance on the GloWbE and the International Corpus of English (ICE; Greenbaum 1996) respectively. Second, the approach to the retrieval of the relevant data – once a corpus has been selected – presents a separate challenge to the extent that the specific method chosen can delimit what metaphorical data is possible for extraction. By means of illustration, this challenge involves an evaluation of the effectiveness of the metaphor retrieval approach outlined by Stefanowitsch (2004, 2006), known as Metaphorical Pattern Analysis (MPA). Third and finally, the identification of conceptual metaphors offers a unique challenge for the analyst in terms of finding adequate conceptual labels for corpus data that can only be initially extracted as linguistic metaphors. Because this step entails a significant amount of intuition-based analysis on the part of the researcher, this final section will introduce a methodological step to corpus-based metaphor study that exploits collocational patterns found in the corpus as a whole – termed “intuition boosters” (see Güldenring 2017) – in order to provide more transparency concerning the formulation of the postulated conceptual metaphors.

References

This paper sets out to analyse some Cameroon English idioms from a Cultural Linguistics perspective. The data for the study was drawn from Kouega’s (2007) Dictionary of Cameroon English and from a corpus gathered by Meutem Kamtchueng (2013). The work is discussed from the vantage point of Cultural Linguistics (Sharifian 2012; Kövecses 2010; Holland & Quinn 1987; Nisbett et al 2001; Goddard & Wierzbicka 2004; Machery et al 2004). The findings reveal that the following factors account for the creation of idioms in this variety of English: African cultural models (e.g. the community-based cultural model), orality and the overwhelming influence of taboos.

References

Looking for metaphors in a specialized corpus of New Zealand Stories

Alexander Onysko
University of Klagenfurt

This paper provides insights into the compilation of a specialized small corpus that gathers short oral narratives by New Zealanders with different linguistic repertoires. The corpus is designed to explore internal variation in New Zealand English dependent on the monolingual or multilingual background of the speakers and their ethnic and cultural affiliations. This is different from the scope of corpora in World Englishes (e.g. the components of the International Corpus of English) and in New Zealand English (the Wellington written and spoken corpora; see Bauer 1994; Holmes et al. 1998), which were compiled to give a representative sample of their target varieties. By contrast, the Corpus of New Zealand Stories is a specialized collection of spoken language which is aimed at allowing an internal comparison of the language data according to the linguistic repertoire and the ethnicity of the speakers. The corpus consists of oral accounts given by 140 university students from both Māori and Pākehā (New Zealand European) ethnicity in response to the same narrative task.

The paper describes the making of the corpus including data collection and transcription of the oral narratives. Particular emphasis will be put on the questions of how conceptual metaphors can be retrieved from the corpus and whether those metaphors might be indicative of internal variation among different speaker groups in New Zealand. Since many of the bilingual Māori-English speakers are also bicultural, it is a further question whether any culturally specific Māori metaphors can be found in the individual stories.

References


The International Corpus of English (ICE) coordinated by Prof. Marianne Hundt and hosted at the English Department of the University of Zurich. http://www.ice-corpora.uzh.ch/en.html
Contemporary communication is becoming an increasingly complex multimodal process. As a result, the task of the linguist to unpack meaning-making is also becoming more complex, requiring multiple methods, e.g. corpus linguistics and (multimodal) discourse analysis, collaboration between researchers with specialist expertise, and the use of visualizations that can present findings in a meaningful manner. Displaying the results of complex multimodal analyses in a way that reveals patterns within and across modes and texts is very difficult to achieve with tables and traditional static visualizations, such as bar- and line-charts. In this paper, we review the state of the field and then introduce a new, freely available, online tool for visualizing the results of complex multimodal analyses: Kaleidographic (Caple & Bednarek 2017).

Kaleidographic is a dynamic and interactive visualization tool that can reveal relations between variables within multimodal texts or the ‘steps’ of a multimodal communication process. The number of variables and multimodal layers of the visualization can be increased or decreased to match the target data. The values of the variables can also be displayed on a nominal, ordinal, interval, or ratio scale. Kaleidographic is able to ‘play’ through target data at different speeds to reveal patterns across multimodal texts or steps of a communication process. As each multimodal text or ‘scene’ is shown, the values of the different variables are activated and ‘light up’, revealing information that is difficult to perceive in a static visualization. Increasing the speed of the playback gives a more holistic view of the data, whilst slowing down the playback or stepping through the playback one text or ‘scene’ at a time allows the user to spend more time with individual data points and to explore the combinations of meanings being displayed at any one time. It is also possible to block out variables and multimodal layers in order to focus on specific variables and layers of interest. After demonstrating how Kaleidographic works, we will reflect upon its current limitations and discuss future developments. We suggest that Kaleidographic is useful not only in multimodal discourse analyses of small datasets, but also in analyses of other datasets including those used in monomodal corpus linguistics studies.
Visualisation is necessary to decide if clustering is appropriate in corpus analysis

Jack Grieve
University of Birmingham

The basic assumption underlying the application of common forms of cluster analysis, including hierarchical cluster analysis and k-means cluster analysis, is that a set of objects can be meaningfully grouped into subsets based on the values of one or more variables, so that the objects in those subsets are similar to each other and different from the objects in the other subsets.

Among other applications, cluster analysis is commonly used in corpus linguistics to group texts based on a series of quantitative linguistic variables. Perhaps most commonly, these groups of texts are interpreted as representing situationally (i.e. registers) or socially (i.e. dialects) defined varieties of language, depending on the composition of the corpus under analysis. Although there is nothing inherently wrong with this approach to corpus analysis, it is often misused, identifying groupings of texts in the corpus that are not distinct and that should not be interpreted as representing varieties of language. Before conducting a cluster analysis it is therefore important to explore the underlying dataset visually, examining the distribution of texts in the corpus across the linguistic variables under analysis, for example based on scatterplots for the variables individually or in the aggregate by using techniques such as multidimensional scaling. This process is not only important for licensing the use of cluster analysis, but, if clustering is deemed appropriate, for deciding which form of cluster analysis should be applied.

The goal of this presentation is therefore to illustrate the potential issues with the application of cluster analysis in corpus linguistics and to discuss how these issues can be addressed through the careful visualisation of linguistic variation the corpus. After briefly introducing standard methods for cluster analysis as well as their implementation and visualisation using the programming language R, examples of both insightful and misleading cluster analyses will be provided based on existing corpora of American English, including the identification of both socially and situationally defined varieties of language, demonstrating how cluster analysis can identify groupings of texts that do not actually exist in the data.

In particular, this presentation will stress the importance of exploratory visualisation of the data prior to conducting cluster analysis, including the use of multidimensional scaling and other dimension reduction techniques to condense the underlying dataset to a small number of aggregated dimensions against which the texts in the corpus can be plotted to see if the extraction of clusters is justified. In addition, to illustrate the necessity and complexity of this process, a series of live, interactive visualisations will be used.
Fractal visualization of corpus data

William A. Kretzschmar, Jr. & Steven Coats
University of Georgia, University of Oulu

Corpus linguists understand that frequency lists of words or collocates in corpora all come out with nonlinear, Zipfian distributions. A chart of results ordered by frequency will have a small number of frequent forms on the left, a moderate number of fairly common forms, and a large number of rare forms, as in Figure 1 (top 1000 nouns in the Academic section of COCA). Corpus linguists understand that the same nonlinear chart will appear for subsamples of the same data, as in Figure 2 (top 1000 nouns from just the Humanities portion of the Academic section of COCA). This scale-free “self-similarity” between and within samples is evidence that words and collocates in corpora are fractal, in the sense introduced by Mandelbrot (1982; see also Kretzschmar 2009, 2015).

Fractal distributions are not well described by visualizations that assume a central tendency, as one would find in a normal distribution. Corpus linguists thus should choose visualizations that display nonlinear distributions, not means and standard deviations. The problem for corpus linguists, after having done so, is how to compare the self-similar charts. When every graph looks the same, how do you make effective visualizations to show the differences in word or collocate frequency between corpora or subsamples within corpora? This paper will demonstrate two basic methods for effective comparisons, clustered/composite graphs and paired graphs. It will also describe a standard method for coping with comparison of scale-free distributions by charting rates of occurrence instead of raw counts of tokens. All of these methods will be illustrated from real data. Appropriate visualizations make for more effective exploration of data by concentrating on the real differences in usage in different samples.

References

Visualization of text corpora – seven years on

Harri Siirtola and Tanja Säily
University of Tampere, University of Helsinki

We argued seven years ago that the information visualization community has a lot to offer to corpus linguistics – once the domain is better understood (Siirtola et al. 2011). We now have some observations on this progress.

Three years ago Kucher & Kerren (2015) made an interactive survey tool for text visualization techniques and tools, and it has been online ever since. The idea was to create a crowd-sourced taxonomy of all text visualization techniques and tools on this planet. Initially, the site listed 141 items, but the current number is about 400 and growing. Only a few items in the taxonomy are specifically for corpus linguistics, but many of them are just waiting to be applied. The survey result is a useful resource and starting point for all text visualization related work.

Another promising development is the tidytext package for Statistical System R (Silge & Robinson 2016). They bring the tidy data analytics approach (Wickham 2017) to text analysis, and provide state of the art text mining tools. The combination of tidytext, ggplot2 (Wickham 2016), and ggraph (Pedersen 2017) is a powerful tool for text visualization, but it is mainly for static or animated graphics and less for the interactive kind.

It is widely recognized that interaction facilitates discovery, and is the most important tool to uncover insights (Yi et al. 2007). Interactive and visual exploration of text corpora is something that corpus linguistics clearly needs, and there are recent developments also in this area. Coquery (Kunter 2017) is a corpus query tool that has a visualization designer module that can produce e.g. ‘barcode plots’ of query results, while #LancsBox (Brezina et al. 2015) allows users to analyse and visualize e.g. collocations and keywords in corpora. Both Coquery and #Lancsbox mainly operate in the stepped interaction mode, where relatively infrequent and discrete user actions are typical. However, the continuous interaction mode – where the system tries to respond continuously to user actions – is the most productive for insight generation (Faconti & Massink 2001).

Khepri (Mäkelä et al. 2016) and Text Variation Explorer (TVE) (Siirtola et al. 2016) are two tools under development that are based on continuous interaction. Designed for the purposes of historical sociolinguistics, Khepri is a corpus query tool that links together corpus texts, metadata, visualizations and statistical analyses and enables users to move fluidly between them. In TVE, users can explore the structure and variation of texts by modifying visualization parameters, and the visualization is kept in synchronization. This approach allows users to explore a large number of visualizations in a short period of time, and to react when something interesting is seen. The key idea behind both TVE and Khepri is to preserve the connection between the text and the visualization: if we see something interesting on the graph, we need easy access to the text and metadata behind it to be able to interpret it. We hope to see more tools of this kind in the future.
References


Measuring and visualizing linguistic distances

Benedikt Szmrecsanyi
KU Leuven

Calculating linguistic distances between varieties, dialects, and languages is a methodological cornerstone in a number of linguistic sub-disciplines including dialectometry (see e.g. Goebl 1982) and quantitative typology (see e.g. Cysouw 2013). In these fields, linguistic distances are typically calculated on the empirical basis of structured data found in surveys and dialect atlases, such as e.g. the World Atlas of Language Structures or the Survey of English Dialects. Calculating linguistic distances on the empirical basis of naturalistic corpus data is, by contrast, a less widespread exercise in the literature. In the first part of this talk I thus sketch a corpus-based, frequency-oriented method for determining distances: linguistic distance is defined as being proportional to the extent to which text frequencies of linguistic features differ across varieties. As a case study, I discuss a dataset that covers linguistic distances between 34 traditional dialects in England, Scotland, and Wales, based on an aggregate analysis of text frequencies of 57 grammatical features in the Freiburg Corpus of English Dialects (see Szmrecsanyi 2013). In the second part of the talk, I move on to discussing four visualization methods to visually depict the resulting distance matrix:

1. Dialectometric cartography, which projects distances to geographic space;
2. Hierarchical agglomerative cluster analysis (Aldenderfer & Blashfield 1984), which groups objects (such as varieties of English) into clusters that are internally as homogeneous as possible, and which outputs family tree-style diagrams (so-called dendrograms);
3. Multidimensional Scaling (MDS) (Kruskal & Wish 1978), a procedure which reduces high dimensional distance matrices (consider that a distance matrix covering nine varieties of English is eight-dimensional) to lower-dimensional plots in which proximity between data points indicates linguistic similarity;
4. NeighborNet diagrams (Bryant & Moulton 2004), which were originally developed in bioinformatics. The NeighborNet algorithm is related to hierarchical agglomerative cluster analysis methods, but unlike classical cluster analysis, NeighborNet diagrams can depict conflicting signals in the underlying distance matrices.

The talk will conclude by cataloging advantages and disadvantages of the visualization techniques.

References

Visualising uncertainty with the Wilson score interval

Sean Wallis
Survey of English Usage

The Wilson score interval (Wilson 1927) is the mathematical inverse of the Normal interval about a population mean. Given an observed mean probability $p$, the upper (lower) bound of the Wilson interval is where the population mean probability $P$ would lie if $P$’s lower (upper) bound was at $p$. It is derived from the Normal approximation to the Binomial model and is therefore mathematically consistent with it. It is also algebraically simple to calculate, unlike the Clopper-Pearson interval (Wallis 2013), which inverts the Binomial interval but can only be converged upon iteratively, or bootstrap methods which require significant computation.

For some sixty years the Wilson interval was forgotten until it was ‘rediscovered’ in the late 1990s (Newcombe 1998). As a result, almost every reference to ‘confidence intervals’ in textbooks refer to the mathematically-incorrect ‘Wald’ interval, where the Normal approximation is applied, wrongly, to $p$ rather than $P$. The mistake is repeated in citations of Standard Error and numerous similar methods. Wald intervals are highly misleading, often exceed probabilistic bounds and generate anomalous results. Plotting data with these intervals is visually confusing and often contradicts the outcome of statistical tests, undermining confidence in testing procedures and inferential statistics more generally.

These problems are not weaknesses of the Wilson interval. Employing the Wilson interval in graph plots allows us

(1) to project a good estimate of the most likely value of $P$ on the basis of an observation $p$, on the same scale as $p$,

(2) to visually compare two values and intervals of observed $p$, using a simple overlap heuristic, limiting the need for step-wise significance testing, and

(3) by adjusting the Binomial model for random text sampling, to adjust Wilson intervals accordingly.

In addition it is possible

(4) to compute difference intervals and difference-of-difference intervals, and

(5) to apply transformations to $p$ and its score interval, to compute more complex confidence intervals, or intervals on more meaningful variables.

In this paper we will simply focus on the visual power of Wilson intervals to demonstrate points 1–3 above. We will argue that incorporating good quality intervals gives researchers a much better
insight into the nature of statistical randomness as it impacts on their experiment than would otherwise be possible and makes the interpretation of graph plots much more straightforward.

References


From the perspective of societal democratisation, the nineteenth century was a key period, characterised by intense industrialisation, urbanisation, massive redistributions of wealth and social upheavals across Europe and in Britain. Concurrently, the human condition was increasingly the subject of art and public discourse, and the significant increase in the volume of newspaper and book publishing gave dialects, both established and newly developing, more visibility and recognition. But did these events and phenomena also translate to linguistic democratization at the top tiers of public discourse in the debates at the two Houses of Parliament and to what extent were the two houses linguistically distinct?

In this paper, we explore how these sociocultural changes were reflected in the parliamentary record, a genre that combines elements of spoken, written and spoken-to-be-written discourses. Our main interests are in the processes of linguistic colloquialisation and democratisation, understood broadly as tendencies towards greater informality and equality in language use (e.g. Leech et al. 2009; Farrelly & Seoane 2012). Previous diachronic studies have established that written language has increasingly adopted features associated with spoken language, although genre and register differences are considerable (Mair & Hundt 1999; Biber & Gray 2012). Our hypothesis, which we explore empirically in this paper, is that the relative frequency of informal language features increases over time also in the parliamentary record. At the same time, profound changes took place in the practices of recording parliamentary proceedings, most importantly the introduction of the Official Report in 1909 and recruitment of a team of shorthand writers (MacDonagh 1913; Vice and Farrell 2017). This resulted in a text typological shift in the primary data which complicates the analysis of frequency developments over time.

Our data on British parliamentary debates comes from the Hansard Corpus (Alexander and Davies 2015–), the best available record of all speeches delivered in the two Houses. We investigate the 60-year period 1870–1930, which includes reports of parliamentary debates and, after 1909, transcriptions (in total ca. 40 million words). Focusing on a pattern-driven approach to changes in n-gram frequencies (Tyrkkö & Kopaczky 2018), we first identify major shifts in parliamentary language using unsupervised grouping methods and then focus on a corpus-based analysis of spoken features (e.g. Biber & Gray 2012). By contrasting the most prevalent linguistic practices in the House of Commons and the House of Lords (see also Mollin 2007 and Archer 2017), we trace
the frequency profiles of individual features in the data, and provide a detailed analysis and contextualisation of the observed changes, with particular reference to the increase of colloquial features.

References


Changes in the modal domain in different varieties of English: Effects of democratization?

*Svenja Kranich, Elizabeth Hampel and Hanna Bruns*

University of Bonn

One well-investigated recent change in English that has been linked to democratization is the ongoing decline of the core modals (such as *may*, *might*, *must*) and the concomitant rise of the so-called semi-modals or quasi-modals (such as *be able to*, *have to*) (cf. e.g. Leech 2003; Mair & Leech 2006).

In this context, it has been suggested that speakers prefer the semi-modals over the modal auxiliaries because of the semi-modals’ lesser focus on the speaker as the deontic source. Thus, if one says “You must go now”, the speaker tends to be identified as the source of the obligation, imposing on the addressee, whereas if one says “You have to go”, it is more likely that the source of
the obligation is speaker-external (cf. Westney 1995: 151; Collins 2009a: 35f., 60f.). This makes the latter use often more appropriate in a society that prefers to avoid stressing power hierarchies and rather aims to reduce markers of social distance.

The process is advanced to different degrees in different varieties of English: American English is in the lead, while Australian English is somewhat less advanced and British English even more conservative. Outer circle varieties, such as Indian English, broadly speaking, tend to exhibit still greater conservatism (Collins 2009a, b).

While the suggested correlation between a society’s decreasing attention to social hierarchies and a changing usage of modality markers is intuitively convincing, the line of argument is still somewhat speculative. Thus, it remains to be conclusively shown that a decrease of must and may goes hand in hand with a shift away from hierarchical structures in society and a general avoidance of linguistic structures that draw attention to power differences.

The present paper aims to fill this gap. In a first step, a detailed corpus-linguistic investigation of the decrease of may and must in COHA allows us to see to what extent the different functions of these modal verbs (dynamic, deontic, and epistemic) are responsible for the decline in frequency, and it will be shown that specific construction types are more responsible for the decline than others. In the second part of the study, we combine the corpus-based approach with data gathered using discourse completion tests (DCTs). The DCTs aim to find out to what extent British, American, and Indian English speakers are inclined to choose different linguistic strategies for the realization of requests depending on the power status of their interlocutor (higher, same, or lower). Participants of this study came from two age groups, young adults (20–30) and older speakers (over 50), in order to spot current change. This combination of methods allows us to see to what extent the quantitative differences in modal auxiliary and semi-modal use in the three varieties correlate with overall tendencies to linguistically deal with hierarchical structures in society. The study is part of a larger project that investigates the impact of recent cultural change on linguistic behaviour.

References

Epicene pronouns, Asian Englishes and register variation: Pulling Ariadne’s thread out of the democratization labyrinth

Lucía Loureiro-Porto and Elena Seoane
University of the Balearic Islands, University of Vigo

The search for gender equality in language use is one of the most often cited cases of linguistic democratization (cf. Leech et al. 2009, p. 263; Farrelly & Seoane 2012, for example). Thus, at the lexical level, masculine nouns such as spokesman or fireman are being replaced by the gender-neutral spokesperson and fire-fighter (e.g. Holmes et al. 2009), and at the grammatical level anaphoric pronouns such as generic he used with epicene antecedents (e.g. no student must forget his rights) are being replaced by singular they (no student must forget their rights) or by composite forms such as he or she (no student must forget his or her rights), as shown, among others by Balhorn (2009) and Paterson (2014). Although the existence of singular they goes back to the Middle Ages, when it was commonly used to refer to antecedents with the quantifier every (e.g. Bodine 1975; Curzan 2003; Balhorn 2004), a use unrelated to changes in social norms and search for gender neutrality, its current increase in frequency is due to a recent social demand, namely a wish to make language more democratic by avoiding the use of generic he when the antecedent refers to a human being whose sex is unknown or not relevant (cf. Leech et al. 2009: 263; Farrelly & Seoane 2012). While the variation between generic he, singular they and the composite form he or she has been widely studied in inner-circle varieties of English (e.g. Bodine 1975; Green 1977; Gastil 1990; Meyers 1990; Zuber & Reed 1993; Laitinen 2002; Balhorn 2004, 2009; Paterson 2011, 2014; Parini 2013; among others), their distribution in postcolonial Englishes has been only preliminarily approached (e.g. Loureiro-Porto 2018). With the aim of contributing to the study of this topic, this paper shows the results from the analysis of the whole ICE components of three Asian varieties, namely Hong Kong (HK), Indian (IND) and Singapore (SIN) English. More than 58,000 examples were retrieved from the corpora and manually filtered, resulting in 2,120 valid cases of epicene pronouns, distributed as follows: singular they (321 cases), generic he (1,595 cases) and he or she (204 cases). All of them were analysed taking into account variables such as type of antecedent and register variation. The results show a very different picture of the three Asian varieties. While in HK the democratic pronoun they and the composite form he or she are the preferred option with epicene antecedents in nearly all written text-types (with a relevant higher frequency in persuasive writing, creative writing, popular writing and letters) and represents nearly 50% of all spoken text-types, the picture in IND and SIN is quite different. These two varieties have an overwhelming preference for generic he, and the more democratic options, they and he or she, are concentrated in creative writing (both IND and SIN), social letters (IND) and private dialogue (SIN). These differences will be interpreted in light of Labov’s changes from above (2007, 2010), as well as from a register perspective which takes into account the communicative purpose, the function and the intended readership of the text.

References


This paper stems from the idea that social changes and changes in language practices work in tandem, and processes such as democratization can be observed in language. By using a combination of corpus linguistic and socio-pragmatic methods, it is possible to track social and cultural developments over long periods of time with an evidence-based approach that would have been impossible only a decade ago (cf. Farrelly & Seoane 2012). Earlier linguistic research has explored for example diachronic changes in the use of modal auxiliaries (Myhill 1995) and masculine and feminine pronouns (Baker 2010) and linked their development with democratization. In this study we propose to focus on diachronic changes in word patterns and explore and interpret findings in the framework of democratization, which we understand broadly as “changing norms in cultural relations” working towards less hierarchical and more equal patterns of social organization (Leech et al. 2009: 259).

Comparing subsequent centuries of British English data in ARCHER-3.2 with keyword analysis in our earlier study, we found out several differences including changes in words referring to people. Although division by centuries is somewhat arbitrary, it seems that pre- and post-1900 data show differences that might be understandable in terms of broad democratization processes through which status differences are levelled out in linguistic expression. Some of the observed differences include the abundance of people words in the pre-1900 data in comparison to the post-1900 data as well as qualitative differences in keywords. On the one hand, the eighteenth- and nineteenth-century people keywords refer to individuals and often signify social status (e.g. lady, sir, captain). The twentieth-century people keywords, on the other hand, include indefinite pronouns (e.g. someone) and words referring to groups of people (e.g. Germans, workers, leaders). Although ARCHER-3.2 has been compiled to represent historical registers and to provide a solid basis for diachronic comparisons, it is of course possible that registers change to the extent that a different set of people words is needed, but it may also be the case that language practices have shifted from individual status to groups.

To complement the keyword analysis, we shall explore the found people keywords in their textual context in ARCHER-3.2 to find out more specifically where the changes stem from. Changes in certain socio-pragmatic processes, as in social labelling and identification, often reflect changes in prevalent societal attitudes towards certain social groups and classes (Bucholtz & Hall 2005). Similarly, these micro-level linguistic processes can trigger macro-level changes on their own through the process of democratization.

References

Non-sexist language policy and the rise (and fall?) of combined pronouns in British and American written English.

Laura L Paterson
The Open University, UK

Using the concept of democratisation, Baker (2010: 69) argues that diachronic changes in the English language, such as the relative increase in female pronouns in the LOB family corpora of British English, could suggest moves towards ‘reductions in gender-based bias’. He argues similarly for American English that the apparent decrease of terms like ‘men’ in the Brown family corpora could be an indication of ‘a decline of male-focused discourse’ (2017: 101).

This paper focuses on the promotion of combined pronouns (s/he, his or her, him/her, etc.) as an example of late-twentieth century non-sexist language reform which had an overt democratising aim: to increase the visibility of women in discourse. Arguably, the most democratic use of pronouns would be to eliminate gender entirely using something like singular they, and indeed existing research has shown singular they to be the epicene pronoun of choice in British English (Paterson 2014). However, singular they has received little formal prescriptive endorsement in grammar guides, whereas combined pronouns – despite being labelled as clunky, cumbersome, or ugly – have been promoted as gender-neutral, or in a sense, democratic, pronouns.

The paper uses the LOB and Brown families of corpora to diachronically and synchronically study the use of combined pronouns in written American and British English from the 1930s to the early 2000s. Although combined pronouns are relatively rare in the corpora overall, there are enough tokens to use concordance analysis to determine particular patterns in their use. This paper determines not only what forms these patterns take, but questions whether the use of combined pronouns is influenced by (a combination of) syntax, semantics, or pragmatic functions.
Conversationalization and democratization in a British radio chat show - one-way traffic?

Nick Smith
University of Leicester

This paper investigates conversationalization/democratization in two samples of a specialized register, the radio chat show, represented by BBC Radio 4’s Desert Island Discs (DID). Fairclough (1995) has argued that recent shifts in broadcast media, towards styles similar to private conversation between equals, are motivated by concerted attempts to mask power disparities between producers and receivers (listeners/viewers) of discourse. Most broadcast media researchers (e.g. Scannell 1989; Tolson 2006; Coupland 2014) tend to view more personable, informal styles not in terms of ideological agendas so much as a genuine need to attract audiences by not appearing too distant. While the boundaries between democratization and conversationalization are blurred, it is generally assumed that the ‘direction of travel’ in British broadcasting discourse has been invariably towards more, rather than less, conversational/democratized over time. Our paper seeks to test this assumption empirically. We chose DID because of its longevity and consistency of format, and a good (albeit not enormous) range of guest speakers. Moreover, BBC Radio 4 is an interesting test case as it has consistently maintained a profile of highbrow entertainment designed for an educated audience (Hendy 2007).

Our research questions are:

1. What evidence is there of democratization/conversationalization of style in DID, regardless of sampling mode?
2. What evidence is there of counter-examples to democratization/conversationalization?
3. What linguistic features – whether pro- or contra-conversationalization – are sociolinguistically sensitive, i.e. vary across demographic groups?

To address these questions we use two samples of DID, a Random sample that broadly reflects the skew in the show towards guests of male, middle-class, over-50s backgrounds, and a Sociolinguistic 'judgment' sample that carves out more balanced and consistent proportions of the categories age, gender, education and occupation (Smith & Waters 2018). Each sample contains 56,000 words, with 20 speakers of English English in each of two periods: early 1980s and early 2000s. We employ a 'key POS-tags' approach (Rayson 2008) to automatically compare grammatical frequencies in each sample over time. Features discovered to be shifting to a significant degree are checked manually, coded in concordances, and (in the sociolinguistic sample) examined for inter-
group differences. Results so far suggest a prevailing shift towards conversationalization in both samples, e.g. increasing use of second person pronouns, cognitive present tense verbs (e.g. think), and negative particles, and decreasing use of features reflecting a ‘nouny’, informational style, e.g. proper nouns, articles. The clearest evidence of democratization per se is a reduction of title nouns across the board. However, the samples differ significantly regarding e.g. generic you vs. one (shifting significantly only in the Sociolinguistic sample) and derived -ly adverbs (increasing only in the Random sample). The latter finding represents an apparent counter-example to conversationalization, implying a tendency for speakers to construct a ‘literate’ persona for the Radio 4 audience. As this tendency is greater among speakers of higher education and occupation groups, we suggest that the extent of conversationalization/democratization in a register is at least partly a function of who makes it into the corpus.

References


“The chairperson said she believes that the worst days are over” – Why gender-neutral terms are not gender-neutral in outer circle varieties

Susanne Wagner
Johannes Gutenberg University Mainz

The title quote, taken from a Nigerian newspaper (NOW 17-05-07, Davies, 2013a), exemplifies the crux of gender neutrality in English (cf. Ehrlich & King 1994; Hellinger 2001; Romaine 1999): while a large class of occupational nouns have been marked for gender in a number of different ways for centuries (e.g. steward vs. stewardess, chairman vs. chairwomen, doctor vs. woman/female doctor), over the past decades, a drive towards gender neutrality in language to mirror a changed perspective of occupational predilections has resulted in the emergence of gender neutral terms such as -person compounds or, ultimately, the complete loss of any signs of a +human form (e.g. chair; e.g. Holmes & Sigley 2002; Holmes et al. 2009). However, while a certain amount
of ‘democratisation’ (Farrelly & Seoane 2012) seems to have taken place with nouns, pronouns in general and personal pronouns in particular are still lagging behind: In the title quote, the gender neutrality of chairperson is immediately revoked by the subsequent use of ‘old-fashioned’ she only two words onward rather than more modern singular they (“the chairperson said they…”).

Based on some 2,000 examples of chair nouns from GloWbE (Davies 2013b), the present paper analyses the patterns of use regarding the correlation between the type of chair noun employed (gender-marked chairman, chairwoman; gender-neutral chairperson, also vis-à-vis spokesperson; non-human chair) and other factors such as the mention of a (gender-identifiable?) name (first, last) and pronoun choices (he, she, they; other gender-markable pronouns such as possessives) in 20 varieties of English world-wide. The major question to be answered is whether ‘democratisation’ patterns differ in inner and outer circle varieties. The results are surprising in that they indicate that the use of chair nouns follows a continuum-like pattern: while only chair is truly gender-neutral in inner circle (L1) varieties and often correlates with singular they, the use of chairperson in the outer circle is not truly gender-neutral, but rather marked as +female, and chairmen are (almost exclusively) male.

A relative lack of instances of chair to denote women in outer circle varieties compared to inner circle varieties supports the suggestion that chairperson is the most likely label to be used amongst outer circle countries to signify female presence. Furthermore, linguistic undermining of the gender-neutral functionality of chair in outer circle varieties – for instance by including +female pronouns or terms of address such as Mrs/Ms – compared to a conscious maintenance of this function in inner circle varieties suggests the label loses its status as a true generic in outer circle varieties of English. Overall, the results indicate that speakers of inner circle varieties are more aware of the gender-neutral functionality of chair as an occupational role, a notion which has not been extended yet to the outer circle.

References


Smith (Eds.), *Comparative Studies in Australian & New Zealand English*. Amsterdam: Benjamins, 183–204.

So, what about ‘so’? – The discourse marker ‘so’ in video-mediated communication in the Expanding Circle

Stefan Diemer, Marie-Louise Brunner and Caroline Collet
Saarland University & Trier University of Applied Sciences

This paper examines the use of the discourse marker (DM) ‘so’ in video-mediated conversations (VMC) by speakers of English as a Lingua Franca (ELF). While ‘so’ has attracted considerable attention in the context of face-to-face conversations, particularly in a native-speaker context, its use in video-mediated conversations between non-native speakers has not yet been documented. The study uses examples from a corpus of 20 Skype conversations (113,677 tokens, 152,467 with annotation) between non-native ELF speakers from different European countries compiled by the CASE project (Brunner et al. 2018). Its multimodal composition allows research on a wide range of verbal and interrelated non-verbal aspects of informal spoken video-mediated discourse.

In addition to its central role, indexing causality, Bolden (2006, 2009) describes sequence-initial ‘so’ as implementation signal for incipient action and a marker of other-attentiveness. ‘So’ signals inference (Blakemore 1988), prefaces topic development questions (Johnson 2002), and acts as a follow-up prompt in ordinary conversation (Raymond 2004). In a study of academic consultation settings in ELF, House (2013) describes ‘so’ as self-attentive, functioning as a hesitation and a transition segment marker (Redeker 2006; Brunner 2015; Brunner & Diemer 2016, forthcoming). She also finds a deictic use, looking both forward to plan new moves and backwards to sum up what has been said. House points to potential interference of the German DM ‘so’ as a possible reason for a relative lack of other-attentiveness.

A quantitative and subsequent qualitative analysis of ‘so’ in our data, shows all functions mentioned above, including numerous instances of other-oriented use of ‘so’, in contrast to House’s (2013) observations. ‘So’ also frequently interacts with the discourse marker ‘okay’ as part of a sequence of two (and sometimes even more) DMs in our data. This co-occurrence has also been observed by Koops and Lohmann (2016) who point out that the combination of ‘so’ and ‘okay’ may actually fulfill different functions than the single items on their own. Our ELF Skype data indeed suggests that the sequence may combine self-attentive and other-attentive functions, pointing to both a high pragmatic competence and a creative ELF use (cf. also Hülmbauer 2013; Mauranen 2012; Widdowson 2015). The use of ‘so’ is also situated within an embodied performance (Goodwin & Goodwin 2000), with paralinguistic aspects (such as intonation, pauses, pitch, vowel length, and
laughter) and non-verbal aspects (in particular gestures and facial expressions), all of which contribute to its situational meaning in our data.

In sum, the use of ‘so’ in spoken ELF Skype conversations illustrates uses established in native speaker research while at the same time also showing the considerable complexity and flexibility inherent in lingua franca communication, broadening the range of functions it fulfills in interaction.

References


Brunner, M.-L., & Diemer, S. (2016). “Mhm, ... okay so u:h, maybe we should start with this topic” – Conversation starts in English as a Lingua Franca (ELF) Skype talks. IWODA’16 – Fourth International Workshop on Discourse Analysis, Santiago de Compostela.


From the Expanding Circle to the Inner Circle: Discourse-pragmatic markers among migrants residing in Ireland and Australia

Chloé Diskin
The University of Melbourne

Recent studies have shown that non-native speakers (NNS), or Expanding Circle speakers of English, use discourse-pragmatic markers (DPMs) differently to native speakers (NS), or Inner Circle speakers. Differences have been observed in terms of frequency, function and distribution (see Müller 2005), which are to be expected considering that DPMs are typically not explicitly taught in English language classrooms of the Expanding Circle (Davydova & Buchstaller 2015: 444). It has been demonstrated that as English proficiency increases, the more likely NNS are to use DPMs at similar rates to NS (Hasselgreen 2004; Neary Sunquist 2014).

However, studies focusing specifically on migrants who have moved from Expanding Circle to Inner Circle countries have also found length of residence (LoR), or degree of acculturation, to have a more significant effect on the degree to which local patterns of use are adopted (Diskin 2017; Hellermann & Vergun 2007). Furthermore, examining frequency alone can obscure underlying differences. The range of discourse-pragmatic functions performed by DPMs has been found to be more limited among NNS, with textual or coherence functions being preferred over intersubjective functions (House 2009). The position of a DPM within the turn can also vary, both among NS (Heritage 2015) and NNS (Corrigan 2015; Diskin 2017; Nestor & Regan 2015).

This paper presents a quantitative analysis of the frequency, position and function of discourse-pragmatic markers (including ‘I mean’, ‘like’, ‘look’, ‘well’, and ‘you know’) in two L2 corpora of sociolinguistic interviews: the first consists of 42 interviews with Polish and Chinese migrants in Ireland, and the second of 14 interviews with Chinese migrants in Australia. Each corpus has a comparator sample of interviews with NS. Approximately 3,000 tokens of DPMs are examined.

Using fixed and mixed effects regression models, results show that in Ireland ‘you know’ is employed significantly more frequently among Polish migrants, and particularly in turn-medial position, as compared to both the Chinese and the NS, with no effect found for proficiency. ‘Like’ is readily used by all migrants in Ireland, but its use as a quotative (‘be like’) is confined to those with higher English proficiency. ‘I mean’ is used more frequently in the Australian corpus than in the Irish one, reflecting similar trends for British English when compared to Irish English (Kallen 2005). Turn-initial ‘look’ is more frequent in the Australian corpus, where it is often used as an opener to introduce information, or to convey intimacy; whereas in the Irish corpus it is relatively infrequent, and usually employed as a conversation closer, or when conveying orders.
This comparative analysis shows that in different geographic locations in the English-speaking world, L2 speakers are fine-tuned to the variety to which they are exposed, and are more likely to accommodate to local norms than to use DPMs in more universally predictable ways. Nonetheless, factors such as LoR and proficiency continue to play a role in these situations of language contact, calling into question more general issues surrounding learnability in a second language, particularly when the input is variable.

References


While much has been written about the unique set of discourse particles in colloquial Singapore English (e.g. Lim 2007), among them *lah* as the most common, the form(s) and function(s) of other discourse markers which belong to the common core of English have been neglected. This paper deals with *basically*, a rather recent addition to the paradigm of discourse markers which can be shown to move along the following clines in their development:

- **Clause-internal adverbial > Sentence adverbial > Discourse particle** (“of which discourse markers are a subtype”) (Traugott 1995, p.1)
- **scope within the proposition > scope over the proposition > scope over discourse** (Brinton 2006: 330, cf. Traugott & Dasher 2002, p.40)
- **epistemic > rhetorical** (Schwenter & Traugott 2000, p.22)

*Basically* occurs more frequently in the spoken ICE-Singapore subcorpus (252.99 per mio. words) than in the parallel ICE-GB section (174.95 per mio. words), and its frequency for Singapore ranks second highest among the varieties represented in GloWbE (71.67 per mio. words against the average of 58.52 per mio. words). This paper will use the evidence from ICE-Singapore to trace the functions and contexts for *basically* and to determine whether the grammaticalization clines proposed on the basis of British English also hold for an Outer Circle variety such as Singapore English with its rich array of indigenized discourse particles. It will further take *basically* in Singapore English into the present by taking into account the GloWbE blog data. A secondary aim of this paper is then to evaluate in how far the GloWbE blog subcorpus represents “informal language” mirroring spoken language as claimed by Davies & Fuchs (2015: 26), thus contributing to the discussion on the advantages and disadvantages of big data vis-à-vis balanced corpora (cf. Lange 2017; Loureiro-Porto 2017).

**References**


In like South Korea: The use of Like in Korean English speech

Sofia Rüdiger
University of Bayreuth

Even though English is learned as a foreign language in South Korea it has a very profound status in the society. Not only is it recognized as the language of prestige and social advancement (see e.g. Park 2009) but it is also very visible within the country (cf. the influences of English on the Korean language as documented for example by Shim 1994). English in South Korea has so far mainly been investigated from a qualitative point of view and despite a plethora of studies on English language use on television, popular music, advertising, and the linguistic landscape, next to a number of studies on language attitudes, large-scale investigations employing thorough corpus linguistic methods are still lacking (for a notable exception see Hadikin 2014). In this study, I employ the Spoken Korean English Corpus (SPOKE) to investigate the use of discourse markers by Korean speakers of English. The corpus consists of 60 hours of spoken conversational material between 115 young educated Korean speakers of English and the researcher which as transcribed text amounts to ca. 300,000 words (excluding interviewer speech). The corpus material allows for a full quantification of the investigated phenomena.

This talk specifically focuses on the use of like by Korean speakers of English. As the first description of discourse marker use in Korean English speech, the study at hand will be fundamental for further inquiries in the subject. The interest in like was inspired by a case study of like used as a preposition (Rüdiger 2017), which also (tentatively) suggested that discourse marker like might be exceptionally popular in the data examined. Like, in general, is connected to a range of uses and functions, which have received variable attention by linguists and the lay public. D’Arcy describes some functions of like (e.g. verb, preposition, or conjunction) as “unremarkable” (2017: 3), while other uses (e.g. discourse marker, discourse particle, quotative) are more innovative and more often “remarked upon in metalinguistic discourse” (2017: 14). The present study investigates the use of like as discourse marker or particle and as a quotative by Korean speakers of English.
Altogether, *like* is a high frequency item in the Korean English corpus (1502 instances per 100,000 words; raw frequency 4,497). The discourse particle function is the most frequently used one but quotative *be like* does also occur in the data set. Expanding Circle English speakers, at least Koreans, therefore draw on a similar functional range than speakers from Inner Circle varieties when it comes to *like*. The frequent use of pauses, repetitions, and hesitation markers around *like* in the spoken Korean English corpus indicates, however, that it might also have an additional functionality as a filler or disfluency marker. This function also exists in other (Inner Circle) varieties of English but has been described as “extremely rare” (D’Arcy 2017: 16).

References


**Periphrastic DO as inherited discourse marker – Use and function in Inner and Outer Circle varieties of English**

*Ninja Schulz*

University of Würzburg

According to Quirk et al. (1985), auxiliary DO in affirmative sentences (henceforth DO+) is used in emphatic or contrastive contexts, where it carries stress. However, studies have shown that DO+ is not always stressed and that it can assume a number of discursive functions as well (Nevalainen & Rissanen 1986; Ranger 2015), which sets it apart from semantically empty periphrastic DO in questions or negative statements. For example, DO+ can be used as sincerity marker (*I do promise...*), to reintroduce a topic in the discourse (*I did mention that...*), or to express epistemic stance (*The results do show that...*). The former functions are more frequent in spoken discourse while the latter one is prevalent in written texts, more specifically, in academic writing. In addition, the different functions tend to occur with specific sets of verbs, so that it is often possible to deduce the function of DO+ from the verb it is used with.

These observations are mostly based on corpora of British and American English, while there is a lack of evidence from other varieties of English. Although the use of auxiliary DO is described for single varieties (Kortmann & Schneider 2004) and although Kortmann (2004) conducted a comparative study of it in World Englishes in general, both kinds of studies focus on grammatical
functions of DO, for example as habitual or progressive marker. The use of DO+ in varieties where it has not acquired an additional function as aspect marker has received little attention so far. As it is a rather infrequent phenomenon, which is not part of formal language education, and as it is not limited to one single function, it is liable to changes and reinterpretations by speakers of new varieties.

The present study focusses on the distribution of DO+ in Inner and Outer Circle varieties of English, using various ICE components as database since they provide a (largely) comparable set of different text types. Distributional differences are analysed by calculating the relative frequency of DO+ (occurrence per simple verb form) for spoken and written language as well as for different subgenres. The results show that DO+ is generally more frequent in Inner Circle varieties with New Zealand leading the field, while the ratio for ICE Philippines is the lowest. Furthermore, ICE India shows a conspicuous preference for the use of DO+ in written texts, while for all other varieties this ratio is lower than for spoken data. In a next step, the types of verbs used with DO+ are categorised to analyse whether the same functions are used or whether some are absent and/or new ones have emerged in different varieties, thus indicating tendencies concerning (dis)preferred functions and possible innovations.

References


“It gives me the confidence and all that” – General extenders world-wide

Susanne Wagner
Johannes Gutenberg University Mainz

The title quote, taken from the Nigerian data in GloWbE (Davies 2013), illustrates one of a host of types of general extenders (GE) in English: and all that here is used “to evoke some larger set” (Dubois 1992: 198) of items, generalising from a concrete (set of) preceding referent(s) to a larger group of similar items (cf. Tagliamonte & Denis 2010).
Previous scholarship has shown that GEs in English are typically composed of a combination of items from a relatively limited class of words: a connector (adjunctive and or disjunctive or), a quantifier (e.g. all, some(thing), anything etc.), a generic term (thing, stuff etc.) and possibly a comparative (like that, kind sort type of; cf. Wagner et al. 2015). The order of elements may vary, resulting in sequences such as short fixed expressions like or whatever but also in long and that kind of thing, as well as innovative variations.

Most studies so far have focused on the general distribution of GEs in different L1 varieties of English (e.g. Cheshire 2007; Palcios Martínez 2011; Pichler & Levey 2011 on British English; Tagliamonte & Denis 2010; Denis 2017 on Canadian English; Overstreet 1999, 2005, 2014; Terraschke & Holmes 2007 on American and New Zealand English); few studies have included data from learner or second language varieties (but cf. Overstreet 2005; Parvaresh et al. 2012; Terraschke 2007; Terraschke & Holmes 2007). The present paper aims at filling this research gap by analysing GEs in “big data”, viz. GloWbE: this corpus contains web-based language collected from over 34,000 websites representing 20 varieties of English world-wide, adding up to some 1.9b words (Davies 2013).

GEs typically occur utterance-finally, like other discourse-pragmatic features. This is helpful when extending research on GEs to “big data”, since punctuation marks can be used to narrow down the number of false positives. Based on a quantitative investigation of 18 of the most common GEs, the following trends can be observed in GloWbE:

a) Concerning the preference of adjunctives (and …) or disjunctives (or …), previous studies showed differences between varieties and/or levels of formality (cf. Cheshire 2007). In GloWbE, disjunctives are generally twice as frequent as adjunctives; however, regional differences are obvious: the ratio is closer to 3:1 for the US data and generally >2:1 for all L1 varieties; the only L2-varieties approximating that are (indigenised?) Singapore and Hong Kong. Interestingly, both West African representatives, Ghana and Nigeria, come in at almost 1:1, warranting closer inspection.

b) In terms of overall frequencies, disjunctives display much less variation than adjunctives, with the top 5 forms ranked practically in the same order worldwide. For adjunctives, however, variation is the rule, not the exception: and stuff figures much more prominently in L1 varieties, competing for first rank with and everything, which dominates in L2s. And all that is the most interesting adjunctive, oscillating between first (Nigeria, Ghana, GB, Ireland) and third rank in most of Southeast Asia.

References

Tag questions in Trinidadian and Philippine English: A corpus-pragmatic analysis of form, function, and text type variation

Michael Westphal
University of Münster

The International Corpus of English (ICE) with its many individual components, diverse text types, and large spoken part is an excellent tool to analyze variation between and within World Englishes on all levels of linguistic variation. However, research using the ICE is generally biased with regard to two aspects: first, most ICE-based research has focused on morpho-syntax (e.g. Hundt & Gut 2012), while there are only some phonetic studies (e.g. Rosenfelder 2009) and hardly any investigations on pragmatic variation (e.g. Aijmer 2013). Second, most cross-variety comparisons either pool text types together or merely use face-to-face conversations (e.g. Hundt & Gut 2012).

This corpus-pragmatic study investigates the use of tag questions (TQ) as one set of discourse markers in two outer circle varieties of English, Trinidadian English (TrinE) and Philippine English (PhiE), by means of the two respective ICE corpora. In the ICE-Philippines and the ICE-Trinidad and Tobago, I analyze the form and the pragmatic function of canonical and invariant TQs in four different text types, which differ with regard to their communicative situation: face-to-face
conversations, phonecalls, classroom lessons, and legal cross-examinations. I investigate the effects of text type on the form and function in the two outer circle varieties by means of mixed-effects modelling.

In both varieties, speakers prefer invariant to canonical TQs across all four text types. However, there are cross-variety differences with regard to the form of invariant TQs: for example, Tagalog tags, such as *ba* or *ano*, are exclusive to PhiE, while *nah* and *not so* are only used in TrinE. Most invariant TQs are shared by both varieties but there are preferential differences: for example, *OK* is more frequent in PhiE, whereas *right* is preferred in TrinE. For both varieties, text type has a strong and similar effect on the frequency, form, and function of TQs: for example, TQs are most frequent in classroom lesson, while they are least frequent in legal-cross examinations. *You know* is preferred in conversations, whereas *OK* is most frequent in classroom lesson. Facilitative functions dominate in classroom lesson, emphatic uses are most prevalent in conversation, and informative functions are preferred in legal cross-examinations.

The analysis shows that corpus linguistic and especially ICE-based research needs to widen the canon of linguistic variables and include discourse markers as well as other pragmatic variables. Similarly, investigations in text type variation are dominated by Biber’s (e.g. 2014) multidimensional analysis, which largely excludes pragmatic variables, but a pragmatic perspective has the potential to add to a better understanding of text type variation. Corpus-pragmatic research, which includes form, function, and text type variation, has the potential to pinpoint differences as well as similarities in World Englishes more clearly and to challenge assumptions of homogeneity and exceptionality in New Englishes: speakers of different varieties use different forms but have the same communicative needs in similar situations.

References


Plenaries
The Hansard Corpus through a telescope

Marc Alexander
University of Glasgow

The Hansard Corpus (1803-2003) contains 7.6 million speeches from the UK Parliament, not verbatim. Its size – 1.6 billion words – means that it is particularly difficult to easily explore and summarise. As a result, in 2015 it was tagged semantically using the tagset of the Historical Thesaurus of English in order to enable semantic queries and aggregation. In this talk, I will discuss what the corpus represents, the overall picture of the Parliamentary record from a semantic point of view (‘through a telescope’), and then describe some recent case studies enabled through the semantic tagging.
What the kids can tell us about language change

Alexandra D’Arcy
University of Victoria

Drawing on findings from dialect acquisition, historical linguistics, and variationist sociolinguistics, this talk addresses one of the most long-standing and central questions in the study of language change: How does change advance across successive generations of speakers? Past research has engaged in post-hoc theorizing about the continuous advancement of change but has rarely addressed it directly. Because children must come to speak differently from their parents for any change to both survive and progress, only extended and continuous real time observation of the same speakers can provide answers to this question. In this presentation I introduce a project that explicitly sets out to observe the onset and early progression of change in order to track the diachronic evolution of specific linguistic features, zeroing in on the period when children begin to participate in change by shifting their language model along an apparently pre-set direction of change: the preschool years. Evidence from the community, caregivers, and children is triangulated, in an attempt to understand the dynamics of language change as they are activated on the ground and as they intersect with social aspects of community structure in daily interactions. I present findings that allow me to propose tentative answers to the following kinds of questions: What is the starting point for language change—the adult community model or one in which the relevant vector is shifted? Is the starting point constant across gender groups? Is inception of change parallel across girls and boys? The results of this work have implications for current epistemologies and models of language change.
Large diachronic corpora such as the COHA or Google Books offer data that is tempting in its promise: In the words of Michel et al. (2010), it can “provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology”. Studies that have pursued this line of thought have received widespread attention in the media, but have been met with reservation in the linguistic community. For example, Greenfield (2013) takes frequency trends of words such as choose and get as evidence for increasing individualism and materialism in American culture. Liberman (2013) comments on the paper as follows: “I’m not arguing that her theory is wrong, or that the Google ngrams datasets don’t contain supporting evidence. But it’s going to take a much more careful and systematic analysis of the lexico-historical data to convince me.” Liberman’s comment raises the central question for this talk: How exactly can diachronic corpus data be analyzed in order to yield reliable insights about social change?

My talk will be structured into two parts. The first, theoretical part surveys a series of problems that need to be controlled for in analyses of diachronic textual data. These problems include the increasing risk of spurious correlations in large datasets (Roberts & Winters 2013), the problem of non-stationarity in time series data (Koplenig 2015), and the difficulty of disentangling cultural change and grammatical change (Szmrecsanyi 2016).

The second part implements these ideas in a study of change in the English make-causative (Kemmer 2001). As a grammaticalized expression of authority, the construction lends itself to a study of social change: Examples such as She made the boys clean up their room verbalize that a causer prompted a causee to perform a coerced action. If American culture becomes less authoritarian, as has been argued by Greenfield (2013) and others, examples such as the one above should recede in favor of uses such as That music made me smile, which involve inanimate causers and non-coerced actions. Data from the COHA is retrieved to track the history of the make-causative in terms of several semantic parameters, including animacy of causer and causee and the semantics of the verb that expresses the caused action. Distributional evidence makes clear that the construction has changed in a way that is consistent with the hypothesis of social change.

The final part of the talk will discuss whether the empirical findings can only be explained in terms of social change, or if processes such as subjectification and intersubjectification (Traugott 2010) can offer an alternative account.
“ Entirely false” or “hardly true”: The socio-pragmatics of intensifiers in the late modern courtroom

Merja Kytö
Uppsala University

The development of intensifiers presents numerous exciting turns in the history of English. The work reported on in this paper will focus on Late Modern English, a period that has been rather neglected in research on intensifiers so far. Variation and change in the use of two groups of intensifiers will be investigated, namely amplifiers scaling upwards (e.g. entirely, perfectly, greatly) and downtoners (e.g. hardly, (a) little, faintly), in British courtroom speech from 1700 to 1900. The inventory of forms and their distributions across various types of speakers will be mapped, both with regard to speakers’ social (e.g. gender, social class) and functional roles (e.g. judge, witness). This setup will enable one to provide a sociolinguistic snapshot of Late Modern English in a speech-related context. Spoken language is where intensifiers have been shown to be particularly frequent in Present-day English and prone to intensive change. The analyses will also include discussion of the targets of modification, i.e. to what extent and how intensifiers are used to modify constituents across the period covered.

The material will be drawn from the Old Bailey Corpus (OBC, version 2.0), which comprises 637 trials totalling 24.4 million words drawn from the Proceedings of the Old Bailey available on the net (over 197,000 trials, or 134 million words, https://www.oldbaileyonline.org/). OBC was created at the University of Giessen to contain speech-related language annotated at utterance level for sociobiographical speaker information (gender, age, occupation, social class), pragmatic information (speaker roles such as witness, defendant, judge, lawyer) and textual information (scribes, printers and publishers of the Proceedings).

Based on historical sociolinguistics, historical pragmatics and corpus linguistic methodology, it will be possible to identify change in progress, innovations as well as obsolescence, in the use of intensifiers within these 200 years. The results presented are part of a project aimed at a comprehensive account of intensifier usage and development in the Late Modern English courtroom context. The project team comprises Claudia Claridge and Ewa Jonsson, in addition to the present author.
Full papers
This study seeks to take up Biber, Gray, and Poonpon’s (2011) suggestion to provide empirical evidence for their hypothesized developmental stages for second language learners by investigating the use of different kinds of noun phrases. Parkinson and Musgrave (2014) analyzed the writing data of students in an English for Academic Purposes (EAP) program (lower group) and students in a Master of Arts (MA) program (higher group) to compare against the academic prose data reported by Biber and Gray (2011) and Biber et al. (1999). They found that adjectives were used heavily in the lower group, while nouns, possessives, participials, prepositional phrases, and appositives were used significantly more often in the higher group, confirming the stages suggested by Biber et al. (2011). However, their study was limited in that their corpora came from different genres and L1 groups, which may have influenced the structures used in the data.

We attempted to further test Biber et al.’s (2011) hypothesis while minimizing the effects of differing genres and L1 by comparing texts of different score levels from a subcorpus of the ETS corpus written by Japanese learners of English. The data totaled 90,968 words and were controlled such that all three levels (low, medium, high) and all the topics were similarly represented. After the data were syntactically tagged by Stanford Parser (Chen and Manning 2003), the noun phrases were extracted through a combination of automatic Tregex (Levy and Andrew 2006) searches and manual pruning. Expressions exactly matching the structures in the essay prompts were removed. A total of 8,337 noun phrases (2,656 low; 2,769 medium; 2,912 high) were extracted. The noun phrase modifications were grouped into nine groups: adjectives, participial premodifiers, noun premodifiers, relative clauses, prepositional phrases, past participial clauses, present participial clauses, appositives, and to-clauses.

Our results demonstrate a similar overall pattern to Parkinson and Musgrave’s (2014), with adjectives and prepositional phrases being the most frequent (22 to 39 occurrences per 1,000 words), noun premodifiers and relative clauses being moderately frequent (9 to 13 occurrences per 1,000 words), and participials, appositives, and to-clauses being least frequent (under 5 occurrences per 1,000 words). Regarding the distributions of each modifier among the different levels, adjectives, noun clauses, and relative clauses appeared at a similar frequency in all groups, while prepositional phrases were produced more frequently in the high group than in the low group. The even distribution of adjectives in our study is notably different from Parkinson and Musgrave’s (2014) results, where the lower group produced significantly more adjectives compared to the higher and academic prose groups. Although the extremely low frequencies of participials, appositives, and to-clauses make it difficult to say whether Biber et al.’s (2011) hypothesis holds true for the later-stage structures, the observed trend is suggestive of their hypothesized developmental stages in that the three groups are represented evenly in the early-stage structures (adjectives, nouns, and relative clauses) while prepositional phrases (a middle-stage structure) were produced most frequently by the high group and least by the low group.
‘That’s absolutely fine’ – An investigation of **absolutely** in the spoken BNC2014

**Karin Aijmer**
University of Gothenburg

Intensifiers are characterized by their rapid turn-over and constant state of renewal and therefore need to be studied in up-to-data spoken corpora. The starting-point for this study which is part of a broader study of recent changes in the intensification system was the observation that the intensifier **absolutely** was more frequent in the sample release of the Spoken BNC2014 (Love et al. 2017) in comparison with its use by previous generations of speakers in the (demographic) spoken component of BNC from the 1990s. The changes undergone by **absolutely** have been given less consideration than recent changes involving **really** and **so**.

The aim of the present study is to describe the rise in frequency and changes which have taken place in the twenty-year period separating the two BNC corpora in a linguistic, discourse and a sociolinguistic perspective. The results will serve to give a picture of how frequency changes correlate with grammatical and semantic changes and contribute to the debate on the relationship between changes observable over a short diachronic period and long-term developments involving grammaticalization. Palacios Martínez & Núñez Pertejo (2014) noted, for example, that **absolutely** was more frequent as a verb modifier or a noun modifier in the Bergen Corpus of London Teenage Language (COLT) than in a comparable adult corpus but they did not discuss the developments in a diachronic perspective. The more specific research questions are:

- What are the changes in the frequency of **absolutely** across the twenty-year period?
- How are the changes reflected in the patterns of modification with different syntactic constituents (in particular nouns and verbs) and the free-standing **absolutely**?
Intensifiers put the spotlight on the adjective combining with absolutely. If absolutely is on the rise in present-day British English this may be indicated by the co-occurrence with trendy, age-related adjectives (brilliant, fucking, insane, knackered), and correlation with the positive or negative value of the adjective (semantic prosody).

How can the paths of change be described in terms of an increase in subjectivity (intensity) and principles of grammaticalization? How can the dependent (intensifying) and independent (response item) uses of absolutely be explained as different stages of grammaticalization?

What is the role of the speakers’ age, gender and social class to explain the grammatical and semantic developments of absolutely (cf Murphy 2007)?

Methodologically, the study uses corpora to compare the use of absolutely at two different sampling points. Given the design of the BNC corpus it is also possible to use the metadata to investigate the importance of a number of sociolinguistic factors such as age, gender and class which can influence variations and change of absolutely.

Preliminary results show that absolutely is becoming less frequent as an adjectival modifier (30% of all the examples of absolutely to be compared with 45.6% in the ‘old’ BNC) and that it is expanding to new syntactic and semantic contexts in conversation where it performs discourse-oriented rather than intensifying functions. In particular it is used as a freestanding lexical item with the function of a response particle marking strong agreement (Tao 2007, Carretero 2010).

References


A diachronic-contrastive approach to phraseology

Gisle Andersen
NHH Norwegian School of Economics

Whether based on translation corpora or comparative corpora, contrastive corpus linguistics has proven useful to ‘investigate etymological or semantic cognates’ (Aijmer & Simon-Vandenbergen 2006: 4) across languages. This approach has shown, interestingly, that the degree of functional equivalence of cross-linguistically similar forms may be substantial (e.g. Foolen 2006 on Dutch toch vs. German doch) or minimal (e.g. Bruti 1999 on Italian infatti vs. English in fact). This paper builds on the contrastive corpus approach, but extends it to the study of linguistic units which are not only structurally similar, but presumed to be the result of borrowing from one language to the other.

In the fashion of Fiedler’s (2012, 2017) recent work on German, this study concerns phraseological units that have emerged in Norwegian but which have structurally equivalent patterns in English, such as tingen er at / the thing is that; når det kommer til / when it comes to; å gjøre en forskjell; to make a difference.

Previous research as well as public opinion seem to take for granted that this parallelism is the result of borrowing, and precisely for this reason, the usage is often subjected to prescriptive criticism (Graedler 1998; Andersen 2014, 2016). Among the strengths of the diachronic-contrastive method is that it discloses relevant facts about the emergence and frequency developments of such structures in the two languages, which in turn informs the issue of whether a structural parallelism is the result of borrowing or separate developments. Phrasal units are typically form/function pairings which extend metaphorically from their literal sense and develop non-referential and often discourse-organising functions. Inspecting concordances from corpora, we can establish timelines and frequency profiles for these units and observe whether a form has a well established function in a source language (SL) before it emerges in the recipient language (RL) and takes on consistent and recurrent use there.

The diachronic-contrastive method will be illustrated by the use of comparable historical and diachronic corpora in both languages. For English, I use the Old Bailey Corpus (1674–1913) and COHA (1810–2009) as historical sources, and COCA (1990–2015) as a contemporary source. For Norwegian, I investigate the Bokhylla Text Archive (1690–2013), the Nynorskkorpus (1870–present) and The Norwegian Newspaper Corpus (1998–present).

The paper focuses on eight phrases which are either direct or indirect borrowings and which display different degrees of functional parallelism in the SL and RL. The study shows that corpus frequency profiles are (only) indicative of borrowing. A borrowed phraseological unit is typically preceded by a significant frequency increase in SL, but a frequency increase is not a sufficient criterion for borrowing. The results also show counter-cases where the presumed SL-RL trajectory is a misconception, either because data suggest that the two forms developed in parallel, or because the origin may not be conspicuously determined. Thus, this corpus approach leads us to reject
preconceived ideas about a form’s alleged origin in English – what could be called the “Anglicism Illusion” (cf. Zwicky’s Recency Illusion & Frequency Illusion; Zwicky 2005a Zwicky 2005 b).

References

http://itre.cis.upenn.edu/myl/languagelog/archives/002407.html
http://itre.cis.upenn.edu/~myl/languagelog/archives/002386.html

Two sides of the same coin? Tracking the history of the intensifying adverbs *deadly* and *mortal*(ly)

**Zeltia Blanco-Suárez**  
University of Cantabria

Given their tendency to rapid change and constant renewal or ‘recycling’ (Tagliamonte 2008), intensifiers have for long been at the forefront of academic discussions on semantic and sociolinguistic change (cf., among others, Macaulay 2006; Méndez-Naya 2008; Nevalainen 2008; Barnfield and Buchstaller 2010; Claridge and Kytö 2014; D’Arcy 2015).

The present paper is also concerned with intensifiers, specifically with forms from the semantic field of death, a topic which has so far remained largely unexplored (cf., however, Claridge 2011 and Margerie 2011). The aim of this study is twofold. On the one hand, it sets out to explore the diachronic development of two such death-related forms, the vernacular adverb *deadly* and its Anglo-Norman counterpart *mortal*(ly). On the other, it looks into the frequency and current usage of both adverbs in the contemporary language.
Our preliminary data suggest that despite having the same meaning originally ('subject to death'), as in (1) and (2), the diachronic development of *deadly* and *mortal(ly)* is not entirely parallel.

(1) *He wonded þe Kyng dedely fulle sore.* (OED, s.v. deadly adv. 1a)
(2) *But this scorpion..The Sowdanesse..Caste vnder this ful mortal(ly) to styngle.* (OED, s.v. mortal B. adv.)

Thus, *deadly* seems to comply to Adamson’s (2000) cline of evolution for intensifiers, originating in descriptive or literal meanings ((1) above). Later on, it underwent subjectification (cf. López-Couso 2010; Traugott 2010) and developed increasingly affective or subjective readings (‘resembling or suggesting death’), as in (3). At a final stage, *deadly* became grammaticalised as an adverb of degree (‘extremely, excessively’) (cf. (4). The same cline, however, does not fully apply in the case of *mortal*, since it was borrowed already with subjective senses. Moreover, while *deadly* developed intensifying readings from the early fifteenth century, the degree function of *mortally(ly)* only arose in the eighteenth century (cf. (5)–(6) below).

(3) *She louede mykel þe slayn broþer, & dedlyk hated sche þat oþer.* (OED, s.v. deadly adv. 2)
(4) *I pat es sa dedli dill ['stupid, foolish'].* (OED, s.v. deadly adv. 4)
(5) *She is mortal fond of the book, and has got it by heart.* (OED, s.v. mortal B. adv.)
(6) *The people are in general mortal(ly) ugly and dwarfish.* (OED, s.v. mortally adv. 5)

Data for this corpus-based study have been drawn from a variety of sources, including the *OED*, the *Early English Books Online Corpus 1.0* (EEBOCorp 1.0), *Eighteenth Century Fiction* (ECF), *Nineteenth Century Fiction* (NCF), and the *Brigham Young University-British National Corpus* (BYU-BNC).

References and sources


BYU-BNC = Brigham Young University-British National Corpus (Based on the British National Corpus from Oxford University Press). Davies, M. 2004–. Available online at: [http://corpus.byu.edu/bnc/](http://corpus.byu.edu/bnc/).


Characteristic elements of verbless sentences: A corpus-based contrastive approach

Antonina Bondarenko
Université Paris 7 – Diderot (Clillac-Arp) & Inalco (Sedyl)

The difficulty of processing the absence of the verb from the sentence automatically has meant that most analyses of verbless sentences have relied on fragmented data. Overcoming the typical problems of fixed annotation and verb-centric syntactic modeling associated with most existing parsed corpora (Landolfi et al. 2010), we develop a new method of automatic verbless sentence extraction. Taking a corpus-based contrastive approach, we aim to uncover the semantic and pragmatic factors associated with the absence of the verb. In this paper, we focus on the grammatical categories, lemmas and forms revealed through statistical characterization of verbless sentences.

We investigate the feature from the perspective of monolingual, parallel-text and third-language translation corpora. Cross-linguistic differences make it particularly relevant to compare Russian, which permits the most liberal use of verbless sentences among the Indo-European family, with English, known for its dependency on the verb phrase (McShane 2000; Kopotev 2007). For monolingual analysis, we use the English play *The Caretaker* (Pinter, 1960) and the Russian dialogue-based *Brothers Karamazov* (Dostoyevsky 1880), selected particularly for the high frequency of direct speech. Relying on Guillemin-Flescher’s (2003) contrastive analysis principle,
we explore re-occurring verbless sentence translation patterns using a bidirectional parallel-corpus that includes several corresponding translations (Pevear & Volokhonsky 1990; Avsey 1994; Doroshevich 2006) selected in line with the parallel-corpus criteria outlined by Stolz (2007) and Nádvorníková (2017). Following Baker (1993; 2000) and Zanetti (2012), we use a third-language subcorpus, consisting of English (Ward 1989; Gilbert 1958) and Russian translations (Adamovich 1966; Gal 1968) of Camus’ French The Stranger (1942), in order to control for source language interference and compare the results with the characteristics of translated verbless sentences regardless of their original language.

The 300,000-word pilot corpus was specially segmented and annotated for automatic verbless sentence extraction with an estimated recall of 95%. Morphosyntactic tagging errors were corrected semi-automatically using Trameur (Fleury & Zimina 2014). The annotation, alignment and statistical text analysis software package was used for automatic verbless sentence extraction, paragraph-level alignment and visualization of multiple translations in their original context, and statistical analysis in terms of characteristic elements. The key forms, lemmas, and grammatical categories that are more frequent within verbless sentences than within verbal sentences were compared across the three subcorpora.

In addition to establishing the significance of expected frequency differences, preliminary results show that verbless sentences are statistically characterized by deixis and informal markers. The high specificity index of the Russian second person singular pronoun, the familiar form ты/ty, suggests that verbless sentences are associated with informal address. Correlation with informal speech is also indicated by the specificity of interjections in the verbless sentences of both languages. Overrepresentation of English deictic this and Russian это/èto (this/it) and вот/vot (here/this) imply a link with the utterance situation. The present analysis of statistically characteristic elements aims to shed light on the common ground between the interlocutors and reveal the pragmatic requirements influencing the use of verbless sentences.

References


Multitasking much? A corpus-based analysis of change in the ‘Sarcastic much?’ construction

Samuel Bourgeois
University of Neuchâtel

This paper is concerned with a construction that is illustrated below with examples from the GLoWbE corpus (Davies 2013):

(1)  a. “There is water at the bottom of the ocean”. Obvious statement much?
    b. What is this conspiracy nonsense that you are going on about? Paranoid much?
    c. We, we, we, me, me, me. Self-centered much?

According to the OED, this construction is a combination of the element much with a preceding adjective, infinitive verb, or noun phrase. Though early uses are attested in 1970s episodes of the TV show Saturday Night Live, the construction gained popularity in the late 1990s through uses in the TV series Buffy the Vampire Slayer (Adams 2014). Its primary function has been characterized as a critical assessment of someone else’s behavior, which is judged to exceed an acceptable limit on a given scale (Adams 2014: 182; Gutzmann & Henderson 2017). Furthermore, the construction has been linked to genres of computer-mediated communication, which often function to invoke both otherness and affiliation (Zappavigna 2012).

This study investigates the ‘Sarcastic much?’ construction using the GLoWbE corpus (Davies 2013), from which different structural types (Adj. + much?; Noun + much?; Verb + much?) were retrieved. The resulting concordance is analyzed with regard to form, function, context, and textual genre of the examples. Although the construction is frequently used with the forms ‘Adj/Noun + Much?’ and with a negative evaluative function, which is evident from the most frequently used antecedent descriptors, it is also demonstrated that the construction is opening up to other formal and functional variants. Specifically, newly emerging uses show that the construction serves to make metatextual comments (example 2), to positively evaluate a situation, or even to brag about an achievement (example 3).
(2) Still I kept at the classic literature because it was important to me that others respect my intelligence (damaged by high school much 😐 -P). Naturally when I decided to write a novel, it came out as literary fiction.

(3) We have a few fixed points: a dinner here, a soccer football game there. Christmas in southern Germany, New Year’s in Paris (jealous much ?!). But apart from that it’s all pretty wibbly-wobbly.

On the basis of the GLoWbE data, it is argued that the changes of the forms, functions, and possible contexts of the construction can be understood through the lens of subjectification and intersubjectification (Traugott 2010). The construction is undergoing a development leading to a generalization of its meaning away from a negative evaluation only to a more subjectified indicator of the speaker/writer’s attitudes in general. This function can be linked to its usefulness in contexts of computer-mediated communication, thus connecting issues of subjectification with considerations of text type and textual media.

References


The development and pragmatic function of a non-inference marker: this is not to say (that)

Laurel J. Brinton

University of British Columbia

The appositive marker that is (to say) is discussed in standard grammars (e.g. Quirk et al. 1985; Biber et al. 1999; Huddleston & Pullum 2002) and in the OED. It originates in Old English, possibly as a calque on Latin (Brinton, 2008).

In contrast, its negative counterpart, this/that/which/it is not to say (that) receives no scholarly attention. Online discussions suggest that it is used “to make sure the person you are talking to does not think something that is not true” (Longman Dictionary) or “for adding a statement that corrects what you have just said” (Macmillan Dictionary). Evidence from COCA points to its use as an introductory clause refuting an inference that could be drawn from the previous discourse; it
frequently occurs in the context of a following negative (receiving a positive reading) and may be accompanied by an explicit denial of the possible inference:

(1) a. **This is not to say** that no wisdom can be gained from this year’s cases. Quite the contrary. (2017 COCA ACAD)
    b. **That is not to say** there have not been chances; Argentina has lost major finals the past two summers. (2016 COCA NEWS)
    c. **Which is not to say** he didn’t work in one. If he said he did, he did… (2017 COCA FIC)

Liberman (2012) discusses the existence of rare “triple negative” forms, which he calls “misnegations”. These are likewise interpreted as positive and call for rather nuanced interpretations, e.g.:

(2) a. **This is not to say** there weren’t some men in the department who did not, shall we say, embrace the concept of having women out on patrol (1992 COCA SPOK)
    b. **This is not to say** that there aren’t times when fires don’t have a mind of their own (https://www.fs.usda.gov/main/hmnf/learning/safety-ethics)

In Present-day English (based on COCA), the variants show some genre differentiation: *that/this is not/isn’t to say* are primarily found in academic discourse, while *which is not/isn’t to say* and *that’s/it’s not to say* predominate in the spoken form.

This paper provides a corpus-based pragmatic analysis of this form as a non-inference marker and traces its development. Historical corpora (COHA, CLMET3.0, OBC) as well as the OED quotation database would seem to point to the rise of the construction in the mid-nineteenth to early twentieth century, though it is a low frequency item: earliest examples are *that’s not to say* (COHA 1840); *it is not to say* (CLMET3.0 1895 or OBC 1900); *which is not to say* (COHA 1901); and *this is not to say* (COHA 1905 or OED 1908). However, the newly available EEBO corpus (Davies 2017), provides sixteenth century examples, which seem to have the same pragmatic function, e.g.:

(3) they … haue recouered their former strength: **which is not to say**, that there haue not hapned many other admirable euents in other seasons; but these are most notable (1594 EEBO)

The gaps in the corpus data are most likely the result of genre differences in the make-up of available corpora.

References


Liberman, M. (2012, Nov. 3). This is not to say that I don’t think that it isn’t illogical. [Blog post]. Retrieved from languagelog.ldc.upenn.edu/nll/?p=4292.


Corpora


A woman who lives in the city has a sister that lives in a town: Subject relativizers in Canadian English

Marisa Brook1 and Sali A. Tagliamonte2

University of Victoria1, University of Toronto2

Subject gap restrictive relative clause markers in varieties of present-day English include a variety of types – predominately who, that and zero, as in (1).

(1a) I had an uncle THAT worked as a firefighter.
   Amelia Hannock (F, 36, Kirkland Lake)

(1b) She may have been the lady WHO cleaned the bank.
   Jeanette Dean (F, 51, Beaverton)

(1c) There’s a little cactus Ø grows there with a berry on it.
   Lachlan Moyles (M, 84, Almonte)

The wh variants – especially the productive who, as in (1b) – were introduced as relativizers in a change from above in Middle English (Romaine 1982: 213). They are well established in urban areas, where they continue to act as prestige forms (Macaulay 1991; Tottie 1997; D’Arcy &
Tagliamonte 2010). In contrast, who tends to be low frequency (or absent) in rural varieties of English in both the United States (Hackenberg 1972; Ball 1996; Tottie & Rey 1997; Bayley 1999; Tottie & Harvie 2000) and the United Kingdom (Romaine 1982: 222; Tagliamonte, Smith, & Lawrence 2005; Tagliamonte 2013: 94–105). In accordance with Romaine (1982), Tagliamonte (2013: 105) argues that wh relativizers in English are “an overlay from outside the variable grammar” (Tagliamonte 2013: 105) that have not necessarily spread to nonurban locales (cf. Tagliamonte et al. 2005: 94).

To probe these claims, this study examines Canadian English, a variety that is reputedly homogeneous (Priestly 1951: 75–76; Chambers 1986, 2010:19–20; Labov, Ash, & Boberg 2006: 217). Our question is this: to what extent have wh forms infiltrated nonurban communities of the country’s largest province, Ontario? Using D’Arcy & Tagliamonte (2010)’s study of relativizers in the largest city, Toronto, as a baseline, we draw on the Ontario Dialects Archive (Tagliamonte 2007–2010; Tagliamonte & Denis 2014; Tagliamonte 2010–2013, 2014). We extract nearly 10,000 subject relativizers across ten smaller towns and cities in the province and use mixed effects statistical modeling (Johnson 2009) and comparative sociolinguistic methods (Tagliamonte 2002) to analyze them.

In Toronto, the proportion of who among subject relativizers is 31.2% (D’Arcy & Tagliamonte 2010: 391). In each of the ten smaller locations it is much lower. Towns that have a population above 10,000 and/or that lie along Ontario’s major north-south highway have rates of who between 25% and 28%. Those that are both smaller and located in more peripheral locations have rates of who between 11% and 17%. This urban-rural divide mirrors the Cascade Model of linguistic diffusion (Labov 2003). The frequency of who is also strongly correlated (Pearson’s r: 0.87) with rates of the incoming prestige form for a different variable: -one in indefinite human quantifiers, e.g. anyone, someone, no one (Jankowski & Tagliamonte 2015).

Taken together, these findings corroborate regional differentiation in relativizer behaviour in nonurban varieties of Canadian English (D’Arcy & Tagliamonte 2010; Levey & Hill 2013; Brook 2014). Indeed, this variable is just the ‘tip of the iceberg’ of many dialect features in Ontario that bear investigation. Moreover, we have discovered that despite changing societies in the 21st century, the Cascade Model remains a viable explanation for regional patterning of variation.

References


In 2011, for the first time, the Census of the United Kingdom (UK) asked respondents to identify their main language and indicate how well they could speak English. The results revealed that, of the 4.2 million UK residents whose main language was not English, 79 per cent (3.3 million) said they could speak English ‘Well’ or ‘Very well’, while 17 per cent (726,000) said they could speak English but ‘Not well’, and 3 per cent (138,000) said they could not speak English at all (ONS 2013). Despite the results showing remarkable linguistic homogeneity, with 99.74 per cent (53.9 million) of residents having some command of English either as a main or additional language, the UK media was nonetheless quick to frame the result, in sensationalist terms, as signifying the decline of the English language and the erosion of British culture more widely. This study examines the right-leaning UK print media’s response to the Census result, in particular its representation of people and groups who can’t speak English, in the five years following the 2011 Census. The analysis takes a corpus-assisted approach to critical discourse analysis (Baker et al. 2008), and is based on a 2.2 million-word specialised corpus of right-leaning tabloid and broadsheet newspaper articles concerned with the topic of ‘speaking English’ published between 2011 and 2016. The analysis evinces the tendency for the press to focus on immigrants not speaking in public services contexts, in particular in educational and healthcare settings. Drawing on the taxonomy set out by Resigil and Wodak (2001), our analysis shows that these groups are represented with recourse to a series of argumentational strategies or ‘topoi’. These topoi help to frame people with limited English skills in problematic terms, for instance as economic burdens, as harming social cohesion and as threatening cultural and linguistic norms in the UK generally. It is argued that the topoi used to represent and problematise the presence of these groups are themselves problematic, as they present a series of paradoxes, overlook the difficulties associated with language learning and cultural assimilation, and generally contribute to a broader UK anti-immigrant media narrative which serves to legitimise exclusionary and discriminatory practices against people from minority linguistic and ethnic backgrounds.

References

Transitions between discourse segments in English native speaker contexts have been associated with various phenomena, e.g. prosody (Hirschberg & Nakatani 1996 name for example pauses), lexical cohesion (cf. Hearst 1997; Ferret et al. 1998), explicit references (cf. Moser & Moore 1995; Hirschberg & Litman 1993), and the use of discourse markers (Bolden 2006; Johnson 2002; Norrick 2001). Passonneau & Litman (1997), among others, have shown that these phenomena tend to occur in combination.

In the context of English as a Lingua Franca (ELF), House (2013) also describes the use of discourse markers as transition segment markers (cf. also Redeker 2006). Mauranen (2012: 191) observes similar effects of discourse markers. Meierkord (2000) notes long pauses between discourse segments. Brunner’s (2015) qualitative analysis of ELF conversation openings and first topic introductions reveals a tendency to signal transitions between the two with the help of several linguistic markers. Brunner and Diemer (2016) find similar phenomena between segments in the smaller scale ‘transition zones’ between segments within the conversation opening of ELF conversations. They document the following transition markers (TMs): discourse markers, hesitation markers, pauses, laughter, repetitions, and lengthening.

The current paper analyzes how transitions are achieved between different discourse segments and how they contribute to structuring ELF conversations. The study is based on BabyCASE (Brunner et al. 2017), a corpus of Skype conversations between speakers of ELF from Germany, Bulgaria, Spain, Italy, and Finland. The corpus is part of the CASE project, compiling a larger ELF corpus between nine European countries and the US. It consists of dyadic Skype conversations taking place in an informal context. It provides video and/or audio, as well as a transcription component including pragmatic aspects such as the use of nonverbal, paralinguistic, and plurilingual resources.

Based on two qualitative studies (Brunner 2015; Brunner & Diemer 2016), the current study takes a quantitative approach with the aim of developing a method for a larger scale corpus analysis of discourse segmentation in which segments are identified through locating clusters of the six types of TMs found by Brunner & Diemer (2016). In our study, we qualitatively identify and mark transitions between discourse segments. Then we quantify the TMs in the annotated corpus by type and identify clusters. We then carry out a significance analysis of TM frequency in transition zones.
against the rest of the interaction, restricting a first analysis to conversation openings. Our findings show that TMs occur significantly more often in transitions between discourse segments in ELF conversation openings and could thus be useful as a transition predictor. We then apply our findings to full conversations in order to test a (semi-)automated method for finding transitions between discourse segments in ELF, in general, which we then verify qualitatively.

In sum, the current study depicts the complex mechanisms used to structure video-mediated communication in ELF through a combination of specific TMs and demonstrates how TMs can serve as external indicators for transitions between discourse segments. The paper thus aims to contribute to the understanding of discourse management and segmentation in ELF conversations.

References


Brunner, M.-L., & Diemer, S. (2016). “Mhm, ... okay so u:h, maybe we should start with this topic” – Conversation starts in English as a Lingua Franca (ELF) Skype talks. IWODA’16 – Fourth International Workshop on Discourse Analysis, Santiago de Compostela.


The HeidelGram project has a twofold aim. Firstly, it makes an essential contribution to historical grammar studies by compiling, analysing, and giving open access to a representative 10-million-word corpus of historical English grammar books from the 16th to the 19th centuries. Secondly, it introduces state-of-the-art network analysis into diachronic corpus linguistics; thus, considerably extending the set of concepts and methods applied in historical linguistics. Our overall aim is to examine discourses in English grammar writing by exemplarily implementing and analysing three networks – a network of grammars and grammarians, a network of evaluative terms associated with verbal hygiene (Cameron 2012 [1995]), and a network of lexemes referring to grammatical phenomena.

While network analytical methods have been applied to historical textual material (e.g. Bergs 2005; Sairio 2009; Fitzmaurice 2010) and fictional texts (e.g. Agarwal et al. 2012; Moretti 2013), the combination of corpus-based diachronic linguistics and network analysis is rather uncharted territory. This new approach poses significant methodological challenges and requires us to come up with new forms of extracting, annotating, and analysing historical linguistic data.

A series of exploratory studies (Busse, et al. 2016a and 2016b; Busse and Gather 2016), based on a systematically compiled and representative corpus of 19th-century British grammar books (40 texts, approx. 2.6 mio. words), has already shown the potential of this approach towards conducting historical grammar studies. In the present paper we want to present initial findings regarding the network of evaluative terms and discuss some of the major methodological and technical challenges associated with this approach. These include expressions like “greatly erred” in Crombie’s 1802-grammar: “Priestley, in defending the other phraseology, appears to me to have greatly erred” (Crombie 1802: 302).

This second network will not only help us to critically reflect upon the concepts of prescriptivism and descriptivism, but also to uncover linguistic practices and patterns that may have led to these discursive turns. Based on an extended and optimized version of our pilot-corpus, containing the most-well known and widely distributed grammars of the 19th century (cf. Leitner 1986, 1991; Linn 2006; Michael 1987; Görllach 1998), we will begin to quantitatively investigate terms associated with verbal hygiene (Cameron 2012 [1995]), i.e. active practices of filtering, evaluating, and modifying normative language usage, and their relationships.

Furthermore, informed by this initial analysis, we will discuss three major challenges associated with historical corpus-based network analysis and potential strategies of mitigating them. We will discuss typical issues with optical character recognition (OCR) and state-of-the-art workflows and procedures and tools, both automatic and manual, to reduce misreadings. Also, we will look at problems and solutions associated with automatically generating meaningful graphs (i.e. networks)
out of unstructured and unannotated linguistic data. Finally, we will present an early approach of visualizing such graphs in a way that allows for visual diachronic analysis.

References


Does address reflect gender relations? The English address system over the twentieth century

Anouk Buyle
KU Leuven

British society has seen an increase in gender equality since 1890, according to different parameters including legal and political rights, life expectancy, the organization of daily life, and access to education and jobs (Thompson 1990; Mahood 1995; Holloway 2005; Davis 2014). Even though these parameters have been firmly established as indicators of gender equality, researchers still struggle to find accurate measures for comparative or historical investigations, as well as measures that more reliably reflect the attitudinal dimension of gender equality (Carmichael et al. 2014: 219). This paper wants to suggest a new, linguistic parameter, viz. that of address terms. Since address terms, as a core resource for realizing social deixis, code aspects of the speaker-hearer relationship, we hypothesize that shifting gender relations must be reflected in changes in the address system.

The connection between address use and gender has been considered before in different varieties of English and different social settings (e.g. Brown & Ford 1961; Kramer 1975; Murray 2002; Walker 2007; Rendle-Short 2009). Ervin-Trip, for instance, observes that in some universities men address each other by surname only, while they address women by the combination of Mrs. or Miss and a surname (1972: 224). However, a systematic diachronic case study of address use depending on the gender of speaker and addressee is lacking.

This paper investigates the address system in twenty-four British English plays, equally distributed over three periods: 1890–1915, 1950–1965 and 2000–2015. All plays offer a portrayal of everyday, contemporary life. The collection is gender-balanced by author for each period, and has been digitized. For each conversational turn, address terms are identified, as well as the relation between speaker and hearer. The speaker-hearer relation is described in terms of the traditional power and solidarity dimensions, but a gender variable and a contextual variable are included as well (Brown & Gilman 1960; Ervin-Trip 1972; Wales 1983; Culpeper 1996; Murray 2002; Nevala 2004), as in (1). Finally, address terms are categorized in terms of general categories, including titles, given names, surnames, endearments, etc.

(1) But then, nurses are but nurses. I must not forget that I am a wife, Harry. (Pinero 1899)

[address term = given name, speaker = Sidonia , addressee = Harry , power = downward, solidarity = close, context = harmonious, gender speaker = female, gender addressee = male]

Using these parameters, address use is described for four types of speaker-hearer dyads, i.e. turns coded as male-male, female-female, male-female and female-male. If the changes in the address system reflect increasing gender equality, we can expect that, across the four types of speaker-hearer dyads, the use of nominal address terms becomes more similar over time. Such a tendency can confirm that the address system offers an independent view on changing social relations.

References

Throwing light on the development of impersonal constructions in Early Modern English: A case study of lust (after)

Noelia Castro-Chao
University of Santiago de Compostela

This paper is concerned with the later stages of impersonal constructions in English, an example of which is given in (1):

(1) Me liketh nat to lye
    me-OBJpleases-3SG not to lie

    ‘I do not like to lie.’ [MED c1425(a1420) Lydg. TB (Aug A.4) 4.1815]

The construction in (1) was frequent in Old and Middle English. Morphosyntactically, it contains a finite verb inflected for the third person singular, but lacks a subject marked for the nominative case
controlling verbal agreement. The impersonal construction began to decrease in frequency between 1400 and 1500 (van der Gaaf 1904: 142; Allen 1995: 267–283, 441–442). Although traditional accounts generally link this decline to the deep morphosyntactic transformations that took place during Middle English (e.g., Jespersen 1961[1927], Allen 1995), recent investigations outline some additional hypotheses (e.g., Trousdale 2008; Möhlig-Falke 2012; Miura 2015). For instance, Trousdale (2008: 302) approaches impersonals from the perspective of Construction Grammar (Goldberg 1995) and suggests that “the loss of the impersonal construction is tied in with the increased productivity and schematicity of the transitive construction”.

Following up the approach taken in these recent studies, my paper aims at: (i) observing the frequency of occurrence of the various morphosyntactic patterns that came to replace the impersonal construction; (ii) tracing diachronic change in the course of Early Modern English; and (iii) assessing the extent to which the corpus evidence in this particular study serves to confirm or not Trousdale’s (2008) hypothesis about the motivation for the loss of impersonal constructions. To this purpose, I will provide a corpus-based diachronic analysis of the verb lust (after) (<OE lystan), an impersonal verb in origin (OED s.v. lust, v. sense ‡2.) which belongs to the semantic class of verbs of DESIRE. This study will take into consideration a range of morphosyntactic patterns documented with this verb, namely:

1. impersonal (in)transitive [PRO_{ACC}-V3_{SG}-∅/X_{CLAUSAL}]; e.g., any thynge that hym lustethe to do
2. personal transitive [S-Obj{NP/CLAUSAL}]; e.g., he shoulde luste those thynges that lawes allow
3. personal intransitive [S-∅/Obl{AFTER/FOR/UNTO}]; e.g., we should not luste after an other mans wyfe
4. personal transitive patterns involving a reflexive pronoun [S-i-V-{PRO}_{RFL}]; e.g., they appoynt and decree what they lust themselves

The focus, as already mentioned, will be on the Early Modern English period (1500–1700), when the impersonal construction is known to have lost productivity. The data have been retrieved from EEBOCorp 1.0 (525 million words) and the examples have been manually analysed.

Preliminary results show that patterns 1, 2 and 4 above decrease dramatically over time. By contrast, pattern 3 seems to gain ground significantly, in particular that with a prepositional complement. It can also be observed that, as lust (after) shifted from impersonal to personal constructions, it also joined the prepositional class of verbs, a change that seems to be affecting a good many verbs of desire in Present-day English. Following Hopper & Thompson (1980: 262–263), this development may be interpreted in terms of low or reduced transitivity, thus running counter to Trousdale’s hypothesis above mentioned.

References


Gender representation in 19th century children’s literature

Anna Čermáková and Michaela Mahlberg

University of Birmingham

The 19th century is often considered the real beginning of English children’s literature and the period is referred to as the ‘Golden Age’. This period produced so many – still widely read – children’s classics, such as Lewis’s Alice in Wonderland, Barrie’s Peter Pan, Kipling’s Jungle Book and many others, that it is intriguing to ask why so many authors of that time chose children’s literature as their medium of expression (Carpenter 1985). The concept of ‘childhood’ itself also radically changed in the course of the 19th century (Hopkins 1994), as did British society itself, with the rising middle class, urbanization and industrialization. This paper aims to look at Victorian society and particularly at gender roles of the time. The focus will be on Victorian society as it was portrayed for its (middle and upper-class) child readers – a picture of a society as the authors perhaps wished it to be, rather than as it really was. The whole notion of children’s books changed dramatically at that time. There was a clear shift from the early morally didactic literature to literature that had entertainment as its main purpose with new emerging genres, such as fantasy.

Based on a corpus of 19th century fiction texts aimed at children, we will sketch a portrait of the society of the time. We will look at the most frequent nouns referring to people and their occupations and positions in the world of the 19th century. We will study the patterns in which these nouns occur (e.g. lexico-grammatical patterns, distributions across books, positions in a text, and in relation to author gender). We will thus combine frequency information with more qualitative analyses of textual patterns specific to children’s fiction. Our research draws on data from ChiLit – the 19th Century Children’s Literature Corpus (4.4 million words, 71 books) (Čermáková 2017).
For comparison we will also refer to other 19th century literary texts (Dickens’s novels and a 19th century reference corpus, all corpora at http://clic.bham.ac.uk).

Our results clearly show the asymmetrical nature of the portrayal of female and male characters in children’s fiction. While women and men are more or less equally represented in the private sphere of the family, the number of occupations they can take up in society is, unlike men’s, very limited and tied to the home sphere. Apart from queens and princesses, women are maids, governesses, nannies, and of course brides. As Walvin (1982) claims, children’s books are “an important source for contemporary social history” in that “they often reflect contemporary concerns and social problems” (Walvin 1982: 127). We will also aim to see to what extent major social problems of the time – particularly those related to gender – are covered, or are absent, in these texts.

References


Epistemic adverbs in the Old Bailey Corpus

Claudia Claridge
University of Augsburg

The gradual and late expansion of epistemic meanings and uses in English has been described by Hanson (1987), Traugott (1989), and Wierzbicka (2006). With regard to speaker-based epistemic adverb uses such as probably important developments seem to have taken place especially in Late Modern English, with the result that today these adverbs exhibit a remarkable frequency and wide sociolinguistic distribution, setting English apart from other European languages (Wierzbicka 2006: 262). Wierzbicka (2006: chap. 8) outlined the emergence of new epistemic types with the help of Literature Online sources, but a full-scale corpus investigation of types, uses, and frequencies is so far missing.

The present study therefore studies the development of these adverbs in the Old Bailey Corpus of courtroom speech, spanning most of the relevant period of their assumed expansion (1720–1913) and providing a spoken context in which degrees of knowledge play an important role. Probably, presumably, likely, evidently, apparently, undoubtedly/indubitably, conceivably, supposedly, allegedly and reported, which denote speaker (un)certainty and source of knowledge, thus
qualifying speaker commitment, are in focus here. Only the first seven types of the above list are attested in the OBC, but the group as a whole gains remarkably in frequency, starting from 26 instances per million words and ending at 173 instances per million. Noticeable jumps in frequency are found in the 1780s and the 1830s. Such a rise points to increasing usefulness of the items, perhaps connected the emergence or establishment of more clearly epistemic uses.

Aspects to be investigated are:

- The distribution of semantic interpretations, from more objective, ‘fact-based’ to more subjective epistemic uses, indicated by collocations and often the presence of circumstantial detail. The former meanings are much more commonly found in the OBC with evidently and apparently, cf. [the purse] was completely behind every thing; it evidently appeared to me, that it must have been put there (18320705) and a little girl standing by the side of her at the time, apparently about four years of age (18800803). Probably as in I do not remember why; probably he wanted them cashed (19110717) yields more and more clearly subjective epistemic uses.

- The distribution across speaker roles: while they occur most frequently in the speech of witnesses and defendants, they are also not infrequent with judges and lawyers, who may use them in a conducive way. Special attention will thus be paid to the latter’s usage, as in Do you think you heard evidently the voice of a child? (17931204, answered involving certainly) and If it had been unbolted before half-past one o’clock, would you not probably have heard it? (18320705, answered by I suppose so).

References


All (of) this is rather interesting

Florian Dolberg and Klaus Heimeroth
TU Dortmund

Speakers of contemporary English choose between all this and all of this, two apparently functionally equivalent NPs. These display a different head-modifier order: according to Huddleston & Pullum (2002: 374–378), all this consists of a demonstrative pronoun or a demonstrative determiner in a fused-head construction as NP-head, predetermined by the universal
quantifier, while all of this is a partitive construction comprising a universal pronoun serving as NP-head, postmodified by a prepositional phrase containing a pronominal NP.

Although no research appears to have focussed on all (of) this yet, such an investigation might yield interesting results, particularly in comparison to other, well-known and – studied cases of variation between pre- and post-modified constructions, such as Saxon genitive vs of-possessive (cf. e.g. Rosenbach 2002, 2003; Stefanowitsch 2003) or morphological vs periphrastic comparatives (cf. e.g. Hilpert 2008; Mondorf 2009). This research identified a number of semantic/cognitive and formal factors at work, which we apply to all (of) this to see how pervasive these factors are across constructions.

One semantic/cognitive factor influencing e.g. comparative alternation is the Complexity Principle, (e.g. Rohdenburg 1996: 151), yielding the hypothesis that the of-version predominantly occurs with more complex (e.g. larger, more abstract) targets of anadeictic (cf. Cornish 2011) reference. Conversely, as a “discourse deictic demonstrative” (Diessel 1999: 100–105), all (of) this refers in “‘strict’ anadeixis” (Cornish 2011: 757–758) to propositional units within the preceding discourse, presumably conceptualised as mass or non-count entities. Mass-concepts vary in their degree of e.g. unboundedness, expansibility, or homogeneity (cf. e.g. Langacker 1991: 73–81). We hence hypothesise that the partitive of-construction is preferred where the target concepts are particularly homogeneous, unbounded, or expansible.

Another well-established, formal factor influencing variation is the principle of rhythmic alternation (cf. e.g. Schlüter 2003): the choice between functionally equivalent variants differing in syllable count is influenced by which of the two better fits the preferred stressed-unstressed rhythm. Thus, we expect stress-patterns such as Áll of this demánds a clóser lóok and All this shóws some influènce of rhýthm rather than vice versa.

Quantitative and qualitative analyses ascertain which of the factors previously identified apply in this case and in which relative strength. Further, comparisons across varieties (in e.g. NOW), between registers (in e.g. BNC and COCA), and over time (in e.g. EEBO and COHA) show how these dimensions differ in the factors underlying the choice between all this and all of this. All (of) this finally relates to (similar) cases of variation and change, thereby elucidating what drives variation and change in general, mapping out commonalities in the interplay of functional and formal; language-internal and -external; semantic, cognitive, psychological, socio-cultural, and stylistic factors.

Preliminary results indicate that both British and American English (i.e. BYU-BNC and COCA) prefer the of-less version. This preference is much stronger in BE, with a ratio of 6.3:1, than in AE with a ratio of 1.4:1. This synchronic frequency difference apparently extends to diachrony: in EEBO, the partitive of-construction occurs several decades later and merely yields two dozen instances throughout 1470–1690, compared to several thousand hits for all this.

References

Exploring online register variation using machine learning methods

Jesse Egbert¹, Veronika Laippala² and Douglas Biber¹
Northern Arizona University¹, University of Turku²

The internet has dramatically changed the way writing is published and disseminated. This has led to the emergence of many new registers, such as blogs, question/answer forums, and social media posts. The internet has also introduced important changes to traditional print registers, such as encyclopedia articles and reviews (Biber & Egbert forthcoming). Recent research has revealed important patterns of linguistic variation across online registers using methods such as Multi-Dimensional analysis (Biber & Egbert 2016; Titak & Roberson 2013) and keyword and key feature analysis (Biber & Egbert forthcoming). The objective of this paper is to add to this growing body of research by applying a new methodology to explore linguistic variation across online registers.

In this paper, we analyze the linguistic characteristics of online registers using the Corpus of Online Registers of English (CORE). CORE is composed of an unrestricted sample of the searchable web,
containing with 48,571 web documents and over 50 million words of running text. CORE is currently the largest corpus of online writing to be manually classified for register information (see Egbert, Biber and Davies, 2015). This manual classification resulted in 8 register and 33 sub-register categories (see Biber, Egbert and Davies 2015).

To study the characteristics of the registers in CORE, we use a novel method: supervised machine learning, more precisely support vector machines (SVMs). SVMs allow for a diverse linguistic analysis of register variation in large data sets. While their primary goal is to identify the category for each document in a data set, the SVM model also identifies the important features of each category that contribute to its identification. Thus, when applied to a register-classified corpus, the features that contribute to the identification of registers in a corpus can be analyzed as key features that reflect the linguistic characteristics of those registers (Baker 2004; Scott and Tribble 2006). In contrast with traditional keyword analysis, however, SVM has three major advantages: (1) it can be applied not only to words, but to any linguistic feature; (2) it analyzes these features at the level of the text, rather than at the level of the corpus; (3) rather than relying only on two corpora, a reference corpus and a target corpus, it simultaneously compares all registers in a corpus and identifies features that are key for each register when compared with all other registers.

This study applies the SVM to the CORE to explore two sets of key features of online registers: lemmas and lexico-grammatical features. The results for both feature sets reveal functionally interpretable patterns of register variation at the lexical and lexico-grammatical levels that can be interpreted functionally. Comparisons with previous studies (e.g. Biber & Egbert forthcoming) show that SVM key feature analysis provides new insights into patterns of linguistic variation across online registers of English that complement previous linguistic research on these registers. Our presentation will provide detailed examples of these patterns and also discuss further applications of this novel methodology to linguistic research on register and other language varieties.

References


**Zooming in on Trumpish: A corpus-based analysis of Donald Trump’s idiolectal language use in the social media**

*Matthias Eitelmann, Britta Mondorf and Ulrike Schneider*

University of Mainz

A public media personality long before his political career, Trump is the first ‘media president’. This raises the question whether the “grandiosity, informality and dynamism” (Ahmadian, Azarshahi and Paulhus 2017: 49) characterising his presidential tweets are part of a register acquired as host and guest of various competitive TV formats. Furthermore, Twitter as a social media channel has unique restrictions but also a wide reach. The latter causing Trump’s iconic phrases to reverberate in the language of pop culture. *Fake news*, for instance, was elected ‘Anglicism of the Year 2016’ in Germany (cf. Stefanowitsch 2017).

Due to the cultural and political significance of his statements, there has been a growing interest in fine-grained analyses of his language use, which has been taken to be crucially different from his predecessors’ and that of other politicians. So far, along the lines of ‘if you don’t like the content, ridicule the vessel’, first publications claimed that Trump’s vocabulary and grammar are reminiscent of a prepubescent child’s linguistic competence (e.g. Spice 2016). However, such an assessment seems to critically underestimate his rhetoric, deeply rooted in a network of conceptual metaphors (Lakoff 2016) and associated with the language of populism (Lakoff 2009; Lakoff and Wehling 2016; Mudde and Rovira Kaltwasser 2017). What is more, corpus-based pilot analyses on differences between Trump’s idiolect before and after the election (reported in Fiedler 2017) indicated that Trump’s way of communication, characterised by repetition, parallelisms and short paratactical sentences, is highly strategic and thus successfully tailored to a political purpose.

With representative social media and political corpora as of yet lacking, descriptive analyses of Trump’s language use are still sparse – which is all the more unfortunate if we assume that ‘Trumpish’ might not simply be a salient idiolect, but rather a reflection of changing social norms. Against this backdrop we provide an in-depth exploration of Trumpish, based on a 350,000-word corpus of his tweets, as well as on an equally large comparative corpus of tweets by other contemporary US politicians, addressing three main issues. First, the characteristic features of Trump’s idiolect are identified in a quantitative and qualitative analysis, with a special focus on the question whether Trump’s language behaviour reflects a particularly high use of linguistic features commonly associated with lower cognitive complexity. Second, we consider media-related issues by investigating to what extent his language use is dependent on the (social) medium he uses and how reminiscent Trump’s current idiolect is of his language use as evidenced in former media appearances. Third, we shed light on issues of political concern by probing into the question of how
populist his idiolect actually is and to what extent his language use may have changed the voice of populism.

References


Constructional interaction and language change: Voice-marking on the into-causative

Susanne Flach
Université de Neuchâtel

One aspect that receives relatively little attention is the interaction between ‘unrelated’ constructions as we tend to focus on onomasiological relationships. For instance, relationships in argument structure constructions are well-studied individually, as is the active–passive alternation. Yet, constructional interaction is trivially inevitable: speakers combine ASCs with passives, if contextually required and/or semantically appropriate. Thus, we may treat an interaction as a coded variable, but often independent of more fine-grained properties or developments in the interacting construction. This paper discusses constructional interaction of voice-marking and the into-causative (to pressure sb into confessing). It asks to what extent interaction between constructions (into-causative and passive-active) sheds light on changes within a construction (i.e., into-causative).
The *into*-causative is a subtype of the caused-motion construction (Goldberg 1995; Rudanko 2011). A causer influences a causee to perform the action encoded in the *-ing*-gerund. The examples in (1) illustrate ‘canonical’ active uses with three participants mapped onto three argument roles. For non-canonical uses in (2), the causer remains unexpressed in passives (cf. [2a]), while causer and causee are co-referential in reflexives, which can be seen as ‘bridge contexts’ between actives and passives (cf. [2b]):

(1) a. They tricked me into buying the machine for $186 on credit. [COHA, 1971]
   b. She can talk people into doing almost anything she wants them to. [COHA, 1961]

(2) a. Booksellers were terrorized into removing it [Rushdie’s book]. [COCA, 1990]
   b. But we shouldn’t fool ourselves into thinking it can’t be done. [COCA, 2011]

Previous work on slot fillers shows that the matrix verbs come from distinct semantic classes, which systematically co-vary with causation events and constructional subsenses (Gries & Stefanowitsch 2004; Kim & Davies 2016; Rudanko 2006, 2011). Synchronic data show similar systematic interaction with voice: passives are associated with force verbs, actives with persuasion verbs, and reflexives with deception verbs.

This type of constructional interaction maps straightforwardly onto diachronic questions about the role of constructional interaction in language change. Using simple association measures, COHA data show that the distribution of voice marking is non-random across time, but instead reflects gradient patterns of constructional change: First, passives decrease from 50% in the mid-19th century to 20% in the 2000s. Second, rare ‘bridge’ reflexives have their ‘heydays’ for a brief period in the early 20th century. These observations provide evidence that the *into*-causative shifts from primarily ‘force’-related towards ‘persuasion’-related meanings. This supports the view that interaction of two or more constructions is non-trivial for (the study of) language change. Third, however, measuring this change is also influenced by construction-external factors, such as the general decrease in passives over time. Thus, we will address methodological aspects that account for these confounds.

References


When syntax connects explicit and invisible language: A corpus-based study of syntactic linking between antecedents and ellipsis sites in Modern English

Evelyn Gandón-Chapela
University of Cantabria

This paper investigates the type of syntactic linking established between the antecedent clause(s) and the ellipsis site(s) in cases of Post-Auxiliary Ellipsis (PAE) in Modern English. PAE (Sag 1976; Warner 1993; Miller 2011; Miller & Pullum 2014) covers those cases in which a Verb Phrase, Prepositional Phrase, Noun Phrase, Adjective Phrase or Adverbial Phrase is omitted after modal auxiliaries, auxiliaries be, have and do, and infinitival marker to. The two subtypes of PAE under investigation are VP ellipsis (VPE) and Pseudogapping (PG), illustrated, respectively, in (1)–(3) and (4)–(6):

(1) That I had received such from Edward also I need not mention; but I do, you see. [VPE: coordination]
(2) They can by no means, therefore, be members of happiness; for if they were, happiness might be said to be made up of one member. [VPE: adverbial subordination]
(3) I can recollect nothing more to say. When my letter is gone, I suppose I shall. [VPE: none]
(4) A skilled florist will produce a finer effect with a few inexpensive blossoms than an unskilled one will with a cartload of choice material. [PG: comparative subordination].
(5) but did not admire the strain of its poetry in general, though I did its morality. [PG: adverbial subordination]
(6) I have been always in the habit of treating my own Horses much after the same manner that I would myself. [PG: relative subordination]

In an attempt to extend the few previous corpus-based works on this topic for the Present-Day English period (Hardt & Rambow 2001; Nielsen 2005; Hoeksema 2006; Bos & Spenader 2011; Sharifzadeh 2012; Miller 2014), I analyse the type of syntactic linking in PAE constructions in Modern English (data retrieved from the Penn Parsed Corpus of Modern British English, 1700–1914) and present a retrieval algorithm of instances of PAE via CorpusSearch 2. This complex algorithm has led to successful recall ratios (0.97) and is applicable to parsed corpora which follow the conventions of the Penn Parsed Corpus of Modern British English. The results show that, regarding PG, the vast majority of cases are comparative constructions (74%; illustrated in (4)), followed by those cases with lack of syntactic linking (15.12%), coordination (4.65%), adverbial subordination (4.65%; as in (5)) and relative subordination (1.16%; as in (6)). The comparison with other studies on PG in Present-Day English (Hoeksema 2006; Sharifzadeh 2012; Miller 2014) has revealed that instances of PG with NP remnants have a stronger preference for comparative constructions in Present-Day English (around 90%) than in Modern English (70%). Regarding VPE, in over 50% of the examples there is no syntactic linking between the source and the target of ellipsis (as in (3)). The second most important type of syntactic linking is comparative subordination (31.51%). However, although the percentage of comparative constructions is high in VPE, it is almost 2.5 times higher in PG (74.42%). Far less common are cases of relative subordination (7.22%), coordination (5.56%; as in (1)) and adverbial subordination (5.37%; as in
(2)). If these findings are compared with Bos & Spenader’s (2011), it is observed that the first three types of linking are the same in both studies: as-appositives, comparatives and lack of syntactic linking. Hardt & Rambow (2001), on their part, found that the different forms of subordination favour VPE, while the absence of a direct relation disfavours its presence. However, this type of linking is the third most frequent one in Bos & Spenader (2011) and in this study.

References


Toward(s) an Americanization of non-native Englishes? ESL and EFL compared

Gaëtanelle Gilquin
Université catholique de Louvain

In their comparison of British and American English, Leech et al. (2009: 253–254) noted that “the evidence is cumulatively persuasive in indicating American ‘leadership’ being one of the major moving forces on BrE [British English]”. Since then, the linguistic influence of American English has been claimed to extend to other varieties of English, leading to a more widespread phenomenon
of Americanization. Thus, Mair (2013) takes American English to be the hub of his “World System of Englishes”, that is, the variety that is relevant to all other varieties of English and is “a potential factor in the[ir] development” (ibid. 261). However, his claim mainly relies on “anecdotal evidence” (ibid. 263) and his model, while including English as an institutionalized second language (ESL), does not take English as a foreign language (EFL) into account.

In this presentation, I will investigate the possible influence of American English on ESL and EFL varieties, and compare it with the influence of British English. My hypothesis is that, ESL varieties being acquired and used in an essentially naturalistic environment, they will be subject to the forces of globalisation which according to Mair (2013) are associated with the dominance of American English. On the other hand, EFL varieties, whose typical context of acquisition and use is that of the classroom, are expected to be more subject to the forces of education, which tend to be more conservative and more oriented towards British English models (cf. Schneider 2007: 172; Trudgill & Hannah 2017: 5).

The study uses data from two large corpora, namely the Global Web-Based English Corpus (GloWbE) for ESL and the EF-Cambridge Open Language Database (EFCAMDAT) for EFL, representing a total of 645 million words and 32 million words, respectively. On the basis of a list of twenty pairs of words or phrases taken from Algeo (2006) and shown to be distinctive of American English vs British English (e.g. movie/film, toward/towards, take a shower/have a shower), it assesses the influence of American English on ESL and EFL varieties. It appears that, contrary to expectations, both ESL and EFL are on average more strongly influenced by American English than by British English. EFL even turns out to display a higher Americanness rate than ESL (58% in ESL, 63% in EFL), which contradicts the initial hypothesis. However, while the American influence on ESL is relatively stable across the different pairs of items, EFL is characterised by a large degree of variation, with an Americanness rate ranging between 6.19% (for have gotten/have got) and 98.44% (for give it a try/give it a go). In an attempt to explain these results, different factors are considered, including the changes in society that have blurred the distinction between ESL and EFL, the functional domains that are more likely to be affected by Americanization (like those of entertainment or technology), and the linguistic features that may favour the use of certain items (e.g. semantic transparency, morphological simplicity, or similarity to the speakers’ L1).

References


Variable cues vs. multiple grammars: Genre specificity in the English genitive alternation

Jason Grafmiller
University of Birmingham

I examine variability in the genitive alternation (*the president’s decision* vs. *the decision of the president*) across five written genres of 20th century American English, focusing on quantitative differences in the constraints on writers’ deployment of this variable in different stylistic contexts. As in speech, variation across contexts in writing is an example of complex ‘style switching’ (Rickford 2014), thus my findings speak to theoretical issues regarding the relation between grammatical representation and the stylistic sensitivity of probabilistic grammatical constraints.

I draw on data from the Brown and Frown corpora and examine a subset of available genres/registers—Press reportage, Non-fiction (memoirs), Academic, General fiction, and Adventure fiction—in the 1960s (Brown) and 1990s (Frown). I extracted 5098 genitive tokens from the two corpora (Brown N = 2497; Frown N = 2601), carefully circumscribing the variable context and annotating for numerous factors known to condition the choice of genitive variant (see e.g. Grafmiller 2014). These include the length (in words) of both the possessor and possessum; the semantic relation between possessor and possessum; the presence of a final sibilant on the possessor; and the animacy, frequency, givenness, and NP type (common vs. proper) of the possessor. I analyze the data using mixed-effects logistic regression to assess the influence of these factors on genitive choice, taking into account author-specific variability in the proportion of usage of the two variants.

I find sizable differences among registers in the quantitative influence of possessor animacy, which has a significantly weaker effect in journalistic writing than in other registers. In addition, I find that the relative importance of this constraint (compared to others) is lower in Press than in other genres, and that in general, the relative constraint rankings vary markedly across genres. I also observe that within genres, there is little intra-author variability in the constraint effects, however, inter-author rates of genitive use vary considerably across genres; Fiction writers vary the most, while journalists vary the least.

I interpret these patterns in press writing as reflections of journalists’ semi-conscious move toward more economical and colloquial modes of expression (Biber 2003), a move that can have social significance. But does the variability I observe imply that we are dealing with distinct, genre/register-specific grammars, a la Guy (2015)? I argue that within an experience-based, probabilistic approach to grammar, the multiple-grammar model is ill-defined, and advocate an alternative model in which situational/stylistic cues can directly shape the influence of internal constraints within a single ‘grammar’.

References


Discipline-specific corpora and the development of disciplinary literacy

Clarence Green and James Lambert
Nanyang Technological University

Disciplinary Literacy is a research area and pedagogical approach in secondary education that focusses on developing awareness in learners of language variation across disciplines. The approach is being adopted by K-12 educational systems worldwide (Shanahan & Shanahan 2017). It shares much in common with recent trends in English for Specific Purposes (Hyland 2017); e.g. both emphasize the language used for conceptual access varies by discipline and question the utility of general literacy/general academic vocabulary. Recent research has moved from general academic wordlists, such as Coxhead’s AWL (2000) or Gardener and Davies’ AVL (2012), towards discipline-specific wordlists. However, most discipline-specific resources have been developed for either ESL learners or the university context, as demonstrated by the corpora that underlie them: e.g. journal articles, course assigned readings etc. (Lei & Lui 2016). We argue that existing wordlists do not represent the lexical demands of secondary students. This presentation reports a large-scale project aimed at developing academic vocabulary lists for secondary school context, thereby expanding the current methods of corpus-based ESP into new contexts. The work is unprecedented in scale for discipline-specific corpus-derived material, and will be publically released. We built a 16.5 million word corpus of current secondary school textbooks with approximately 1.5–2 million words per discipline, and we cover with a consistent methodology eight disciplines: biology, chemistry, economics, English, geography, history, mathematics and physics. Following Lei and Lui (2016) and Gardner and Davies (2013), discipline-specific lexemes were extracted according to state-of-the-art metrics and statistical measures: minimum frequency of 28.57 p/m (Coxhead 2000); range of more than 50% of texts; Oakes’ Dispersion of 0.5; range ratio of 20% of an item’s expected frequency in 50% of texts; and new metrics such as a discipline-specific frequency Ratio and Keyness (i.e. a lemma needed to be a keyword in the discipline at a ratio 3 times higher than the rest of the corpus). We also POS-tagged and required the lemma to be a major part of speech (noun, verb, adjective, adverb). Going beyond previous studies, we simultaneously developed collocational and word family lists for each target lemma. The final lists to be presented distill from the corpus the core lexical items that need to be mastered by students to read successfully in a content area (reaching lexical coverage within disciplines of > 20%). The lists are aimed to assist in the development of disciplinary literacy, an approach to education that is changing many education systems worldwide.

References

Analysing change in a corpus of patents using move analysis and edit distance

Jack Grieve and Nicholas Groom
University of Birmingham

Move analysis is a widely used methodology for describing the rhetorical structure of texts (e.g. Biber et al. 2007; Upton & Cohen 2009; Tardy & Swales 2014). In this system, a text is seen as a sequence of ‘moves’, which are parts of the text that have distinct communicative functions. For example, Swales (1990) has famously proposed that academic research article introductions can be seen as a sequence of three moves, which function together to ‘create a research space’ that the article will occupy. Furthermore, by conducting a move analysis for a sample of texts drawn from a given genre, it is possible to make generalisations about what types of rhetorical structures are typical of that genre, including which moves are obligatory, which are optional, and how they tend to be ordered.

Although most move analyses have been synchronic, there is growing interest in how the rhetorical structure of genres changes over time (e.g. Berkenkotter 2009; Gillaerts 2014). Such an approach to the study of language change, however, brings up interesting issues about how to analyse evolution in rhetorical structure as identified through move analysis, including how to measure the rate at which the structure of move sequences changes over time. The goal of this paper is therefore to introduce a novel method for quantifying the degree of change over time in move sequences based on string edit distance (Navarro 2001), a commonly used technique in natural language processing and genetics. In particular, our analysis is based on a corpus of 277 British patent texts dating from 1734–2011 (one randomly selected patent from each year aside for 1739, for which no data was available) and consists of three steps.

First, we coded each of the 277 patents for 21 different rhetorical moves, which we identified through a manual analysis of our corpus, following standard procedures, including inter-rater reliability testing. These moves fulfil a wide range of functions and include ‘filing information’, ‘declaration of invention’, and ‘statement of claims’. In this way, each patent was reduced to a string of moves, with each move being representing by a unique character in our coding system.

Second, using string edit distance, we measured the dissimilarity between each temporally adjacent pair of patent move sequences (e.g. 1734 vs. 1735, 1735 vs. 1736). String edit distance is a
relatively simple way of measuring how dissimilar two strings are from by counting how many changes are necessary to convert one string into another. For example, the edit distance between the string “ABC” and “ABD” is 1, because “ABC” can be converted to “ABD” by replacing “C” by “D”. By applying this technique to strings representing move sequences, we are able to quantify change in rhetorical structure in our corpus.

Finally, we graph change over time by plotting these string-edit distances, finding that the rhetorical structure of British patents changes relative consistently and gradually over time, aside from occasional rapid periods of transformation caused by legal and technological changes.

References


Watching the future unfold: A longitudinal analysis of future temporal reference

Tomoharu Hirota and Alexandra D’Arcy
University of British Columbia, University of Victoria

The future temporal markers will and be going to emerged at different times through distinct developmental pathways. Will, originally a verb of desire, developed in Old English and spread during Middle English (Warner 1993). Be going to is a fifteenth-century development, originating in the progressive construction accompanied by a purpose to-infinitive clause (Danchev & Kytö 1994). In contemporary spoken English, they are the primary variants of future temporal reference, as in (1).

(1) a. She just said, “I am not going to faint. I will not faint.” (CatherineMaclure_1)

   b. I was like, “Oh next year we’re gonna be grade seven. We’ll be old.” (GD33m)
Indeed, studies of North American varieties (e.g. Szmrecsanyi 2003; Torres Cacoullos & Walker 2009; Denis & Tagliamonte 2017) demonstrate that they are used with near equal frequency. This rigorous competition appears to have entrenched in the early twentieth century—apparent time perspectives suggest stable functional partitioning (cf. Tagliamonte & D’Arcy 2009; Denis & Tagliamonte 2017). At the same time, corpus-based investigations suggest that stability is relatively recent: be going to underwent a major leap in text frequency in the late nineteenth century (e.g. Krug 2000; Mair 2004; Seggewiß 2012). These observations suggest that the last 150 years have marked a critical period in the development of the current state of future temporal reference; what is lacking is evidence from vernacular practice that extends the window beyond the living speech community. Is change limited to frequency, or are systemic shifts relevant also?

To answer these questions, we investigate future temporal reference in Victoria, Canada; our data come from the Victoria English Archive, a large corpus of speech data that covers the mid-nineteenth century to the present (years of birth 1865–1998; D’Arcy 2017). This longitudinal view enables us to assess variation and change in this grammatical sector over 133 years in real and apparent time to uncover its diachronic trajectory in vernacular language.

Preliminary analysis of these materials reveals the upswing and saturation phases of the S-curve (N = 500). In the second half of the nineteenth century, be going to accounts for less than 25% of the future sector. It then rises rapidly during the early twentieth century, and among those born after 1950 it stabilizes at roughly 50%. Such results suggest that after a protracted period of slow and incipient change (roughly 400 years), be going to experienced a short but vigorous burst of frequency expansion. However, as be going to increases in all contexts, the well-known constraints on future temporal reference exhibit continuous parallelism (e.g. clause type, verb type). Only the animacy/person predictor reorganizes: as predicted by Denis & Tagliamonte (2017), the semantic distinction between 1st person and inanimate subjects begins to level over time, eroding a former foothold of will within the system.

We interpret these findings as support for the independence of frequency and probabilistic conditions, while at the same time echoing Denis & Tagliamonte’s (2017: 23) assertion that semantic-pragmatic factors may disrupt the course of change.

Selected References


Exploring the potential and the limitations of the audio edition of the 1994 Spoken BNC

Sebastian Hoffmann¹, Sabine Arndt-Lappe¹ and Valérie Keppenne²
Trier University¹, Penn State University²

Our paper will present a resource that is still fairly unknown in the corpus-linguistic community but that – we believe – has considerable potential for the quantitative study of spoken language phenomena: the audio edition of the 1994 Spoken British National Corpus (BNC). The audio recordings of a sizeable proportion of the spoken component of the BNC (approx. 5.5 million words) were made freely available through a digitisation project carried out by the British Library Sound Archive in collaboration with Oxford University Phonetics Laboratory (see Coleman et al. 2012). The sound files were aligned with the original transcriptions in the form of Praat TextGrid files, making it possible to integrate the resource in BNCweb, a web-based interface to the BNC hosted by Lancaster University (Hoffmann et al. 2008), giving its users direct access to the audio version of individual query results.

While the links to the audio data have been available in BNCweb for some time, a further output of the digitisation project has only recently attracted our attention: the data are also available in a phonemically transcribed format, with time alignment information that allows direct access to realisations of individual phonemes in millions of words of natural speech. This is a phenomenal resource that greatly extends the traditional repositories available to corpus-linguistic research of speech phenomena in British English.

In our paper, we will focus on the potential – but also the limitations – of the phonemic transcriptions of the spoken BNC for corpus-linguistic analyses. We will show how the information has been integrated into BNCweb, making it possible to perform searches that combine
lexical/morpho-syntactic features with phonological phenomena. We will also present the results of a sample study conducted on the basis of the data.

Our sample study is concerned with so-called ‘intrusive r’. The term refers to non-etymological [r] appearing in intervocalic contexts in non-rhotic varieties of English (1). A related phenomenon is ‘linking r’, which occurs in the same phonological contexts, but is etymological. Neither of the two processes are categorical.

(1) intrusive r
   a. Now [pause] the idea [r] is that er they put a couple of monitors and monitor your brainwaves. (BNC FLY 608)

Existing research suggests that appearance of intrusive r is conditioned by a combination of structural-phonological, phonetic, sociolinguistic, and lexical factors (e.g. Pavlik 2016; Keppenne 2015; Soskuthy 2013; Hay & MacLagan 2010).

In our sample study we use the phonemic transcriptions to extract all hiatus contexts from the spoken section of the BNC, and coded the presence or absence of intrusive r on the basis of the corresponding audio files. We then show how the variation is correlated with the different factors that have been proposed in the literature. A preliminary analysis of a subset of the data (N = 272; Keppenne 2015) already provides evidence for both sociolinguistic and lexical factors. Specifically, women use more intrusive r than men in the BNC data. Also, the frequency of similar linking r contexts involving etymological r seems to play a role.

References


This paper discusses the rise of so-called -ingly adverbs in late Middle and early Modern English by investigating some selected texts in the *Early English Books Online* (and the *Early English Books Online Corpus*, which has recently been released to the public by Mark Davies). A special focus will be placed upon some major authors in the Middle and early Modern English periods: John Trevisa, Geoffrey Chaucer, Thomas Malory and William Caxon in the Middle English period; and Thomas More, Francis Bacon, Richard Baxter and John Locke in the early Modern English period. The main objective of this study is to contextualize their linguistic behaviours and idiosyncratic features within the framework of the overall development of -ingly adverbs, and ultimately the development of -ly adverbs, in the history of English. The methodology employed in this study is a variationist one.

Previous studies show that a notable growth of -ingly adverbs takes place from the later Middle English period onwards, and this is largely confirmed by the present study. Even in the early Modern English period, however, the inventory of -ingly adverbs is still unstable. Items such as accordingly, willingly, and everlastingly are fairly well established in early Modern English, whereas other items are rather ad hoc, occurring only once or twice or in particular authors’s texts only. The path of the lexicalization of -ingly adverbs, therefore, seems to be motivated by an individual’s creativity in forming words, which may or may not become fully established in the history of English.

This paper also discusses the functions of the -ingly adverbs observed in the data. Most examples modify verbs or verb phrases, although here again there are some notable idiosyncrasies. Malory, for instance, is quite peculiar in this respect, in that his use of -ingly adverbs is virtually confined to the modification of adjectives. Otherwise, examples modifying adjectives locally are not common from late Middle English through early Modern English. Likewise, modification of the whole sentence by -ingly adverbs is not yet common, suggesting that this use is a later development in the history of English. All in all, -ingly adverbs in the early Modern English period are still at a seminal stage as their usage does not show all-around functions as found in the present-day usage.

Finally, this paper discusses what Broccias calls Harry Potter adverbs as illustrated by: “‘I’m not sure this is going to work, you know’, said Hermione warningly” (*Harry Potter*, cited in Broccias 2012a; see also Broccias 2012b). Broccias notes that the main verb say in this example conveys an objective fact while the manner adverb warningly gives the “subjective evaluation” of the event. He also states that -ingly adverbs of this particular type is attested from the late Modern English period onwards. The present paper shows that there are some preliminary examples of this type in the data investigated.

References

Since Crystal’s (2001) early subsumption of the language in various written modes of digital communication under one variety labeled “Netspeak,” linguists have challenged the assertion that all such communication is collectively different from speech and writing. By contrast, they have increasingly recognized the heterogeneity of communication in various modes, tracing differences to be effects of varying synchronicity of interaction (e.g. Herring 2004, 2007; Tagg 2012; Jonsson 2015). Several investigators have thus assumed synchronous (i.e. real-time) digital communication to be more conversational (or, “oral”) than asynchronous communication (e.g. Condon & Čech 2010; Georgakopoulou 2011; Jonsson 2015), without necessarily defining “orality” in their approaches. In this paper, I attempt to narrow down conversationality (or, orality) in written, screen-mediated communication by relating online chats and mobile text messages, modes popularly held to be conversational, to spoken conversations, using multidimensional analysis.

Conversationality is here defined as lexico-grammatical similarity to spoken conversations, the degree of conversationality in a mode being the degree of similarity to spoken conversations. Data for the study is drawn from a corpus of online chats, UCOW (Jonsson 2015), and a corpus of SMS text messages, CorTxt (Tagg 2012). The discourse in the two corpora is compared to the discourse in 23 registers of speech and writing (Biber 1988), including face-to-face and telephone conversations, taking into account 67 lexico-grammatical features (e.g. pronouns, tense and aspect markers and measures of lexical specificity). By positioning the online chats and mobile texts on Biber’s five established dimensions and systematically exploring their relationship to spoken conversation, it is possible to assess their degree of conversationality.

Previous research has shown that online chat (synchronous and supersynchronous communication) displays a high degree of stylistic and lexico-grammatical similarity to oral conversations (cf. Georgakopoulou 2011; Herring 2013; Jonsson 2015), leading Jonsson (2015) to label online chat “conversational writing.” In the present paper, I investigate whether SMS text messages (prototypically asynchronous communication) yield a comparable high degree of conversationality. The study addresses the following research questions:

- How do asynchronous mobile text messages compare lexico-grammatically to spoken conversations and online chats?
- What are the distinguishing lexico-grammatical and situational features of the modes of communication investigated?
What conclusions can be drawn as regards the definition of conversation/ conversationality on the basis of the results?

In addition to the multidimensional analysis, the paper presents the result of a cluster analysis of the registers studied. Among the findings is a relationship more distant to oral conversations for asynchronous than for synchronous communication, even though the text messages generally range in the vicinity of conversations on the dimensions, approximating certain online chats. The cluster analysis, however, yields Jonsson’s (2015) term “conversational writing” not immediately applicable to the text messages, as they cluster more readily with, e.g., interviews than with face-to-face conversations. Taken together, the lexico-grammatical and situational features of the text messages suggest that the register is more of a cousin than a sibling of spoken conversations.

References

CorTx = A corpus of more than 11,000 SMS text messages in English. 2004–2007. Compiled by C. Tagg (see Tagg 2012).
Using the BNC and the Spoken BNC2014 to study the syntactic development of *I think* and *I’m sure*

Henrik Kaatari¹ and Tove Larsson²
University of Gävle¹, Université catholique de Louvain²

The grammaticalization of *I think* has received considerable attention in recent years (Thompson & Mulac 1991; Kaltenböck 2011; Van Bogaert 2011). However, far less attention has been paid to the related epistemic marker *I’m sure*, despite the fact that this construction has been shown to exhibit similar behavior (Kaatari forthcoming). The present study aims to investigate the degree to which *I’m sure* is on the same grammaticalization trajectory as *I think*, as explained below.

Following Traugott & Heine (1991), we view grammaticalization as both a diachronic and a synchronic phenomenon to be studied “at a synchronically segmented moment in time” (Traugott & Heine 1991: 1). In addition to a propensity for *that*-omission in clause-initial position, one of the main arguments put forth to support the claim that *I think* is grammaticalized is that it has developed an ability to occur in clause-medial (1) and clause-final position (2), that is outside its canonical clause-initial position (3) (Thompson & Mulac, 1991). Contrary to Hooper’s (1975) claim, a recent empirical study on *I’m sure* has indicated increased flexibility in this respect, as exemplified below (Kaatari, forthcoming).

(1) He is, *I think*/*I’m sure*, an interesting person.
(2) He is an interesting person, *I think*/*I’m sure*.
(3) *I think*/*I’m sure (that)* he is an interesting person.

Nonetheless, the question remains whether the development of the two constructions can be accounted for in the same way, despite the fact that these constructions have different frequency entrenchment and that the predicates belong to two different word classes.

The aim of the present study is to investigate whether *I’m sure* follows the same grammaticalization trajectory as *I think*. The research questions are as follows (see Lehmann 1985: 303, for a discussion of the methodological parameters of grammaticalization used):

- What is the frequency distribution across the clausal positions (syntagmatic variability)?
- To what extent is the complementizer *that* omitted (paradigmatic variability)?
- Are there any differences across time such that the development of *I’m sure* could be considered to mirror that of *I think*?

The study uses comparable subsets from the spoken component of the BNC (Burnard 2007; Lee 2001) and the newly compiled Spoken BNC2014 (Love et al. 2017). The results show that *I think* and *I’m sure* exhibit remarkable similarity, especially in the most recent data, not only in terms of their proportional distribution across clausal position, but also in terms of their propensity for *that*-omission. Even though the time span covered is relatively short, a clear increase of *that*-omission can be noted for *I’m sure*, for which the frequency increased from 93 to 98 percent, thus mirroring the frequencies for *I think* (99 percent) very closely. In order to reconcile the fact that *I think* and *I’m sure* thus exhibit similar behavior, despite differences in frequency entrenchment, we argue that
both constructions are part of the same constructional grammaticalization schema in which the frequency of *I think* seems to reinforce the grammaticalization of both *I think* and *I’m sure*.

**References**


---

**New light on -ing complements of prevent, with recent data from the NOW Corpus**

*Mark Kaunisto and Juhani Rudanko*

University of Tampere

Consider the sentences in (1a–b), from the NOW Corpus.

(1)  
  a. Denayer stuck to Coutinho like glue and prevented him from pulling the strings. (*Liverpool Echo*, 2016)
  
  b. … Darren Randolph prevented him opening the scoring by tipping his half-volley just over the bar. (*The Guardian*, 2016)

In both (1a–b) the matrix verb *prevent* selects a sentential -*ing* complement, with an object NP intervening between the verb and the -*ing* clause. In (1a) the complement construction includes the preposition *from*, and it may be labelled “NP *from*-*ing*,” and in (1b) it is non-prepositional, and may be
labelled “NP -ing.” The alternation is of interest for instance because it has been observed in earlier work (Rohdenburg 1995; Mair 2002) that the non-prepositional pattern is much more frequent in British English than in American English, with the difference constituting a major grammatical difference between the two major core varieties of English.

The present study sheds new light on aspects of the variation between the two alternants with very recent data from the NOW Corpus. The new data are seen to confirm the difference between British and American English, with the non-prepositional pattern being much more frequent in the former. The present study then extends the range of data to Indian, Pakistani and Philippine English, again with the NOW Corpus as the source, to gain a view of the variation, or lack of it, in selected non-core varieties of the language, in relation to the difference between British and American English. Going beyond descriptive objectives, the study also examines the question of what factors may promote the incidence of the non-prepositional pattern. The role of complex NP objects of prevent is a well-known consideration in view of Rohdenburg’s (1996) Complexity Principle, but is set aside in the present study with the help of search strings with non-complex pronominal object NPs, as in (1a) and (1b), and the objective is instead to investigate whether a semantic difference may be identified between the two variants when they are viewed as separate constructions. The study also investigates whether new data from the NOW Corpus may shed fresh light on the two types of constructions from the point of view of the distinction between control (see Sag and Pollard 1991, where prevent is listed as a control verb) and NP Movement (see Postal 1974, where prevent is discussed as an NP Movement verb).

References


Adverbial fronting in learner language: Production and perception

**Kathrin Kircili**
Justus Liebig University, Giessen

As being one central non-canonical syntactic feature, fronting has been shown to fulfil various functions in the speech and writing of native speakers of English, such as the emphasis (Biber et al. 1999; cf. (1)) as well as the linking (cf. (2)) or contrast (cf. (3)) of sentences or sentence elements (cf. Callies 2009):
As examples (1)–(3) already indicate, one constituent which can fulfill all of these functions is the adverbial, a sentence element which has, however, mainly been neglected in descriptive analyses so far (cf. Callies 2009). In the field of learner corpus research, to the best of my knowledge, adverbial fronting has not played a major role either, with the laudable exceptions of Granger & Tyson (1996), Milton (1998) and Leńko-Szymańska (2008) who have, however, mainly focused on linking expressions or connectors.

Against this backdrop, this study investigates both the production and perception of adverbial fronting in German learner language by employing a multi-method approach. In the respective corpus analysis, approximately 1,500 sentences (75 essays) taken from the International Corpus of Learner English (ICLE; Granger et al. 2009) were manually analyzed, taking both the form and function of the individual adverbials into consideration. It has been revealed that there is a highly significant overuse between the frequency of forms and functions of fronted adverbials in essays produced by German learners compared to those by their native peers, with the use of single-word conjuncts as well as clusters of up to three or four adverbials being particularly prominent, e.g.:

(4)  <ICLE GEAU1053>:
  [Every time] NP,a [when I have arrived [sic] the middle of the street,] cl,a
  [suddenly] A,a a car seems to come out of nowhere and drives right in my direction.1

Further findings include the fact that, in German learner essays, fronting is significantly more frequently predicted when the contained information is discourse-old while in the LOCNESS data the opposite is the case.

In order to assess learners’ perception, an elicitation experiment based on sentences from the corpus data was carried out to evaluate the degree of acceptability as well as the preferred choices with regard to the placement of adverbials or the use of adverbial clusters.

It has been found, for example, that students of English, compared to those of other fields, exhibit a significantly higher acceptance for more complex structures, such as inversion, which proved to be more difficult to recognize for inexperienced learners of English:

(5)  <GEAU1036>:
  But then only, the problems in question are solved. [sic!]

These findings, among others, will be discussed in the light of the frequent call for a change in teaching materials as well as the development of more appropriate pedagogical approaches which are supposed to enable students to explore the syntactic, semantic and stylistic functions of conjuncts (or adverbials in general) necessary to employ them in an adequate manner.

References

1 The small letter following the underscore indicates the function of the adverbial. In this case, all three, the noun phrase, the clause as well as the single-word adverbial operate as adjuncts (cf. Quirk et al. 1985).
The pragmatics of intonation: Fall-rise tones in the SPICE-Ireland Corpus

John M. Kirk
University of Vienna

Prosody does not have any propositional meaning and thus functions pragmatically. The fall-rise – one of a small number of nuclear tones – conveys some kind of implied meaning. When statements are made with fall-rises, the speaker states one thing but implies something further.

The present data for come from the SPICE-Ireland Corpus (Kirk et al. 2011), which comprises 15 discourse situations encompassing private and public speech, monologues as well as dialogues and multilogues, and totals 626,597 words. This paper will concentrate on the fall-rise tone. The corpus has 1051 fall-rise tones distributed across each of the 15 text categories. In the SPICE-Ireland Annotation Scheme (Kallen & Kirk 2012; Kirk 2016), a fall-rise tone is indicated by capitalisation of the vowel concerned and by prefixing the number 5 to the word concerned. Comparison is made with British data from the Lancaster/IBM Spoken English Corpus (Knowles et al. 1996), which has 23,855 words and 1221 fall-rise tones (Mindt 2001).

Building on the research literature (see References), I have devised a taxonomy of two mega-pragmatic functions, each with several subfunctions, and applied the taxonomy to a preliminary study of 200 fall-rise tones (100 each from Northern Ireland and the Republic of Ireland).

The first mega-pragmatic function of fall-rise tones, shown by the great majority of examples in the pilot study, is textual and shows how information is both ordered and focused upon. Attached to adverbial structures (such as clauses, phrases or simple stand-alone adverbs), fall-rise tones indicate
that the discourse is still unfolding and ongoing. One textual function is the indication of the status of the information as old, given or negotiated information. Another textual function is to indicate that the information to which the tone is attached is to be interpreted simply as one of a choice of references or expressions or formulations from a range of possibilities which could have been invoked by the speaker in the situation. In these two main ways, fall-rise tones pragmatically provide some textual coherence.

Here is an example showing fall-rise tones at the end of two conditional if-clauses:

<#> <rep> No 11prOblem with that Alistair%@ </rep> <#> <rep> I- mean* if you 1thInk about 1thAt in the 1eOntext of 1sOciolinguistic 5lIterature% it would be a good 1mEthod% or 1shOUld be a good method to 2Use% because uh you-know* if you use uh a 5nEtwork% that you 're 1pArt of% 1yOU are more likely to be less 1intImidating to them% and to get better 1spEEch% </rep> <#> <rep> More 1naturallIstic speech% </rep> [ICE-NI-CLD-P1B-001$A]

The second mega-pragmatic function of fall-rise tones, as evidenced by the remaining examples in the pilot study, is interpersonal and shows a range of attitudes which the speaker conveys towards what he or she is saying. Such attitudes range from reservations, doubt or uncertainty through to polite or partial corrections of what a previous speaker has just said. In this function, fall-rise tones can tentativise a speaker’s choice of word indicating contrast or emphasis or simply express politeness. Such pragmatic uses are sometimes reinforced by the accompanying syntax, especially negatives, yes-no questions and feedback responses.

Here is an example from a news broadcast of two fall-rise tones indicating interpersonal tentativeness:

<#> <rep> It 's 2nIne minutes to 2elEven% </rep> <#> <rep> You 're 1wAtching a 1UTV 1Llve 1spEcial% 1GrAssroots 1LOyalists gave an 1Upbeat reaction% to the IRA 1cEAse_2fIre% in a 1spEcial 1phOne-in 2prOgramme% on the 2ShAnkill 1ROAd tonight% </rep> <#> <rep> The 1consEnsus 5opInion% was a 5cAlm% 1And a 1considered one% </rep> [ICE-NI-BRN-P2B-001$A]

Although fall-rise tones are predominantly attached to nouns in just over half the examples, the present study also shows that most grammatical classes may be accompanied by fall-rise tones, including names.

A working hypothesis is that, given the apparent universality of fall-rise tones, and notwithstanding the predominance of rise tones for which Northern Ireland is renowned, the pragmatic functionality of fall-rise tones should actually turn out not to be dissimilar between both Northern Ireland and the Republic of Ireland, or between them and Great Britain.

Further research questions concern the relationship between fall-rise tones and the text categories of spoken discourse, and the distribution among speakers in text categories.

References

The choice between –ing and to complement clauses in English as first, second and foreign language?

Daniela Kolbe-Hanna and Lina Baldus
University of Trier

This paper examines the choice between –ing and to complement clauses in English as a first (ENL), as a second (ESL) and as a foreign language (EFL) to reveal whether and how speakers from these different backgrounds employ different complementation patterns. Some matrix verbs licensing non-finite complement clauses can control both to and –ing clauses. While the choice of clause type results in a semantic difference with some verbs, e.g. remember or try (remember to do...
something has a different sense than remember doing something), this is not the case with other, mostly aspectual, verbs such as start (Biber et al. 1999: 758–759).

In a dataset including the complement clauses of this latter type of verbs, we investigate speakers’ choice of clause type across different varieties of English and in particular whether this choice is influenced by the way they acquired English. Due to the lack of comparable corpora providing a sufficient amount of data for the study of grammatical variation in EFL vs. ENL vs. ESL, our paper consists of two studies. The first study compares the use of –ing and to clauses by learners of English in German-speaking countries in the International Corpus of Learner English (ICLE) with the use of –ing and to clauses by native speakers in LOCNESS and in a subcorpus of the BNC. As ICLE, LOCNESS comprises argumentative student essays; the data from the BNC will be restricted to the corresponding genres W:essay:school and W:essay: university.

In the second study, complementing previous research by Deshors and Gries (2016), we draw on written data from the following ICE-components: Canada, Ireland and New Zealand as ENL varieties and Jamaica, Nigeria and Philippines as ESL varieties (applying the distinction between L1 and L2 as in, e.g., Kortmann & Lunkenheimer 2012).

Using a MuPDAR approach in both studies, we conduct regression analyses to first identify the predictors in the ENL varieties, and then compare these with the predictors identified in the EFL and ESL data. The regression analyses will be complemented by decision tree and random forest analyses. The independent variables included pertain to morphosyntactic features of matrix and complement clause, cognitive factors (priming, length of embedded clause, frequency of individual matrix verbs) and speakers’ social background (country of origin, sex, age).

Preliminary results show that the overall proportion of –ing and to clauses is similar in ENL and ESL varieties. The EFL data exhibit a higher proportion of –ing clauses. This result is somewhat unexpected, because German lacks a similar grammatical structure, but it reflects that learners know that this construction exists, even if they do not use it as a native speaker would. Furthermore, different matrix verbs are similarly frequent across the ENL and ESL varieties, while there is a difference in their proportions (e.g. of start vs. begin) in the EFL data.

As different varieties of English become increasingly visible and begin to mix in the wake of globalisation, this paper illustrates how change in society may affect the face of English across the world.

References


Possession in the Australian Hansard: Stative have (got) in the 20th century

Minna Korhonen
Macquarie University

Previous research has revealed differences in the development and usage patterns of stative possessive have (got) between the North American varieties of English and the British and Antipodean varieties of English in that by and large British English and New Zealand English are moving towards have got (I have got a cat) while in American English and Canadian English the preference for have (I have a cat) is increasing (e.g., Biber et al. 1999; Jankowski 2005; Quinn 2004, 2009; Tagliamonte 2003; Tagliamonte et al. 2010). The rise of do-support in negation (I don’t have a cat) and questions (Do you have a cat?), with American English reportedly leading in its use especially in negative contexts (Hundt 2015: 80), is of special interest in relation to the overall development of this construction.

Thus far studies on stative possessive have (got) and its development in historical perspective in Australian English (AusE) are rare, and they mainly discuss the use of do-support with negatives (and interrogatives) (e.g., Hundt 2015). To account for the diachronic changes in the overall use of this construction, this paper presents a case study on the use of stative have (got) in AusE in the 20th century. The study is based on material from the Australian Diachronic Hansard Corpus (Kruger & Smith 2018 in press) covering four data points at approximately 30-year intervals in the period of 1901 to 1995. By being a “substantially verbatim” written record of parliamentary discourse, the Australian Hansard provides a comparable dataset for investigating linguistic changes occurring during the time period in question.

In order to investigate the changes in the usage patterns, all instances of stative have (got) in simple present tense were searched for and subsequently coded and analysed with reference to the linguistic factors of subject type (pronoun/noun phrase) and object type (concrete/abstract) as well as the use of do-support in negatives and questions. The study shows how the different forms of have (got) as used for possessive meaning vary across the 20th century in AusE parliamentary discourse and indicates an increase in the use of the innovative forms (i.e., have got as well as do-support) towards the latter part of the time period.

In addition to giving an overall picture of the increasing variability in the use of stative have (got) in AusE, the study also provides a close examination of the contexts in which the innovative have got emerges. Furthermore, the diffusion of do-support in negative and interrogative contexts is investigated in detail.
References


A sociolinguistic corpus study of *actually* in current spoken British English

**Manfred Krug and Lukas Sönning**

University of Bamberg

*Actually* ranks among the most frequent adverbs in present-day spoken British English (Aijmer, 2015). According to recent analyses (Traugott & Dasher 2002; Defour et al. 2010), it was originally an adverb indicating realness which has developed the additional meaning of an epistemic adversative and, most recently, the multifunctionality of a discourse marker. As such, *actually* has received considerable attention in previous research, both from a theoretical (e.g. Traugott & Dasher 2002; Taglicht 2001) and an empirical perspective (e.g. Aijmer 2002, 2013, 2015). Existing real-time and apparent-time studies suggest an ongoing, steady increase in frequency in present-day English (Waters 2011; Aijmer 2013). Furthermore, corpus-based work has identified the following systematic usage patterns: (i) *actually* is a characteristic feature of speech: in spoken British English, it occurs 6 to 9 times more often than in writing (Aijmer 2013); (ii) regional differences surface in spoken genres, with the text frequency in British English exceeding that in American English by a factor of 2 to 3 (Oh 2000; Aijmer 2002, 2013); (iii) its prevalence in spoken texts also
appears to be sensitive to the presence of interlocutors, as the incidence of actually in dialogues is about twice that in monologues (Aijmer 2013).

The present study aims to enhance our understanding of usage patterns of actually by offering new corpus-based insights into the sociolinguistic dimensions of its variation. With previous work having identified actually as a typical feature of spoken, dialogic, present-day British English, we use the Spoken BNC2014 (Love et al. 2017) to shed light on the relationship between usage rate and social factors. It has been claimed that a key characteristic of discourse markers is their indexicality (Silverstein 1976), which, among other things, refers to the marking of socio-cultural dimensions of the communicative situation, including classic sociolinguistic categories such as age and gender of the speaker. We support this claim empirically by modeling the frequency of actually as a function of four sociolinguistic variables: age, gender, qualification level, and region. A negative binomial regression model (Hilbe 2011) is used to analyze the utterances of \( n = 446 \) speakers (i.e. the subset of native speakers of English currently residing in England, with complete information on age, gender, and educational background in the Spoken BNC2014). Considering the development into a discourse marker as a change from below, we find the expected apparent-time increase, with younger speakers showing higher usage rates. In line with Labovian principles, this change in progress appears to be led by females. In addition, actually seems to be used more frequently in the south of England than in the Midlands and the north, suggesting a spatial diffusion pattern, which may originate in the populous parts of the south. Finally, the rate of actually is found to increase with level of education, a pattern typically associated with change from above. We therefore find an interesting change-in-progress scenario in which, among the factors studied, the rate of actually appears to be most sensitive to qualification level and age, followed by region and gender.

References


**Informativity in spoken corpora: Contextual predictability affects word order**

*Gero Kunter*
Universität Siegen

From an information-theoretic perspective, the amount of new information that a word adds to the discourse is determined by its predictability in a given context, as shown in (1) and (2):

(1) *the aunt found she couldn’t have guests in wheelchairs in the dining room*  
(COCA_MAG_1992)

(2) *when there was nothing to do I could sack out behind his desk in the reading room*  
(COCA_FIC_2006)

In (1), *room* is highly predictable: according to COCA (Davies 2008–), there is a probability of 52.3 percent that the word following *dining* is *room*. In contrast, the probability of *room* following *reading* as in (2) is much lower (0.3 percent). Accordingly, *room* adds less new information to the discourse in (1) than in (2).

Recently, a number of studies has investigated the effect of informativity in diverse linguistic areas such as articulation (Bell et al. 2009), processing (Frank & Willems 2017), or discourse organization (Xu & Reitter 2018). Additionally, informativity has been considered as a determinant of syntactic structure. Levy and Jaeger (2007) and Levy (2008) suggest that speakers tend to arrange linguistic units within an utterance so that informativity is distributed as uniformly as possible, as this is assumed to facilitate processing of the utterance by listeners (Jaeger 2010; Jaeger & Tily 2011). It follows from ‘uniform information density’ principle (UID) that the informativity of neighboring words in a corpus should not be independent. Instead, words with relatively high informativity should be followed by words with relatively low informativity, and vice versa. However, as yet there is only indirect support for this prediction, for instance in Schneider and Grigonytė (2016) who show that the overall distribution of word informativity in spoken corpora is compatible with UID.

The present paper explicitly tests the prediction that informativity of adjacent words is not independent. Using data from two corpora of spoken American English (Buckeye Corpus, Pitt et al. 2007; Switchboard Corpus, Godfrey & Holliman 1997), it is shown that the informativity of a word is inversely related to the informativity of both the preceding and the following words. Words with relatively high or low informativity are typically surrounded by words with low or high
informativity, respectively. No such effect is visible for words with an average informativity. As this relation persists if word-class is accounted for and even if function words are excluded from the analysis, the effect cannot be reduced to a sequential alternation of function and content words.

Thus, the present results supports the central claim of UID insofar as in spoken corpora, words with high or low informativity are surrounded by words which compensate this informativity excursion, thus approximating a uniform distribution within an utterance. However, as other factors such as word class of the target words show an even stronger effect, the distribution of informativity is apparently subordinate to syntactic structures. Nevertheless, the findings imply that within these constraints, speakers choose to arrange lexical items so that informational load is balanced, possibly for the benefit of listeners.

References


Keyword analysis has become a common tool in corpus linguistics to examine large text collections. For instance, the method has been applied to explore discourses of refugees (Baker et al. 2004), and the characteristics of different text varieties (Xiao & McEnery 2005; Kilgarriff 2012). Typically, keywords are defined as words occurring ‘with unusual frequency’ in the corpus. They are said to be informative about the corpus topic and style (Scott 1997; Baker 2004). This article develops this notion by 1) proposing a novel method to measure the keyness and 2) extending the analysis beyond keywords to cover also other key features. In the study, key features and keywords estimated with this novel method are compared to the results of a traditional, frequency-based keyword analysis (Scott 1997).

As a method for estimating key features, we suggest supervised machine learning: a support vector machine (SVM). SVM uses information from training data consisting of example documents with register labels to automatically identify (1) the register for new documents (Vapnik 1998) and (2) the important features of the register classes that contribute to their identification. These can be analyzed as key features reflecting their linguistic characteristics (Baker 2004).

SVM has several advantages when compared to traditional keyword analysis. First, SVM does not treat corpora as unitary wholes but seeks to identify features that best characterize each text. Consequently, it suits data where words or other features are not evenly dispersed across texts. Further, SVM qualifies well for large data sets, and instead of relying only on a reference corpus and a target corpus, it identifies features that are key for a register when compared with all other registers. Finally, SVM can use as features a wide range of linguistic elements: words, but also, e.g., syntactic information.

We apply this novel method to analyze English online registers, i.e. text varieties associated with a particular situation and linguistic features (Biber 1988). As data, we use the Corpus of Online Registers of English composed of 48,571 documents, each manually classified for register information (see Biber, Egbert & Davies 2015; Egbert, Biber & Davies 2015).

Previous studies have shown that registers display important grammatical and syntactic variation (Biber & Egbert 2016). Because keyword analysis focuses on topics and accounts for grammar only indirectly, it has limited usefulness as a method for analyzing register differences. We compare traditionally counted keywords to keywords and key features estimated by an SVM. As key features, we apply unlexicalized syntactic n-grams: subtrees of dependency syntax analysis without the lexical information (see Figure).
Figure: The formation of unlexicalized syntactic n-grams. Nsubj: nominal subject, xcomp: clausal complement, punct: punctuation.

Previous studies (Laippala et al. 2015; 2017) have shown that syntactic n-grams describe diverse text characteristics, such as narrative constructions and interjections, features that are identified as predictors of register variation (see Biber 1988). Our presentation provides examples of these constructions as key features of online registers, thus giving new information on their characteristics.

References


Grammatical stance marking in student and expert production: Revisiting the informal-formal dichotomy

Tove Larsson
Université catholique de Louvain

Stance-marking devices help us position ourselves in relation to our claims, and as such are of central importance in both speech and writing (e.g. Biber et al. 1999). While frequent grammatical stance-marking devices such as stance adverb (e.g. probably), stance complement clause constructions (e.g. it is probable that) and stance noun followed by a prepositional phrase (e.g. the probability of) have received ample individual attention in previous studies, little is known about the interrelation of these morphologically and semantically related lexical sets and to what extent the use is register-specific. This paper aims to investigate the distribution of such sets across registers in expert production and in non-native-speaker (NNS) and native-speaker (NS) student writing to further explore previous claims of informality, as outlined below.

The examples above represent some of the ways in which the base form PROBAB* can be realized. Other, semantically similar sets include CONCEIVAB* (conceivable, conceivably, conceivability) and LIKEL* (likely, likely, likelihood). What pragmatic and syntagmatic relations each of the members of such sets participate in might, however, be unknown to student writers who may view these as fully synonymous variants from which to choose freely. In fact, previous research on POSSIB* has noted that learners tend not to be able to use such markers in a target-like manner (Larsson 2017). As student writers also have been shown to exhibit somewhat insufficient register awareness (Larsson & Kaatari under review), further investigation of how these variants pattern across registers will shed further light on the often-cited claim that students, in particular NNS students, are overly informal in their writing (e.g. Altenberg & Tapper 1998).

Against this background, this study maps out the distribution of semantically and morphologically related stance markers across registers to investigate the following research questions:

- What differences and similarities in use can be noted across registers in the expert writing?
- Which of the experts’ registers is the student writers’ use closest to, and what can this tell us about (in)formal uses of such stance markers?

The study uses subsets from one expert corpus, BNC-15, two learner corpora, ALEC and VESPA, and two NS student corpora, BAWE and MICUSP. Biber et al.’s (1999: 969–970) framework of grammatical stance marking is used, along with inferential and explorative statistics, such as
Multiple Correspondence Analysis. The results show that the lexico-grammatical preferences vary across registers; for example, the base form PROBAB* (in particular in its stance adverb form: probably) is most strongly associated with spoken data, whereas POSSIB* (in particular when realized as it is possible that) is strongly associated with the academic register. The results show further that while the use by the students is most like that of the academic experts, all student groups share some preferences with the non-academic registers, suggesting somewhat insufficient register awareness. The results will hopefully contribute to nuance the informal-formal dichotomy, thereby benefitting both EAP instruction and theories of stance marking.

References

ALEC (the Advanced Learner English Corpus). Corpus compiled by Tove Larsson at Uppsala University in 2013.

Using corpora to probe shared grammatical constraints: Evidence from restrictive relativization

Stephen Levey and Heike Pichler
University of Ottawa, Newcastle University

Variation in restrictive relativization strategies is well documented in diachronic and synchronic varieties of English (Ball 1996; Nevalainen & Raumolin-Brunberg 2002; Romaine 1982;
Previous research addressing this variation reveals a preoccupation with two unresolved questions: To what extent have WH-relativizers infiltrated modern vernaculars? Is there a vernacular norm with respect to the distribution of relativizer variants? Observing that earlier responses to both questions are largely based on overall variant rates (e.g. Ball 1996), we extend the research focus to the underlying constraints governing relativizer selection, as revealed by multivariate analysis of spontaneous usage data.

We apply the framework of comparative variationist sociolinguistics (Poplack & Tagliamonte 2001) to five varieties differing in terms of geographical location as well as their relative participation in mainstream linguistic developments. Over 3,800 tokens of restrictive relative constructions were extracted from 200 speakers representing: (i) Berwick English, a peripheral variety spoken in the far north-east of England; (ii) Tyneside English, spoken in the northeastern Tyneside conurbation including the city of Newcastle; (iii) Salford English spoken in Greater Manchester, a large metropolitan area in the north-west of England; (iv) London English, a rapidly innovating cosmopolitan variety in the south-east of England; and (v) Ottawa English, a mainstream variety of urban Canadian English.

We coded these data for: extra-linguistic variables, linguistic predictors widely implicated in relativizer choice (e.g., syntactic function of the relativizer; animacy and definiteness of the antecedent head NP; adjacency and length of the relative clause); and lesser-investigated linguistic predictors (e.g., lexical specificity of the antecedent head NP; subject of the relative clause; internal syntax of the matrix and relative clause).

The results turn up a number of key findings: (i) who is the only productive WH-relativizer in all varieties; (ii) there is evidence of apparent-time change in the use of who (BwE, OttE), the zero-relativizer (BwE, TyE, LnE), and that (SE, LnE); (iii) relativizer choice in non-subject relative clauses is strongly influenced by surface-level processing constraints, as indicated by the direction of effect for adjacency and subject of the relative clause.

Volatility in relativizer usage rates across varieties points to evolving differences in the social conditioning of variant selection. But deeper correspondences across varieties, arguably linked to the processing complexity of filler-gap dependencies and the avoidance of syntactic ambiguity, suggest that varieties share insufficiently acknowledged structural affinities. Taken together, the results indicate that the quest for a vernacular norm in restrictive relativization should be refined by moving beyond superficial usage rates and examining the underlying constraints on variability.

References

Acronyms are prevalent and increasingly frequent both in English (Leech et al. 2009: 212–213) and other languages, such as German (Steinhauer 2000: 1), a development which mirrors the increasing societal prominence of science/technology and politics/business outside specialized domains (Kobler-Trill 1994: 200). Although acronyms allow brief and unambiguous communication among experts, they also decrease transparency for non-experts both when it comes to retrieving the full form of the acronym (e.g., LSD) or its referent (UNFCCC). The potential lack of transparency is further compounded in translations due to cultural differences. However, few previous studies have addressed the translation of acronyms and none from a corpus-based perspective.

This study investigates the use of acronyms in English originals and their translations into German and Swedish, comparing forms, functions and distributions across the languages. A major outcome will be a typology of translation strategies and acronym use in the three languages. The material consists of an English-German-Swedish popular non-fiction parallel corpus currently being compiled by the authors. This genre covers, for instance, popular science and biographies, and the texts are aimed at informing and entertaining non-specialist audiences. Therefore, writers and translators need to strike a balance between brevity and transparency without compromising accuracy or alienating readers.

Preliminary results suggest that acronyms most often occur as noun phrase heads (When IBM introduced…), but they are also frequent in more complex structures such as English premodifiers (PGP encryption) and German (UN-Klimakonvention) and Swedish compounds (NKVD-officer) (cf. Ström Herold & Levin in preparation). They also occasionally form part of new words (NAFTA-style). This flexibility is likely facilitated by the simplex forms of acronyms (Fleischer & Barz 2012: 284).

The first-time mentions of acronyms in texts are of particular interest. Based on our popular non-fiction corpus, knowledge of some frequent acronyms is presupposed (e.g., DNA tests), others are
given as appositive noun phrases alongside the full form (The chemical dichlorodiphenyltrichloroethane (usually known as DDT) […] (cf. Biber & Gray 2016: 202–207), while some receive more extensive meta-linguistic comments (WYSIWYG, pronounced “wiz-ee-wig,” an acronym for “What you see is what you get.”). This is also found in translations, which can be either more or less explicit than the original:

(1a) Complete the CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart), and you’re in.

(1b) den CAPTCHA […] (den “Completely Automated Public Turing test to tell Computers and Humans Apart”, also den “vollautomatischen öffentlichen Turingtest zur Unterscheidung von Computern und Menschen”) [‘i.e. the “completely-automated…”’]

(1c) captcha-rutan (ett robotfilter för att skilja människor från datorer) [‘the captcha-box (a robot-filter to tell …’]

The translations of first-time mentions vary greatly between German and Swedish target texts. Important factors are the target audience’s (assumed) culture-specific knowledge and their knowledge of English. Our acronym typology will consider structural and pragmatic features and their relevance to translation.

References


Searching for parallels between ontogenetic and diachronic grammaticalization: The case of want to/wanna

María José López-Couso and Raúl Río-Fernández
University of Santiago de Compostela

Over the last few decades a number of similarities in developmental patterns between diachronic change and first language acquisition have been identified both for English and for other languages in different grammatical domains. One of these is modality, which has been shown to constitute a particularly suitable testing ground for the identification of potential parallels between ontogeny and diachrony. Stephany (1986), for instance, demonstrates that the deontic meanings of obligation and permission emerge earlier than the epistemic meanings of possibility and necessity in Child English, a development which seems to replicate the rise of epistemic meanings out of deontic ones in the history of the language (cf., among others, Traugott & Dasher 2002: 105ff). More recently, Schmidtke-Bode (2009) has examined the similarities between the historical evolution of the ‘emerging modal’ going to/gonna and its acquisition by English-speaking children (see also López-Couso 2017).

The aim of this paper is to explore the potential parallels between the ontogenetic and the diachronic developments of another ‘emerging modal’, namely want to/wanna, which has not been investigated in detail so far. Along the lines suggested by Schmidtke-Bode (2009) for going to/gonna, we compare Krug’s (2000) account of the historical evolution of want to/wanna with the results obtained from the examination of longitudinal data of two children (Adam and Sarah) from the BROWN corpus in the Child Language Data Exchange System (CHILDES) archive. The variables selected for analysis include: (i) the variant forms of the want to-construction (want + bare infinitive, want + to-infinitive, want(t) (ta), and wanna); (ii) subject type (first/second/third person personal pronouns, proper names, NPs, and omitted subjects); (iii) clause type (declarative vs. interrogative clauses and main vs. subordinate clauses); and (iv) whether the use of want to (and variants) is self-initiated by the speaker or is a reaction to a previous occurrence of the construction.

Our analysis considers not only the children’s production, but also that of their input (parents, caregivers) in conversational interaction, with a view to identifying whether there exist inter-individual differences between the two children which may be due to the nature and the frequency of the input they receive in the process of acquisition.

The preliminary results show that the overall developmental patterns in Child English are consistent with the diachronic facts, with the full form want to being acquired earlier than the reduced wanna, via an intermediate stage (in the case of Adam) with the variant want(t) (ta), which represents the beginnings of the process of erosion of want to. Moreover, the two children seem to acquire first “specific subject + auxiliary combinations” (Theakson et al 2005), those more frequently used in the ambient language becoming lexicalized units in the children’s production (e.g. I want to in Adam’s speech). As regards the relevance of clause type to the distribution of the variant forms of the want to-construction, the data show that the most complex clause types (interrogative and subordinate clauses) typically favour the presence of the infinitival marker to in the production of the two children.
Foreign elements in L2 writers’ research papers – Communicative strategy or display of academic literacy?

Antorlina Mandal and Leonie Wiemeyer
University of Bremen

This study explores the use of foreign elements in linguistic research papers written by L1 German EFL learners from the Corpus of Academic Learner English (CALE; Callies & Zaytseva 2013). By definition, learner corpora contain texts produced by multilinguals. As a result, learner texts are likely to contain elements from other languages (Callies & Wiemeyer 2017). Research into codeswitching has shown that even advanced learners resort to their mother tongue for bridging lexical gaps and for self-repair (Liebscher & Dailey-O’Cain 2005). Analysing interviews from the LINDSEI (Gilquin, De Cock, & Granger 2010; Nacey & Graedler 2013) established that recourse to the L1 was a typical communication strategy. In another study based on the LINDSEI, De Cock (2015a, 2015b) found that aside from lexical bridges, the learners frequently used cultural bridges, e.g. place names, and L1 pragmatic markers. Lexical gap-filling strategies are also found in L2 writing. A longitudinal study of young learners (Augustín-Llach 2010) showed that beginners tend to resort to their L1, while with increasing proficiency the same learners employ strategies that rely on the target language.

In academic contexts, language learners also exploit their multilingual competence. The specialised nature of academic texts is likely to bring about different multilingual practices. Thus far, however, there is no corpus-linguistic evidence of learners’ use of foreign elements in their academic writing. The present contribution aims at addressing this research gap with regard to two research questions:

References

• Which kinds of foreign elements do L2 writers use in research papers and how are they embedded?
• What are the functions of these foreign elements?

In order to explore the multilingual strategies of advanced L2 writers, two coders analysed 93 research papers from the CALE. They had been written by undergraduate English students in linguistics courses at three German universities. Occurrences of foreign elements were identified via the tag <mentioned item>, single and double quotation marks, and additional manual analysis. They were then categorised along the following parameters: language of origin, type of element, and textual integration.

It was found that the majority of foreign elements were individual words and phrases, usually from the writers’ L1 German. The learner texts in the CALE, like those in the LINDSEI, also contained cultural bridges, though they were generally a minor phenomenon. Unlike in spoken language, foreign elements were not employed to fill lexical gaps. Instead, in accordance with the specialised nature of the texts, the learners used discipline-specific terminology, examples, and illustrations from languages other than English. They exploited their multilingual skills to review and cite academic publications in their L1 German, and other L2s. Foreign-language citations showcased their ability to discuss linguistic phenomena encountered in a variety of languages. The learners’ use of foreign elements was thus indicative of their academic literacy in several languages and their professional and comprehensive approach to academic writing. The findings show that L2 academic writers’ use of foreign elements at advanced levels of proficiency is not a communicative strategy, but fulfils academic goals.

References


The participle in Old English translations from Latin. Adjectival inflection and text type

Javier Martín Arista
Universidad de La Rioja

This paper deals with the participle in a selection of translations from Latin into Old English. More specifically, it focuses on the inflectional morphology of the Old English participle and analyses the variation between adjectivally inflected and uninfl ected participles that is illustrated in example (1).

Whereas the past participle *afliemde* ‘banished’ in (1a) is inflected for the nominative masculine plural, *afliemed* ‘banished’ in (1b) shows the verbal inflection corresponding to the past participle but does not have any adjectival ending.

(1)  
(a. [Or 1 029600 (10.29.14)]
*On þære ilcan tide wurdon twegen æþelingas afliemde of Sciphtian...*
At that same time two noblemen were banished from Scythia...

(b. [OrHead 001000 (1.10)]
& *hu ii æþelingas wurdon afliemed of Sciphiu.*
And how two noblemen were banished from Scythia.

Previous research raises the questions of the loss of inflection and the role of Latin influence. Wojtyś (2016) finds variation in the prefixation of the past participle as well as remarkable regularity in the verbal morphology attached to this non-finite form by means of suffixation. As regards inflectional suffi xes, the loss of the adjectival inflection of the past participle is considered evidence for the reanalysis of the adjectival construction into the syntactic passive (Traugott 1992; Denison 1993). On the role of Latin influence, Timofeeva (2010) regards the absolute participle construction as a direct translation from Latin or an imitation. Against this backdrop, the aim of this paper is to quantify the adjectival inflection of the participle in order to answer two questions: (i) do Latin translations differ from other text types as to the adjectival inflection of participles? and (ii) are Latin translations consistent in presenting similar rates of the adjectival inflection of the participle? The scope of the analysis comprises the participle as modifier in the noun phrase and the participle agreeing with the verb *bêon* ‘to be’. The corpus consists of text files that contain at least 10,000 words in the York-Toronto-Helsinki Parsed Corpus of Old English Prose, including Bede’s History of the English Church, Boethius, Consolation of Philosophy, Byrhtferth’s Manual, Chrodegang of Metz, Cura Pastoralis, Gregory’s Dialogues, Herbarium, Bald’s Leechbook, Orosius, and St Augustine’s Soliloquies. The results indicate that if Latin translations are compared with other types of texts, it turns out that Ælfrician texts throw higher rates of adjectival inflection in participles, while vernacular texts like the Anglo-Saxon Chronicle evince lower figures. At the
same time, the vast majority of the participles functioning as modifiers in Latin translations display adjectival morphology, whereas less than one half of the participles in verbal function are inflected as adjectives. Bald’s Leechbook and Herbarium, two scientific texts with Anglian influence, yield totals of adjectival inflection over the average (56 and 73.2 percent, respectively).

References


Lexical co-occurrence and linguistic meaning in historical texts: Novel approaches and new findings

Seth Mehl
University of Sheffield

This paper presents new findings from a major project in relation to lexical co-occurrence and linguistic meaning in historical English texts. This project maps semantic and conceptual change in Early Modern English, using a data-driven approach based on computational analysis of lexical co-occurrence in Early English Books Online (specifically EEBO-TCP). The project innovates by analysing lexical co-occurrence in two unprecedented ways: first, in its use of grammatical baselines when calculating statistical probability of co-occurrence; and second, in its identification of co-occurring lexical trios in wide co-occurrence windows.

We analyse lexical co-occurrence data for every noun, verb, and adjective lemma in every text in EEBO-TCP, including calculations of chi-square, PMI, and w-index statistics (cf. Sheskin 2004). These statistics are calculated based on a grammatically defined statistical baseline, comparable to Bowie et al. (2013), rather than a traditional baseline of all words in the data set. We have argued that this grammatical baseline reflects linguistic probability more meaningfully than traditional measures (cf. Wallis 2012; Fitzmaurice et al. 2017) by dramatically reducing invariant Type C terms (cf. Wallis 2014). In the first portion of this talk, I address the research question: How do lexical co-occurrence measures based on this grammatical baseline differ from a traditional baseline? I present results of both baselines across multiple co-occurrence window sizes. I compare
the two approaches, and I describe and evaluate differences between them in relation to linguistic features, with a focus on lexical semantics.

We also apply an innovative technique to identify not just pairs of co-occurring words, but also trios of co-occurring words within the given proximity windows. Fano (1960), in his original formulation of mutual information, discussed conditional probabilities for up to three data points. In linguistics, pointwise mutual information (PMI) has only been applied to two data points at a time, in the measurement of co-occurring lexical pairs. PMI measures are calculated for co-occurring lexical pairs across wide proximity windows of 50 tokens to either side of the node. These pair measurements are then expanded to measure co-occurring lexical trios within the same proximity window. In the second portion of the talk, I address a second research question: How can we interpret co-occurring lexical trios in terms of semantic, pragmatic, or discursive meaning? I provide examples of such lexical trios and relate these examples to linguistic meaning. I also map distributions of co-occurring pairs and trios across multiple sizes of co-occurrence windows. I identify typical and atypical distributions of such pairs and trios, and investigate whether such typical and atypical distributions can be linked to observable linguistic features or trends, with a focus on semantic, pragmatic, and discursive meaning.

References


The subjunctive in main clauses: A diachronic corpus study

Lilo Moessner
RWTH Aachen University

Present-Day English (PDE) reference works distinguish three types of subjunctive: the mandative subjunctive, the formulaic subjunctive, and the past or were-subjunctive (Quirk et al. 1985: 155–
Since the last two are characterized as archaic, it is no wonder that only the mandative subjunctive has attracted the attention of linguists (for a discussion of recent publications on this construction cf. Waller 2017: Chapt. 5). The mandative subjunctive seems to be the only type with a recent frequency rise, which started in the first half of the 20th century in American English and was followed by other diatopic varieties (Övergaard 1995: 89).

In this paper I propose to look at the underresearched category of the formulaic subjunctive. It is used in main clauses, where it expresses a wish, and it is allegedly restricted to set phrases like God save the queen in PDE. It can be assumed that like the other subjunctive types the formulaic subjunctive, too, was a productive pattern in the earlier periods of the English language. I will explore if it lost its productivity at all, and if so, when this happened and which construction(s) took over its function.

My paper focuses on the Old English (OE), Middle English (ME) and Early Modern English (EModE) periods and includes an outlook on later developments. I will adopt a research method, which combines a quantitative and a qualitative approach, and my data come from the three chronological parts of the Helsinki Corpus.

The varying competitors of the subjunctive in main clauses (OE: modal constructions, uton + infinitive, is + to + infinitive, imperatives; ME: modal constructions, uten + infinitive, periphrastic constructions, imperatives, let-constructions, ambiguous verb form + 1st ps pl. pronoun; EModE: modal constructions, periphrastic constructions, imperatives, let-constructions) will be established, and all occurrences of the relevant verbal syntagms will be analysed with respect to the influence of the parameters period, format (prose vs verse), and text category on the distribution of their realisation possibilities.

The quantitative results will allow me to trace a precise frequency development of the construction, complementing earlier statements on the subjunctive in general (Fischer 1992; Mustanoja 1960) and providing evidence against the hypothesis that the frequency of the formulaic subjunctive dropped only after the OE period (Ogawa 1989). Additionally I will claim that modal constructions were the strongest competitor of the formulaic subjunctive as early as the OE period. Finally I will argue that the most interesting result of the analysis is the big influence of the parameter text category on the history of the formulaic subjunctive. Of all categories, statutory texts will turn out as the main stronghold of the construction until the EModE period.

The qualitative analysis of the statutory texts will show that changes in the genre conventions of legislative texts were not only instrumental in the preservation of the construction, but also contributed to the development of one of the so-called formulaic frames, namely be it ... enacted ... that..., in which the PDE formulaic subjunctive is productive and not restricted to a finite set of expressions (Huddleston and Pullum 2002: 944; Waller 2017: 24).

References


Interaction and culture – conversational patterns in Caribbean and Southeast Asian varieties of English

Theresa Neumaier
University of Regensburg

Conversational interaction is inevitably shaped by the cultural context it is situated in. This has long been claimed in literature on intercultural communication (e.g. FitzGerald 2003), and has also been described in World Englishes (e.g. Kachru and Smith 2008). In fact, as postcolonial and global Englishes are situated in culturally diverse contexts, conversational structures are likely to differ between varieties. Given the abundance of anecdotes on culturally varying speaking styles and clashes in intercultural communication, it certainly seems surprising that detailed and quantitative analyses of interactional patterns are still lacking for World Englishes (an exception being Sidnell 2001).

In my study, I aim at closing this research gap, employing a mixed-methods analysis to investigate if culturally distinct varieties of English differ in their turn-taking routines, and if they do so in a systematic way. To answer these questions, I expanded the scope of traditional Conversation Analysis (CA) from the study of “talk-in-interaction” to analysing “talk-in-interaction-in-culture” (Carbaugh, 2005, p. 2). Using data and audio files from the Asian Corpus of English (ACE) and the Jamaican component of the International Corpus of English (ICE), I created detailed CA transcripts of four hours of unscripted face-to-face conversations. These transcripts were then tagged with respect to the type and scenario of speaker change at Transition Relevance Places (TRPs) as well as to turn-claiming or -holding strategies used by the interactants. This allowed me to combine quantitative and qualitative analyses and to compare the data systematically.

In this presentation, I will show that the turn-taking model described by CA can be observed in all the interactions analysed. However, even though speakers of different varieties use a similar set of strategies to accomplish turn allocation, they differ in which strategies they prefer and how they realise them in conversation. Almost half of all TRPs in Caribbean interactions lead to self-selections of a next speaker, for instance, often involving passages of simultaneous talk. Southeast
Asian speaker changes, on the other hand, typically occur “in the clear”, without gap or overlap. Speaker groups also differ in how they defend or claim their right to speak: Caribbean interactants, for example, frequently use “machine-gun-style” repetitions or changes in volume to prevent others from starting up. Speakers in ACE, however, prefer more indirect strategies, such as cut-offs, “points of maximum grammatical control” or topicalisation (Schegloff 1996: 93; Sidnell 2010). These observations strongly support the hypothesis that the turn-taking system as described by CA might be a culturally independent framework (cf. Sidnell 2001) which is then shaped by a number of culturally sensitive features.

References


Synonymy revisited: Some diachronic notes on the concept SWEET-SMELLING in American English

Daniela Pettersson-Traba
University of Santiago de Compostela

Despite extensive research, synonymy remains a perplexing linguistic phenomenon: although synonyms share core denotational meaning, they differ as regards other semantic dimensions such as style and connotation, as well as in usage-patterns (Murphy 2003). Recent distributional approaches to semantics (e.g. Behavioral Profiles [Gries & Divjak 2009]) have successfully uncovered subtle distinctions in meaning between synonyms by analyzing, for instance, their collocational and stylistic behavior. So far, most studies have dealt with the semantic structure of sets of synonyms from a synchronic perspective (e.g. Divjak & Gries 2008; Liu 2010, 2015), while their diachronic evolution generally has been neglected. Against this backdrop, the present paper sets out to examine the historical development of the attributive uses of the adjectival synonyms fragrant, perfumed, and scented, which symbolize the concept SWEET-SMELLING, as in (1)–(3):

(1) In England there is the luxuriant foliage, the fragrant blossom, the gay flower […] (COHA, 1870, FIC, LadyIceANovel)
(2) Over his shorts he wore shirts of scented silk or pongee. (COHA, 1947, FIC, Gallery)
(3) McClintock poured tonic on Vridar’s hair and then saturated the hair with **perfumed oil**. (COHA, 1934, FIC, PassionsSpinPlot)

The aim is to compare the collocational behavior of these synonyms in Late Modern and Contemporary American English, using **COHA** (Davies, 2010–), which spans from 1810 to 2009. Since attributive adjectives modify nouns, special attention is paid to the semantic types of their collocates. To this end, head nouns are classified into nine categories extracted from the *Historical Thesaurus of the Oxford English Dictionary*, among others, ‘plants & flowers’, ‘sensation’, ‘aesthetics’, and ‘cleaning’. A multinomial logistic regression analysis is carried out, with synonym (i.e. *fragrant* vs. *scented* vs. *perfumed*) as response variable, and semantic category and period as predictors. The main effects of the two predictors and their interactions are tested, thus considering both the synchronic semantic structure of the synonyms and their diachronic development.

Preliminary results show that, although the individual collocational patterns of the three synonyms do not change considerably over the period analyzed, there is a clear division of semantic labor, especially between *fragrant*, on the one hand, and *scented* and *perfumed*, on the other. Whereas *fragrant* modifies nouns which denote entities with a natural pleasant smell (e.g. those belonging to ‘plants & flowers’), *perfumed* and *scented* collocate with nouns referring to man-made, artificial objects (e.g. those in ‘aesthetics’ and ‘cleaning’). This difference might explain the significant rise of *scented* and the decline of *fragrant* through time. Interestingly, the concept SWEET-SMELLING experiences a major change over the time span examined, since nouns designating a natural smell are modified significantly less frequently by the three synonyms at the end of the period, while the probability of those denoting an artificial smell increases significantly. This semantic development could be understood as the reflection of a changing society, whereby industrialization and mass production have led to an ever-increasing need to refer to *artificially* scented soaps and oils rather than to *naturally* fragrant flowers and aromas.

**References**


Divjak, D., & Gries, S. Th. (2008). Clusters in the mind? Converging evidence from near synonymy in Russian. *The Mental Lexicon*, 3(2), 188–213. [https://doi.org/10.1075/ml.3.2.03div](https://doi.org/10.1075/ml.3.2.03div)


On the history of permissive *get* in American English: New quantitative evidence

Florent Perek and Martin Hilpert
University of Birmingham, University of Neuchâtel

This paper investigates the diachronic development of permissive uses of the English verb *get*, as illustrated below:

(1) In the movies the prisoners always get to make one phone call.

Different views exist on how such uses emerged. Gronemeyer (1999: 30) suggests that the permissive meaning derives from causative uses (*I got him to confess*); van der Auwera et al. (2009: 283) view permissive *get* as an extension of its acquisitive meaning (*I got a present*). In this paper, we argue that permissive *get* evolved out of inchoative uses of *get* in a similar construction (e.g., *You're getting to be a big girl now*), which invited the idea of a permission, which eventually conventionalized.

Drawing on data from the COHA (Davies 2010) between 1860 and 2009, we find a substantial diachronic increase of permissive *get*, which is driven by verbs such as *see*, *be*, and *meet*:

(2) I guess we won't get to see Colonel Morrison after all. (1910s)
(3) Some day she'd get to be an editor herself. (1930s)
(4) Oh thank you and you'll get to meet our new minister then sure! (1900s)

Early examples are non-agentive and compatible with the inference of a permission. Later uses include a wider set of lexical verbs with agentive meanings, which thus encode actions that can be permitted.

We report further evidence for the diachronic relation between the permissive and inchoative uses from a characterization of the semantic domain covered by the two constructions by means of a distributional semantic model, which captures semantic similarity between verbs through their co-occurrence frequency with other words (Lenci 2008). The pairwise distributional distance scores calculated from the model are used to place verbs in a visual representation. By creating different plots for the permissive and inchoative uses at several points in time, it is possible to compare the distribution of the two constructions from a semantic perspective, and to highlight whether and when they overlap.

We find that in the late 19th century the semantic domain of the inchoative use still largely overlaps with that of the permissive use, although the latter is already wider. Over time, the overlap becomes relatively smaller, as the permissive use sees a sharp increase in type frequency and expands into increasingly diverse semantic areas. We argue that these results illustrate the common origin of these constructions as well as the later emancipation of the permissive use, marked in particular by an increase of productivity, as is typically found for newly grammaticalised constructions.

References

Determinants of Object-Verb word order in Modern English predicates: From core to edge

Javier Pérez-Guerra
University of Vigo

Whereas Verb-Object (VO) is the unmarked design of the predicate in Present-Day English clauses, in older stages of the development of English surface Object-Verb (OV) was the preferred option at least in certain syntactic contexts. The frequency of OV predicates was significant in Old and Middle English, and they were attested until 1550, when they “appear to dwindle away” (Moerenhout & van der Wurff 2005: 83). Previous empirical studies on OV (Pérez-Guerra & Martínez-Insua 2015) revealed that OV is statistically marginal in Late Modern English and a mere archaism in Present-Day English.

In light of the scarcity of data from Late Modern English onwards, this study deals with OV in Early Modern English (EModE) and, from a corpus-based perspective, explores a number of variables that may account for what was at that time a marked alternative. The data are retrieved from The Penn-Helsinki Parsed Corpus of Early Modern English (1500–1710, 1,737,853 words) and The Parsed Corpus of Early English Correspondence (c. 1410–1695, 2.2 million words). The analysis diverges from previous accounts in terms of methodology and data: first, a logistic regression analysis of the data have made it possible to determine the textual, syntactic and semantic/discoursive variables with greater explanatory power; second, the corpora on which this study is based constitute the largest electronic collection of EModE texts that has been used to investigate word order in predicates.

Findings reveal that OV has proved to be favoured in speech-based/purposed/like text types (as in Culpeper & Kytö 2010), when verbal groups contain auxiliaries, with short, so-called ‘average’-length and pronominal objects. By contrast, OV is disfavoured by the occurrence of lexical material between the auxiliary and the object. These conclusions are related more to performance or processing than to systematic syntactic principles.

The order of objects and verbs in Old English was triggered basically by morphosyntactic variables (Pintzuk 1996; Fischer & van der Wurff 2006; Moerenhout & van der Wurff 2010), in Early Middle English (Allen 2000; Kroch & Taylor 2000; Koopman 2005; Elenbaas & Kemenade 2014) also by determinants such as end-weight, quantification and definiteness of the objects, and in Late Middle English (van der Wurff 1997; Ingham 2002; Elenbaas 2007) by genre. This study reveals, on the
one hand, an upward trend of the relevance of textual and processing facts to the justification of OV versus VO and, on the other hand, a lessened role for morphosyntactic factors in EModE. In other words, this investigation provides evidence in support of the decline of OV in EModE and the account for the remaining OV examples exclusively by way of two major determinants: the ‘speechy’ nature of the texts and end-weight, that is, non-morphosyntactic factors.

References


“Why don’t you go and do something different for a year?” Tracking change in constructions expressing advice in informal conversation

Nele Põldvere¹, Rachele De Felice² and Carita Paradis¹
Lund University¹, University College London²

Recent decades have experienced an upsurge of research into various types of hearer-oriented directives, such as advice, suggestions and recommendations (e.g., Butler et al. 2009; Heritage & Sefi 1992; Paradis 2009, fc). This research, however, has been mainly concerned with advice in professional contexts, while advice in everyday conversation among family and friends has received very little attention (but see, Adolphs 2008; Shaw et al. 2015), and we have very limited insights into how expressions used to give advice have changed over time. The goal of this study is to describe the constructions that interlocutors use to give advice in conversation in Present-Day British English, and how these constructions have changed over the past 50 years.

Couper-Kuhlen and Thompson (2017) identified five major formats for suggestions and advice in conversation, illustrated here with do: (1) imperatives (do it), (2) interrogatives (why don’t you do it), and three types of declaratives: (3) you should do it, (4) you could do it and (5) I would do it. The formats are ordered on a cline of deonticity; imperatives are strongly deontic while declaratives are modalised and therefore come across as less direct. This is especially true for I would do it that construes the advice as a hypothetical action carried out by the advisor. These five format types will be adopted as the analytical model for this study.

The data for the analysis are two comparable corpora of Present-Day spoken British English, the London-Lund Corpus (LLC-1) and the London-Lund Corpus 2 (LLC-2). LLC-1 contains data from the 1950s–1980s, and LLC-2 from 2015–2018. This difference of time allows us to track changes in spoken English over more than half a century. The study focuses on casual conversation among family and friends and the sample contains approximately 500,000 words, equally distributed across LLC-1 and LLC-2. Access to the raw text files of these corpora allows us to use text processing tools such as part-of-speech taggers, parsers, and speech act taggers to extract the full range of constructions of interest. We also annotate the constructions for several formal and functional features (e.g. collocates, speaker commitment, addressee response) in the actual sequential and interactional contexts in which advice is given.

A preliminary comparison of constructions of advice in the two corpora suggests that although declarative constructions are still the most common format, there is an increase of advice expressed in more direct ways. For example, compared to LLC-1, there is a slight increase in the use of interrogative constructions in LLC-2, especially why don’t you (just). Among declaratives, there is a clear shift from less deontic to more deontic constructions, such as you should, as well as the semi-modals you need to and you have to (Leech et al. 2009). This suggests that, in everyday conversation among family and friends, there is an increased preference for constructions that make the act of giving advice explicit and that do not warrant a high degree of caution and tact.
References


On the notion of non-concord in existential interrogative *there*-clauses

*Jirina Popelikova*

Charles University

While the notion of number concord between the subject and the predicate in finite clauses belongs to one of the most basic grammatical features of Standard English, a small group of constructions exists in which the compliance to this rule is observed with a noticeably lower level of strictness than what the prescriptive tradition might indicate. One of the most frequent and frequently-quoted examples in this regard are the so-called existential *there*-clauses.

In Standard English, the verb is supposed to choose its form in accordance with the number expressed by the immediately following noun phrase. *There*, in this case, is interpreted as a special kind of a morpho-syntactic mirror, automatically reflecting and copying any relevant grammatical features carried by the post-verbal noun phrase. Cf. e.g.:

(1) In the pond there lives a grass snake, which will emerge at some random point on the surface, and eat whichever frog is closest to it. (AE7 793)

(2) In other words, there are no aesthetic objects, only physical objects, which, when observed, are capable of stimulating an aesthetic event.’ (A04 157)

The corpus evidence over the recent years, however, shows a tendency, especially, but not exclusively, in spoken communication for the verb to take a singular, not plural form, even though
the notional noun phrase is expressly referring to more than one entity. Perhaps the most startling examples of this discord include clauses with NPs containing specific numeral quantifiers, e.g.:

(3) ‘There’s two vehicles, not one,’ he said. (AT3 1591)

Features which are recognized as contributing to the frequency of non-concord include reduced verbal form ‘s, a lexical material intervening between the predicate and the notional noun phrase, such as a disjunct, length and/or complexity of the post-verbal sequence. Recognizing the role of the clitic ‘s as influential to the resulting concord, the aim of the study presented in this paper is to try and isolate the remaining factors by focusing on non-concord in syntactic structures in which the use of reduced forms is not possible, i.e. in polar questions. This allows the author to weigh the role of these other factors separately from and against the influence of the presence or absence of the reduced there’s.

The paper is based on data obtained from the original spoken component of the British National Corpus in comparison with the newly released BNC2014. The preliminary results suggest a significantly lower number of discord structures in questions than in declarative clauses, but also seem to point to an overwhelming preference in there-constructions for singular post-verbal noun phrases and present tense predicates in general.

Selected references


Dimensions of deception: Using multi-dimensional analysis to detect fake online reviews

Olumide Popoola and Jack Grieve
University of Birmingham

A hotchpotch of linguistic cues has been used in computational linguistics to distinguish fake and authentic online reviews including part-of-speech features (Ott et al. 2011), pronouns (Fornaciari & Poesio 2014), hedges (Banerjee & Chua 2016) and psycholinguistic lexical categories e.g. emotion
or cognitive words (Li et al. 2014). This domain’s focus on accuracy “no matter how” has, however, left the bi-directionality and context-dependency displayed by these linguistic features unexplored. Using a qualitative corpus approach, Popoola (2017; 2018) linked review veracity to discourse structure and salience finding that the contrastive coherence relations typically used in the review genre to downplay positive evaluation (Taboada, Carretero & Hinnell 2014; Ikatura 2013) were largely absent from deceptive reviews.

Building on this, the current research deploys a statistical corpus linguistic approach – Biber’s (1988) multidimensional analysis (MDA) – to test the hypothesis that fake online reviews are ‘genre violations’. Specifically, it asks:

(1) Are deceptive reviews a distinct language variety from authentic reviews?

(2) Do deceptive and authentic reviews vary systematically on a primary language dimension?

The study uses the Deception in Reviews corpus (DEREV; Fornaciari & Poesio 2014), a forensic collection of 6819 Amazon book reviews of 68 books written by 4811 different authors. The corpus is a mix of known fake, highly suspicious and likely authentic reviews compiled from the investigative research of renowned ‘sock puppet hunter’ Jeremy Duns [Owen 2012] and journalist David Streitfield [Streitfield 2011; Streitfield 2012] in addition to evidence from a self-confessed professional review writer [Parker 2011]. DEREV consists of 1175410 tokens in total.

Analysis was conducted on the positive (5-star) reviews with the highest and lowest truth values and a minimum text length of 200 words. All reviews were tagged using the Multidimensional Analysis Tagger (MAT; Nini 2015) which replicates the tagger used in Biber (1988) and computes relative frequencies. A preliminary analysis mapped these to the existing Biber dimensions; fake reviews scored significantly higher on Involvement (Dimension 1) and Overt Persuasion (Dimension 4) with authentic reviews more Informational (Dimension 1) and Abstract (Dimension 5) (see Figures 1 and 2). An exploratory factor analysis was conducted to build a bespoke MDA model of the dimensions underlying this data set. It found fake reviews used a more persuasive register (e.g. modals, infinitives, susasive verbs,) while authentic reviews were more informational (e.g. lexical diversity, attributive adjectives, place adverbials), suggesting variation on a telling vs. selling dimension.

Figure 1: MAT Dimension 1: Involved-Informational

Figure 2: MAT Dimension 4: Overt Persuasion
References


‘Nothing can we call our own but death’: A diachronic analysis of Sri Lankan English morphological features and move structures of obituary notices in newspapers

Mahishi Ranaweera
University of Kelaniya

Studies on Sri Lankan English newspaper genre have revealed that newspapers are a rich source of Morphological examples of Sri Lankan English (Gunesekere 1989; Gunesekera 2005; Ranaweera 2007). However, the register of obituary notices is a neglected area in studies that focus on Sri Lankan English morphology. Although the main purpose of an obituary notice is to announce the details of the funeral and the memorial service, obituary notices display distinctive morphological features expressing socio cultural practices unique to a country or an ethnic group. A diachronic study of obituary notices furthermore uncovers how society has changed over the years. This study initially attempted to find whether there are morphological features unique to Sri Lankan English in obituary notices belonging to four ethnic groups in Sri Lanka viz. Sinhalese, Tamil, Muslim and Burgher. The study, moreover, observed the move structures to determine whether there is a typical structure in obituary notices that is specific to the ethnic identity of the deceased. The study also observed whether the morphological features and move structures have changed over time reflecting the changes of the society.

In order to study the morphological and structural features of obituary notices, a corpus based approach is used. To facilitate the diachronic analysis, the corpus includes a total of 120 obituary notices from year 1977 and year 2017. The texts were selected through purposive sampling selecting 60 notices from each year. An in depth structural and linguistic analysis was made using both qualitative and quantitative methods. Each notice was first manually read to find morphological features unique to Sri Lankan English. Then the morphological features were categorized according to each ethnic group. Next the vocabulary items were electronically searched in the corpus to determine the presence and the frequency.

The moves in each notice were examined to observe emerging patterns. The moves were identified manually and were coded according to the function. A list of the move patterns in each notice was created. Frequencies of the pattern types were manually counted and were classified according to ethnic group to find similar or different move structures across ethnicities. Finally a diachronic analysis was made comparing both morphological features and the move structures.

The data reveal unique Sri Lankan English morphological features which disclose the reciprocity between obituary notices and socio cultural practices that evolve through time. Terms such as ‘family burial grounds’ and ‘interment’ have petered out during the 40 years under observation and new terms such as ‘funeral parlour’ and insertion of phone numbers appear, signaling socio cultural changes the country has undergone over four decades. Additionally, the move analysis reveals that there is a typical move pattern which is similar in notices of all ethnic groups. Broadly, the study unveils how morphological features and move patterns in Sri Lankan English obituary notices across four decades reflect the changing society.
Revisiting the ENL-ESL-EFL continuum: A multifactorial approach to grammatical aspect in spoken Englishes

Paula Rautionaho, Sandra C. Deshors and Lea Meriläinen
University of Eastern Finland, Michigan State University

At a time when the global spread of English has already started to blur the sociolinguistic divide between English varieties worldwide, corpus linguists have recently engaged in a scholarly conversation on the ‘native–foreign–second language continuum’, assessing whether native and non-native Englishes are dichotomous types of English or form a gradient continuum (see e.g. Mukherjee and Hundt 2011; Edwards and Laporte 2015; Gilquin 2015; Gries and Deshors 2015; Meriläinen and Paulasto 2017). In this context, and as scholars are currently revisiting the validity of the ‘EFL’ and ‘ESL’ labels in the continuum debate, progressive marking is a linguistic phenomenon that is likely to shed light on the extent to which the ENL–EFL–ESL distinction remains relevant today. Although recent work on the progressive in World Englishes has shown how \textit{BE} + \textit{Ving} is used in different functions and linguistic contexts across individual English varieties (e.g. Rautionaho 2014), it remains to be shown whether differences also exist across the three different types of Englishes and to what extent those differences can support the tripartite categorization of Englishes and its validity. To explore such differences, unlike traditional studies that have investigated progressive marking independently of its unmarked counterpart, we examine (i) how the grammatical contexts of both constructions systematically affect speakers’ constructional choices in ENL (American, British), ESL (Indian, Nigerian and Singaporean) and EFL (Finnish, French and Polish learner Englishes) and (ii) what light speakers’ varying constructional choices bring to the continuum debate.

Based on 4,101 (non-)progressive constructions extracted from the \textit{International Corpus of English}, the \textit{Louvain International Database of Spoken English Interlanguage} and the \textit{Santa Barbara Corpus of Spoken American English}, we start by clustering Englishes based on contextual features such as \textit{AKTIONSART}, \textit{ANIMACY}, \textit{SEMANTIC DOMAIN} (of aspect-bearing lexical verb), \textit{TENSE}, \textit{MODALITY} and \textit{VOICE} to assess the validity of the ENL–ESL–EFL classification for our data. Then, we integrate the three types of English as a predictor of potential linguistic variation directly into a logistic regression model to assess whether linguistic patterns can be reliably predicted to belong to

References


a specific type of English and explore how grammatical contexts influence speakers’ constructional choices differently across English types.

Overall, our cluster analysis supports the ENL–ESL–EFL classification as a useful theoretical framework to explore cross-variety variation: with non-progressives especially, we observe a clear-cut divide between ENLs, ESLs, and EFLs. However, the regression model ($C=0.82$) reveals that, when we dig into the specific linguistic contexts of progressive and non-progressive constructions, this classification does not uniformly transpire in the data. More specifically, in certain linguistic contexts, EFL and ESL speakers use progressives more than native speakers, in other contexts they use progressives less than natives and in yet other contexts they both differ from one another while being altogether different from ENL. These results emphasize the importance of carefully considering the grammatical contexts in which linguistic phenomena are assessed as those contexts may lead to different conclusions on the validity of the continuum. Ultimately, by exploring the continuum using more than one statistical technique, scholars could avoid potential methodological biases.

References


Genitive variation in Spoken British English – a multi-method diachronic approach

*Susan Reichelt*
Lancaster University

This paper presents a multi-method approach to nominal genitive variation in Spoken British English using diachronic data sets compiled from the original BNC1994 and the newly collected BNC2014. The genitive variable describes two competing contexts of two NPs, “one of which is syntactically subordinated to the other” (Altenberg 1982: 22) that can be defined as either
The study adds to a wide range of previous work on genitive variation between the *s*-genitive (*a farmer’s field*) and the *of*-genitive (*the field of a farmer*) that focused on the historical trajectory of the variable (Altenberg 1982), cognitive explorations (Heine 1997; Rosenbach 2003; Stefanowitsch 1998), and variable contexts (Hinrichs & Szmrecsanyi 2007; Rosenbach 2002, 2014; Röthlisberger & Schneider 2013).

The genitive system is undergoing widespread change with the *s*-genitive replacing the *of*-genitive across various contexts. Hinrichs and Szmrecsanyi (2007: 439) point out that much of the previous research, while consenting to this shift, do not provide inclusive factor analyses and that “what is most needed in this debate is a comprehensive corpus-based study determining the contribution of factors, and their relative strength, for a broad range of constraints that purportedly influence genitive choice”.

With regards to that request, we are presenting a detailed sociolinguistic account including linguistic constraints (end weight, final sibilant, and semantic relationship between possessor and possessum), as well as social constraints (social class, gender, region, age). The data consists of balanced subsets of the original BNC from 1994 (250 speakers), and the new BNC2014 (250 speakers) and offers analyses comparing both points in time (real time), as well as comparisons across speakers of different ages (apparent time). Speakers included in the corpus subsets ranged, in their years of birth, from late 1890s to early 2010s.

The research focus is on language change generally with two research questions in particular.

1. Is the above mentioned shift observable?
2. What are the factors contributing to this shift (to what extent, empirically)?

Our findings suggest that while the overall shift towards an increased use of *s*-genitive is only minimally visible between the two points in time, a shift across generations is significant and might point to a generational shift of formality registers. We find that both apparent and real time analyses are needed for a full picture of how possession is marked in spoken British English. Previous suggestions that the *s*-genitive is specifying particular possessive contexts are also confirmed. The paper further provides clear evidence that the type of possessive relation between both NPs is stratified from most typical to most abstract. In summary, this paper briefly contextualizes the current study within existing work before presenting new insights into our knowledge of a changing variable. We highlight the usefulness of multi-method approaches into issues of language change and what each approach contributes to our findings.

References


Semantic Neology: Challenges in matching corpus-based language change to real-world societal change

Antoinette Renouf
Birmingham City University

The conference theme is ‘Corpus Linguistics and Changing Society’. In this paper, ‘corpus linguistics’ is interpreted as the study of language change in a diachronic corpus of news text, and ‘changing society’ as real world events reflected in changing language in the corpus. Our focus is on the emergence of semantic neology in response to societal change.

In the 1990s, we developed the automated systems AVIATOR (Renouf 1993) and APRIL (Renouf 2010) which identify candidates for semantic neology according to the criterion of collocational change. That is to say, a statistically significant change in the collocational profile of a word is deemed to indicate a new sense. Taking collocation, together with word frequency, as the criteria was an experimental compromise, to exploit readily identifiable surface textual features and thereby streamline the identification process, while disregarding less accessible features such as distributional semantics, lexical priming, or onomasiological strategies which also play a role. The surface approach has been shown to be surprisingly successful, though it cannot differentiate between subtleties of meaning change such as use and reference; or stylistic uses such as irony or euphemism (Stockwell 2000).

Only recently has interest in the automatic identification of semantic change begun to emerge within other research groups (Boussidan & Ploux 2011; Cabre & Nazar 2012; Reutenauer 2012;
Cartier 2016). These groups are still grappling with the conceptualisation of semantic change criteria, and its operationalisation by relevant software on a diachronic corpus.

Against this background, our research aim here is to reassess our approach, with particular reference to the problem of matching software and statistics to meaning in text. The tools involved are AVIATOR and APRIL, located within the WebCorp LSE toolset. The data comprise an accumulating flow of Independent and Guardian newspapers, around 1.4 billion words of news text. These data are fed in monthly time chunks through software filters, which compare them to identify ongoing significant collocational change, into an evolving collocate bank. The output is accessed via WebCorp Linguist’s Search Engine tools.

The research method involves the selection of words deemed to reveal various change patterns, as informed by earlier work (Renouf 2012) on the life-cycle of words across time. These include words denoting events which suddenly capture the public imagination, such as natural disasters (e.g. tsunami) or scandals (e.g. flipping, i.e exploiting tax allowances on second homes), and spur sudden increases in corpus frequency sufficient to identify clearly a new sense. Other words reveal a more gradual or less evident change in the corpus. For example, the new use of bandwidth to mean ‘capacity’ takes time to establish itself; the use of silo to refer to ‘managerial isolation’ remains rare; the new sense of toxic to mean ‘risky’ is co-mingled with earlier senses; the volatility of troll leads senses to compete; while the euphemistic sense of gaunt as ‘symptomatic of HIV positivity’ is unidentifiable at surface level.

What is revealed by this study is that, while our automated sense change detector achieves considerable success given its conceptual simplicity, its weaknesses lie in depending on the changing fortunes of real-world referents, the degree of promptness and clarity with which societal change is reflected in newspaper reportage and comment.

References


The impact of substrate influence on sentential verb complementation across World Englishes

Raquel P. Romasanta
University of Vigo

Many innovative patterns of World Englishes tend to occur in the verbal complementation system, where grammar and lexis intersect (Schneider 2007: 86). Olavarría de Ersson and Shaw (2003: 138) describe it as “an all-pervading structural feature of language and thus likely to be more significant in giving a variety its character than, for example, lexis.” Similarly, Mukherjee and Hoffmann (2006: 149) refer to verbal complementation as an area “in which regional differentiation figures prominently.” Studies on this topic in World Englishes are confined to transfer-cause-motion constructions and ditransitive verbs, which involve noun phrases and prepositional phrases as complements (cf. Olavarría de Ersson & Shaw 2003; Mukherjee & Hoffmann 2006; Mukherjee & Schilk 2008; Mukherjee & Gries 2009; Schilk, Bernaisch, & Mukherjee 2012; Schilk, Mukherjee, Nam, & Mukherjee 2013; Bernaisch 2013; Nam, Mukherjee, Schilk, & Mukherjee 2013; Gries & Bernaisch 2016 for discussion); within sentential verbal complementation, only the variation between to-infinitive and -ing-clauses after verbs has received attention so far (Deshors 2015; Deshors & Gries 2016). In contrast, there is a plethora of in depth studies on the verbal complementation system in the metropolitan varieties of English, from both synchronic and diachronic perspectives. Cuyckens, D’hoedt and Szmacsanyi (2014) in particular study the complementation profile of three retrospective verbs (REMEMBER, REGRET, and DENY) in Late Modern British English. This group of verbs is characterized by a functional distribution between the to-infinitive and -ing patterns. However, the choice between that-clause and the gerundial -ing is non-categorical or probabilistic, which means that the speaker’s choice between both structures seems to be independently motivated. In this presentation, I want to extend the study of the complementation profile of one of these retrospective verbs, REGRET, to include present day British and American data as well as data from two ESL (English as a Second Language) varieties, Hong Kong English (HKE) and Nigerian English (NigE). The aims are (i) to determine whether or not
HKE and NigE show any innovative trends regarding the complementation profile of the verb REGRET, and whether these can be attributed in any way to influence of substrate languages (Cantonese, Hausa, Igbo, Yoruba, and French) and (ii) to ascertain, by means of statistical tests, which linguistic internal factors are relevant for the choice between the finite *that/zero* patterns and the non-finite *-ing* pattern in the four varieties (e.g. the complexity of the complement clause, the presence of negative markers, the voice and meaning of the verb in the main clause, and so on). A preliminary overlook at the data retrieved from *Corpus of Global Web-Based English* (GloWbE, Davies 2013, 1411 examples) shows that both HKE and NigE present a relatively higher proportion in the use of finite patterns, which might be partially explained by the fact that none of the substrate languages has a non-finite pattern that corresponds to the English *-ing*. Other potential explanations are also taken into consideration, such as the fact that finite patterns involve an increase in isomorphism and transparency, and their frequent use could be derived from SLA processes and language development contact-language situations.

References


Silly much? Tracing the spread of a new expressive marker in recent corpora

Patricia Ronan
TU Dortmund

In recent times, a new expressive marker has been developing, and the rise of this marker is increasingly manifest in new, large-scale corpora of media and online language. This structure consists of a semantically mostly negatively connotated adjective, noun or infinitive postmodified by much. Most frequently, the syntactic structure is that of an elliptical question. Examples are the following:

(1) A person could fake salary figures as well. Paranoid much UK??? (NOW Corpus, 10-06-10 US. Marketplace)
(2) You need to go as far as space to cover your countrys [sic] lies?
   Desperate much? (NOW Corpus, 12-08-13 HK. Wall Street Journal)
(3) Kate: Why don’t you just shut up? Sami: Overreact much, kate [sic]? (SOAP Corpus 2002 DAYS)

Knowledge of and research on it is scarce so far. It is considered to be an elliptical, frequently ironic, comment or question, with first use observed for 1978 (OED, s.v. much, III.B.1.h.). In an early linguistic note, Liberman (2010) calls the structure X much. Gutzmann and Henderson (2017) investigate this structure, here called “expressive much”, in a formal semantics approach. They find that X much is colloquial, most typically used in elliptical question structures, and that it does not act as a quantifier, but instead has the pragmatic force of an expressive. As such, it is a “shunting expression”, it indicates disdain by adding a negative evaluative attitude to an utterance (Gutzmann and Henderson 2017: 4).

With awareness of this structure just emerging, a number of research desiderates are arising. While the semantics have been described, what has not been determined are the usage patterns of this structure throughout different genres, quantitative elements of distribution and collocation preferences.
This paper discusses and quantifies the distribution of \textit{X much} data and to determine possible variation in its use in different genres and varieties of English, addressing the theme of the conference by connecting linguistic changes to usage of new media and illustrating the emergence of neologism in and spread through new media during the early years of the 21st century.

So far, only few examples are visible in older corpora. Therefore corpus data is collected from the constantly updated, more than 5.4 billion word NOW Corpus (Davies 2013) and from the 1.9 billion word GloWbe Corpus (Davies 2013), as well as from the Corpus of American Soap Operas, SOAP (Davies 2011–). The collection is semi-automatic: examples of nouns, adjectives and infinitives + \textit{much} are collected by means of the search interface of the BYU corpus page and false positives are discarded. Data evaluation is both quantitative and qualitative. The paper provides an overview of uses, genres and comparative frequencies of the structure.

\textit{References}


\textbf{Reporting talk: Why are continuers absent from constructed dialog?}

\textit{Christoph Rühlemann}
Marburg University

It is widely agreed in research on constructed dialog that the ‘verbatim assumption’ is a fallacy (e.g., Wade & Clark 1993). When reporting talk in direct speech, speakers \textit{depict} that talk rather than replicate it, selecting some elements of the original utterance—if there was one—while deselecting others (Clark & Gerrig 1990). Clark & Gerrig’s (1990) unit of observation was the quote compared against the individual utterance it constructs. This talk will be concerned with selection and deselection as processes determining the transformation, not from individual utterance to individual quote, but from sequences of utterances to sequences of quotes. It will address two
research questions: what sequence types are present in and, respectively, absent from constructed dialog?

The research presented is based on quantitative and qualitative analyses conducted on the Narrative Corpus (NC; Rühlemann & O’Donnell 2012). The NC contains two major types of sub-corpora: spontaneous conversational storytellings as well as the turn-by-turn (non-narrative) conversational contexts sequentially preceding and following the storytellings. The storytelling sub-corpus is exhaustively annotated for occurrences of constructed dialog (N = 1,784), while constructed dialog typically does not occur in the non-narrative sub-corpus. Based on this difference, turn lengths in the non-narrative sub-corpus are compared with lengths of constructed dialog in the narrative sub-corpus. The greatest difference in length is found for 1-word units. Permutation tests carried out on the intersecting items in the two subsets show that continuers such as *mm, mhm*, and *yeah* are significantly under-used in constructed dialog.

This is a suggestive finding in that continuers are key components of ‘tellings’ sequences, which are realized in a multi-unit turn “that has places in it for others’ talk” (Sacks 1992: 526). These ‘places’ are typically occupied by the listener’s continuers acknowledging the “structural asymmetry” (Stivers 2008: 34) implied in the sequence type, and exhibiting “an understanding that an extended unit of talk is underway by another [speaker] and that it is not yet, or may not yet be (...) complete” (Schegloff 1982: 81). In response to the second part of the research question—what sequence types are absent from constructed dialog?—the quantitative analysis, then, suggests that tellings are deselected from constructed dialog.

In response to the first research question—what sequence types are present in constructed dialog?—a qualitative analysis is carried out on a random sample of storytellings with four instances of constructed dialog (four being the average number per story). This analysis suggests that the most commonly present sequence type in constructed dialog is the constructed point-counterpoint sequence. The defining structural feature of this exchange is the quick succession of constructed turns. Its interactional hallmark is the confrontation, suggesting the quoted speakers’ difficulty agreeing on a common course of action.

References


Diachronic and synchronic trends in BrE mandative constructions – A multivariate analysis

Juho Ruohonen
University of Helsinki

Though many scholars agree that an unexpected revival of the mandative subjunctive (MS) brought this once-obsolescent English mood back into productive use in the 20th century (Weekley 1952: 36–37; Barber 1964: 133; Visser 1966: 825; Leech et al. 2009: 52), no standardized methodology exists for its quantitative analysis. British English has been identified as the variety most resistant to the resurgence of the MS, with high reported rates of a periphrastic should-construction (Quirk et al. 1985: 1013f). The indicative has been described as another characteristically British option (Algeo 2006: 263). Modally harmonic periphrases with auxiliaries other than should can also be found (Huddleston and Pullum 2002: 997–998). Owing to methodological diversity, reports differ on the situation between the MS and the should-variant in BrE (Övergaard 1995: 52; Crawford 2009: 262; Serpollet 2001: 541), and many studies ignore the indicative altogether. The present analysis seeks to ascertain the distribution of the three variants, as well as its diachronic trend, in contemporary BrE. Secondly, it tries to identify which intra- or extralinguistic variables are statistically significant predictors of variant choice. Little work has hitherto been dedicated to the latter question, with prior analyses largely limited to the lexical item governing the that-clause.

Answers are sought through rigorous variationist methodology and statistical multivariate analysis. The corpus consists of two 9.5-million-word subcorpora of identical design, representing British news from around 1990 (supplied by the BNC) and post-2015 (compiled by the author). The analysis first empirically identifies the pool of over 50 most frequent triggers, i.e., subjunctive-licensing expressions, that collectively account for >90% of all MS’s. Then all inflectional forms of each trigger are searched for mandative that-clauses in the corpus. Relevant observations are coded for the observed variant and explanatory variable values. The subsequent analysis is performed using traditional bivariate crosstabulations combined with multivariate Poisson and multinomial regression (Agresti 2015: 236–243; 203–207).

The dataset consists of 3800 observations. Bivariate cross-classification of diachrony and the dependent variable shows a rise in indicatives from 30% to 40% and a decline of should from 29% to 16%, with the rates of other periphrases and the MS essentially unchanged. This overall diachronic change is highly statistically significant, and the multivariate analysis corroborates it. The final quaternary model was chosen through likelihood-ratio tests between nested models (Agresti 2015: 207), comparisons of predictive power in leave-one-out cross-validation (Stone 1974), comparisons of McFadden’s $\chi^2$ (McFadden 1979: 306–307), and comparisons of different models’ concordance indexes in binary predictions (Agresti 2007: 143–144). The most significant predictor in the model is trigger group. The triggers seem to form a number of fairly homogeneously behaving clusters, corresponding approximately to shared semanto-syntactic characteristics. Monosemous verbs (demand, order) favor the subjunctive. Extraposé-subject adjectives (it is important that…) favor the indicative. Predicative adjectives (he is determined that…) favor periphrases, while nouns and polysemous verbs (requirement, suggest) display the most even distribution. Other highly significant predictors include matrix tense, sub-clause polarity,
sub-clause voice, and *that*-omission. Surprisingly, syntactic contexts where the subjunctive-indicative distinction is lost do not favor periphrases.

References


Data-driven approaches are frequently used for automated analysis of media content (Grimmer and Stewart 2013, Wüest et al. 2014) to detect trends and shifting topics in society, but only rarely in historical linguistics (Hilpert and Gries 2016). Our study shows that societal and linguistic changes can be detected equally well using historical corpora. In line with this year’s special topic of the ICAME conference, we particularly focus on using corpora to detect societal changes. We show two case studies: first, the development of medicine, based on the EMEMT and LMEMT corpora, second, the topics of poverty and industrial revolution, based on EEBO, ECCO, and COHA.

We do so by mapping the spelling variants of the period to their present-day counterparts (Baron and Rayson 2008), then POS tagging and syntactic parsing. To analyse the linguistic data, we use a rich array of state-of-the art statistical approaches to uncover change: measures of overuse, document classification with logistic regression, keyword detection, collocation analysis, topic models, Kernel density estimates, and distributional semantics.

Measures of overuse show spikes of keywords and the advent of new technological and social concepts. Document classification delivers the most salient features of the various periods in a bottom-up fashion. We show that they can be interpreted as linguistic and societal features. Besides classical keyword measures such as TFIDF we use collocation detection to discover multi-word terms and trace the career of multi-word nouns to denote new concepts and compress information (Biber 2003). Topic Models (Blei 2012) allow us to detect changing patterns of associations. We can see how scholastic thinking in medicine is giving way to empirical methods, and how poverty, industrial revolution and urbanisation are associated. Kernel density estimates (McClure 2015) are a method used in Digital Humanities to automatically draw conceptual maps of novels. We extend the approach to periods and show periods and its central topics. Distributional semantics (Sahlgren 2006) allows us detect near-synonyms and closely related terms.

As no gold standard of social history and the history of thought exists, our approach needs to stay exploratory. The aim of our presentation is to give an overview, point out strengths and weaknesses, and to show the complementary nature of the approaches when applied to our case studies. We also discuss the importance of data size and cleanliness, the temptations of distant reading and the necessity for validating the discovered patterns in close reading and distant reading in interaction. Our approach can also complement quantitative approaches to literary content and style (Burrows 2004, Mahlberg 2015).

References


**Processing of corpus-derived interference collocations in intermediate and advanced learners of English: Evidence from eye-tracking and event-related potentials**

*Marco Schilk*
Stiftung Universität Hildesheim

The production of L1 influenced interference collocations by learners of English is one of the areas of interest in the field of Contrastive Interlanguage Analysis (CIA). Frequently investigated types of L1 interference are comparative overuse and underuse of specific collocations and multi-word units when compared with native-speaker data (see e.g. Granger 1998; De Cock 2000), as well as the interference-based misuse of collocational patterns (see e.g. Nesselhauf 2005; Gilquin & Shortall 2007).

While there is also a growing interest in the psycholinguistic study of the processing side of collocations for both native speakers and EFL-learners (Ellis et al. 2009; Ellis & Frey 2009), relatively little research exists at the interface of non-native speakers’ production and processing with regard to interference phenomena (an exception being Siyanova & Schmitt 2008).

Since the collocational range and preference of lexical items varies between different languages, and EFL-learners produce interference collocations as well as native-like collocations, it seems reasonable to hypothesize that these differences are also reflected on the processing side. The present paper explores differences in language processing between different groups of learners in relation to native-like collocations, interference-based collocations and non-collocations/lexical
repulsions. Methodologically, this study is a combination of corpus-based and experimental methods.

We use the German components of the International Corpus of Learner English (ICLE-Ger) as well as the Louvain International Database of Spoken English Interlanguage (LINDSEI-GER) to identify a set of interference-based learner collocations (V+N and Adj.+N), i.e. pairs that would collocate in the learners’ L1 but do not in the L2. Additionally, we use the British National Corpus to create a comparable set of statistically significant native-speaker collocations as well as a set of lexical repulsions (Renouf & Bannerjee 2007).

These corpus-derived sets are used as input stimuli for the experimental setup. Here we use a combined eye-tracking/event-related potential (ERP) approach to test for differences between groups of learners on the levels of eye-movements and ERP-components (N1, N300, N400) that have been shown to be sensitive to initial attention and semantic mismatch (Friederici et al. 1993). Participants are 40 intermediate/advanced German learners of English (20 first-year BA students and 20 first-year MA students).

Results show that:

a) Both learner groups experience a processing advantage from significant English collocations.

b) While both groups display similar total fixation times for the left parts of collocations, our data indicate a difference in reading styles with BA students focusing longer on the relevant areas of interest initially, while the MA students fixate these areas more briefly but revisit them in the case of the interference and incongruous conditions.

c) The group of MA students seems to be more aware of potential interference which is reflected by a very early increase of initial attention (N1) and a somewhat delayed N400 component in the case of the transfer collocations.

Abstracting from the specific experiment design, result c) in particular might be an indication of semantic processing of interference collocations displaying similarities to syntactic processing where it has also been shown that an earlier component (ELAN) is frequently followed by a later “repair component” (P600) (Friederici & Klotz 2003).

References


---

**The use of corpora in EAP textbooks: An analysis of the treatment of academic verbs**

*Natassia Schutz*

Université catholique de Louvain

Despite the positive feedback corpus-based teaching methods (e.g. Data Driven Learning, Johns 1991) have received in the context of EAP (English for Academic Purposes) teaching, these methods had still not “crossed over into mainstream practice or been taken up by major publishers” some ten years ago (Boulton 2008: 38). Harwood (2005), for example, showed that many EAP textbooks proved unsatisfactory and suggested that publishers make use of corpora and research findings to improve teaching material. Similarly, Boulton (2008: 40) suggested that publishers integrate corpus-based activities in their textbooks.

The aim of the present paper is to investigate the use of corpora in relatively recent EAP textbooks by analyzing the treatment of academic verbs. More specifically, I examine how EAP textbooks use corpora to (1) describe the use and phraseology of academic verbs and (2) create exercises on verb use. To do so, this study focuses on three corpus-based textbooks (i.e. textbooks explicitly based on a particular corpus and a corpus-based academic vocabulary list): *Academic Vocabulary in Use* (McCarthy & O’Dell 2008), *Focus on Vocabulary 2: Mastering the Academic Word List* (Schmitt & Schmitt 2011) and *Inside Reading 4 (advanced): The Academic Word List in Context* (Richmond 2012). For the purpose of this study, I examined 4 units in each textbook. This resulted in the analysis of 12 units and c. 500 exercise items. The analysis was carried out so as to create a database describing (1) the focus of the section/exercise (e.g. semantics and/or phraseology), (2) the task (e.g. read and observe, or fill in the gap) and (3), for the exercises, the type of item the students have to retrieve (e.g. a phrase, a collocation or a derivation).
The results reveal that, while there are some good practices in the textbooks under focus, there is also some room for improvement as regards the use of corpora and research findings. For example, the textbooks mainly use corpora to provide relevant examples and illustrations. There are no real corpus-based exercises aiming to help learners become independent learners and explore potential cross-disciplinary variation. As regards phraseology, little attention seems to be given to lexical bundles. These multi-word units have however been shown to be systematically functional in academic prose (Biber and Barbieri 2006) and important for academic proficiency (Hyland 2008). This presentation will end with a number of suggestions for improving the treatment of academic verbs in EAP textbooks.

References


Constructional change in British English concessives: Form, function and frequency

Ole Schützler
University of Bamberg

This paper investigates variation and change in concessives involving the subordinators although, though and even though in British English in the period 1931–2006, using four corpora from the extended Brown family (c.f. Baker 2009). Concessives are treated as constructions (Goldberg 2003; Trousdale 2012) and the focus of this study is therefore on combinations of functional (semantic) and formal (syntactic) parameters rather than isolated features (e.g. clause structure and syntactic position).
In this paper, the functional envelope of variation of concessives corresponds to the three semantic types discussed by Sweetser (1990; cf. Crevels 2000; Hilpert 2013), here called anticausal, epistemic and dialogic concessives, which describe the semantico-pragmatic relationship between propositions in the construction. As to formal variation, subordinate clauses can be in initial, medial or final position relative to the matrix clause, and the clause complementing the subordinator can be finite, nonfinite or a case of though-inversion (cf. Culicover 1976). Based on these functional and formal parameters, ‘constructional spaces’ are set up for each conjunction in each time period. These are essentially contingency tables aiming to quantitatively assess certain properties of constructions. Most analyses in this paper are conducted on the basis of such constructional spaces. In particular, four variables and their potential interrelatedness are explored:

(i) frequency of occurrence,
(ii) semantics,
(iii) constructional specialisation, i.e. the degree of limitation to only few of the theoretically possible (formally and functionally defined) variants in a constructional space,
(iv) constructional divergence and convergence, i.e. the tendency of constructions involving different conjunctions to become more similar or dissimilar over time, and
(v) constructional focusing, i.e. the formal differentiation of semantic types.

Apart from a short discussion of the relevant parameters of variation, the paper will explain the rationale behind constructional spaces, with a focus on their usefulness for diagnostics such as specialisation, divergence/convergence and focusing. Furthermore, the following findings will be discussed for the period under investigation:

- *Although* increases in frequency, *though* decreases, and the least frequent marker, *even though*, displays a slight increase.
- *Although* and *though* become more strongly associated with dialogic semantics; no such trend is found for *even though*.
- *Although* and *though* specialise, i.e. they lose functional and formal flexibility over time; no such tendency emerges for *even though*.
- There is constructional divergence, or diachronic dissimilation, between the three markers; this, however, reflects developments of the more frequent conjunctions *although* and *though*.
- *Although* and *though* undergo constructional focusing in that the most frequent semantic types (anticausal and dialogic concessives) become formally more distinct over time, mostly due to developments of the dominant dialogic type.

Thus, using the example of English concessives, this paper not only proposes and tests specific, quantifiable subprocesses in constructional variation and change, but highlights their correlation with (and possible dependence on) frequency. It is particularly the more frequent markers and the more frequent semantic types that are affected. Those findings and the new diagnostics based on constructional spaces can inform future quantitative approaches to constructional variation and change.

**References**

Contrastive negation: Mapping near-synonymy in a schematic construction family

Olli O. Silvennoinen
University of Helsinki

Contrastive negation refers to expression that combine a negative and an affirmative part that are construed as alternatives. In English, the main ways of expressing contrastive negation are the following (the notion ‘expanded’, taken from McCawley 1991, refers to constructions that consist of two full clauses):

(1) [not X but Y]
    Not stirred but shaken.

(2) [not X, Y]
    Not stirred, shaken.

(3) Expanded Negative-Affirmative
    The drink is not stirred – it is shaken.

(4) [X not Y]
    Shaken, not stirred.

(5) [X and not Y]
    Shaken and not stirred.

(6) Expanded Affirmative-Negative
    The drink is shaken. It is not stirred.

Previous research (Gates Jr. & Seright 1967; McCawley 1991; Toosarvandani 2013) has focused on only a subset of these constructions, using mostly introspective data. Reasons for selecting one construction over the others have been left unexplored. To fill this gap, this paper considers the constructional variation of contrastive negation from the perspective of quantitative corpus...
linguistics. The research questions are: (i) what factors are associated with which construction, and 
(ii) how different or similar the different constructions are in terms of their usage.

The meaning of contrastive negation is highly schematic, and the differences between the 
constructions are hard to pin down. For this reason, this study utilises multiple correspondence 
analysis (MCA) to gauge and visualise the differences in a multidimensional space (Greenacre 
2017; Glynn 2014). The data comes from the broadsheet newspaper component of the British 
National Corpus (BNC) and it comprises 1,831 tokens of contrastive negation that have been 
annotated for a number of variables, such as the type of denial (e.g. metalinguistic negation, see 
Horn 1985; Geurts 1998), the negator used, and the weight of the contrasted elements.

The quantitative analysis shows that the most important factors behind the constructional variation 
of contrastive negation include constructional semantics, information structure and the weight of the 
contrasted elements. The variation can be explained by three dimensions. The analysis shows that 
there are clear differences among the constructions. The \([not \ X \ but \ Y]\) construction is frequently 
used when the contrast is additive (\(not\ only\ X \ but \ Y\)) while the \([not\ X, \ Y]\) construction prefers uses in 
in which there is a restrictive element in the affirmative part (\(not\ X, \ just\ Y\)). The \([X\ not\ Y]\) and \([X\ and\ not\ Y]\) constructions pattern quite similarly, which is to be expected, given their formal similarity. 
The results of the study suggest that despite the schematicity and semantic similarity of the 
constructions, there is a great deal of functional specialisation, in accordance with the Principle of 
No Synonymy (Goldberg 1995).

References


J. A. Robinson (Eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and 

Chicago: University of Chicago Press.

& Francis.


Toosarvandani, M. (2013). Corrective \(but\) coordinates clauses not always but sometimes. *Natural 
Variation and change in the past perfect in British spoken English: An exploration of Spoken Demographic BNC1994 and BNC2014

Nick Smith
University of Leicester

This paper considers variation and change in the past perfect in recent conversational British English (BrE), as in They had literally poured it off (BNC1994 KB7). The research is motivated by recent studies suggesting a decline of the past perfect in two corpora of spoken BrE. Bowie et al. (2013) report a significant decrease of the past perfect in the DCPSE, a multi-register corpus spanning the 1960s to the 1990s. Although the authors do not provide frequency information at the register or speaker level, they usefully demonstrate the decline by two frequency measures (per million words and proportion of past contexts). Smith & Waters (under review), using a socially-balanced corpus of guest speakers on a British radio chat show, find a decline of past perfect between the early 1980s and early 2000s, based on percentage of words, finite verbs, and past tense constructions, respectively. Social group analysis indicated that by the early 2000s older speakers, and those in higher educational groups (who are disproportionately represented in the show), were a significant factor in ‘propping up’ the past perfect. The present paper seeks to explore more recent development in the less situationally specialized domain of everyday conversation and across a more diverse array of speakers, represented by the Demographic Spoken components of the British National Corpus 1994 (Burnard 2007) and 2014 (Love et al. 2017) (henceforth BNC1994DS and BNC2014DS). These resources both profit from rich sociolinguistic metadata and search interfaces (e.g. BNCweb/CQPweb).

Our research questions are:

(1) Has the past perfect declined in everyday conversational BrE?
(2) What social group variation can be found in the BNC in each period, both at an overall frequency level and in respect of different linguistic environments (e.g. past perfect progressive, inverted word order, temporal clauses)?

At the time of writing, a number of methodological complexities prevent a direct frequency comparison of BNC1994DS and BNC2014DS. Issues include differences in transcription practice and tagging processes between the components (e.g. in 1994, greater use of punctuation in transcripts, and the use of more refined POS-tagging). Thus our main focus will be on RQ2. We outline a query strategy in CQPweb for past perfect that has high recall and precision, including cases containing interpolating adverbials and simple NPs (e.g. Oh had he actually collapsed...? BNC1994DS:KP1). We also explain how we apply two frequency measures, count per million words and proportion of past tense constructions. In considering the ‘envelope of variation’, we focus particularly on contexts where the simple past may be replacing the past perfect, e.g. reported speech and thought, conditional clauses, and temporal clauses with after/when/before (cf. Declerck 2006; Bowie et al. 2013). Results indicate that the past perfect in 2014 is significantly more prevalent among older and more educated speakers (as with the 2000s radio data mentioned above), but lacks clear sociolinguistic differentiation in 1994. Other aspects of sociolinguistic variation, and indicators of real-time change, will be examined in the talk.
Visual inference in corpus data analysis: Dot plots of effects sizes with confidence intervals

Lukas Sönning
University of Bamberg

Current efforts to improve statistical practice in the quantitative sciences emphasize the role of effect sizes (ES) and confidence intervals (CI) (e.g. Wilkinson et al. 1999; Cumming 2012; Kline 2013). In combination with visual representations, these ‘new statistics’ (Cumming 2012; Cumming & Calin-Jageman 2016) facilitate understanding and communication of research findings. This paper aims to demonstrate how quantitative corpus studies can profit from this approach to data analysis.

Statistical inference assists in making generalizations based on sample data and has become commonplace in corpus-based research – many studies rely on null hypothesis significance testing (NHST) and p-values to learn from data. As has been convincingly argued in the literature, however, an overreliance on these techniques has a debilitating effect on the growth of scientific knowledge and understanding (e.g. Schmidt 1996; Kline 2013). Apart from its confusing mechanics, NHST offers little information about the focal comparison, conflating ES and sample size into a single score, which – due to the semantics of the adjective ‘significant’ – is susceptible to overinterpretation. Reform efforts therefore propagate a methodological shift towards effect sizes and confidence intervals, which offer several advantages over NHST: The former quantify the strength of an association between variables, the latter not only imply the result of a significance test but also reveal the precision of the estimate of interest. In applied corpus research, two common outcomes are frequencies and proportions. Among the association measures proposed for such quantities are absolute differences (i.e. differences between relative frequencies such as percentages, proportions, pmw, etc.) and relative differences (i.e. ratios in plain, log, or odds form). The literature on uncertainty intervals for these measures is vast (e.g. Altman et al. 2000; Grissom...
and proposed construction methods include simple procedures requiring only a hand-held calculator. Their application in applied research thus requires a minimum of software literacy.

An effective tool for visualizing effect sizes and their CIs is the dot plot (Cleveland 1984), which offers several advantages over alternative chart types (Sönning 2016). In this paper, I will argue that their combination with the ‘new statistics’ bears potential for a number of recurrent quantitative settings in corpus-based work, such as the comparison of frequency profiles, keyword analyses, and (subgroup) comparisons across multiple features or lexical items. This approach will be illustrated in the context of word-level and text-level analyses (Lijffijt et al. 2016) of binary and frequency outcomes. By means of a reanalysis of data sets from the literature, this paper will also highlight how interpretations may change compared to analyses relying solely on $p$-values or test statistics (e.g. the likelihood-ratio or the chi-squared test).

References


He, he or she, or they?: Epicene pronouns in EFL academic writing

Charlotte Stormbom
Åbo Akademi University

Since the 1970s, gender-neutral language has become a much-debated topic in many language communities. Although English no longer has grammatical gender, a problem still remains in the choice of so-called ‘epicene pronouns’, i.e. singular personal pronouns that are used in reference to antecedents of irrelevant or unknown gender (Baron 1981). A bulk of previous research has shown that the use of epicene pronouns has undergone a change over the past forty years: The previously prescribed pronoun ‘generic he’ is no longer the most common epicene pronoun in English; instead, the previously proscribed pronoun ‘singular they’ has taken on this role in L1 speech and writing. However, despite the very global nature of English, the use of epicene pronouns in EFL is still largely unexplored territory.

The present paper investigates the use of epicene pronouns in advanced EFL writing. The data are extracted from the BATMAT Corpus (Lindgrén 2016) and from the Swedish component of the Varieties of English for Specific Purposes database, VESPA-SE (Larsson 2015). These corpora consist of academic texts produced by Swedish-speaking university students of English in Finland and Sweden. The purpose of the study is twofold: (1) to account for diachronic patterns in the use of epicene pronouns in EFL academic writing, and (2) to examine how syntactic and semantic features impact on the choice of these pronouns. This includes investigating different aspects of the antecedent, such as semantic plurality (see e.g. Newman 1992), and gender stereotyping (see e.g. Paterson 2014).

The preliminary results indicate a change in the use of epicene pronouns in EFL academic writing: The use of ‘generic he’ has decreased in the past decades, whereas the use of ‘singular they’ has become more frequent. However, variation within individual texts suggests that this language change is still very much in progress. The results also show that ‘singular they’ is particularly likely to occur with semantically plural antecedents (e.g. everyone), whereas ‘generic he’ and coordinations like ‘he or she’ are more common with antecedents that are both grammatically and semantically singular (e.g. the student). The study increases our knowledge of feminist language planning and its effects on advanced EFL writing.

References

Linguistic traces of L2 processing difficulties found in learner corpus data: A quest of discriminant function between L1 and L2

Masatoshi Sugiura, Daisuke Abe and Yoshito Nishimura
Nagoya University

This study aims to investigate the L1/L2 differences found in corpus data. With the measures Mean Length of T-unit (MLT) and Measure of Textual Lexical Diversity (MTLD), our discriminant function achieved 96% discriminant accuracy. The validation of the function with another dataset proved 90% accuracy. We believe that the most significant linguistic features which discriminate between L1 and L2 should represent the fundamental differences between L1 and L2 processing.

Crossley and McNamara (2009) showed that based on 10 lexical features, L1 and L2 texts could be distinguished with 79% accuracy. Sugiura, Abe, and Nishimura (2017) took both syntactic and lexical features into consideration, distinguishing between L1 and L2 essays with 95% accuracy using Mean Length of T-unit (MLT) and VocD measurements.

Our study refines the approach by simplifying the variable selection process and adding Measure of Textual Lexical Diversity (MTLD) (McCarthy & Jarvis 2010). We first selected 14 syntactic (MLT, MLS, MLC, Clause/Sentence, Verb Phrase/T, C/T, Dependent C/C, DC/T, T/S, Complex T/T, Coordinate Phrase/T, CP/C, Complex Nominals/T, CN/C) and 5 lexical (MTLD, the number of word types in the first 50 words of sample, mean number of word types of 10 random 50-word samples, mean number of word types of 10 random 50-word sequences, Mean segmental TTR) measures, all independent of text length. To calculate a discriminant function, the corpus data from the Nagoya Interlanguage Corpus of English Reborn (NICER), comprising 185 L2 essays (57,418 words, M: 313 words, SD: 102) and 36 L1 essays (35,642 words, M: 990, SD: 328) was used. All the essays in NICER were written in 60 minutes without any references. Since the covariances of the data are not equal, we conducted quadratic discriminant analysis using Mahalanobis’s distance.

To remove highly correlated measures, initial training was conducted using the caret R package to find the variable importance values (computed by ROC curve analysis) of all the 19 measures. Then, the measures with the highest variable importance were chosen, one from the syntactic features and one from the lexical.

The remaining measures were MLT and MTLD. The discriminant function obtained was: $f(x,y) = .1282x^2 + .0010xy - .0031y^2 - 2.4919x + .6545y - 24.1025$, where $x$ is MLT and $y$ is MTLD, achieving 96% accuracy. For the validation, the discriminant function was tested with another dataset from the International Corpus Network of Asian Learners of English (ICNALE), comprising 400 L1 essays (88,658 words) and 492 L2 essays (107,396 words), and these were discriminated with 90% accuracy. These results suggest that MLT and MTLD can be used to fairly reliably distinguish between L1 and L2 writing data. Psycholinguistic studies on non-native language (L2)
speakers have revealed that L2 processing can be native-like in some domains, but L2 speakers have difficulties in processing structurally complex sentences (Clahsen & Felser 2006). Our results support this claim, as the variables distinguishing L1 and L2 writings suggest a difference in the complexity of the productions.

References


Variation in the productivity of adjective comparison in Present-day English

*Tanja Säily¹, Victorina González-Díaz² & Jukka Suomela³*

University of Helsinki¹, University of Liverpool², Aalto University³

English adjective comparison has received a great deal of attention in corpus-based research, particularly in the functional competition between inflectional (*-er*) and periphrastic (*more*) strategies (e.g. Mondorf 2003; González-Díaz 2008; Matsui 2010). There is, however, a key area of competition that remains relatively unexplored, namely, the productivity of either comparative strategy. The received wisdom is that inflectional affixes are fully productive, which would suggest lack of variation within the productivity of *-er*. However, recent research using novel methodologies (Säily 2014) crucially shows sociolinguistic variation in the productivity of extremely productive derivational suffixes. Whether the same variation applies to the productivity of inflectional processes remains therefore an open question.

Our study explores intra- and extra-linguistic variation in the productivity of comparative strategies in the early-access subset of the *Spoken BNC2014* (Love et al. 2017). To provide our study with a diachronic dimension, we compare the corpus with data extracted from the original *British National Corpus* (BNC1994). The intra-linguistic factors considered include syntactic position, the presence of premodifiers, complements and a second term of comparison, and the length of the adjective. The extra-linguistic determinants focus on gender, age, socio-economic status, conversational setting and roles of the interlocutors. Rather than limiting ourselves to the relatively small number of adjective types in which both inflectional and periphrastic comparison can occur, we take a holistic approach and consider the entire range of types within each strategy using the methodology
developed by Säily and Suomela (2009) and Suomela (2016). The data analysis shows that while some intra-linguistic factors seem to have an impact on the productivity of the periphrastic comparative strategy, the overall picture is that of stability in its productivity over time and across social categories. By contrast, the inflectional strategy appears to have become significantly more productive in the recent history of British English, and some of this change is clearly influenced by extra-linguistic factors such as gender and social class (Säily et al. forthcoming).

Our research constitutes a timely contribution to current knowledge of adjective comparison and morphological theory-building. It not only provides more information on the factors shaping the growth and development of the English comparative system, but also deals with much-debated issues concerning analytic vs. synthetic trends in the history of English (e.g. Szmrecsanyi 2012). Past empirical work on analyticity and syntheticity has often excluded derivational morphology, partly because of the strict compartmentalisation of inflectional and derivational morphology. The idea that there is a derivation-to-inflection cline rather than a sharp divide between the two has of course been expressed in previous literature (e.g. Brinton and Traugott 2005; Bauer 2004; Gaeta 2007). However, more empirical evidence is still needed to support this hypothesis. Our results show variation and change in the productivity of inflectional comparison (as an example of inflectional morphology) similar to that previously observed in derivational morphology (e.g. Säily 2014), which in turn provides further support for the ‘cline’ view and therefore suggests, in line with Danchev (1992), that both derivation and inflection contribute to syntheticity.

References


---

**Measuring the variational homogeneity of English as a world language: Probabilistic indigenization effects in four syntactic alternations**

*Iván Tamaredo¹, Benedikt Szmarcsanyi², Jason Grafmiller³, Benedikt Heller² and Melanie Röthlisberger²*

University of Santiago de Compostela¹, KU Leuven², University of Birmingham³

Szmrecsanyi, Grafmiller, Heller, and Röthlisberger (2016: 133) define probabilistic indigenization as the process “whereby stochastic patterns of internal linguistic variation are reshaped by shifting usage frequencies in speakers of post-colonial varieties”. Put differently, probabilistic indigenization is about the extent to which probabilistic constraints come to shape variation patterns in different ways. The outcome of such indigenization will manifest in more heterogeneity, and a concomitant loss of homogeneity in the factors governing syntactic variation within different national varieties of English. The present study extends our previous knowledge of probabilistic grammars’ homogeneity (see, e.g., Bernaisch, Gries, & Mukherjee 2014) by sketching a corpus-based variationist method for calculating the similarity between varieties: what counts is not if and/or how often people use particular constructions, but how – that is, subject to which probabilistic constraints – they choose between “alternate ways of saying ‘the same’ thing” (Labov 1972: 188). As a case study, we discuss similarity patterns between three varieties of English around the world (British English, Indian English, and Singapore English), fueled by a variationist analysis of four grammatical alternations:

1. The dative alternation (see, e.g., Bresnan & Hay 2008) \((N=3,012)\)
   a. *I’d given Heidi my T-Shirt* (the ditransitive dative variant)
   b. *And I’d given the key to Helen* (the prepositional dative variant)
The genitive alternation (see, e.g., Rosenbach 2014) \( (N=3,108) \)

a. the country’s economic crisis (the s-genitive)
b. the economic growth of the country (the of-genitive)

The particle placement alternation (see, e.g., Gries 2003) \( (N=2,480) \)

a. you can just cut the tops off (verb-object-particle order)
b. cut off the flowers (verb-particle-object order)

The subject omission alternation (see, e.g., Torres Cacoullos & Travis 2014) \( (N=1,225) \)

a. I will call again this week (overt subject)
b. Ø will call again this week (null subject)

Based on materials from the International Corpus of English, we ascertain the degree of homogeneity versus heterogeneity of syntactic variables using the following procedure: relevant observations of the (a) and (b) variants above were annotated for approximately 10 probabilistic constraints including e.g. the principle of end weight (longer constituents tend to follow shorter constituents; see, e.g., Wasow & Arnold 2003) and animacy effects (animate constituents tend to occur early; see, e.g., Rosenbach 2008), and then submitted to variationist modeling. To evaluate the similarity between region-specific variation patterns, the method draws inspiration from the comparative sociolinguistics literature (e.g. Tagliamonte 2001): are the same constraints significant across varieties? Do the constraints have similar effect sizes? Is the overall ranking of constraints similar? Regression and conditional random forest analysis indicate that of the four alternations under study, the dative alternation is the most homogeneous one, followed – in increasing order of heterogeneity – by the genitive alternation, the particle placement alternation, and the subject omission alternation. Against this backdrop we will evaluate claims in the literature (Szmrecsanyi et al. 2016: 133) according to which the extent of probabilistic indigenization is proportional to the lexical specificity of the alternation under study.

References


To each their own – The role of social factors in adjective intensification, 1994–2014

*Susanne Wagner and Ulrike Stange*

Johannes Gutenberg University Mainz

Intensified adjectives – phrases like *so good* or *really nice* – are ubiquitous, especially in colloquial English (e.g. Tagliamonte 2016; Fuchs 2017). Recent studies indicate rapid changes in the intensifier system, which introduces new forms and recycles old ones at relatively short intervals (e.g. Tagliamonte 2008; D’Arcy 2015). Research has shown that the oldest, still most widely-used intensifier generally, *very*, is now retreating to more formal and written registers, while more modern *really* and – more recently – *so* are taking over in younger people’s colloquial speech (and women’s in particular; Aijmer 2017; Fuchs 2017, Fuchs & Gut 2016; Hessner & Gawlitzek 2017). Moreover, the general public seems to be subconsciously aware of these patterns and associates certain intensifiers with certain (groups of) people (cf. Beltrama & Casanto 2017), while popular culture is adequately adopting ongoing changes, visible in ‘appropriate’ usage patterns in e.g. American TV shows vis-à-vis spoken language (e.g. Tagliamonte & Roberts 2005, Quaglio 2009).

The present paper focuses on recent changes in the intensifier system in British English in real and apparent time, based on the ten most frequent adjectival collocates of *very*, *really*, and *so* serving as input for conditional inference trees and regression analyses. In almost 15,000 tokens from the spoken samples of the two versions of the *British National Corpus* (BNC1994 & BNC2014), changes in collocational preferences are traced across age, gender, and socioeconomic class. Importantly, the results indicate that one should not generalise pathways of change across adjectives, as each follows its own trajectory. *Good* is taken as example here; the paper will discuss results for all Top10 adjectives. The results show that

a) *good*, the #1 in terms of overall frequency of adjectives occurring with either *very*, *really* or *so*, already shows a very clear pattern of change in the 1994 data: age, gender and socioeconomic class all play a significant role, with age as the strongest predictor: young speakers (<24) have already adopted *really* as the most common intensifier of *good*, while everybody else (still) prefers *very*. Women from one of the inner socioeconomic classes (C2) seem to be at the vanguard of the change, displaying the highest rates of *really* in the older
groups, too, while older men show no such class distinction, preferring *very* in over 80% of all cases.

b) *good* in the 2014 data shows a direct continuation of the 1994 trends, but also displays new and maybe unexpected developments: most importantly, while age is still the most significant predictor, the split is now between the two younger age groups (<45) *vis-à-vis* the two older ones (45+). Within the younger group, class is then the determining factor for higher rates of *so*, the ‘new’ intensifier: it is the youngest speakers (in the next split) from the lowest social class (DE) who seem to introduce *so* in *good* environments, indicating that – different from the change towards *really* – this might be a change from below rather than one from above (cf. Labov 1966).

**References**


ICEweb 2 – A new way of compiling high-quality web-based components for ICE corpora

Martin Weisser
Guangdong University of Foreign Studies

Recent years, as well as ICAME conferences, have seen a renewed interest in the compilation of next-generation or new ICE sub-corpora (cf. Nelson 2017), possibly also including web-based genres or data from Outer Circle varieties (Edwards 2017). As today corpus compilation via the web has become a much more convenient method than the traditional sampling employed in creating the original ICE corpora, it thus makes sense to try and compile as much as possible of the materials for new or updated written ICE materials from online sources. Hundt et al. refer to this strategy as using the ‘‘Web for corpus building’’ (2007: 2), and, despite certain issues described there (ibid. 3) and in other publications (e.g. Schäfer & Bildhauer 2013 or Gatto 2014), this method represents a useful way of generating high-quality corpus data via visual inspection. And even though such an approach may stand in contrast with increasingly popular options for retrieving such data fully or semi-automatically using only seed terms, it is more appropriate for creating smaller-scale corpora constructed according to the principles and categories employed in the compilation of the ICE corpora.

ICEweb 2 is a tool that makes it easy for the user to create new sub-corpora based on these criteria by automatically creating the relevant data structures (which can also easily be extend to new genres), and providing assistance in constructing and running queries through a number of different search engines to create lists of suitable web page addresses for the user to inspect. Any potential bias introduced by using a single search engine only can be avoided by compiling lists produced by these different engines. Pages identified in this way can then be downloaded fully automatically, storing the original URL and other meta information, and cleaned up inside the tool prior to converting them to plain text and/or a dedicated form of XML that allows for later pragmatic annotation, similar to the one suggested in Weisser (2017) for the spoken components of ICE. In addition, ICEweb 2 also contains facilities for PoS tagging, concordancing, and n-gram analysis, including adjustable frequency norming, turning it into an all-round tool for working with new ICE data.

References


Cross-linguistic effects vs. universal learning mechanisms: A corpus-based case study on temporal expression

Valentin Werner¹, Robert Fuchs² and Sandra Götz³
University of Bamberg¹, University of Hamburg², Justus-Liebig-University Giessen³

While researchers from the domains of learner corpus and second language acquisition research widely agree that the linguistic expression of temporal relations represents a central and highly complex area, thus deserving persistent interest (see, e.g., contributions in Ayoun 2015; McManus et al. 2017), opinions diverge as to the impact of cross-linguistic effects (e.g. Leńko-Szymańska 2007; Shirai 2009) vs. universal learning mechanisms (e.g. Klein 1995; Bardovi-Harlig 2000) in this domain. We aim to contribute to the wider discussion by exploring the alternation between the Present Perfect (PP) and the Simple Past (SP) in EFL learner data (cf. Fuchs et al. 2016), reflecting typologically different first-language backgrounds. More specifically, we tackle the following research questions:

(1) The PP (and its alternation with the SP) as a challenging structure: Which rates of uptake do learners show? How and where (i.e. for which variables) do they differ from native speaker usage?
(2) SLA principles: Is the acquisition of temporal expression and the difference to native speaker usage influenced by (i) cross-linguistic effects, (ii) guided by universal principles (irrespective of the learners’ L1), or rather by an interaction of (i) and (ii)?
(3) Linguistic principles: Can the use of PP/SP be predicted by linguistic variables, such as priming effects or lexical aspect of the verb?

To this end, we will present the findings of a Contrastive Interlanguage Analysis (Granger 1996, 2015), applying a modified version of regression-based multifactorial prediction and deviation analysis (MuPDAR; Gries & Deshors 2014) to a sample of c. 25,000 time-reference forms produced by advanced learners of English. Our database are the German and Chinese components of the Louvain International Database of Spoken English Interlanguage (Gilquin et al. 2010) and the International Corpus of Learner English (Granger et al. 2009) as well as native-speaker control material from the Louvain Corpus of Native English Conversation (De Cock 2004) and the Louvain Corpus of Native English Essays, thus also allowing us to assess differences between the spoken and written mode.

Through this approach we are able (i) to disentangle cross-linguistic effects from general learner patterns and (ii) to identify the specific factor(s) where the cross-linguistic effects are salient. A fine-grained picture emerges, showing that learners approximate to native choices in some contexts (e.g. SP choice with certain verbal semantics in conjunction with definite time adverbials), but not in others (e.g. SP choice with indefinite time adverbials). Overall, it further transpires that, while error rates differ between the two typologically different learner groups, actual linguistic conditioning (i.e. the influence of individual linguistic variables on the learners’ choices) does not. This suggests that, when factors playing a part in native speaker data are taken as a baseline, cross-linguistic influence occurs in the larger picture (i.e. error rates) rather than in the details (i.e. linguistic conditioning of these errors). Our results further imply that a focus on particular linguistic
factors (such as temporal adverbials) in teaching may prompt learners to make a native-like choice, even though other factors (such as priming) need to be explored further.

References


L2 writers’ strategies of paraphrasing, quoting, and textual borrowing in linguistics assignments

Leonie Wiemeyer
University of Bremen

Intertextuality is a defining characteristic of academic writing (Hyland 2004). Novice L2 writers are often unaware of the functions of paraphrases beyond avoiding plagiarism and have been found to patchwrite, to overrely on direct quotation, and to generally lack knowledge of academic citing conventions (Davis 2013; Hirvela & Du 2013; Keck 2006; Shi 2004; Verheijen 2015). Ineffective incorporation of sources is likely to arouse suspicions of plagiarism (Crocker & Shaw 2002; Pecorari 2003). However, inappropriate source use may be a sign of developing academic literacy (Keck 2006).

Despite the fact that students are frequently assessed based on their use of intertextual strategies in academic writing (Shaw & Pecorari 2013), there is little corpus-linguistic research into the ways in which L2 writers combine intertextual strategies in acceptable ways (Keck 2015). This contribution aims to address this gap in research.

In this corpus-based study, paraphrasing, direct quotation, and textual borrowing are explored in L2 writers’ linguistics assignments from the Corpus of Academic Learner English (CALE; Callies & Zaytseva 2013). The goal of the study is to shed light on how advanced German learners of English create intertextuality in summaries of academic research articles. The research questions are as follows:

- How do L2 writers paraphrase and quote from source texts? How are instances of intertextuality documented, attributed, and embedded in reporting structures?
- How do L2 writers reuse source text material and which lexical and grammatical alterations do they make?

In order to investigate the strategies of source use employed by L2 writers, an annotation scheme was developed for the identification of intertextual links such as paraphrases and direct quotes. Instances of intertextuality were manually identified and coded by two annotators in 25 summaries of research articles written by German first-year linguistics students. Intertextual links were coded for the section of the source text from which the information stemmed and for referencing, attribution, and reporting structures. Inter-annotator agreement reached more than 80% for each category. In addition, textual borrowing of strings of 3+ words and strategic reuse of vocabulary were explored. The data were analysed quantitatively and qualitatively with a view to identifying aspects that have been mastered and others that require attention in teaching.

The study found that the L2 writers used a range of different intertextual strategies. Paraphrases were far more common than other strategies and frequently based on individual sentences. Unlike in previous studies, direct quotation was peripheral, pointing to text type-specific uses of intertextuality. Though referencing was used sparingly, the students made intertextuality explicit in intricate ways via attribution and reporting structures. There was considerable inter-learner variability with respect to textual overlap. Though some patchwriting occurred, much textual
borrowing was of source text phraseology and thus desirable. While the formal aspects of intertextuality had generally been mastered, insecurities remained regarding how to rephrase the source text and what to quote. The findings suggest that despite general tendencies, L2 writers’ strategies of creating intertextuality are individually different, establishing the need for individual pedagogical advice instead of generalised warnings of plagiarism.

References


Work-in-progress reports
“I don’t have communicate ability”: Deviations in an L2 multimodal corpus of academic English from an EMI University in China – Errors or ELF?

Yu-Hua Chen and Simon Harrison
University of Nottingham Ningbo China

The UNNC Corpus of Academic Written and Spoken English (UNNC-CAWSE) is an ongoing corpus project which aims to build a large collection of students’ L2 English samples from The University of Nottingham Ningbo China (UNNC), one of the EMI (English medium instruction) universities in China. Since 2016, the project has collected 1,887 exam scripts, 883 pieces of coursework, 142 oral interviews, 262 presentations, 7 student-led sessions of spoken English practice, 88 video clips of seminar discussion, among others, from the preliminary-year programme at UNNC. Although data processing (including transcription and annotation) and validation work such as double-checking is still a work in progress, some patterns have started to emerge from preliminary annotation of deviation forms at lexico-grammatical and pronunciation levels. This paper will report those preliminary findings arising from samples of this EMI corpus across registers, modalities, and settings.

Deviations in language forms which are different from the norm, or commonly recognised as native-speaker standards, are often labelled as ‘errors’ in learner corpus research or second language studies (e.g. Dagneaux, Denness, & Granger 1998; Nicholls 2003; Hemchua & Schmitt 2006), but many examples of such deviations identified in the UNNC-CAWSE corpus appear to share similarity with instances found in one of the most well-known ELF corpora VOICE. For example, incorrect use of part-of-speech forms (e.g. “I don’t have communicate ability” or “they will lead to the bad influence on the economic”) is one of the most common deviation types we have identified so far. Similar ‘non-codified’ examples (e.g. “do you arrived there”, “the rest are protect area”) in the VOICE corpus are, however, considered natural in English as a Lingua Franca instead of errors (Osimk-Teasdale & Dorn 2016; Pitzl, Breiteneder, & Klimpfinger 2008). As Hirschmann et al. (2007: 14) pointed out, “in learner language, a deviation might be analysed as an error; in other varieties, it might be analysed as a feature”. Furthermore, terms like ‘deviant’ and ‘natural’ reflect language ideologies, such as the ‘English only’ policy that impacts how teachers and students negotiate pronunciation in interactive seminar discussions as demonstrated in our classroom recordings. By presenting a selection of such examples extracted from the written, spoken, and multimodal components of UNNC-CAWSE, this paper will discuss the options regarding how we, as researchers and practitioners, can reconcile different views towards deviation forms and the implications for teaching, learning and assessment.

References


Distribution of synthetic diminutives across varieties of English: A corpus-based study

Alexandra Chudar

A. Wierzbicka (1985, 1992) dwells on the peculiarities of Australian English, including synthetic diminutives (primarily with -ie and -o suffix) in the analysis. As a result, she claims that the number of diminutives in English is limited to some words she calls “isolated baby forms” (e.g. doggie, birdie). Further, she distinguishes a separate class of words she calls “abbreviations” (e.g. barbie, lippie), typical primarily for Australian English. She states that these words differ from classical diminutives both in their form and function – they express “not endearment, but good humour” (Wierzbicka 1985: 169).

K. Schneider (2003) criticizes the main points of the article by Wierzbicka, speaking about analytically formed diminutives employed equally with synthetic ones and the derivational possibilities of -ie suffix that are not restricted to a bunch of words used in communication with children. As for “abbreviations”, Schneider notes that the process of their formation is similar to the process of any diminutive formation, and that they can also be used in conversation with children. Therefore they can neither be viewed as a specific class of words, nor be considered particular for Australian or any other variety of English. Nonetheless, these assumptions haven’t been proved empirically.

One point that is worth paying special attention to is the way Wierzbicka deals with the diminutives in -o. She considers all such diminutives as ones formed by suffixation, though ¼ of them are formed rather by truncation (e.g. demo for ‘demonstration’, gastro for ‘gastroenteritis’). To my mind, the truncated items deserve special attention as a separate group of diminutives and the peculiarities of their formation and functioning merit further examination.

My corpus-based study, conducted with the help of Global Web-Based Corpus of English (GloWbE, corpus.byu.edu/glowbe), as opposed to the introspective study done by Wierzbicka, has shown that the majority of synthetically formed diminutives, found in her publications (69 items), are not restricted to Australian or any other variety of English. In my study I compared ipm of Wierzbicka’s diminutives in different regions (using GloWbE subcorpora). According to maximum
ipm index, just one fifth (14 items, 20.3 % of the items under consideration) of all “Australian” diminutives presented in her papers can be considered specific to Australia (maximum ipm in Australia only). Some of the items are characteristic primarily of other varieties – American, Irish, South African, Singaporean, Philippine English (maximum ipm in other varieties only – 52.2 %); some are common in several countries (in Australian English and in 1–2 other varieties; maximum ipm in Australia and in other varieties – 10.1 %). Several diminutives are not present in Australia (5.8 %) and in GloWbE at all (11.6 %). Yet, the reasons for this distribution are not clearly defined. It’s apparent that both derivational preferences of each variety of English and discourse factors should be taken into account here. Therefore the main reasons that could explain such distribution of diminutives across varieties need further investigation.

References


Quantifying specificational it- and there-clefts

Ngum Meyuhnsi Njende & Kristin Davidse
KU Leuven

This paper focuses on specificational it-clefts (1) and there-clefts (2) in which the clefted NP, all of us (1), most of the motorways (2), is quantified. No doubt has ever been expressed about examples like (1) being clefts, but examples like (2) are not viewed as clefts in the general literature. According to Davidse (2000, Davidse & Kimps 2016), (2) is a true specificational cleft, which like (1), specifies the size of the set of values corresponding to the variable conveyed by the cleft relative clause and the gap in it, i.e. ‘x that are passable with care’. Formal indications of the cleft status of (2) are: (i) the main information focus is on the clefted NP, (ii) the zero relative marker is used for the subject (cf. Huddleston 1971: 325).

(1) Their cause must be our cause. … Because it is not just Negroes, but really it is all of us who must overcome the crippling legacy of bigotry and injustice. (WB_USNews)
(2) there’s still quite a large number of er roads blocked. … there are most of the motorways in the Cheshire and Lancashire area Ø are passable with care (WB_BrSpok)

In this paper, we bolster the case that it- and there-clefts with quantified clefted NPs are different subtypes of the same macro-construction by interrogating lexico-grammatical patterns in corpus concordances.
First, we support the case for their parallel grammatical and information structure by (i) verifying the attestation of zero subject relatives, and (ii) carrying out a corpus-based study of the discourse-familiarity of the cleft relative clause in relation to the preceding context (Kaltenböck 2004; Gentens 2016). We hypothesize that the propositional material in the cleft relative clause is either textually evoked or inferable from the preceding discourse, e.g. our cause – overcome the crippling legacy of bigotry and injustice (1), or anchored to it, as in (2) where roads blocked contrasts with motorways passable. We predict it will never be brand-new, as this does not tally with the requirement that the variable has to be pragmatically presupposed in the sense of discourse given (Lambrecht 2001). This aim will be realized by carrying out qualitative and quantitative corpus study on extractions from brnews, brspok, usnews and usspok in WordbanksOnline on it|there + be + lexical quantifiers, but without specification of relative markers so as to net zero-relatives as subject.

Second, we hypothesize that different selection restrictions apply to the quantifiers (Milsark 1977; Langacker 1991: 82–3) allowed in the two cleft types. We hypothesize that it-clefts allow universal relative quantifiers like all but bar negative quantifiers like no, while there-clefts bar the former but allow the latter. To verify this hypothesis, we will trawl the whole WordbanksOnline corpus for it- and there-clefts containing the quantifiers all, everything, each, no(ne), not any. If our hypothesis is confirmed, we will interpret it in light of the different semantics and (non-)exhaustiveness implicatures of the identifying and existential specificational matrices.

References

What can a diachronic corpus of modern British fiction tell us about changes in (British) society?

Jarle Ebeling & Signe Oksefjell Ebeling
University of Oslo

“Literature reflects society and society shapes literature.”
– Oscar Wilde

Against the backdrop of the quotation above, this work in progress report explores the potential of a diachronic corpus of British fiction to reflect changes in British society between 1900 and the present. More precisely, we seek to find out in what ways a keyword analysis of the data mirrors changes in society throughout the century. This will be the first exploration of the Corpus of British Fiction (CBF), currently under construction at the University of Oslo. At the time of writing, the CBF contains approx. 37 million words from (mainly) complete novels published between 1900 and 2018. For the purpose of this study we divide the corpus into three roughly equal parts in terms of number of decades, 1900–1939 (P1), 1940–1979 (P2) and 1980–2018 (P3).

Inspired by previous studies using similar corpus linguistic techniques (e.g. Leech & Fallon 1992; Scott 2010; McEnery & Baker 2017), we use WordSmith tools (Scott 2012) to generate lists of keywords from the three periods. For each period, we classify the keywords into broad semantic categories, such as religion, technology and leisure, which may throw light on changes in British society.

Preliminary observations show that the pre-WW2 period (P1) reflects a class society pre-occupied with religion, family relations and (correct) forms of address (faith, divine, cousin, marry, servant(s), captain, madame). P2 is dominated by keywords associated with the military and war (incl. WW2-specific words): enemy/ies, military, shells, as well as several types of weapons, e.g. machine-gun and tanks. More interesting, perhaps, are keywords associated with (the intake of) food and drink (café, whisky) and technology (telephone(d), wireless). The final four decades throw up keywords quite different from the previous two periods, namely words to do with technology/inventions, indulgence and travel, e.g. nuclear, computer, video, pub(s), shopping, weekend and airport. Keywords associated with food and drink also show up here, e.g. lunchtime, pub(s) and wine.

As pointed out by Leech and Fallon (1992: 34), “because of […] multiple meanings, it is sometimes impossible, looking at the comparative frequency lists, to judge to what extent a contrast is due to a particular meaning”. In a more qualitative part of the study we therefore focus on an ambiguous keyword – poor – in an attempt to shed light on potential (socio-cultural) reasons for its keyness in P1. As poor is a highly frequent word in the CBF overall, we first analysed a random sample of 200 concordance lines per period to establish to what extent all its uses and meanings were present in all periods (attributive, predicative and nominalized adjective with meanings reflecting poverty, pity/sympathy or quality). Initial scrutiny of the data suggests that although the poverty-related meaning of poor is more prominent in P1 than in P2 and P3, the pity/sympathy meaning is by far the most frequently attested one in all periods. In fact, further analysis suggests that the decrease in the use of poor throughout the century seems to be related to a marked drop in the attributive
sympathy/pity use. An attempt will be made to interpret this observation in the light of changes in society over the past 100 years.

References


Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking our trails… In M. Bondi & M. Scott (Eds.), *Keyness in text*. Amsterdam: Benjamins, 43–57.


**Beyond the boilerplate: Analysing staff responses to patient comments on the website NHS Choices**

*Craig Evans*

Lancaster University

The personalisation of the means and manner of communication between individuals and organisations is a techno-social phenomenon of the internet age. Yet there is little evidence of this in online communication between patients and the British National Health Service. In an 11-million-word corpus of staff responses to patient comments from the review website NHS Choices, impersonal stock responses, or texts with impersonal boilerplate elements, constitute the majority of the corpus. However, unique texts (i.e. not stock or part-stock responses) do occur in the corpus, which provides evidence of how staff use language to communicate with patients when they are not reproducing formulaic messages represented by boilerplate text.

Building on previous corpus-based research on online health communication (Brookes and Baker 2017; Harvey 2013; Hunt and Harvey 2015), this presentation reports on a keyword analysis of these ‘unique’ texts, with the aim of answering the following research questions: What do staff say to patients when producing texts that are not mediated by boilerplates; how do they use language; and what does this reveal about how staff view patients?

The data used in this analysis is a subcorpus of ‘unique’ texts filtered from the staff responses corpus. Here, ‘unique’ refers to texts that do not match other texts in the corpus by more than 30%, the threshold above which plagiarism is judged to have occurred (Olsson 2004). These texts have been filtered using the ‘duplicate content’ function in WordSmith version 7 (Scott 2017), resulting in a subcorpus of approximately 11,000 texts and 1.4 million words. A reference corpus of stock and part-stock responses was used to highlight keywords that distinguish ‘unique’ staff responses from those that tend to be duplicated.
In this analysis, keywords are grouped thematically and these themes used to direct a closer collocation analysis of individual words of interest, as well as analyses of sample concordances and texts. Preliminary findings of this work-in-progress include the narrative-indicating keyword ‘then’, which indicates how staff are freer to tell a story when not constrained by stock responses. One of the top keywords is ‘patients’, suggesting a greater inclination to talk about healthcare users in general when staff write individual responses. The collocate ‘who’ highlights the frequent construction ‘patients who’, where analysis of the subsequent clauses reveals the variety of ways patient identity is constructed in staff responses. Overall, unique individually written responses to online patient comments provide insights about NHS staff perspectives that were inaccessible in the pre-internet age.

References


**“The nature of the essays”: The colligational pattern ‘the N of the N’ in L1 and L2 academic English**

*Hilde Hasselgård*

University of Oslo

This study explores the use of complex noun phrases in advanced, Norwegian-produced, learner English through the lens of the colligational framework ‘the N1 of the N2’, as in *the nature of the essays* (Renouf & Sinclair 1991). Noun-modifying prepositional phrases are a complexity feature typical of academic written English (Biber & Gray 2016). While phrasal complexity may represent a challenge to L2 users, it is nonetheless a key feature in a register the students should learn to master.

The first step of the investigation concerns the frequency of the pattern in L1 and L2 novice academic English in terms of the number of types and tokens as well as its distribution across texts. The Norwegian learners are expected to use this pattern less frequently than their native peers. There are three potential reasons for this: (i) Noun phrase complexity is generally a challenge for L2 writers (Parkinson & Musgrave 2014). (ii) A cross-linguistic study (Hasselgård 2016) showed that
the English colligation ‘the N1 of the N2’ corresponds to several patterns in Norwegian besides ‘N1(def.) + prep. + N2(def.)’, particularly compound nouns and the s-genitive. (iii) Norwegian has no obvious counterpart to the preposition of, whose main role is to combine “with preceding nouns to produce elaborations of the nominal group” (Sinclair 1991: 83). Learners may thus prefer patterns that are more familiar from their L1.

Secondly, I will examine the nouns involved in the ‘N1 of the N2’ construction, focusing on the most recurrent N1s and their meaning relationship with the N2, e.g. part of N2 (the start of the novel), possession of N2 (the voice of the poet), nominalization + N2 (the depiction of the city). Hasselgård (2016) found that both individual N1 lexemes and meaning relations between the N1 and the N2 had an impact on the preferred type of Norwegian counterpart of the colligation. Such features are therefore also expected to affect the Norwegian learners’ use of the pattern ‘the N1 of the N2’.

The study is based on the Norwegian component of the VESPA corpus (Varieties of English for Specific Purposes dAtabase) and the BAWE corpus (British Academic Written English). The selected texts are from the literature discipline. This part of VESPA is still under construction, and the final analysis will be based on a larger sample than what is currently available. A preliminary analysis suggests that the pattern is more common in L1 than in L2, as hypothesized. The most frequent N1 in both corpora is end (the end of the N2). Other N1s that indicate a ‘specialized part of the N2’ (Sinclair 1991: 88), e.g. beginning, end, seem to be used similarly in both corpora. However, there seem to be more recurrent abstract and nominalized N1s in BAWE than in VESPA, e.g. importance, nature, use, but this needs to be confirmed when the two corpora are more comparable in size.

References


Corpora

BAWE (British Academic Written English Corpus): http://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/

Building an infrastructure for linguistic big data research: The case of Reddit Comment Corpus

Aatu Liimatta
University of Helsinki

For big data approaches in linguistics, when the size of the corpus climbs to millions of texts, the traditional way of storing the data in text files and Excel sheets will not be enough. The human-readability of such files is an advantage, but searching through every text file for every linguistic query is slow and unnecessary. The bigger the dataset, the bigger this problem becomes. Commonly used concordance software such as AntConc (Anthony 2017) also have their limitations when dealing with extremely large datasets, and may not be very useful for statistical big data research.

One dataset of these proportions is the Reddit Comment Corpus, a full dump of every comment posted on the social media website Reddit, collected and updated monthly by Jason Baumgartner and freely available through his open big data website (http://pushshift.io). Reddit is the third-largest English-speaking social media website (after Facebook and Twitter), made up of thousands of user-started and run “subreddits”, subforums which concentrate on a huge range of different topics. Consequently, Reddit provides interesting opportunities for linguistic inquiries of various kinds (see e.g. McEwan 2016; Stewart & Eisenstein 2017; among others). However, the comments in Reddit Corpus Comment contain no linguistic annotation such as part-of-speech tagging. Moreover, they are formatted as JSON objects which include various kinds of metadata (e.g. author, posting time, subreddit), and are thus difficult to search using traditional means.

In this work-in-progress report, I will describe my solutions for converting the multi-terabyte Reddit Comment Corpus into a format more useful for linguistic studies and queries and, in particular, big data driven linguistic research.

I will provide a rationale for choosing to store the data in a MongoDB document database (http://www.mongodb.com) instead of a more commonly used relational database, exploring potential benefits this database choice has for a technically-minded linguist. Importantly, while a relational database may be a good choice when the structure of the data is fully known in advance (see e.g. Davies 2009), a document database facilitates exploration and trying out different approaches. Moreover, from a technical standpoint, it is “easier to get the data in than out” of a relational database (Jacobs 2009: 4). For example, a document database is considerably faster for highly order-dependent queries (Jacobs 2009: 5), which are necessary for many big data studies.

I will also describe the processes and tools I use for language recognition, tokenization, and part-of-speech tagging, and the data structures I use to include this linguistic metadata in the database. Furthermore, I will show examples of the results of the kinds of linguistic queries which can be made using this infrastructure.

The ideas presented are by no means limited to the Reddit Comment Corpus or the specific implementation details presented, but provide one possible model which can be adapted and expanded to be used for any kind of massive linguistic dataset using almost any modern programming language and set of natural language processing tools.
Non-standard functions of like in spoken British English: A diachronic view

Veronika Raušová and Gabriela Brůhová
Charles University

The present paper focuses on non-standard functions of like in spoken discourse. Like has received wide attention from scholars over the past 30 years and with the recent publication of Alexandra D’Arcy’s exhaustive work “Discourse-Pragmatic Variation in Context: Eight hundred years of LIKE”, the subject appeared to be depleted. Nevertheless, non-standard like is a phenomenon that is in motion. The launch of the new Spoken British National Corpus 2014 presents a chance to compare the current use of non-standard like (data collected in 2012–2016) with data from the original Spoken BNC1994 (data collected in 1985–1993). Their comparable architecture renders them suitable for diachronic research.¹

The aim of the paper is to compare the frequency of use of the following non-standard functions in the two corpora: hedge², focus marker, discourse marker, and a quotative marker. Hedging like provides speakers with the ability to mitigate the force of their propositions, to make them “fuzzier“ (Lakoff 1972: 458), like as a focus marker helps the speakers mark new or important information in a proposition (Underhill 1988: 238), and a discourse marker may serve as a filler, discourse link, etc. (Andersen 2001; D’Arcy 2005). Finally, the quotative marker allows speakers to introduce direct speech, thoughts, or enactments (Tagliamonte and D’Arcy 2004; Fox and Robles 2010).

Comparing the relative frequency of both standard and non-standard like in BNC2014 (13,781.87 instances per million words) and the BNC1994 (3,668.45 instances per million words) the results clearly imply a steep rise in use, most likely caused by the spread of non-standard like. The paper examines two hundred examples of non-standard like (i.e. one hundred examples from each corpus), which had to be extracted manually from randomly ordered results. Only 167 hits were needed in BNC2014 to obtain 100 relevant tokens in comparison to 426 hits in BNC1994. The results suggest

¹ See Love et al. (2017)
² Also referred to as discourse particle (D’Arcy, 2005)
that more than half of the tokens of *like* in the BNC2014 corpus are non-standard uses, while in the BNC1994 it is approximately every fourth token.

Preliminary results suggest that *like* used as a hedge, a focus marker, and a discourse marker are devices that are consistently employed by speakers of British English in both corpora. However, the relative frequency of non-standard *like* rapidly increases in the BNC2014 data. The quotative marker is of particular interest, since it can be shown to have quadrupled its frequency in the new BNC2014 compared with the older BNC. The high frequency of its use in the BNC2014 is then consistent with the high frequency of quotative *like* observed in other varieties of English (Tagliamonte and D’Arcy 2004), from which it has likely spread into British English (Buchstaller and D’Arcy 2009).

**References**


**Analyzing diachronic change in the American English amplifier system**

*Martin Schweinberger*

Universität Hamburg, Universität Kassel

This study takes a corpus-based approach to examining amplifying intensifiers across four genres in American English (AmE) using the *Corpus of Historical American English* (COHA) that comprises data from 1810 to 2000. From a language variation and change perspective intensifiers are
particularly interesting as they play a crucial part in how speakers express themselves socially and emotionally and because intensifier systems are prone to change (Brinton & Arnovik 2006: 441).

(1)  
a. I feel assured that I looked very terrific. (COHA:FIC:1846)  
b. Pamela was really cool in the midst of all this. (COHA: MAG:1993)  
c. That is so awesome. (COHA: FIC:2007)

The study compares historical changes in the intensifier systems of American English (AmE) to findings from diachronic analyses of the New Zealand English (NZE) intensifier system (D’Arcy 2015). D’Arcy (2015) has shown that very used to be the sole dominant intensifier in NZE but, during the 20th century, really has successfully replaced very as the dominant intensifier. The study shows that there are two dominant intensifiers in AmE (very and so) while the NZE system has been dominated by only a single variant (very and later really). As a consequence, the replacement observable in NZE does not occur in the AmE system where very remains a highly frequent form but is replaced by so as the primary intensifier. The findings of this study thus show that the AmE intensifier system differs substantially not only from the NZE system (D’Arcy 2015) but also from the intensifier systems of other varieties such as Tyneside (Barnfield & Buchstaller 2010) and Toronto English (Ito & Tagliamonte 2003) which exhibit parallel developments to the NZE system.

To understand the diverging pathways of change, the current study investigates frequency distributions of intensifiers across time as well as collocation patterns of amplifier-adjective bigrams. During the statistical analysis cross-tabulations (configural frequency analyses) and semantic vector space models are used to investigate similarities and differences in the collocational profiles of intensifiers across varieties. The proposed hypothesis is that lexical replacement takes place in both varieties but that the mechanism affects both systems differently as so replaces very in AmE – thereby blocking potential rivals from playing a more dominant role – while really replaces very in NZE due to semantic similarity.

References


An n-gram-based analysis of Czech and English parliamentary debates: In search of optimal n-gram length

Denisa Sebestová and Markéta Malá
Charles University

The study focuses on a crucial methodological issue in n-gram-based research: determining the most informative length of n-grams for examining a given genre (Biber 2009; Hyland 2008). As shown in contrastive studies, the optimum length of n-grams is not only genre- but also language-specific (Granger 2014; Cvrček & Václavík 2015; Hasselgård 2017). Applying the n-gram method to the study of typologically different languages, such as English and Czech, seems to be particularly “challenging” (Čermáková & Chlumská 2017: 76).

We examine the relatively well-defined genre of parliamentary debates, relying on comparable corpora of English and Czech parliamentary discourse (Hansard, CzechParl corpus). The study is corpus-driven: it proceeds from the identification of 2–10-grams (i.e. continuous “recurring strings, with or without linguistic integrity (Lindquist & Levin 2008: 144)) to the qualitative functional description of n-grams of various lengths. The optimal n-gram length was determined for each language separately, considering the frequency and the amount of genre-specific information (structural and functional) obtained based on n-grams of a particular length.

The optimum length of n-grams appears to be different in Czech and in English. While the differences can be accounted for primarily by the typological differences between the languages (Czech is predominantly a synthetic language with rich inflection), some specific features of English and Czech parliamentary debates can also be pointed out.

The number of n-grams (min. frequency 20 pmw) is similar in both languages, except for bigrams. These occur more frequently in English. The structure of the English bigrams reveals the analytic character of the language: they typically comprise a combination of function words (of/in/to/on/for the, that/and the, it is, I am). Lexical words are few, mostly honorifics (hon friend, member for). In Czech, the representation of lexical words is much higher, comprising more varied forms of address (vážená paní, pane ministře), discourse-organizing and stance bigrams (děkuji za ‘I-thank for’, myslím že ‘I-think that’) and aboutness words (e.g. numerals). Among the function words, demonstratives and other deictic words are particularly frequent, indicating ties to the immediate co(n)text. The structure of trigrams displays higher variability in Czech than in English, where the structures a * of, of the *, the * of together form about 7% of 3-gram tokens. When exploring the content of parliamentary debates, longer n-grams seem more informative (at least 3-grams for Czech and 4-grams for English).

While n-gram based research often focuses on n-grams of up to 5 words, the study shows that the register of parliamentary debates relies on longer n-grams to a large extent (6–8 words in Czech, 5–6 in English). The English parliamentary discourse contains highly specific honorifics (My hon. Friend the Member for) and other politeness markers, while the Czech debates were characterised by a high frequency of discourse-specific performative formulae (zahájuji hlasování ptám se kdo je pro kdo je proti – ‘I-open the-vote who is for who is against').
References


Sources and tools


Collocational networks, keywords, and n-grams in a study of road deicing salts on surface and groundwater contamination

*Susana M. Sotillo and Manali Pradhan*
Montclair State University

This study presents preliminary results for the corpus analysis part of an interdisciplinary study which includes specialists from hydrology, geochemistry, and linguistics. Using LancsBox and AntConc, we explore how deicing impacts surface and groundwater based on published material found in various Environmental Science sources. Informed by the work of Brezina, McEnery, and Wattam (2015), we examine collocation networks. Sodium chloride or road deicing salts have been
used widely to remove snow and ice from roadways in urbanized areas of the northeastern US since the 1950s. Studies show that road deicing helps to reduce accident rates and road delays (Keummel 1992). While beneficial for highway safety, road salts have also been shown to negatively impact surface water and groundwater quality. These effects include loss of plant and biodiversity, nutrient depletion of soils, release of toxins, infrastructure damage, corrosion of pipes, and aquifer stratification and stagnation. To examine how road deicing and contamination are discussed in the relevant literature, we compiled corpora consisting of a reference and specialized corpus. The reference corpus includes texts which examine various contaminants of surface water and groundwater in scholarly articles, dissertations, magazines, press releases, and a few government reports. The specialized corpus more specifically discusses the effects of road salts on water contamination. The specialized corpus consists of 60 files from various sources (139,462 tokens), while the reference corpus includes 83 files and 413,366 tokens. Using LanesBox and AntConc, we examined collocates of specific words and phrases used by specialists in environmental science, as well as collocation networks. We selected the following association measures (AMs): DeltaP to measure directionality, MI3 to reduce low frequency bias, LogDice to measure the strength between collocate and node, and Loglikelihood, a significance statistic (Brezina et al. 2015; Evert 2008 Gries 2013; Hardie 2014). The following research questions were addressed: 1. What do visualizations of nodes and collocates in our corpora reveal with respect to deicing and water contamination using different AMs? 2. What environmental insights can we gain from examining positive keywords, lockwords, and negative keywords via GraphColl in these corpora? And 3. Are 3 and 5 n-grams generated with AntConc typical of the language used by environmental scientists? Preliminary results show visualizations of collocational networks around the nodes contamination, deicing, soil, sodium chloride, drinking water, surface water, and groundwater. These are possible factors that ultimately impact the quality of our drinking water. Next, simple math significance scores >10 are shown for the first 25 positive words, negative words, and lockwords when comparing our specialized corpus with the reference corpus. This gives us insight into possible factors and contaminants that have high frequencies and levels of statistical significance in our corpora. Lastly, the n-grams obtained show that while there are no unique phrases that specifically characterize environmental specialists’ language, their writing does contain formulaic sequences that are typical of academic research writing (Simpson-Vlach et al. 2010).

Partial References


Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next… International Journal of Corpus Linguistics, 18(1), 137–166.


“We did know exactly what to do” – Syntactic variation in Cape Flats English

Niklas Steih
University of Innsbruck

As part of an ongoing PhD project (and hence very much work-in-progress), this report presents the first results of a small pilot study on syntactic variation in Cape Flats English (CFE). While other varieties of South African English have been investigated in greater detail (most notably both White and Black South African English(es)), CFE remains a largely under-researched variety, especially with respect to empirical language data and additional statistical analysis of these data. Being spoken by the largely Coloured population that was relocated from Cape Town’s District Six to the outskirts of the city under the Apartheid regime, CFE can be placed on a continuum between the two big European settler languages of South Africa, Afrikaans and English. Previous research shows distinct syntactical features for CFE (cf. Malan 1996 or McCormick 2008) such as do as an unstressed tense marker, the use of simple past forms where Standard English would take present perfect, associative plural in noun phrases, absent distinction of number in demonstratives, and a variety of other features, all of which sometimes overlap with those of other South African varieties of English. However, the perceived salience of these features is largely the result of single researcher’s impressions, whereas frequency analyses, statistical testing and, moreover, actual language data remain scarce. While the final goal of this project is to collect a larger (and ideally CFE-only) data set via sociolinguistic interviews, some tentative results could already be found with the help of the Corpus of Black and Coloured South African English in Contact (Meierkord et al. 2009). Utterances produced by speakers with the regional and linguistic background that (loosely) fit the profile were manually extracted from the corpus, coded for the syntactic features in question and subsequently analysed for frequency in order to answer the question whether these features actually are to be taken as common as they have been described in the existing literature. At present, this does not seem to be the case, however it is to be discussed whether these results might be attributed to the corpus’ size. In addition to this statistical approach, it seems fruitful for our understanding of the variety to place CFE within a theoretical sociolinguistic variationist framework. As the report will show, the often-used Schneider’s dynamic model of World Englishes seems to be not applicable to CFE’s evolution without major problems, a new framework had to be found. At this moment, Buschfeld and Kautzsch’ recent extension of Schneider’s model (2017) to a framework of intra- and extraterritorial forces seems most applicable to CFE and its status as a variety and a first attempt will be outlined in the report.

References

Towards a genre map of Late Modern English medical texts: Essays and their neighbouring genres

Irma Taavitsainen
University of Helsinki

The aim of this empirical study is to cast new light on eighteenth-century medical writing with special attention to its genre dynamics. I shall adopt the prototype approach to genres, as individual texts exhibit genre features to different extents and there are overlaps. Genres can be defined as inherently dynamic cultural schemata, used by discourse communities to organize knowledge and experience through language. They unfold with more or less conventionalized linguistic expressions that display variation and undergo change when the functions change. My focus is on how a group of core genres of medical writing, as labeled by their users with different names, relate to one another in the eighteenth century: on the one hand, it is of interest what their common features are, and, on the other hand, what distinguishes between the neighbouring genres.

The data comes from a subcorpus of Late Modern English Medical Texts 1700–1800 (forthcoming) compiled for this study. It comprises texts marked by the labels “essay”, “dissertation”, “treatise”, “enquiry” and “pamphlet”. My focus is on essays. They are particularly interesting, especially as subjectivity marks them from the start (Montaigne 1580) and a personal point of view also prevails in essays by Francis Bacon, who introduced the genre into English. The label became fairly common in the seventeenth century, meaning ‘a draft or unfinished first attempts’ (OED), although Experimental essays in The Philosophical Transactions (1665–) deserve a very different characterization. In the eighteenth century, the label was frequently used in medical writing and texts may have several genre names in their titles. Overlaps are found with the new genre of “dissertations”, and “treatises” and “pamphlets” from the older genre repertoire are also mentioned in this connection. Some labels are more ad hoc like “enquiry” and “attempt”.

Linguistic features creating involvement and subjectivity are of particular interest for this assessment. They include personal pronouns, emotive expressions such as exclamations and expressive adjectives, interactive features like questions, and politeness phrases revealing text-participant relations and period style. Their frequencies vary a great deal in the material. A
quantitative study employing unigrams, bigrams and POS-bigrams for an unsupervised hierarchical clustering is employed to find out similarities and differences between texts. The results are then discussed in the light of qualitative textual analysis.

References


The transitivization and reflexive uses of sit (someone/oneself) down in Early and Late Modern English

Turo Vartiainen and Mikko Höglund
University of Helsinki, Stockholm University

In Present-day English, the act of ‘being or remaining in that posture in which the weight of the body rests upon the posteriors’ is typically expressed by the intransitive verb construction to sit (down) (OED, s.v. sit). While this intransitive structure has a long history in English, the same meaning has also been expressed, though much more rarely, by using different reflexive strategies in the past. In the simple strategy, personal pronouns are used coreferentially with the subject (e.g. I sat me down), while in the SELF-strategy, reflexive pronouns are used (e.g. I sat myself down). Contrary to the general trend in the history of English, whereby the SELF-strategy became more frequent of the two reflexive strategies in the sixteenth century (Peitsara 1997: 288), data from the Corpus of Historical American English suggest that the simple strategy remained more common with sit down until the late nineteenth century (examples (1) and (2)).

(1) I sat me down in the shadow of the portico and waited once more. (COHA, 1884)
(2) My mind grew clearer, and I sat myself down to deliberately shape some plan of action. (COHA, 1867)

Interestingly, our data show that the decline of the simple strategy is accompanied by the gradual emergence of a new transitive pattern, as exemplified in (3) and (4), raising the question of whether there is a connection between these contrasting developments; it might be that the emergence of the transitive use was discouraged by the fact that the object pronouns were already used in the reflexive pattern.

(3) She took me to her room and sat me down on the edge of her bed. (COHA, 1959)
(4) They sat him down against a tree at the edge of the field. (COHA, 1993)
In our work-in-progress report, we will first establish a precise timeline for these changes in American English and chart the linguistic contexts in which the transitive pattern (e.g. *he sat me down*) emerged, taking stock of the semantic/pragmatic factors that have been shown to be relevant for expressions of transitivity in previous literature (e.g. Hopper and Thompson 1980). We will then study the use of the two reflexive strategies in seventeenth- and eighteenth-century English by consulting large historical databases, such as *Early English Books Online* (EEBO) and *Eighteenth Century Collections Online* (ECCO), paying particular attention to the interchangeability of *sit* and *set* in the Early Modern data and considering the potential relevance of their overlap to the transitivization of *sit down*.

**References**

ECCO-TCP = *Eighteenth Century Collections Online. Text Creation Partnership*. [https://quod.lib.umich.edu/e/ecco/](https://quod.lib.umich.edu/e/ecco/).  
GloWbE = *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries* (GloWbE). [http://corpus.byu.edu/glowbe/](http://corpus.byu.edu/glowbe/).  

---

**A multifactorial analysis of the choice of anaphoric devices in English-Chinese translation**

**Xiuling Xu**  
Beijing Foreign Studies University

In the past two decades, an increasing amount of evidence has shown that translated language is a “third code” (Frawley 1984), which is different from both source and target languages. These differences are supposed to occur at all linguistic levels. Nevertheless, corpus-based translation studies have been largely confined to lexical and grammatical levels (e.g. Laviosa 1998; Olohan & Baker 2000; Xiao & Dai 2014). Textual features, however, remain under-explored in corpus-based translation studies. The present study therefore aims to explore the use of anaphora, a linguistic feature at the textual level, in Chinese translated from English as compared with non-translated native Chinese.

Anaphoric devices in both English and Chinese can be classified into three types, i.e. nominal, pronominal, and zero anaphora. However, the two genetically distant languages differ considerably:
English is characterized by a heavy use of pronominal anaphora while Chinese makes a more frequent use of zero anaphora. How the choices are made from the three types of anaphoric devices has been extensively investigated in English and Chinese linguistics. Previous studies have revealed that anaphoric choice is not at random but instead is influenced by a constellation of factors, e.g. animacy of the referent, syntactic roles of the antecedent and the anaphor, referential distance, potential interference, text type, etc (Ariel 1990; Chen 1986; Givón 1983; Xu 2004).

Although the choice of anaphoric devices has been widely studied in both English and Chinese linguistics, little research has been done to compare anaphoric choice in Chinese translated from English with that in native Chinese. Though some studies have been undertaken on pronouns in Chinese translated from English (e.g. Wang & Hu 2010; Xiao & Dai 2014), most of them primarily focus on pronouns as isolated lexical items rather than a textual feature. Such research is mainly concerned with counting the frequencies of pronouns, paying little attention to the semantic, syntactic and discourse features that co-occur with them. Furthermore, nominal and zero anaphora are rarely dealt with in corpus-based translation studies. It is therefore worthwhile to compare the factors affecting anaphoric choice in translated Chinese with those in native Chinese, so as to give a full picture of the distinct patterns of anaphoric use characteristic of translated Chinese.

The present study utilizes an English-Chinese parallel corpus and a corpus of native Chinese, both following the sampling frame of the Brown Corpus. Three thousand instances of anaphora in each corpus are annotated for a number of semantic, syntactic, discourse and translation-related features that co-occur with the three types of anaphoric devices. The annotated data will be subjected to sophisticated statistical treatment. Multifactorial statistical approach, or more specifically multinomial regression analysis (cf. Gries 2013) will be employed in the present study to model the concurrent influence of multiple factors that work together to trigger the different choices of anaphoric devices in Chinese translated from English and non-translated native Chinese.

References

Posters
Measuring syntheticity

Helena Filipová and Lucie Gillová
Charles University

Language typology is a common tool for describing, analyzing and comparing languages. The concept is very often used rather intuitively, especially in contrasting languages and their structures and it is rather problematic to quantify the degree of a language type present in the language under scrutiny.

The poster is going to present verification, or falsification, of the methodology for measuring morphological syntheticity proposed in “Measuring typological syntheticity of English diachronically with the use of corpora” (Tichý & Čermák 2014). The methodology is based on morphological behaviour of high-frequency nouns, adjectives and verbs, the basic assumption being that the more varied the inflectional system of a language, the more synthetical its nature. Based on the distribution of morphological markers across paradigms, authors computed syntheticity indices for these parts of speech in a given period (Old English, Middle English, Early Modern English, Present-Day English) and an overall index for the language of this period. A morphological marker is understood to be a formal expression, so it is not described on functional basis. The index expresses the choice a speaker has in creating forms of a word, that is how well the markers are distributed throughout the system. To compute this index the authors used the informational entropy, capturing the degree of randomness in the system underlying a speaker’s choice.

If this methodology is applicable to different languages and yields any comparable results, the description of English and its development can be considered functional and reliable. To see if this is really the case, the authors replicate the research on Modern Czech, using Czech corpora (SYN series, Institute of the Czech National Corpus). Czech language is commonly described as a synthetic language, thus in terms of typology it is closer to Old English than it is to Present-Day English.

Firstly, only probes of high-frequency words are carried out, in a similar way as in the original study. If the syntheticity index is closer to the value of Old English than to the values of later stages of English, it is reasonable to assume that the index is representing the same property of the two language systems and to consider this methodology valid.

The second step attempts to assess whether the fact that the calculation of the syntheticity index is based on a probe, i.e. on a limited set of data, may in itself skew the results and the value of the index. Having the advantage of working with a fully lemmatized and tagged corpus, the whole procedure is repeated for the entire corpus and the final values compared with those obtained in the first step. Similarly, the procedure is repeated for the BNC and so that the results can be contrasted with those from the original study.

Selected References

Recent diachronic change in the sociolinguistics of amplifiers in spoken British English

Robert Fuchs
University of Hamburg

Age, gender, social class and dialect are known to influence amplifier frequency and choice, but previous research has insufficiently investigated the complex interactions that can occur between such variables and/or is restricted to small, tight-knit communities (Cameron 2009, Mills 2006, Murphy 2010). Moreover, a change in attitudes towards the different social roles that women and men are expected to fulfil in society might have contributed to the language of males and females becoming more similar, as suggested by Hancock et al. (2015). Corpus-based research on recent diachronic change in intensification in spoken British English has shown that age- and class-based differences in amplifier frequency have decreased in the last 20 years, while the female lead in intensifier frequency has hardly diminished over time (Fuchs 2017). At the same time, age- and gender-based preferences for specific amplifier variants (e.g. so, totally) have been shown to persist (Hessner and Gawlitzek 2017), although an analysis of the whole amplifier system, the role of social class and the interaction of sociolinguistic variables is still to be accomplished.

In order to address this research gap, the present study investigated how age, gender, social class and dialect influence which amplifiers (sometimes also simply called ‘intensifiers’) speakers of British English use in private conversations and whether this has changed over the last two decades. In addition, the meaning of the amplified adjective (determined based on USAS tag) is taken into account. The analysis was restricted to amplifiers (boosters and maximisers). Logistic regression models were run in R to analyse a total of 41,226 amplifier tokens occurring before adjectives that were extracted from the British National Corpus 1994-demog and 2014-sample, with over 600 speakers, covering 134 amplifier variants. The analysis then focussed on the nine most frequently occurring amplifiers.

These results reveal a complex picture, which will be illustrated here by focusing on so, whose frequency increased from 1.6% to 2.1%. As Fig. 1 shows, in 1994, 40–49 year-olds used this amplifier least frequently, and both older and younger speakers used it more often, with the highest frequency among the youngest. In 2014, the 40–49 year olds from 1994 belong to the oldest age...
group in the analysis (60+), and all other generations use so more frequently than they do. At the same time, preferences in social class and gender in the usage of so have changed substantially. Working class female speakers used so more frequently than middle and upper class women in 2014, with no significant difference in 1994. By contrast, working class men used so less frequently than other men in 1994, and use it slightly more frequently than upper class men.

Together with the analysis of the nine other most frequent amplifier variants, the results suggest that gender differences in preferences for particular variants have overall not become smaller in the last two decades, in contrast to Hancock et al.’s (2015) hypothesis. Gender, together with other sociolinguistic variables such as age and social class, continues to shape the patterns of amplifier usage in (British) English.

References


Murphy, B. (2010). *Corpus and Sociolinguistics. Investigating Age and Gender in Female Talk*. Amsterdam/Philadelphia: John Benjamins.
Fig. 1. Logistic regression analysis of the frequency of *so*. 
Fluency in ENL, ESL, and EFL: Introducing a contrastive approach to English as a first, second, and foreign language

Sandra Götz, Christoph Wolk and Katja Jäschke
Justus Liebig University Giessen

Kachru’s (1985) Three Circles Model and Strang’s (1970) distinction between English as a native language (ENL), English as a second language (ESL), and English as a foreign language (EFL) has exerted an enormous influence on the modelling of Englishes worldwide. However, this distinction is far from clear-cut and newer research tends to emphasize EFL, ESL, and ENL rather as a continuum (e.g. Gilquin & Granger 2011; Housen et al. 2011; Mukherjee & Hundt 2011; Deshors et al. 2016). Corpus-based studies in that vein have mainly focused on written English so far (few exceptions being Ballier & Martin 2011 or Götz & Schilk 2011). One crucial peculiarity of spoken language is that speakers need to balance the cognitive demands of producing speech with the need to maintain a consistent flow of the conversation, i.e. fluency. However, a comparative approach to spoken fluency in EFL vs. ESL vs. ENL has not yet been undertaken; indeed, in the description of post-colonial varieties of English, fluency has only played a role in some first pilot studies so far (e.g. Götz 2015; Gut & Fuchs 2017).

Against this background, our poster introduces our recently started project in which we investigate the fluency of speakers of ENL, ESL, and EFL. As a database for the corpus analysis, we will use several components of the International Corpus of English (ICE; Nelson 1996) and the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al. 2010). In order to systematically assess how speakers of these three different types of Englishes establish fluency, we analyze these corpora for linguistic variables that can potentially have an effect on a speaker’s fluency (i.e. “fluencemes”; Götz 2013: 8–9). For this purpose, we analyze and describe fluency on three different levels: (1) temporal variables in speech production (e.g. filled and unfilled pauses) as well as speakers’ use of fluency-enhancing strategies (e.g. discourse markers or prefabricated units), (2) the (potential combinations of) different fluencemes to overcome planning difficulties (fluenceme chunking) as well as their positions in the utterance (fluenceme positioning), and (3) correlations of fluencemes with extralinguistic variables (e.g. gender, age) that can predict the type and use of fluencemes in different types of Englishes (fluenceme preferencing).

In this project, we will test whether speakers of different types of English (1) are more or less fluent than others, (2) establish fluency in different ways (e.g. by using different fluencemes), (3) need planning phases in different positions in the utterance and use different strategies to overcome them and further (4) if fluency and fluency enhancing strategies can be predicted by extra-linguistic parameters, such as age or gender. Our proposed poster will outline our project and present the results of a first pilot study: (1) From a quantitative point of view, there is a distinct cline in the frequency of dysfluency features in spoken English from EFL to ESL to ENL. (2) The qualitative analysis of the functions of fluency enhancement strategies shows some variety/variant-specific functional preferences for how fluency is established in the three speaker groups (e.g. a more frequent use of discourse markers in ENL, a preferred use of repeats in ESL, a higher number of filled pauses in EFL).
References


Relaxing the verbal dress code? The Hansard and language change in Australian, British and South African English (1900–2015)

Haidee Kruger¹, Bertus van Rooy², Pam Peters³, Adam Smith³ and Minna Korhonen³
Macquarie University/North-West University¹, North-West University², Macquarie University³

This poster reports on an ongoing project to develop a comparable diachronic corpus of the Australian, British and South African Hansard, covering the period 1900 to 2015, to allow investigation into how language, stylistic and register change occurs across these three varieties of English. The Hansard is the “substantially verbatim” or “edited verbatim” written record of
parliamentary proceedings, and as such is a specialised hybrid (sub)register of “written speech” (Slembrouck 1992) which offers unique opportunities to investigate the interplay between key forces in short-term language, stylistic and register change, specifically colloquialisation and densification (Biber & Gray 2012; Leech, Hundt, Mair, & Smith 2009). Moreover, the Hansard is a British institutional register that has been transplanted to other Commonwealth countries, allowing for the investigation of English usage in different locations within the frame of a reasonably comparable and clearly defined subregister with a continuous historical record. Changes in the language of the Hansard may reflect changes in the speech of parliamentarians, but may also reflect changes in conventions for representing formal spoken language in writing.

In this poster, we focus on the corpus resource developed in this project, and outline first findings on language and register change in the three varieties. We summarise the completed first stage of the corpus compilation: the data collection for and construction of the comparable Hansard corpus (covering five periods from 1900 to 2015 at approximately 30 year intervals and consisting of approximately 8 million words), using Australian and British English materials from electronic records that are publicly available, and South African data from electronic records and archival materials. Subsequently, we outline the ongoing development of a smaller corpus of recordings of speech, approximating 200,000 words per variety, with the largest chronological span possible on the basis of available audio recordings for each set of parliamentary proceedings. The transcribed speech and the official Hansard will be aligned at sentence level to constitute a parallel corpus. The comparison of the audio recordings with the official Hansard enables insight into the extent to which linguistic and register changes in the Hansard are due to changes in the original speech, or changes in the norms for recording speech in the written record (see Mollin 2007).

The Hansard, as “written speech”, is particularly suitable to investigate the interplay between colloquialisation and densification. Furthermore, as record of parliament, it also reflects changes in the composition of parliaments and the sensibilities of the day, which are expected to influence the language used to talk about current events. The poster will offer a brief summary of some of the early results from this longer-term project, focusing on evidence of colloquialisation, densification, and other rhetorical changes, and relating these to changes in the larger society, the composition of parliaments (Spirling 2015), and the changing editorial norms of the Hansard itself (Edwards 2016; Gravlee 1981).

References


“*Here comes the police! Here they come!*” A diachronic corpus-based sociolinguistic study of agreement with collective nouns

*Alexander Lakaw*
Linnaeus University

English collective nouns and their agreement patterns, as illustrated in (1)–(3) below, have received a great deal of attention in corpus linguistics.

(1) …and the police *has* not yet been aroused from *its* lethargy. (COHA; 1822; Magazine)

(2) the police *were sent* for, and on their arrival he said there was no need of so much fuss. (OBC; 1841)

(3) “*Here comes the Police! here they come!*’ shouted the boys (COHA; 1859; Fiction)

Previous research has found evidence of variability within and across the different varieties of English (e.g. Levin 2001; Hundt 2006), and relates the differences between the varieties partially to the lexical characteristics of the collective nouns themselves (e.g. Depraetere 2003: 124; Bock et al. 2006: 101; Levin 2006: 339). However, the diachronic developments of this variation have not been investigated in detail, and there is a lack of research that focuses more on qualitative analyses of the collective nouns themselves by highlighting their lexical characteristics and their historical development from a sociolinguistic and semantic perspective.

In an attempt to fill this gap, this poster presents a study that investigates the agreement patterns of the collective noun *police* and its conceptual predecessors *watch* and *patrol* in 19th-century AmE and BrE. It combines diachronic corpus linguistics and historical sociolinguistics by discussing the linguistic as well as the social history of the concept of *police* in America and Britain. The results are drawn from the *Corpus of Historical American English*, the *Old Bailey Corpus* and the *Corpus of Late Modern English Texts*. They indicate that the change in the agreement patterns of the noun *police* was conditioned by lexical and historical sociolinguistic factors. One finding suggests, for instance, that the agreement pattern of the noun *police* changed from being variable to a preference of plural agreement due to specific changes made in the organisation of the police patrols prior to the 1870s. In his sociological study of the history of policing, Thale (2004: 1053) argues that, because of these changes, “[t]he public experienced not ‘the’ cop on the beat, but ‘the cops’”, as they resulted in a shift from singularity towards plurality with regards to the public perception of ‘the patrolling police officer’.
References


Aspects of the lemmatisation of an aligned parallel corpus of Old English

**Javier Martín Arista**
Universidad de La Rioja

This paper addresses some questions related to the lemmatisation of a parallel corpus. On the basis of a 10,000 word pilot corpus (Figure 1), an aligned parallel corpus of Old English prose is being compiled within the *Nerthus* Project, which carries out linguistic analysis of this historical language with lexical databases.
The implementation of the parallel corpus crucially depends on the source databases, including a dictionary database called *Nerthus* (ca. 30,000 files), which is geared to morphological and lexical analysis; and a dictionary database called *Freya* (ca. 35,000 files), oriented to the indexation of secondary sources. The source databases provide the information necessary for lemma annotation, which comprises alternative spellings, inflectional class, inflectional paradigm, derivational paradigm and secondary sources (Figure 1). This, in turn, requires that the inflected forms in the texts are lemmatised, or attributed to one of the headword entries in the source databases. The lemmatiser *Norna* (Figure 2) contains ca. 190,000 files and is based on the attestations of the *Dictionary of Old English Corpus* (DOEC). *Norna*, which is currently focused on the verbal lexicon, searches the texts for preverb, stem and inflectional ending, thus reaching a maximal accuracy of 80 percent. In its present state, *Norna* can deal with the choice of headword spelling (as ābyrnan vs. ābiernan for abyrneð), the spelling variants of prefixes (like the alternants of the prefix æfter-: æft-, æftyr-, æfter-, æftyr-), the changes to the word stem caused by apophony (as in seoðan vs. sudon), the weakening of unaccented inflectional syllables (like -an instead of -en) and co-existing inflectional endings (such as bīetð and bīett).

Other areas require manual revision with the index and the concordance to the DOEC, including unpredictable spellings (<k>, for instance, as in kymð instead of cymð), the disambiguation of members from different lexical categories such as scēotan ‘to shoot’ and scēot ‘ready’ as representing the attestation scēot; or items from the same lexical class as, for example, brūcan ‘to use’ vs. brýcian ‘to be useful’ for brecað. Results are checked against lexicographical and textual sources. With respect to lexicographical sources, the *Dictionary of Old English A–H* (DOE) permits the direct comparison of lemmas and inflections, although the headword spelling adopted in the
source databases does not coincide exactly with the DOE, which prefers later forms with <y> rather than <ie>, among other differences. As for textual sources, The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE) is unlemmatised, but it is tagged morphologically and identifies the inflections of a given category, such as the verb. The main conclusion of the paper is that the search options of database software enhance the lemmatisation process, although the lack of fixed spelling precludes the complete automatisation of the lemmatisation of a corpus of Old English. Hence, the comparison with the existing lexicographical and textual sources constitutes an invaluable tool for verification and feedback.

References

Open science for English historical corpus linguistics: Publishing the Language Change Database

Terttu Nevalainen¹, Turo Vartiainen¹, Joonas Kesäniemi², Agata Dominowska¹
University of Helsinki¹, CSC – IT Center for Science²

In this work-in-progress report, we discuss an open-access resource that can be used as a baseline for corpus-linguistic research into the variation and change of the English language: the Language Change Database (LCD). The LCD draws together information extracted from hundreds of corpus-based articles that investigate the ways in which English has changed in the course of history. The database includes annotated summaries of the articles, as well as numerical data extracted from the articles and transformed into machine-readable form. It therefore provides scholars of English with the opportunity to study fundamental questions about the nature, rate and direction of language change. It will also make the work done in the field more cumulative and sustainable by ensuring that the research community will have continuous access to existing results and research data.

Each of the ca. 300 LCD entries is annotated for several features according to which the database can be queried, including grammatical and sociolinguistic keywords, the periods studied, and bibliographical details. The quantitative research data from the articles are included in the entries as
reformatted and standardized Excel files, which the end users can easily combine into larger bundles, download and reanalyze on their own computers. Each entry also includes an abstract and a summary of the main findings of the study. The LCD has an intuitive search interface which allows the user to select and filter the results with parameters such as corpus, time period, grammatical category, and a variety of search keywords (Nevalainen et al. 2016).

With this kind of open-access tool and cumulative data, the reproducibility of studies becomes a more attainable goal (cf. King 2017). As the number of articles in the LCD grows, researchers will be able to conduct large-scale meta-analyses of multiple changes in the English language, which could result in significant advances in descriptions of the history of English as well as theoretical accounts of language change. Indeed, we envisage that in addition to historical corpus linguists, the database will be of use to a wider audience interested in, for instance, statistical modelling and sociolinguistic typologies. Furthermore, the database will be useful for teaching purposes.

We are now ready to proceed from the development phase to piloting the Language Change Database. The search interface will be made available to the research community at the ICAME conference. We will also invite colleagues who are interested in inputting the results of their own work in the LCD to test the back-end of the database. All feedback is warmly welcome.

References


Finding evidence for changing society: A diachronic study of medical discourse in 1500–1800

Maura Ratia
University of Helsinki

In terms of scientific advancement and medical writing, the period from 1500 to 1800 can be described as intriguing. Whereas the early modern period brought about new scientific methods of observation and experiment as well as new genres of writing, the spirit of Enlightenment of the eighteenth century saw groundbreaking new methods of cure, such as smallpox inoculation, accompanied by a better understanding of diseases. In the latter period, novel cultural ideals also developed. This is manifested in, for example, the establishment of philanthropic societies and especially the emergence of “polite society”. Politeness, benevolence and compassion were among the traits that were expected of ‘medico-gentility’ of the time (see Brown 2011: 74), and the newfound concerns about the humane treatment of patients entered in tandem the medical discourse. On the discourse level, the form and structure of medical discourse were clearly linked to an assessment of what is socially acceptable behaviour.
In my study I wish to explore how the new and changing ideals, specifically those of the polite society, are represented in medical discourse from 1500 to 1800. My initial findings suggest that from the late seventeenth century onwards significant changes take place in the overall attitude of physicians in relation to their patients and to the public suffering from various ailments. I will use collocation analysis as a method (see e.g. McEnery and Hardy 2013) to gather evidence of ethical concerns and growing compassion of physicians towards their patients. Two corpora are used as material: EMEMT, the Corpus of Early Modern English Medical Texts (Taavitsainen et al. 2010), totalling over 2 million words, and LMEMT, the Corpus of Late Modern English Medical Texts (Taavitsainen et al. forthcoming) with 2.2 million words. LMEMT as a new tool allows for studies of medical writing in a long diachrony, which were previously difficult to conduct due to a lack for corpus material suitable for linguistic study.

References


Introducing the Linnaeus University English-German-Swedish corpus (LEGS): rationale, composition and first results

Jenny Ström Herold
Linnaeus University

Translation studies have been largely based on fiction texts (Bassnett 2014) and literary analysis. Fiction texts nevertheless contain many author-specific features and are therefore less suitable for many research purposes. Moreover, some translation corpora, such as the Oslo Multi-Lingual Corpus (OMC), largely consist of idiosyncratic and fairly old fiction texts.

In order to address the lack of present-day non-fiction translation corpora, we are in the process of compiling the Linnaeus University English-German-Swedish corpus (LEGS) containing, for instance, biographies and texts about popular science and history. The results of the studies carried out on this corpus will help translators solve actual problems in their work, make language-specific text norms visible, and allow for more research-based translator training.

LEGS consists of recently published (2000s) popular non-fiction texts in English, German and Swedish, and it is balanced for the three languages, every original always being accompanied by two target texts. Each author and translator is represented only once. The corpus currently
comprises about 350,000 words in each source language with translations. The aim for the first compilation phase is half a million words.

The main advantage of the corpus is that there are always two translations available for every source-text segment. This makes it possible to compare how the very same instance has been translated into two target languages, allowing identification of language-specific and translation-specific features. Moreover, the corpus provides translations from two source languages into each language.

The current research questions concern German and Swedish correspondences of various English lexico-grammatical structures. One recent study concerns English supplementive *ing*-clauses in contrast (e.g., *Hitler exploded, demanding examples.*) (Ström Herold & Levin submitted), and another deals with proper nouns used as modifiers (Ström Herold & Levin in preparation). The findings indicate that German translations of proper noun modifiers produce more compound nouns than Swedish ones, which instead prefer prepositional phrases (e.g., *the Norway fiasco > das Norwegen-Fiasko (GE) / fiaskot i Norge (SW)*). In contrast, German and Swedish prepositional phrases (*Laboratorien in New York (GE) > a New York laboratory*) and genitives (*FBI:s högkvarter (SW) > FBI headquarters*) are very rarely rendered as English proper noun modifiers.

This poster will mainly focus on the progression of the corpus compilation process and the benefits of working with a trilingual translation corpus.

**References**


**A diachronic analysis of the distribution of the distributive quantifier *each***

*Tomohiro Yanagi*

Chubu University

In this presentation, on the basis of data collected from the YCOE and the PPCME2, I quantitatively show the distribution of the distributive quantifier *each* in Old and Middle English and discuss how its distribution changed diachronically. In Present-day English the quantifier *each* can be used with a *of*-genitive phrase, as in *each of the rooms*. It can follow the subject noun phrase, as in *The rooms each have a telephone*, but it cannot follow the object, as in *I called the mem/them each*.

The quantifier *ælc* ‘each’ in Old English, on the other hand, exhibited a slightly different distribution. It could precede a genitive noun phrase, as in *ælc ðæra manna* ‘each of those men’, or
it could follow a genitive pronoun, as in *heora ælc* ‘each of them’. In addition, the quantifier *ælc* ‘each’ could be separated away from the subject noun phrase it is associated with, as in *Godæ gecorenan scinað on heofonlicum wuldre ælc* (*ÆCHom I 30:434.149*) ‘God’s chosen shine in heavenly glory, each according to his merits’. The quantifier could be also floated from the object noun phrase, as in *his suna swa ilce ælcne hi hæfdon for god* (*Bo* 38.115.27) ‘likewise they had each of his sons for a god’, although the number of such examples is quite small. In this example, the quantifier *ælcne* is the accusative form of *ælc*. In Middle English, the *each* construction with a genitive nominal such as *ælc ðæra manna* ‘each of those men’ and that with a genitive pronoun such as *heora ælc* ‘each of them’ were both getting lost and the *of*-genitive phrase instead started to be used in this construction with a noun phrase as in *ech of tho thingis* ‘each of those things’ and with a pronoun as in *ech of hem* ‘each of them’.

As for the position of the distributive quantifier *each*, the quantifier can precede a singular noun phrase and follow a plural noun phrase. This contrasts with the universal quantifiers *all* and *both*. Those universal quantifiers can either precede or follow a noun phrase, whether that noun phrase is singular or plural. I argue within the generative framework that this distributional difference between the distributive and universal quantifiers can be attributed to the difference of the syntactic structures containing those quantifiers and the difference of their positions within the structures. Furthermore, I suggest that the development of *each other* may be related to the availability of the floating *each* and the loss of the case endings. At the beginning, the two words *each* and *other* were separated, as in *ant euch an luueð oðer ase muchel as him-seoluen* (*Sawles Ward 183.247*) ‘and each one loves other as much as himself’. Later, they came to be adjacent to each other and finally they started to be used as a single unit.

*References*


Corpora


Software demonstration
In recent years, data visualization has become a major area in Digital Humanities research, and the same holds true also in linguistics. The rapidly increasing size of corpora, the emergence of dynamic real-time streams, and the availability of complex and enriched metadata have made it increasingly important to facilitate new and innovative approaches to presenting and exploring primary data. This demonstration showcases the uses of Virtual Reality (VR) in the visualization of geospatial linguistic data using data from the Nordic Tweet Stream (NTS) project (see Laitinen et al. 2017). The NTS data for this demonstration comprises a full year of geotagged tweets (12,443,696 tweets from 273,648 user accounts) posted within the Nordic region (Denmark, Finland, Iceland, Norway, and Sweden). The dataset includes over 50 metadata parameters in addition to the tweets themselves.

We demonstrate the potential of using VR to efficiently find meaningful patterns in vast streams of data. The VR environment allows an easy overview of any of the features (textual or metadata) in a text corpus. Our focus will be on the language identification data, which provides a previously unexplored perspective into the use of English and other non-indigenous languages in the Nordic countries alongside the native languages of the region.

Our VR prototype utilizes the HTC Vive headset for a room-scale VR scenario, and it is being developed using the Unity3D game development engine. Each node in the VR space is displayed as a stacked cuboid, the equivalent of a bar chart in a three-dimensional space, summarizing all tweets at one geographic location for a given point in time (see: https://tinyurl.com/nts-vr). Each stacked cuboid represents information of the three most frequently used languages, appropriately color coded, enabling the user to get an overview of the language distribution at each location. The VR prototype further encourages users to move between different locations and inspect points of interest in more detail (overall location-related information, a detailed list of all languages detected, the most frequently used hashtags). An underlying map outlines country borders and facilitates orientation. In addition to spatial movement through the Nordic areas, the VR system provides an interface to explore the Twitter data based on time (days, weeks, months, or time of predefined special events), which enables users to explore data over time (see: https://tinyurl.com/nts-vr-time).

In addition to demonstrating how the VR methods aid data visualization and exploration, we will also briefly discuss the pedagogical implications of using VR to showcase linguistic diversity.

References

List of delegates and participant

Abe, Daisuke
Aijmer, Karin
Alexander, Marc
Alissandrakis, Aris
Alosaimi, Saleh
Andersen, Gisle
Anthony, Laurence
Arndt-Lappe, Sabine
Baldus, Lina
Bednarek, Monika
Biber, Douglas
Blanco-Suárez, Zeltia
Bondarenko, Antonina
Bourgeois, Samuel
Brinton, Laurel
Brook, Marisa
Brookes, Gavin
Brůhová, Gabriela
Brunner, Marie-Louise
Bruns, Hanna
Busse, Beatrix
Buyle, Anouk
Callies, Marcus
Caple, Helen
Castro-Chao, Noelia
Cermakova, Anna
Chen, Yu-Hua
Chudar, Alexandra
Claridge, Claudia
Collet, Caroline
D’Arcy, Alexandra
Davidse, Kristin
De Felice, Rachele
Degani, Marta
Deshors, Sandra
Diemer, Stefan
Diskin, Chloé
Dolberg, Florian
Dominowska, Agata
Dossena, Marina
Ebeling, Jarle
Ebeling, Signe Oksefjell
Egbert, Jesse
Eitelmann, Matthias
Evans, Craig
Filipová, Helena
Flach, Susanne
Fuchs, Robert
Gandón-Chapela, Evelyn

Nagoya University
University of Gothenburg
University of Glasgow
Linnaeus University
Imam University
NHH Norwegian School of Economics
Waseda University
Trier University
University of Trier
University of Sydney
Northern Arizona University
University of Cantabria
Université Paris 7 Diderot
University of Neuchâtel
University of British Columbia
University of Victoria
University of Nottingham
Charles University
Trier University of Applied Sciences
University of Bonn
Heidelberg University
KU Leuven
University of Bremen
University of New South Wales
University of Santiago de Compostela
University of Birmingham
University of Nottingham Ningbo China
Minsk State Linguistic University
University of Augsburg
Saarland University
University of Victoria
KU Leuven
University College London
University of Verona
Michigan State University
Trier University of Applied Sciences
University of Melbourne
TU Dortmund
University of Helsinki
Università degli Studi di Bergamo
University of Oslo
University of Oslo
Northern Arizona University
Johannes Gutenberg-University
Lancaster University
Charles University
Université de Neuchâtel
University of Hamburg
University of Cantabria
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geers, Sarah</td>
<td>University of Bremen</td>
</tr>
<tr>
<td>Gillová, Lucie</td>
<td>Charles University</td>
</tr>
<tr>
<td>Gilquin, Gaëtanelle</td>
<td>University of Louvain</td>
</tr>
<tr>
<td>González-Díaz, Victorina</td>
<td>University of Liverpool</td>
</tr>
<tr>
<td>Grafmiller, Jason</td>
<td>University of Birmingham</td>
</tr>
<tr>
<td>Green, Clarence</td>
<td>Nanyang Technological University</td>
</tr>
<tr>
<td>Grieve, Jack</td>
<td>University of Birmingham</td>
</tr>
<tr>
<td>Groom, Nicholas</td>
<td>University of Birmingham</td>
</tr>
<tr>
<td>Götz, Sandra</td>
<td>Justus-Liebig-University Giessen</td>
</tr>
<tr>
<td>Güldenring, Barbara Ann</td>
<td>University of Marburg</td>
</tr>
<tr>
<td>Hampel, Elizabeth</td>
<td>University of Bonn</td>
</tr>
<tr>
<td>Harrison, Simon</td>
<td>University of Nottingham Ningbo China</td>
</tr>
<tr>
<td>Hasselgård, Hilde</td>
<td>University of Oslo</td>
</tr>
<tr>
<td>Heimeroth, Klaus</td>
<td>TU Dortmund</td>
</tr>
<tr>
<td>Heller, Benedikt</td>
<td>KU Leuven</td>
</tr>
<tr>
<td>Higashizumi, Yuko</td>
<td>University of British Columbia</td>
</tr>
<tr>
<td>Hilpert, Martin</td>
<td>Université de Neuchâtel</td>
</tr>
<tr>
<td>Hiltunen, Turo</td>
<td>University of Helsinki</td>
</tr>
<tr>
<td>Hirota, Tomoharu</td>
<td>University of British Columbia</td>
</tr>
<tr>
<td>Hoffmann, Sebastian</td>
<td>Trier University</td>
</tr>
<tr>
<td>Huber, Magnus</td>
<td>University of Giessen</td>
</tr>
<tr>
<td>Höglund, Mikko</td>
<td>Stockholm University</td>
</tr>
<tr>
<td>Iyeiri, Yoko</td>
<td>Kyoto University</td>
</tr>
<tr>
<td>Jäschke, Katja</td>
<td>Justus Liebig University Giessen</td>
</tr>
<tr>
<td>Kaatari, Henrik</td>
<td>University of Gävle</td>
</tr>
<tr>
<td>Kaunisto, Mark</td>
<td>University of Tampere</td>
</tr>
<tr>
<td>Kehoe, Andrew</td>
<td>Birmingham City University</td>
</tr>
<tr>
<td>Keppenne, Valerie</td>
<td>Penn State University</td>
</tr>
<tr>
<td>Kesäniemi, Joonas</td>
<td>CSC – IT Center for Science</td>
</tr>
<tr>
<td>Kireili, Kathrin</td>
<td>Justus Liebig University Giessen</td>
</tr>
<tr>
<td>Kirk, John</td>
<td>University of Vienna</td>
</tr>
<tr>
<td>Kleiber, Ingo</td>
<td>Heidelberg University</td>
</tr>
<tr>
<td>Kolbe-Hanna, Daniela</td>
<td>Trier University</td>
</tr>
<tr>
<td>Korhonen, Minna</td>
<td>Macquarie University</td>
</tr>
<tr>
<td>Kranich, Svenja</td>
<td>University of Bonn</td>
</tr>
<tr>
<td>Kretzschmar, William A.</td>
<td>University of Georgia</td>
</tr>
<tr>
<td>Krug, Manfred</td>
<td>University of Bamberg</td>
</tr>
<tr>
<td>Kruger, Haïdee</td>
<td>Macquarie University / North-West University</td>
</tr>
<tr>
<td>Kuehmstedt, Patrick</td>
<td>University of Potsdam</td>
</tr>
<tr>
<td>Kunter, Gero</td>
<td>Universität Siegen</td>
</tr>
<tr>
<td>Kytö, Merja</td>
<td>Uppsala University</td>
</tr>
<tr>
<td>Laippala, Veronika</td>
<td>University of Turku</td>
</tr>
<tr>
<td>Laitinen, Mikko</td>
<td>University of Eastern Finland</td>
</tr>
<tr>
<td>Lakaw, Alexander</td>
<td>Linnaeus University</td>
</tr>
<tr>
<td>Lambert, James</td>
<td>Nanyang Technological University</td>
</tr>
<tr>
<td>Lange, Claudia</td>
<td>Dresden University of Technology</td>
</tr>
<tr>
<td>Larsson, Tove</td>
<td>Université catholique de Louvain</td>
</tr>
<tr>
<td>Leuckert, Sven</td>
<td>Technische Universität Dresden</td>
</tr>
<tr>
<td>Levey, Stephen</td>
<td>University of Ottawa</td>
</tr>
<tr>
<td>Levin, Magnus</td>
<td>Linnaeus University</td>
</tr>
<tr>
<td>Liimatta, Aatu</td>
<td>University of Helsinki</td>
</tr>
</tbody>
</table>
López-Couso, María José  
Loureiro-Porto, Lucía  
Lundberg, Jonas  
Mahlberg, Michaela  
Malá, Markéta  
Mandal, Antorlina  
Martín Arista, Javier  
Mehl, Seth  
Meriläinen, Lea  
Meutem Kamtchueng, Lozzi Martial  
Moessner, Lilo  
Mondorf, Britta  
Mäkinen, Martti  
Neumaier, Theresa  
Nevala, Minna  
Nevalainen, Terttu  
Nishimura, Yoshito  
Njende, Ngum Meyuhnsi  
Nokkonen, Soili  
Onysko, Alexander  
Palander-Collin, Minna  
Paradis, Carita  
Perek, Florent  
Pérez-Guerra, Javier  
Peters, Pam  
Pettersson-Traba, Daniela  
Pichler, Heike  
Pöldvere, Nele  
Popelikova, Jirina  
Popoola, Olumide  
Pradhan, Manali  
Ranaweera, Mahishi  
Río-Fernández, Raúl  
Ratia, Maura  
Rausova, Veronika  
Rautionaho, Paula  
Reichelt, Susan  
Renouf, Antoinette  
Reski, Nico  
Rice-Whetton, John  
Riihimäki, Jenni  
Romasanta, Raquel P.  
Ronan, Patricia  
Rudanko, Juhani  
Ruohonen, Juho  
Rüdiger, Sofia  
Rühleman, Christoph  
Rørvik, Sylvi  
Röthlisberger, Melanie

University of the Balearic Islands  
Linnaeus University  
University of Birmingham  
Charles University  
University of Bremen  
Universidad de La Rioja  
University of Sheffield  
University of Eastern Finland  
University of Maroua  
RWTH Aachen University  
University of Mainz  
Hanken School of Economics  
University of Regensburg  
University of Tampere  
University of Helsinki  
Nagoya University  
Katholieke Universiteit Leuven  
City of Helsinki  
University of Klagenfurt  
University of Helsinki  
The Open University  
University of Birmingham  
University of Vigo  
Macquarie University  
University of Santiago de Compostela  
Newcastle University  
Lund University  
Charles University  
University of Birmingham  
Montclair State University  
University of Kelaniya  
University of Santiago de Compostela  
University of Helsinki  
Charles University  
University of Eastern Finland  
Lancaster University  
Birmingham City University  
Linnaeus University  
University of Melbourne  
University of Tampere  
University of Vigo  
TU Dortmund  
University of Tampere  
University of Helsinki  
University of Bayreuth  
Philipps University Marburg  
Inland Norway University of Applied Sciences  
KU Leuven
Schilk, Marco  
University of Hildesheim

Schmidt, Christa M.  
RWTH Aachen University

Schneider, Gerold  
University of Zurich

Schneider, Ulrike  
University of Mainz

Schulz, Ninja  
University of Würzburg

Schutz, Natassia  
Université catholique de Louvain

Schweinberger, Martin  
Universität Hamburg

Schützler, Ole  
University of Bamberg

Sebestova, Denisa  
Charles University

Seoane, Elena  
University of Vigo

Siirtola, Harri  
University of Tampere

Silvennoinen, Olli O.  
University of Helsinki

Skaffari, Janne  
University of Turku

Smith, Adam  
Macquarie University

Smith, Nick  
University of Leicester

Sotillo, Susana  
Montclair State University

Stange, Ulrike  
Mainz University

Steih, Niklas  
University of Innsbruck

Stormbom, Charlotte  
Åbo Akademi University

Ström Herold, Jenny  
Linnaeus University

Sugiura, Masatoshi  
Nagoya University

Suomela, Jukka  
Aalto University

Szmrecsanyi, Benedikt  
KU Leuven

Säily, Tanja  
University of Helsinki

Sönnig, Lukas  
University of Bamberg

Taavitsainen, Irma  
University of Helsinki

Tagliamonte, Sali  
University of Toronto

Tamaredo, Ivan  
University of Santiago de Compostela

Tyrkkö, Jukka  
Linnaeus University

van Rooy, Bertus  
North-West University

Vartiainen, Turo  
University of Helsinki

Vlasova, Ekaterina  
University of Helsinki

Wagner, Susanne  
JGU Mainz

Wallis, Sean  
Survey of English Usage

Wardell, Andrew  
Bloomsbury Publishing

Weisser, Martin  
Guangdong University of Foreign Studies

Werner, Valentin  
University of Bamberg

Westphal, Michael  
University of Münster

Wiemeyer, Leonie  
University of Bremen

Wolk, Christoph  
University of Giessen

Woolford, Kaleigh  
Newcastle University

Wright, David  
Nottingham Trent University

Xu, Xiuling  
Beijing Foreign Studies University

Yanagi, Tomohiro  
Chubu University