English Language and Linguistics, page 1 of 7.

Review

doi:10.1017/S1360674323000424

Susanne Flach and Martin Hilpert (eds.), *Broadening the spectrum of corpus linguistics: New approaches to variability and change* (Studies in Corpus Linguistics 105). Amsterdam and Philadelphia: John Benjamins, 2022. Pp. vi + 321. ISBN 9789027212665.

Reviewed by Laetitia Van Driessche, University of Zurich

The edited volume *Broadening the Spectrum of Corpus Linguistics* aims to bring together new perspectives, new datasets and new or revised tools that allow for different approaches to old discussions and materials. In the introduction, the editors emphasise the importance of the fact that the papers in the volume come from the 2019 ICAME conference, as just like the conference, they come from a wide variety of different fields within both synchronic and diachronic research (phonetics, computer linguistics, research on English as a Foreign language…), with the common denominator being the use of corpora. The volume consists of the introduction followed by ten papers, which are divided into three parts, namely 'New perspectives', 'Revisiting old debates' and 'Refinements and innovations'. The first two parts contain chapters that would fit well in either part: Olaf Mikkelsen and Stefan Hartmann's chapter (in 'New perspectives') revisits the well-researched future construction, and the focus of both Gaëtanelle Gilquin's and Gerold Schneider's chapters (placed in 'New perspectives' and 'Revisiting old debates' respectively) is to discuss the advantages and disadvantages of new corpora. The final part is the most technically dense, with a lot of descriptions of coding and statistical analyses.

Olaf Mikkelsen and Stefan Hartmann's 'Competing future constructions and the Complexity Principle: A constructive outlook' (pp. 8–39) is the first chapter in 'New perspectives'. The authors test the validity of Rohdenburg's (1996) Complexity Principle for the future construction by a comparison between English and Norwegian. The principle states that the more explicit form (*BE going to* for English and *kommer til å* for Norwegian) will be preferred in cognitively more complex environments (e.g. negation). The alternative that the authors posit for the Complexity Principle is a difference in meaning (or more specifically in intentionality vs prediction) between *BE going to* and *skal* on the one hand, and *vil* and *kommer til å* on the other (*will* fits into both categories). Using binary logistic regression, they find that the Complexity Principle cannot account for the variation in the Norwegian data, which casts doubt on its explanatory power for English as well. The authors themselves mention the limitations of their research, namely the exclusion of an 'animacy' predictor (which seems important in a discussion on the distinction between intentionality and

prediction) and the inclusion of both *BE going to* and abbreviated versions (e.g. *gonna*) in the same category of explicitness.

In her chapter on 'Diachronic learner corpus research: Examining learner language through the lens of time' (pp. 40–67), Gaëtenelle Gilquin presents a diachronic corpus for EFL (English as a Foreign Language) speakers with a French L1. She argues convincingly that a diachronic overview of Learner Englishes, alongside longitudinal studies of specific speakers, could prove informative (e.g. as regards the question whether learners – and thus perhaps teachers – have become more 'Americanised' over time). For her analysis, she compares the original French subcorpus of the *International Corpus of Learner English* (ICLE), which dates from 1991–3, with a corpus she calls ICLE-FR + 25, which spans the period 2016–19 (i.e. 25 years later). As the new corpus also contains untimed essays (which was not the case in the original corpus), she makes sure to include comparisons with only the timed sections. She finds that there has been a shift towards American words and expressions (e.g. *flat* vs *apartment*, *have a shower* vs *take a shower*) and a trend of colloquialisation in the decrease of *no*-negation to the advantage of the more informal *not*-negation (e.g. *no longer* vs *not any longer*). The core modals are generally used less in the new corpus, with the exception of *can*, which shows a sharp increase in usage. Finally, in line with the evolution in ENL (English as a Native Language) varieties, there has been an increase in bare infinitival complementation after HELP. Gilquin acknowledges that EFL varieties are different from other varieties in that they are not acquired naturally and that language change thus works differently. Instead, she offers several other factors that may have led to the changes: the norm that is followed by the EFL varieties may have changed, the ways in which students come into contact with English has changed over time (e.g. via Netflix), and teachers and school curricula may have changed what they see as the norm. The final factor she mentions is less ideal: she discusses how teachers at the Université Catholique de Louvain (one of the Belgian universities in the corpus) have used findings from the old ICLE version to point out common mistakes to their students, and that this may have led to students avoiding certain types of mistakes because they were pointed out so often by their teachers. This means that the students from Université Catholique de Louvain may not be entirely representative of EFL learners with a French L1 elsewhere. This also raises the question whether 'university undergraduates specialising in English' (Granger 2003: 539) are a reliable representation of EFL speakers in general.

The final chapter in 'New perspectives' is Marco Schilk and Lena Pickert's 'Rhoticity in Southern New Zealand English: An acoustic analysis of the QuakeBox database' (pp. 68–89), which looks at the potential of a corpus that was not created for linguistic purposes for the analysis of New Zealand English (NZE). Their corpus is the 'UC QuakeBox Project' (p. 70), which consists of testimonies of the Canterbury earthquakes. The authors use these to study rhoticity in southern New Zealand (Otago and Southland), an area which has retained this feature due to a difference in colonial history from the northern parts of the country. The authors provide a very clear description of the formants they looked at and how these contribute to rhoticity, which

https://doi.org/10.1017/S1360674323000424 Published online by Cambridge University Press

makes the chapter accessible to any readers who do not have an extensive knowledge of phonetics. Unfortunately, the QuakeBox only contains eight participants from the areas under study (and not all of these participants have lived exclusively in the south). They find that southern speakers show a trend towards rhoticity and that the degree of rhoticity correlates with age: older speakers have a higher degree of rhoticity. In some cases, however, there seems to be some collinearity, in that the older speakers also came from more rural areas. However, the authors are aware of the limitations of the study: they describe it as a 'pilot study' (p. 84) and state that they would like to add several improvements (e.g. the addition of speaker variation in the cut-off point for rhoticity). The focus of this chapter is therefore more on the potential of the corpus for further study. The corpus definitely has potential, although future research may want to extend the speakers they include to the northern areas as well.

The next part, 'Revisiting old debates', begins with Elen Le Foll's study on '"I'm putting some salt in my sandwich" – The use of the progressive in EFL textbook conversation' (pp. 92–131). The author compares the form, use and collexemes of the progressive construction in the Spoken BNC2014 with forty-three French, German and Spanish textbooks from different stages in the school curriculum. She finds that, contrary to previous research, 'progressives are not overrepresented in EFL textbook conversation; if anything, they are, in fact, underrepresented compared to native BrE conversations' (p. 121). She finds several differences in frequency and use between BrE conversations and textbook conversations, but argues that in several cases this makes sense: the fact that textbooks start out with the present progressive and only introduce the past and perfect progressives at more advanced stages is justified by the much higher frequency and contexts of use of the former tense. In other cases, such as the exclusion of the use of the progressive in 'repeated actions or states' (p. 113) and the underrepresentation of progressives with stative verbs, she believes changes to textbooks might be necessary.

In the chapter 'Determinants of exaptation in Verb–Object predicates in the transition from Late Middle English to Early Modern English' (pp. 132–71), Javier Pérez-Guerra uses random forest and logistic regression analyses to look at predictors for the SVO vs SOV word order from Late Middle English to Early Modern English. He starts the chapter with a concise overview of typical and atypical word orders from Old English to Early Modern English. The chapter focuses on the period where V2 was losing ground and the pre-verbal slot was not yet claimed by the syntactic role of the subject. On the basis of 'Variability-based Neighbour Clustering' (Gries & Hilpert 2008), Pérez-Guerra divides this period into four subperiods: M4 (1442–79), E1 (1500–69), E2 (1570–1639) and E3 (1640–1710). He describes the predictors under study at length, as well as his reasons for leaving out or merging several factor levels (mostly because of collinearity). His findings show that the SOV word order in M4 and E1 is still conditioned by factors associated with pragmatics and processing, namely the principles of end-weight and given-new. In E1, genre has also become an important predictor. In E2 and E3, on the other hand, the choice between the word orders was based (almost) purely on genre. Pérez-Guerra concludes that there has been a shift from

https://doi.org/10.1017/S1360674323000424 Published online by Cambridge University Press

'systematic predictors … to an essentially stylistic alternative' (p. 164). He briefly touches on the term 'exaptation', but since it is mentioned in the title of the chapter, it would have been interesting to have had a longer discussion about how the variation discussed in the chapter is an example of this phenomenon.

Gerold Schneider's chapter on 'Recent changes in spoken British English in verbal and nominal constructions' (pp. 172–95) looks at the Spoken BNC2014 corpus and compares this with the spoken section of the original BNC (which he calls BNC1994). He looks at changes in word frequencies in 'the verbal domain' (p. 175, more specifically aspect and voice) and in noun compounds (as a proxy for neologisms). He finds changes for all three categories. For his analysis of word frequencies, the author looks at differences in gender and class, and compares his findings for the BNC2014 with previous work on BNC1994 (Rayson *et al*. 1997), although differences in normalisation processes and in transcription practices make comparison difficult. His findings indicate that there is a decrease of swearwords in men and lower social classes and that women lead the increase of quotative *like*. His conclusion that the gender differences he finds in the BNC2014 (e.g. a higher use of pronouns and adverbs of appreciation in women) corroborate Tannen's (1991) claim that women and men live in 'different worlds' (p. 181) and that 'women see the world more as a network of connections and social relationships' (p. 185) seems a bit of a stretch. For the verbal domain, he finds that *get*-passives and progressive forms have increased, but that, contrary to what would be expected from colloquialisation processes, *be*-passives have increased as well. Finally, he finds an increase in noun compounds, which is mostly due to the increasing use of terms describing (new) technology. However, the type–token ratio shows that the increase is not due to an increase in creativity.

The final chapter of this part is Samuel Bourgeois' '"Oh yeah, one more thing: It's gonna be huge" – On the use of *oh yeah* in journalistic writing' (pp. 196–225). Bourgeois discusses the discourse marker *oh yeah* in the context of a broader change in journalistic practices in the second half of the previous century, namely the 'blurring [of] the border between … written and … oral language' (p. 198). For discourse markers, this led to a rise in intersubjective functions. He finds that *oh yeah* is not common in general and that despite a strong increase from the 1980s to the 2000s (*Corpus of Historical American English*), its use started declining in the 2010s (*News on the Web* corpus, NOW). He finds several interesting uses, the most common of which are 'faux spontaneous additives, faux utterance launchers' and, in the NOW corpus, 'meta-textual comments' (p. 214). He describes each category at length. The first category refers to those cases where the journalist uses *oh yeah* as a pretend (faux) afterthought in a list. The faux utterance launchers are used to pretend to react to a remark by the reader. The meta-textual comments are reactions to 'imagined reactions of the readership' (p. 216). In all cases, the discourse marker is used to have pretend conversations with the readers, making the article seem more like dialogue. The author emphasises that *oh yeah* and its innovative functions are used in the 'soft news' sections of journalism, e.g. the entertainment sections (p. 217). The chapter would have benefited from a more extensive methodology section, as Bourgeois does

https://doi.org/10.1017/S1360674323000424 Published online by Cambridge University Press

not discuss in detail how he annotated his data for function and whether there were any unclear cases.

The final part, 'Refinements and innovations', begins with Nathan Dykes, Philipp Heinrich and Stephanie Evert's chapter on 'Retrieving Twitter argumentation with corpus queries and discourse analysis' (pp. 228–55). The authors present an impressive technique to retrieve arguments about Brexit in Twitter discourse. They compare tweets from before (Brexit2016) and after the Brexit referendum – but before the original date that Brexit was supposed to take place (Brexit2019). They look for abstract patterns, such as '$experts $say X' (p. 242) where the 'experts' and 'say' slots can have several instantiations. This allows them to see, for example, which experts Twitter users cite and whether this has changed before and after Brexit. Due to the huge amount of information that their technique provides, it is not straightforward to find general trends. The results section would have benefited from more examples with more context in order to feel less overwhelming. The authors find that the use of expert citations has decreased and that, instead, it has become more likely for Twitter users to give direct suggestions. In their case study on the *as NP I VP* pattern (e.g. '*as a normal human, I can accept a compromise*', p. 247), they also find that the purpose of the NP slot has shifted away from positioning the user as an expert. They conclude that tweets from 2019 '[reflect] a more nuanced discourse' (p. 252) in which users generally have a more negative attitude towards Brexit. The chapter shows that the authors' technique is incredibly useful for the study and tracking of changes in argumentation patterns, which makes it all the better that they have made their code publicly available.

In 'MuPDAR for corpus-based learner and variety studies: Two (more) suggestions for improvement' (pp. 256–83), Stefan Th. Gries discusses two refinements to MuPDAR(F) ('Multifactorial Prediction and Deviation Analysis using Regression/Random Forests', p. 259). The chapter is also a good introduction to MuPDAR(F) and its usefulness for researchers who want to compare ENL and ESL/EFL varieties and who want to see more specifically in which contexts the latter group deviates from choices made by native speakers. This is done by using two regression analyses or random forests, one that predicts which choice the native speaker would have made or preferred, and one that compares this to the actual choice made by the ESL/EFL speaker. The improvements Gries discusses in this chapter are firstly a way to include more dependent variables (i.e. 'multinomial alternations', p. 262) and secondly a way to solve how, in many cases, the independent variables do not have enough data points for each level or that 'very few levels … or very small value ranges … account for most of the data' (p. 262), making robust data analysis difficult. The first improvement is made by including 'multi-class logloss' (p. 266), which are values that indicate the goodness of fit of the first regression analysis or random forest, and using these values as the dependent variable in the second regression analysis or random forest. The second improvement consists of a 'casewise-similarity approach' (p. 275). In this approach, the researcher calculates values that indicate the degree of similarity between ENL and EFL/ESL speakers and then calculates the mean distance of these values. This in turn allows one to calculate predictions of the choices speakers would have

https://doi.org/10.1017/S1360674323000424 Published online by Cambridge University Press

made, which can be used to fill in those cells for which no data exists. This chapter is extremely technical in nature and asks for at least a basic understanding of logistic regression and/or random forest analyses as well as statistical analysis more generally.

In 'A data-driven approach to finding significant changes in language use through time series analysis' (pp. 284–317), Andrew Kehoe, Matt Gee and Antoinette Renouf discuss a bottom-up way to uncover changes in word frequency across time. Their *WebCorp Linguist's Search Engine* (WebCorp LSE) provides an interface for tracking language change in a newspaper corpus that includes articles from the *Independent* and the *Guardian* spanning the period 1984–2017 (the chapter focuses on the period 1989–2017). The authors describe three types of language change: 'trends', 'seasonality' and 'sudden jumps' (p. 290). The first type refers to a consistent increase or decrease in word frequency, seasonality refers to an increase in word frequency that is purely seasonal, and sudden jumps refer to those increases in word frequency where a word becomes popular very suddenly. The authors discuss the tests they use for each type and provide a nice way for users of WebCorp LSE to visualise the language change under study. They also discuss a way to visualise all types of language change for one specific word in one graph. In applying these techniques to their newspaper corpus they find that newspapers have become less formal and that they have started giving more 'live' news (hence a decrease in the word *yesterday*). They also find that there has been a shift in how newspapers refer to 'specific groups within society' (p. 312, e.g. from *gays* to *gay people*). The authors make it clear that this is still a work in progress and that they plan on fine-tuning several of the tests they use.

An issue that comes up several times in this volume is the continuity between two corpora when a newer corpus has been created for the comparison with an older corpus (ICLE-FR + 25 in the case of Gilquin and BNC2014 in the case of Schneider). Should corpus compilers learn from mistakes made in the older corpus and change the design of the new corpus, or should they strive to keep the two corpora as similar as possible? The chapters in this volume provide a good argument for the second option. Finally, although the chapters differ in which topics and research fields they address, Susanne Flach and Martin Hilpert find a common thread between them and present a volume that is more coherent than it may appear at first glance. The editors emphasise how all the chapters bring something new or improved to their research area and to the field of corpus linguistics in general.

Reviewer's address:
English Department
University of Zurich
Plattenstrasse 47
CH-8032 Zurich
Switzerland
laetitia.vandriessche@es.uzh.ch

https://doi.org/10.1017/S1360674323000424 Published online by Cambridge University Press

References

Granger, Sylviane. 2003. *The International Corpus of Learner English*: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly* 37 (3), 538–46.

Gries, Stefan Th. & Martin Hilpert. 2008. The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora* 3(1), 59–81.

Rayson, Paul, Geoffrey N. Leech & Mary Hodges. 1997. Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the *British National Corpus*. *International Journal of Corpus Linguistics* 2(1), 133–52.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2), 149–82.

Tannen, Deborah. 1991. *You just don't understand: Women and men in conversation*. New York: Ballantine Books.

(Received 11 July 2023)