DE GRUYTER OPEN

Review in ICAME Journal, Volume 39, 2015, DOI: 10.1515/icame-2015-0012

Lieven Vandelanotte, **Kristin Davidse**, **Caroline Gentens** and **Ditte Kimps** (eds.). *Recent advances in corpus linguistics. Developing and exploiting corpora.* Amsterdam/New York: Rodopi. 2014. ix + 349 pp. ISBN13 9-789-042-038714. Reviewed by **Leonie Wiemeyer**, University of Bremen.

The conference volume *Recent advances in corpus linguistics. Developing and exploiting corpora* is a collection of peer-reviewed papers presented at the 33rd ICAME conference held in Leuven, Belgium, in 2012. The conference provided a forum for the discussion of the importance and implications of careful corpus compilation and of sensible data analysis for corpus linguistics in general. It put an additional focus on the intersections between corpus linguistics and other fields of (linguistic) research. The present collection of fifteen studies is divided into three thematic sections. The first part, "Corpus development and corpus interrogation", consists of five studies devoted to the design of corpora and related tools. In the second part, "Specialist corpora", five studies focussing on the interrogation of corpora designed for various specific linguistic research questions are presented. Finally, "Second language acquisition" contains five articles reporting on corpus research of English as a foreign language (EFL) and English for academic purposes (EAP).

In their introduction, **Lieven Vandelanotte, Kristin Davidse, Caroline Gentens,** and **Ditte Kimps** briefly state the overarching focus and goals of the volume and provide a summary of the contents and findings of each paper. It may be reverted to by readers interested in a more detailed overview of each study than can be provided here.

Part 1. Corpus development and corpus interrogation

The corpus of Letters of Artisans and the Labouring Poor (LALP) introduced by **Anita Auer, Mikko Laitinen, Moragh Gordon,** and **Tony Fairman** fills a gap in corpus availability by allowing for the analysis of the writing of the lower classes of the 18th and 19th centuries and thus for a new perspective in corpus-based historical linguistic research. The coding process, which is based on the

transcription codes used in the *Helsinki Corpus of English Texts* (Kytö 1996), as well as the handling of problematic cases, is made highly transparent and illustrated by images. This should be of great usefulness to users of this corpus as well as to fellow corpus compilers. It will be interesting to see how the challenges of normalisation are tackled once the project advances to the next stage. The interdisciplinary approach to corpus compilation is a fine example of how insights from other disciplines can be of benefit and of how corpora can be effectively designed so as to be of use for other specialists.

Joan S. Beal and **Ranjan Sen**'s paper proposes a database of 18th century English phonology which would allow researchers to directly compare a variety of phonological descriptions of Late Modern English and thus provide a unique resource for the study of phonological variation and change in this period. The authors emphasise its usefulness by presenting a study on the variation in the pronunciation of <wh> and the /hw ~ w/ contrast based on 18th century pronunciation dictionaries which considers geographical, lexical, and phonological factors. The study documents the usefulness of the proposed database, though an illustration of corpus design, i.e. of the envisaged structure and search functions, would have been useful here.

In their article, **Gregory Garretson** and **Henrik Kaatari** outline the semi-automated 'shared evaluation' approach in which a computer is used in the first step to sort and classify data, which is then, in a second step, reviewed by a human to ensure its accuracy. Their article doubles as an introduction to the use of their tool, System for Variable Extraction of Patterns (SVEP), which implements the method. It is characterised by automatic classification based on scores assigned to each token and by batch searching which prevents tokens being assigned to several patterns. The 'shared evaluation' method opens up exciting new avenues of corpus investigation and decreasing manual analysis time. The case study makes the programme accessible even for users with no experience with command lines. It would be interesting to answer the question of how the token scores might be calculated in studies of other structures, especially those without intervening material.

Marco Schilk compares currency-annotated PoS-tagged ICE corpora to explore register variation in four native varieties and five institutionalised second language varieties of English. His analysis is based on text-type clusters which were empirically defined based on their statistically significant similarities in lexis and currency. The study is based on several clever ideas of how to exploit corpora in new ways. Schilk's approach is sophisticated and comprehensive and reveals important problems in terms of corpus design and comparability. It will be a fruitful endeavour for future studies to determine if Schilk's

empirically derived text-type clusters based purely on lexical properties have advantages over a functional distinction, if it can be generalised to other ICE subcorpora, and if the approach has to be expanded to other features for a more complete picture.

An exciting way to exploit corpora – for the investigation of historical changes in stress patterns – is showcased in **Frank Zumstein**'s contribution. His survey of data from a lexico-phonetic corpus containing dictionary entries and detailed comparison of historical stress patterns to contemporary ones reveals stress changes as caused by a process of regularisation. In this process, those variants which do not adhere to the Normal Stress Rule are dispreferred. His argumentation is supported by ample examples from the corpus and orthoepic dictionaries, providing a very specialised account of the development from variant to main stress patterns in a variety of word groups. Zumstein's analysis is well-documented and immensely insightful, and certainly a valuable diachronic perspective in this less investigated field of corpus linguistics.

Part 2. Specialist corpora

In an investigation of ICE Philippines and "Phil-Brown", a corpus of Philippine English parallel in design to the Brown corpus, **Peter Collins, Xinyue Yao,** and **Ariane Borlongan** shed light on the diachronic changes in relativisation strategies in Philippine English in a process of alignment with BrE and, more importantly, AmE. The authors skilfully make use of the possibilities afforded by short-term diachronic corpus linguistics in combination with grammatical research in a commendable approach which focuses not only on the development of relative clauses in PhilE as compared to the core varieties, but also on the speed of this development and the alignment tendencies across genres. The study, which is preceded by a very well-written theoretical part, is an important point of departure for studies of *that*-relativisation in other ESL varieties.

Marco Schilk and **Marc Hammel**'s paper contributes to the study of inter-variety variation in tense and aspect, in this case the use and distribution of the progressive. Their cluster analysis allows them to identify groups of varieties which are significantly similar to each other, which gives the study a fruitful perspective of 'overuse' and 'underuse' between these groups and not just between individual South Asian varieties and BrE. In the analysis, Schilk and Hammel establish that the extension of the progressive to habitual actions and the combination of two present participles can be identified as majorly responsible for variation. They also find that – contrary to expectation – stative verbs rarely occur in the progressive in ESL varieties. The qualitative analysis could

have, however, additionally addressed the potential influence of subcorpus size and composition and considered further semantic, contextual, topic-induced or functional factors which might be at play.

Antoinette Renouf explores neologisms in a large corpus of UK news texts and how they are bolstered by other topical words and phrases, which eventually together form a collocational, register-like network. In her article, Renouf expertly unveils the multi-dimensionality and diachronicity of neologistic activity in a series of case studies accompanied by illustrative excerpts from her data. She shows that the invention of lexical neologisms frequently leads to the invention of other new words which often occur in their vicinity, the result of which is an inter-collocational network. She also observes that the frequencies of neologisms increase steadily during the currency of the cultural aspects they denote, but then often fall out of use. The study combines collocation analysis, diachronic frequency and extensive qualitative analyses and is both linguistically and socio-historically captivating.

The senses of *amid(st)* and *among(st)* as evident in their Norwegian and Swedish translation equivalents are traced by **Thomas Egan** and **Gudrun Rawoens**. Their contrastive approach, accompanied by detailed analyses and statistics of correspondences and divergences in translation, yields noteworthy variation in the choice of forms used as equivalents for both prepositions in the two target languages. By taking into account the predication types that occur with the prepositions, their study offers new insights into the semantics of *amid(st)* and *among(st)*. Some interesting questions for further research might be which semantic predications are generated in the translations and whether the original predications and senses are kept intact in the target languages.

Another contrastive study is presented by **Kerstin Kunz** and **Ekaterina Lapshinova-Koltunski**, who compare English and German cohesive conjunction strategies in an investigation based on complex multi-layered annotations of the written components of the *German-English Contrasts in Cohesion* (GECCo) corpus. The parallel and comparable corpus design (original texts in German and English and corresponding translations) enables a direct comparison of the cohesive conjunction frequencies in original and translated texts in both languages. Through this, inter-language and translation-induced differences in the inventories of conjunctive devices and in their properties are made visible. Despite the fact that this is an initial study in a larger project, it offers important insights into the language-dependent variation in the use of cohesive conjunctions.

Part 3. Second language acquisition

Katrien L. B. Deroey focuses on the lexico-grammatical means of marking relevance and lesser relevance employed in lectures. The markers of (lesser) relevance were identified in a combination of corpus-driven and corpus-based analyses of lecture transcripts from the British Academic Spoken English (BASE) corpus. The most frequent markers of relevance, which are used by lecturers to structure and evaluate their lectures, are imperatives and structures containing metalinguistic nouns. Markers of lesser relevance, such as note-taking directives, are often multifunctional and rely quite heavily on pragmatic interpretation. The principles of categorisation are described transparently, embracing the necessity of replicable linguistic research advocated by many (see, for example, Stubbs 2001: 123ff.). Deroey made a laudable decision to include enough detail to make the patterns accessible for lecturers. Her findings have practical implications for future experimental studies, especially in EAP and EFL contexts.

False friends, a key topic in EFL research, are explored by **María Luisa Roca-Varela** in the written and oral production of Spanish learners of English in order to fill a gap in the corpus-linguistic investigation of these lexical items. The study has practical implications in that it shows that false friends are a problem even for advanced learners. More detailed quantitative analyses could have been conducted, for example in order to determine which false friends contribute the most to the overall incorrect usage, and some of the generalisations drawn from the qualitative analysis of some low-frequency items are debatable. The influence of learner variables and task specification on the production of false friends are, in my opinion, likely to provide important additional insights for future studies.

Thomas Gaillat, Pascale Sébillot, and **Nicolas Ballier** test the automatic classification of non-native and native uses of demonstrative pronouns in learner writing. In two experiments using NLP tools, the authors employ native speaker and learner data to first investigate the features that lead to the distinction between expected and unexpected use of demonstratives, and second to derive the features that lead to the selection of an incorrect form in learner English. By analysing the context of the pronouns, the authors can explain the learners' choices in reference to learner-specific features and go beyond a mere description of deviations from the native speaker norm. At this initial stage, the performance of the classifier still leaves room for improvement, but given more accurate results on larger samples, this approach could prove an important further step towards reliable automatic error tagging of learner corpora.

A pilot study into measuring students' progress in spoken learner English over two years is at the heart of **Monique van der Haagen, Pieter de Haan,** and **Rina de Vries**' contribution. The research objective, as it is often the case in applied linguistics, arose from practical necessity, as the authors' university did not assess students' entry level and progress in English language proficiency. The authors analysed recordings of unprepared speech recorded at three different times during the students' first year at university. The analysis of crude measures yielded significant increases in speech rate, mean word length and type-token ratio between the first recording and the last one two years later, while lexical density and sophistication did not provide conclusive evidence but pointed towards increased syntactic complexity. In a short but well-documented study, the authors give an insightful perspective on the usefulness and lack thereof in establishing students' language proficiency.

The final article in this volume is by **Pieter de Haan** and **Monique van der Haagen** and documents a longitudinal study of the syntactic development in the writing of very advanced Dutch learners of English. It addresses the use of sophisticated language as reflected in the choice of specific word classes and their combinations in first-year writing assignments. The study reveals inter-learner variability in terms of general proficiency and word-class distribution, and overuse of personal pronouns and speech-like features by learners, some of which may have resulted from the fact that the tasks did not necessarily call for the use of academic language. The authors also observed dramatic changes in the use of tag bigrams, which point towards a growing awareness of the syntactic structures typical of academic writing. The corpus data is diligently quantified; yet it seems desirable that the study be expanded to a larger set of texts and compared to a larger reference corpus for a more generalisable account.

The ICAME 33 conference volume contains an array of high-quality studies in some of the exciting areas of linguistic research where corpora can provide new and meaningful insights. As a reflection of the character of the ICAME conferences, the scope of the contributions is very broad – from introductions of new corpora and new methods of corpus interrogation to studies which are corpus-linguistic only in the widest sense: from phonology to neology to applied linguistics.

A clear strength of the volume is in the choice of studies with very different research foci, which complement each other very well and provide points of departure, tools and methods for interdisciplinary research and for practical applications. The first and second part contain studies using a variety of different corpora and employing diverse methods of corpus interrogation and statistical evaluation. They provide the reader with valuable insights into the impor-

tance of intelligent corpus design. In the third part, the intersections of corpus-linguistic and educational research are explored, highlighting the practical implications of such studies. A point of criticism is the subdivision into parts, which seems somewhat forced. Some articles from the first part, e.g. Auer *et al*.'s and Zumstein's, which are based on specialist corpora, may have well been allocated to the second and vice versa, and Deroey's contribution, included in part 3, is not a study in second language acquisition, though with practical implications for that field. Furthermore, a figure is curiously absent from the second paper and there are some incomplete sentences and missing commas in the volume which affect readability.

This volume edited by Vandelotte, Davidse, Gentens, and Kimps offers a collection of studies which make use of a wide range of quantitative and qualitative methods. They emphasise the added value and potential practical implications of corpus-linguistic methods in several linguistic disciplines. The studies should thus be of great interest to both fellow linguists and other experts alike: they are well-presented, well-balanced, answer central questions of corpus-linguistic research and, in many cases, break new ground for corpus-linguistic explorations at the crossroads with other disciplines.

References

Kytö, Merja. 1996. *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts*. 3rd ed. Helsinki: University of Helsinki, Department of English.

Stubbs, Michael. 2001. *Words and phrases*. Oxford: Blackwell.