O'Keeffe, A. and Farr, F. (Eds) (2011) “Introduction”. *International Journal of Corpus Linguistics*, [special issue on “Applying Corpus Linguistics”]. 16(3): 151-156.

Introduction

Fiona Farr and Anne O’Keeffe

University of Limerick / Mary Immaculate College, University of Limerick

This special issue brings together work in areas where corpus linguistics has been applied as a methodology. Many of these papers are developments of work presented at conferences and symposia of the *Inter-Varietal Applied Corpus Studies* research network. Our aim in this special issue is to present a snapshot of how CL has been taken up and applied to diverse research questions. This lauds the versatility of CL in its applicability to a wide range of areas while also posing new and interesting theoretical and practical challenges. The range of topics covered here is but a sample. Many other established areas have, to varying degrees, adopted and adapted CL approaches and methods, including literature and stylistics, lexicography as well as media, forensic, healthcare, workplace and other domains of discourse.

In tracing the development of corpus linguistics, technology has been the central factor in its growth and evolution. As hardware developed, more and more data could be stored on relatively small computer drives and servers. This meant that corpora could be as big as one wanted. The drive for larger and larger corpora was led by the field of lexicography in order to get the widest semantic range, coverage and contexts of use for as many words as possible in a language. As McCarthy & O’Keeffe (2010) note, the early COBUILD corpora were measured in tens of millions of running words, other publishing projects soon competed and pushed the game up to hundreds of millions of words and, by the middle of the first decade of the 21st century, we move into a billion running words of text, not to mention the endless potential of the entire world-wide-web as a corpus, with its trillions of words (see Lee 2010). However, concomitant with the ever-increasing size of general corpora, we see the development of many small context-specific corpora (i.e. fewer than one million words of data). This trend is driven by the technological developments which make it easier for individuals or research teams to gather data and also by the use of corpora in a broader range of research areas which require in-depth quantitative and qualitative analysis of smaller specialised amounts of data. This special issue, in particular, brings together work on smaller corpora from a range of areas and research frameworks which are availing of CL tools and methods to better understand empirical data from their field of research (areas such as pragmatics, language learning and spoken discourse analysis). In the application of corpus linguists to other areas, it raises a number of interesting issues for our field as well as enriching the areas to which CL has been applied. In the selection of papers for this special issue, we have tried to find balance between programmatic treatments and case studies in order to best display a range of approaches that have been taken in the application of CL. The papers that we have selected provide initial focus on the application of CL to a particular area or theoretical framework and then exemplify this using corpus data.

One of the greatest challenges in the development of corpora, especially in the case of spoken data, is how to comprehensively capture the context of use. Technology now enables the creation of multi-modal corpora, in which various communicative modes (e.g. speech, body-language, writing) can all be part of the corpus, all accessible at one go. This development means a moving away from spoken corpus linguists having to rely only on transcripts of a speech event. With multi-modal corpus tools, the speakers in video and audio stream will be aligned to the transcript and as Adolphs, Knight and Carter (this issue) illustrate, they can also be linked by GPS to location and even individualised recording headsets which show exactly what the speaker is looking at as they speak. Adolphs et al. note that the use of video enables a more informed examination of prosodic, gestural and proxemic features of the talk that occurs at a specific time and in a specific place. The development of their multi-modal corpus, Adolphs et al. note, was motivated by a shared interest in the analysis and associated tool development by computer scientists, applied linguists and researchers in the area of social psychology (see Knight & Adolphs 2008 and Knight et al. 2006, 2009), thus leading to advancement for all of these areas, as well as an enriching dividend for CL through this collaboration with other disciplines. Adolphs et al. push the boundaries of multi-modality by attempting to create a ‘heterogeneous’ corpus. These, they define as “emergent multi-modal datasets which comprise a variety of different records of everyday communication, from SMS/MMS messages to interaction in virtual environments and from GPS data to phone and video calls”. Essentially, they argue that by tracking a person’s specific (inter)actions over time and place, the analysis of such “ubiquitous” corpora fosters more detailed investigations of the interface between different communicative modes from an individual's perspective. They add that the compilation of such corpora may enable us to extrapolate further information about communication across different speakers, media and environments, helping to generate useful insights into the extent to which everyday language and communicative choices are determined by different spatial, temporal and social contexts.

This issue contains two papers which bring CL into contact with other analytical frameworks for the study of spoken discourse. Walsh, Morton and O'Keeffe explore the compatibility of CL with Conversation Analysis (CA) in the context of small group interactions in Higher Education contexts, using the *Limerick Belfast Corpus of Academic Spoken English* (LIBEL). They consider how the two approaches can be combined in an iterative process to account for features of spoken discourse at both micro (word and pattern) and macro (text) levels. Their process begins with CL, focusing on lexis and multi-word units in the data. They then use CA to highlight pertinent interactional features. This leads to an iterative process: from CL to CA, back to CL, using the more top down findings from CA to meet bottom up outputs from the corpus analyses. This approach to analysis allows for powerful insights because it helps account for the data from the perspective of the interaction as well as from the corpus data. In particular, it highlights the inter-dependency of words and patterns, utterances and text in the co-construction of meaning in context. In applying CL to CA, it brings into relief the limitations of each (see also O'Keeffe & Walsh forthcoming). Santamaría-García also looks at spoken corpora using a combined methods approach using data from the *Santa Barbara Corpus of Spoken American English* (*SBCSAE*) and the *Corpus Oral de Referencia del Español Contemporáneo* (*CORLEC*). She valorises an eclectic methodology for cross-linguistic comparison at the level of discourse by drawing on CL, CA and Discourse Analysis (DA). Obviously, this presents challenges and limitations which are discussed and exemplified. Among her insights from this eclecticism is the need for spoken corpora to include complete conversations, discourse annotation, sound files and detailed contextual information, or as she puts it, a step forward from "corpora of spoken language to discourse corpora". The development of multimodal corpora, as detailed by Adolphs et al., make this a leap rather than a step in terms of progress.

Another application of CL to a framework which we have included here straddles the fields of sociolinguistics and pragmatics, namely variational pragmatics, a concept coined by Schneider and Barron (see Schneider & Barron 2008). Variational pragmatics is a framework which specifically addresses the intersection of the fields of dialectology and pragmatics, where Schneider and Barron see a research gap that existed. According to Schneider & Barron (2008: 1), variational pragmatics "investigates pragmatic variation in (geographical and social) space". Variational pragmatics (VP) has as its primary concern how the choice of one pragmatic strategy over another encodes macro-social indices of region, socio-economic status, ethnicity, gender or age in everyday language use. Schneider and Barron work with discourse completion task data and Clancy, in this issue, makes a cogent case for the application of CL. He points out that CL offers a methodology which benefits variational pragmatic analysis in a number of ways. Firstly, the fact that most corpora are constructed to be representative of a particular language variety facilitates an accurate account of language-use differences across various social categories. Clancy provides a case study of CL working with VP in the form of a comparison between two corpora representing spoken language recorded in the home environment, one from a middle class Irish family and one from a family from the Irish Traveller Community. He shows that the variational distribution of the occurrences of hedges across these two distinct cultural groupings differ due to socio-pragmatic factors. At a broader level, we can say that because pragmatic analysis relies heavily on context for its interpretation, the future availability of multi-modal data will have even greater application to VP research.

The issue includes two papers which relate to language teaching in specific contexts. One is in the context of English Language teaching and the other relates to the teaching of French. Rodgers, Chambers and Le Baron illustrate the use of a corpus of articles on biotechnology in French with university students of biotechnology. They discuss the issues involved in designing and creating an appropriate corpus for this specialised pedagogical context. They also evaluate the learners' reactions through questionnaires, semi-structured group interviews, and teacher observation. Their paper makes a valid contribution to the application of corpus linguistics to language teaching but, very importantly, it seeks feedback on how the learner views this method and its effectiveness as a learning tool. While much has been written about data-driven learning and the use of corpora in language teaching, without an evidence base for its effectiveness, its actual usefulness will go uncharted.

Rankin and Schiftner also look at the use of corpus data in terms of how it can inform language teaching in a university setting. They conduct a comparative interlanguage analysis of a specific class of complex and marginal prepositions in a range of L1 German learner corpora. Their analysis shows that, in native speaker English, prepositions in the semantic field of "reference" and "aboutness" are used in distinct structural and collocational environments, while the learner data shows a greater degree of interchangeability. Their analysis is drawn from a small "in-house" corpus of student writing, the Vienna *Database of English Learner Texts* (DELT) (cf. Rankin & Schiftner 2009) and they subsequently compare their findings with the German component of the *International Corpus of Learner English* (ICLE, Granger et al. 2002) and the *British National Corpus* (BNC Baby 2005). This gives a fine-grained insight into the specific areas in which learner production diverges from target native production. They point out that using a large native reference corpus permits the establishment of norms of native production, which can be used to inform the subsequent development of pedagogical materials to address the specific issues highlighted by the corpus study. Their work is a good example of CL being applied to solve a real pedagogical issue, where data is collected locally from learners and compared with other data, native and non-native, in order to inform practice. Rankin and Schiftner provide examples of the types of materials that resulted from their extensive research.

This issue, we hope, showcases an interesting sample of how CL has been and is being applied to a variety of areas and how it is complementing and influencing other methodologies. These papers also illustrate the trend towards the development of small localised corpora which meet specific user needs and which are highly concentrated in terms of richness of context and language use. In applying CL to all of these different and differing contexts, we have also learnt a lot. In each of these papers, challenges for CL are exposed. The washback for CL, in its wider application, is very positive because it means that we are forced to think of new solutions, either in terms of how we approach data or in terms of how we record, store or format it. We are forced to consider, using the most up-to-date technology, how we can better represent language and its context of use. We are forced to think about how we can better teach it. All of these challenges are important for a vibrant and evolving field such as CL.

Acknowledgements

In the compilation of this special issue, we are indebted to the many reviewers who we enlisted: Jozef Andor, Andrew Caines, Winnie Cheng, Karen Corrigan, Jane Evison, Lynne Flowerdew, Gaëtanelle Gilquin, Sylviane Granger, Michael McCarthy, Tony McEnery, Bróna Murphy, Kieran O'Halloran, Randi Reppen, Christoph Rühlemann, Klaus Schneider, Scott Thornbury, Paul Thompson, Ivor Timmis, Elaine Vaughan and Martin

Warren. Theirs is a time-consuming voluntary task, without which a journal cannot function. We are also sincerely grateful for the support, insight and help of the general editor, Michaela Mahlberg. We thank her for giving us this opportunity to showcase this aspect of CL. Fanie Tsiamita, Assistant Editor, cannot go without high praise: Fanie, thank you for your consistent and helpful guidance throughout the process. And finally thank you to those who contributed to this special issue, from the initial call for abstracts to the various rounds of selection. There are many more papers that we would like to have included but this is again a sign of the thriving and vibrant ongoing research in the application of CL.

References

BNC Baby, version 2. 2005. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Granger S., Dagneaux, E. & Meunier, F. 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Lee, D. Y. W. 2010. “What corpora are available?”. In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 107-121.

Knight, D. & Adolphs, S. 2008. “Multi-modal corpus pragmatics: The case of active listenership”. In J. Romeo Trillo (Ed.), *Corpus and Pragmatics*. Berlin/New York: Mouton de Gruyter, 175-190.

Knight, D., Bayoumi, S., Mills, S., Crabtree, A., Adolphs, S., Pridmore, T. & Carter, R. A. 2006. “Beyond the text: Construction and analysis of multi-modal linguistic corpora”. *Proceedings of the 2nd International Conference on e-Social Science*, *Manchester, 28 - 30 June.*

Knight, D., Evans, D., Carter, R. & Adolphs, S. 2009. “Redrafting corpus development methodologies: Blueprints for 3rd generation ‘multimodal, multimedia’ corpora”. *Corpora,* 4 (1), 1-32.

McCarthy, M. J. & O’Keeffe, A. 2010. “Historical perspective: What are corpora and how have they evolved?”. In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3-13.

O'Keeffe, A. & Walsh, S. Forthcoming. "Appropriate methodologies for investigating classroom discourse". In J. Cutting & B. Murphy (Eds.), *Spoken Corpora and Education.* Cambridge: Cambridge University Press.

Rankin, T. & Schiftner, B. 2009. "The *Vienna Database of English Learner Texts* – A resource for language research and teaching". Paper presented at *Sprachendidaktik: Der wissenschaftliche Nachwuchs im Dialog, University of Klagenfurt, 25 April.*

Schneider, K. & Barron, A. 2008. "Where pragmatics and dialectology meet: Introducing variational pragmatics". In K. Schneider & A. Barron (Eds.), *Variational Pragmatics: A Focus on Regional Varieties in Pluricentric Languages*. Amsterdam/Philadelphia: John Benjamins, 1-32.

Editors' addresses

Fiona Farr
School of Languages, Literature, Culture
and Communication
University of Limerick
Ireland

fiona.farr@ul.ie

Anne O'Keeffe
Department of English Language and
Literature
Mary Immaculate College
University of Limerick
South Circular Road
Limerick
Ireland

anne.okeeffe@mic.ul.ie