Utilising corpus linguistic technologies: Quantifications in conversation analysis

Kazuki Hata (Tokyo City University, Japan)

This study proposes an interdisciplinary approach utilising conversation analysis (CA) in accordance with corpus linguistic (CL) techniques. CA stands as a micro-analytic approach designed to investigate the structure of social interaction (Sacks, Schegloff & Jefferson, 1974). Although CA is well-known as a qualitative approach, it inevitably incorporates a quantitative feature of research in that investigations of interactional organisation typically derive from the collections of samples (Schegloff, 1993). Thus, CA studies can certainly benefit from using CL approach as a methodological tool, as findings in corpora help identify the remarkable pattern(s) that can allow further investigations through a qualitative approach (Walsh et al., 2011).In the CA field, however, an application of CL methodology has not been well-explored in terms of how CL methods can be utilised in alignment with CA disciplines, and in particular, the significance of integration between computer-aided techniques and systematic micro-analytic procedures. The present study highlights that an application of CL can extend a CA in a way that manual analytical practices cannot achieve.

Although it has been continuously suggested that actual pragmatic features of language are not well-pursued by purely quantitative CL due to its inherent limitation: a lack of contextual information (Adolphs, 2008, p. 6–7; Mautner, 2007, p. 65–66; Sinclair, 2008; Widdowson, 2000), language corpora help studies across different disciplines in many ways by attaining significant numbers of language samples (Carter & McCarthy, 2006; Mautner, 2007, p. 54; McCarthy, 1998; Sinclair, 2008, p. 30).Spoken corpora offer tremendous insights from the authentic data, which becomes a reliable departure point for further case-by-case investigations. A utilisation of CL techniques thus has potential to help identify language samples to be analysed in the strict 'data retrieval model' (Leech, 1991, p. 20) for qualitative studies. That is, CL in this sense is a powerful tool to highlight language-in-use that is in accordance with interactional properties of target token in a particular context; investigating not only which word is used but also how the token is used and recognised by participants in a given interaction(Mautner,2007, p. 54;Teubert, 2005, p. 8). Regarding a methodological combination, in particular between CL and CA, one prominent work is Walsh et al. (2011). They suggest a data-driven CL–CA approach, in which descriptive results (i.e. frequency lists) operate to be a good starting point for CA practices. They claim that the proposed CL–CA approach has a potential to provide detailed insights of interactions and bridge methodological limitation for each approach. However, one limitation in their current CL–CA approach is that the focus was considerably made on linguistic-centred aspects of interaction, providing partial proofs for the sequence organisational method of the participants and its systematic features on a turn-by-turn basis.

This work, utilising two corpora: the BNC Spoken Audio Sampler (Coleman et al., 2012) and Newcastle University Corpus of Academic Spoken English (NUCASE), tackles an application of CL techniques into strict 'ethnomethodological' CA perspectives on talk-in-interaction by suggesting two integrated approaches between CL and CA. The first approach considers the verbal production of guys in goal-

oriented interactions through the top-down CL–CA approach, wherein the research focus is characterised through a statistical result. A keyword list, generated from a comparison between the BNC and NUCASE data, identifies that a noun guys is predominantly found in NUCASE when compared to BNC Sampler, which becomes a base for a qualitative CA analysis. The following turn-by-turn exploration suggests that the token guys is used as an interactional resource to achieve a re-entry to the goal-oriented sequence from a diverted part of talk. In this top-down approach, the quantitative result can operate to identify a potentially distinctive and contextual feature of linguistic token used in social interaction; which can also be substitution for a manual 'unmotivated-looking' procedure in CA.

The second approach also deals with quantification in CA utilising CL techniques, yet is carried out in the opposite 'bottom-up' direction. This investigation is focusedon the turn-final placement of erm (or um): a token placed at a possible turn transition point with ambiguous syntactic information (Jefferson, 1983; Local & Kelly, 1986). During the CA analytical process, an annotation is made on the target phenomenon by the researcher, which allows the comparison for particular social actions in line with the production of the target token between different datasets: ordinary and institution talk. Interestingly, the quantitative features of erm are differently seen between those contexts, implying that a turn-final erm is context-dependant and a simple classification (e.g. filler, conjunctional or pragmatic marker) is insufficient to describe how the token is produced and recognised in social interaction. Unlike the top-down CL–CA procedure, this bottom-up approach can be reciprocal and can proceed to a further CA analysis when needed.

This study discusses the significance of quantification in CA studies made through applying CL techniques. Considering that organisation of talk in interaction is truly contextual, one would suggest that the linguistic patterns would not be able to be captured through truly quantitative procedures. Nevertheless, the utilisations of CL can provide clues for forthcoming case-by-case analyses, indicating the significant patterns for further investigations. Whereas the current project has not been complete, the results generated to date have already highlighted implications for such interdisciplinary approach.

References

Adolphs, S. (2008). *Corpus and context: Investigating pragmatic functions in spoken discourse*. Amsterdam: John Benjamins.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.

Coleman, J., Baghai-Ravary, L., Pybus, J., & Grau, S. (2012). *Audio BNC: The audio edition of the Spoken British National Corpus*. Phonetics Laboratory, University of Oxford.

Jefferson, G. (1983). On a failed hypothesis: 'Conjunctionals' as overlap vulnerable. *Tilburg Papers in Language and Literature*, 28, 1–33.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer, & B. Altenberg, (Eds.),*English corpus linguistics* (pp. 8–39). London: Longman.

Local, J., & Kelly, J. (1986). Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies*, 9(2), 185–204.

Mautner, G. (2007). Mining large corpora for social information: The case of elderly. *Language in Society*, 36(1), 51–72.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, 26(1), 99–128.

Sinclair, J. (2008). Borrowed ideas. In A. Gerbig, & O. Mason, (Eds.), *Language, people, numbers: Corpus linguistics and society* (pp. 21–42). Amsterdam: Rodopi.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), 1–13.

Walsh, S., Morton, T., & O'Keeffe, A. (2011). Analysing university spoken interaction: a CL/CA approach. *International Journal of Corpus Linguistics*, 16(3), 325–345.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1), 3–25.