Annotating a corpus of spoken English: the Engineering Lecture Corpus (ELC)

Academic staff and students are increasingly moving from country to country to receive and deliver academic lectures. However, although English is often used as a lingua franca in higher education, and although lecture topics and syllabuses for disciplines such as engineering and medicine tend to be similar around the world, it is likely that different cultural norms and expectations will result in different lecture styles and structures in different local academic contexts. This suggests that staff and students may need to adjust the way they deliver and receive lectures in unfamiliar academic contexts, and that they may benefit from corpus linguistic insights when making these adjustments.

The corpus annotation of features other than syntax and part of speech is extremely time-consumingand encumbered by questions of subjectivity (Meyer 2002; Leech 2005; Smith 2008).. Some spoken corpora such as the London-Lund Corpus (LLC) (Garside et al. 1997: 10) and the spoken component of the HKCSE business corpus (Warren 2004, Cheng 2004) have been manually encoded for prosodic features such as tone units, pitch and stress, but very few corpora have been annotated from a functional perspective, because of the labour intensive nature of such work, and because of the degree of interpretation it requires. A number of small written corpora have been marked up in terms of generic moves and steps (see, for example, Durrant and Mathews-Aydinli 2011), and classroom interaction in the *Singapore Corpus of Research in Education* (SCoRE) has been marked for pragmatic and pedagogical features (Peréz-Paredes and Alcaraz-Calero 2009), but as far as academic lectures are concerned, progress with pragmatic mark-up has been very slow. Young (1994) identified a sort of generic move structure in academic lectures, consisting of various 'phases', each with a different communicative function, and Maynard and Leicher (2007) experimentally tagged a small subcorpus of 50 MICASE transcripts for pragmatic features such as 'advice' and 'disagreement', but there does not seem to have been any prior attempt to mark up an entire corpus of lectures to reflect their structure or purpose. The largest British lecture corpus, the British Academic Spoken English (BASE) corpus (Nesi 2001), is only encoded for part of speech, pausing, and contextual information. The BASE corpus annotation follows TEI (Text Encoding Initiative) conventions so that it can be compared with other similarly encoded corpora, but TEI has not traditionally been used to signal the function of larger stretches of discourse, and appropriate coding strategies are still under development.

This paper describes an approach to the pragmatic annotation of the Engineering Lecture Corpus (ELC), which contains over 60 English-medium engineering lectures from across the world, currently including Malaysia, New Zealand and the UK (www.coventry.ac.uk/elc). The lectures are in the form of videos, transcripts and XML files encoded using traditional TEI methods, but also marked for a limited number of functions which shed light on the specific nature of lecture discourse. These pragmatic features include 'storytelling', 'housekeeping', 'summarising' and 'defining'. Attributes have been assigned to some of these functions; for example storytelling is divided into narratives of either 'personal' or 'professional' experience, and summarising is comprised of four subcategories which distinguish between reviewing and previewing material from current, past and future lectures. By encoding these features in the corpus we are able to compare their location and their relative frequency in lectures delivered by local lecturers in different cultural contexts.

As annotation of the ELC is yet to be finalised, results at this stage remain tentative. However, some interesting patterns are beginning to emerge. The use of 'defining', for example, appears to be consistent across the corpus, indicating that this function is fundamental to the English-medium engineering lecture regardless of cultural context. On the other hand, early findings suggest that the use of 'storytelling' occurs consistently in the

lectures from the UK and New Zealand, but infrequently and to different effect in those from Malaysia. Ultimately, if significant variation of this sort is identified a model of the lecture structure can be established. We also hope to be able to describe the linguistic features that realise the various typical purposes of lecture discourse. Our annotation system will be of interest to other corpus developers who intend to apply pragmatic mark-up, and our comparative findings will be of interest to EAP and ESP practitioners, staff developers, and all academics on the move.

Bibliography

Cheng, W. (2004) '//→ did you TOOK //↗ from the miniBAR// What is the practical relevance of a corpus-driven lanugage study to practitioners in Hong Kong's hotel industry?' In: U. Connor & Upton, T. A. (eds.) *Discourse in the Professions*. Amsterdam: John Benjamins. pp. 141-166

Durrant, P., & Mathews-Aydınlı, J. (2011) 'A function-first approach to identifying formulaic language in academic writing'. *English for Specific Purposes* 30 (1) 58-72

Garside, R. and Rayson, P. (1997) 'Higher-Level Annotation Tools'. In: Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman. pp.179-193

Leech, G. (2005) 'Adding Linguistic Annotation'. In: Wynne, M. (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books. pp. 17-29

Maynard, C. and Leicher, S. (2007) 'Pragmatic Annotation of an Academic Spoken Corpus for Pedagogical Purposes'. In: Fitzpatrick, E. (ed.) *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodophi. pp. 107-116

Meyer, C. (2002) *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press

Nesi, H. (2001) 'A corpus based analysis of academic lectures across disciplines'. In: Cotterill, J. and Ife, A. (eds.) *Language Across Boundaries*, London: Continuum Press. pp. 201-218

Peréz-Paredes, P. & Alcaraz-Calero, J. M. (2009) 'Developing annotation solutions for online data driven learning'. *ReCALL* 21 (1) 55-75

Smith, N., Hoffmann, S., and Rayson, P. (2008) 'Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations'. *Literary and Linguistic Computing* 23 (2), 163-180

Warren, M. (2004) '//↓so what have YOU been WORKing on Recently //: Compiling a specialised corpus of spoken business English'. In: Connor, U. and Upton, T. (eds.) *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins. pp. 115-140

Young, L. (1994) 'University Lectures – Macro-Structure and Micro-Features'. In: Flowerdew, J. (ed.) *Academic Listening*. Cambridge: Cambridge University Press. pp. 159-176