The pragmatic annotation of a corpus of academic lectures

Alsop, S. and Nesi, H.

The Pragmatic Annotation of a Corpus of Academic Lectures

Siân Alsop, Hilary Nesi
Coventry University
Priory St., Coventry, CV1 5FB, UK
E-mail: alsops@uni.coventry.ac.uk, h.nesi@coventry.ac.uk

Abstract

This paper will describe a process of 'pragmatic annotation' (c.f. Simpson-Vlach and Leicher 2006) which systematically identifies pragmatic meaning in spoken text. The annotation of stretches of text that perform particular pragmatic functions allows conclusions to be drawn across data sets at a different level than that of the individual lexical item, or structural content. The annotation of linguistic features, which cannot be identified by purely objective means, is distinguished here from structural mark-up of speaker identity, turns, pauses etc. The features annotated are 'explaining', 'housekeeping', 'humour', 'storytelling' and 'summarising'. Twenty-two subcategories are attributed to these elements. Data is from the Engineering Lecture Corpus (ELC), which includes 76 English-medium engineering lectures from the UK, New Zealand and Malaysia. The annotation allows us to compare differences in the use of these discourse features across cultural subcorpora. Results show that cultural context does impact on the linguistic realisation of commonly occurring discourse features in engineering lectures.

Keywords: pragmatic annotation, lecture, engineering

1. Aim

We have identified a set of pragmatic features in a corpus of academic engineering lectures. Our project has two aims: 1. to identify features that are typical of English-medium engineering lectures, and 2. to shed light on how these lectures are delivered in different parts of the world.

2. Data and annotation categories

The data set, the Engineering Lecture Corpus (www.coventry.ac.uk/elc), contains 76 transcripts of approximately one hour videos of English-medium lectures given at three universities: Coventry University in the UK (UK), Universiti Teknologi in Malaysia (MS), and Auckland University of Technology in New Zealand (NZ). In our identification scheme and analysis we treat these components as sub-corpora. The video data was locally transcribed in plain text, and then mark-up was added to create a corpus of TEI-compliant XML files. Using the video data as a guide, we then added the pragmatic annotation utilising a simple system of inline XML tags (Figure 1).

```
<u who="nm2001"><summary type="preview content of future lecture">you're going to need to be able to do all of those moment questions that are in the book<gap reason="pause"/> because we're going to start using them next week to work out beam reactions</summary><gap reason="pause"/><humour type="sarcasm">thank you for the yawn</humour></u>
```

Figure 1: Example of ELC XML

Currently, the annotation is inline as we found that the transcription data was not completely accurate. For efficiency, we are making minor corrections during the annotation process and so do not yet have a completely stable text with which to work. Ultimately, the annotation for each separate pragmatic category will be exported and stored as an individual stand-off layer of XML. The corpus will then be made public.

Our intention is to identify features that are typical and interesting in the academic discourse. Our final set of annotation tags includes five elemental categories, with various sub-categories attributed (see Table 1).

element	attribute	definition
explaining	equating, defining, translating	where lecturers demonstrate, define or translate concepts and terms
housekeeping		where lecturers talk about academic commitments or events external to the lecture
humour	black, bawdy, denigrating, ironic/sarcastic, joke, playful, self-deprecating, teasing/mock-threatening, wordplay	where lecturers employ various types of humorous discourse
storytelling	anecdote, exemplum, narrative, recount	where lecturers provide stories about their personal or professional experience, c.f. Martin's (2008) storytelling genres
summarising	review past lecture content, review current lecture content, preview current lecture content, preview future lecture content	where lecturers review the content of previous and current lectures, or preview the content of current and future lectures

Table 1: The ELC pragmatic annotation categories 2014

All annotation was done manually, with regular inter-coder reliability checks and advice from local language experts when the L1 of the lecturer was not English, or where brief instances of code-switching occurred. We are currently on the third pass at annotation. Significant evolution of the content and architecture of our original instinct/experience-based preliminary tag set has occurred at each pass as we gain a more coherent picture of the overall data.

The purpose of this 'middle ground' annotation system is to provide a layer of description to the text that allows more accurate conclusions concerning discourse features to be drawn. For example, in order to look at humour, research into the main corpora for academic speech has largely relied on structural markup for the vocal description of laughter; it is limited to a discussion of humorous discourse that elicits an audible laughter response (e.g. Lee 2006 on MICASE and Nesi 2012 on BASE). This limitation is significant because we know that laughter and humour are related, but not coextensive (Attardo 2003: 1288). Also, that laughter does not necessarily indicate humour (Ross 1998; Swales 2006).

Likewise, little work has been done to investigate discourse features that could perform a structural function, such as summarising. At the micro-level of the lecture, predictive devices such as 'enumeration' and 'advance labelling' (Tadros 1985) can be identified qualitatively. At the macro-level, Young's (1994) recognition of 'discourse structuring' and 'evaluation' within a phasal analysis aids consideration of the use of summary in the lecture as a contained unit. We are not aware of any systematic identification of these features, however, particularly across cultural contexts, nor are we aware of any systematic identification of similar pragmatic categories in any academic speech corpora. The presence of some pragmatic features in the header information of a small, non-random subset of the MICASE academic speech events has been identified (Maynard and Leicher 2007). The precise location of these features within the text, however, is not indexed.

By indexing the start and end of stretches of text that perform certain pragmatic functions within the ELC, we are able to linguistically describe discourse features that are typical of the lectures. The annotation also facilitates the analysis of cultural difference/similarity in lecture delivery. Most simply, indicators such as frequency of discourse feature and average token length across each cultural sub-corpus can be quantified and conclusions can be drawn concerning usage in different cultural settings.

Isolating chunks of text that perform specific discourse functions allows more fine-grained quantitative and qualitative analysis to be performed. In this case, we look at indicators such as keyness (c.f. Scott 2012) to examine changes in the lexis used when lecturers are doing something other than delivering academic content (for example, telling a story). Indexing the duration and dispersion of pragmatic features (where in the lecture, and for how long they occur) also allows the data to be mapped onto visual structures, which further enables the identification of patterns - in this case, across the discourse features rendered and across cultural sub-components.

The pragmatic annotation of the ELC has enabled firstly the identification of commonly occurring discourse features, and then the conclusion that cultural context does impact on their linguistic realisation. The normalised figures show that, as an umbrella category, speech that performed the humour function was most commonly employed in the UK lectures. Significant cultural variation between humour types also emerged as a variable across the cultural components (see Figure 2). The average token length of each instance showed significant variation between humour types, but not between the cultural sub-corpora.

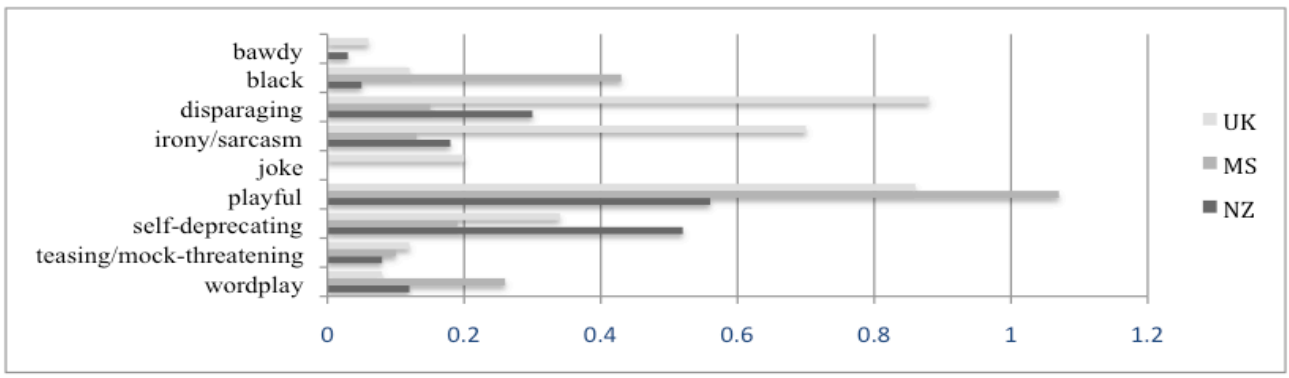

Figure 2: ELC humour types normalised (by tokens) as a percentage of 3 ELC subcomponents

Cross-referencing the annotation of humour in the ELC against the presence of structural laughter tags reveals two significant patterns. Firstly, that not all humour elicits laughter. Its co-occurrence occurred in approximately only one quarter of cases, showing that laughter in the ELC does not necessarily indicate a reaction to humour (it can be prompted by, for example, anxiety or relief). Secondly, that typical laughter response varied across cultural components. When we looked at the summarising aspect of each of the three components of the ELC, it was clear that different strategies were impacting on the lecture structure.

The greatest amount of content performing a summative function occurred in the lectures from New Zealand (an average of 8.8% per lecture). Within this, over 3% was dedicated to previewing current content and over 3% to reviewing previous content. The lectures from the UK contained the least amount of summarising (2.3%), which predominantly involved previewing current content (0.8%) and reviewing previous content (1.3%). In the Malaysian lectures, the different types of summary were evenly distributed (at approximately 1% per type), with the largest token total for previewing future content and reviewing current content.

This emphasis on reviewing and previewing in the lectures from New Zealand and Malaysia might be due to a more rigid syllabus, which requires that lecturers ensure specific content has been covered and understood on a weekly basis. This would indicate that these lectures serve more of a content-delivery function, rather than introducing concepts for students to investigate independently.

Storytelling was a particularly interesting discourse function to investigate because it offers a mode through which lecturers can convey something that students cannot find in their written materials: an insight into real-world engineering problems. Significant differences emerged in the use of the four types of storytelling identified. On average the least commonly occurring, yet subject to most culture-specific variation, types were anecdotes and exempla. Anecdotes occurred significantly more often in the UK component, and exempla were used heavily in the Malaysian lectures and were significantly uncommon in the lectures from New Zealand. As with the use of summarising as a structural device, and the differing extents to which humour types were employed, we hypothesise that the use of storytelling may reflect differing concepts of the role of lectures in different cultural contexts.

The function of exempla is the exemplification of information, which ties in with a lecturing style that is geared towards conveying information. The function of anecdotes is more entertaining and has greater appeal to the emotions (Martin 2008). Their use may be linked to modelling attitudes towards professional engineering experience. The comparative lack of extemporisation in the Malaysian lectures may reflect a heavier reliance on pre-prepared materials, perhaps due to lesser confidence in knowledge of English (both the lecturer and their students), or a stricter expectation that a programme will cover specific ground, regardless of who delivers the lecture.

Within the narrative type of storytelling, we found that lecturers from the UK largely relied on personal experience, whilst the Malaysian lecturers narrated the experience of others. The lecturers from New Zealand drew on both types of experience equally (see Table 2).

narrative type	UK		MS		NZ	
	raw	%	raw	%	raw	%
personal experience	13	68	1	9	8	50
experience of others	6	32	10	91	8	50

Table 2: Narrative types in the ELC

Dyer and Keller-Cohen (2000) suggest that lecturers may use personal narratives to model an expert identify. The notable difference in experience type used to inform the UK and Malaysian narratives may relate, as a Malaysian colleague suggested, to differences in the career paths undertaken by the lecturers. Engineering lecturers in Malaysia tend to enter academia pre-experience, whilst the UK lecturers have most commonly spent years in industry and so accumulated more 'on site' stories. In combination with attitudes to lecture function and other cultural factors, this difference in experience may also account for the different types of humour identified in each of the components. Specifically, the greater use of 'bawdy', 'sarcastic/ironic' and 'disparaging' humour in the UK lectures may be more indicative of the humour used on site, and so reflective of lecturer experience.

We can infer that there is less emphasis on content-delivery in the lectures from the UK, with more space given to expressing thoughts and opinions more loosely related to the curricula, and more freedom for 'off-the-cuff' experiences to be shared. Use of the discourse strategies identified suggests that students are expected to discover key information for themselves; that there is more emphasis on student autonomy in the UK lectures compared to those from Malaysia and New Zealand.

3. Conclusions

Lectures are increasingly being delivered in English-medium in universities around the world, especially in disciplines such as Engineering where global language skills are emphasised. ELC data has been used to investigate whether when language medium, syllabus and education level are constant academic lectures are roughly the same at the level of discourse, regardless of *where* they are delivered.

Our results indicate that cultural context does impact on the linguistic realisation of commonly occurring

discourse features in engineering lectures. The impact of these findings is largely pedagogic. The results offer a step towards being able to compare across cultures subjective linguistic phenomenon. The annotation of the ELC also provides a resource from which authentic examples of types of spoken data that have a strong pedagogic function can be extracted. Both factors may be of interest to EAP and ESP practitioners supporting staff and students who move between cultural contexts and would benefit from greater understanding of differences in lecturing style.

4. Acknowledgements

The Engineering Lecture Corpus (ELC) is under development at Coventry University. The project is directed by Hilary Nesi and has received funding from the British Council (RC 90). www.coventry.ac.uk/elc.

5. References

Lee, D. (2006). Humour in spoken academic discourse. *NUCB JLCC*, 8(3), pp. 49--68.

Martin, J. R. (2008). Negotiating values: Narrative and exposition. *Bioethical Inquiry*, 5, pp. 41--55.

Nesi, H. (2012). Laughter in university lectures. *Journal of English for Academic Purposes,* 11(2), pp. 79--89.

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.

Ross, A. (1998). *The Language of Humour*. London: Routledge.

Scott, M. (2012). *WordSmith Tools version 6*. Liverpool: Lexical Analysis Software.

Simpson-Vlach, R. and Leicher, S. (2006). *The MICASE Handbook: A Resource for Users of the Michigan Corpus of Academic Spoken English*. Ann Arbor: University of Michigan Press.

Swales, J. M. (2004) *Research Genres: Explorations and Applications*. Cambridge: Cambridge University Press.

Tadros, A. A. (1985) *Prediction in Text.* Birmingham: ELR.

Young, L. (1994). University lectures – macro-structure and micro-features. In Flowerdew, J. (ed.) *Academic Listening*. Cambridge: Cambridge University Press, pp. 159--176.