Coventry University

DOCTOR OF PHILOSOPHY

The pragmatic annotation of academic lectures
an exploratory study applied to engineering disciplines

Alsop, Sian

Award date:
2016

Awarding institution:
Coventry University

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of this thesis for personal non-commercial research or study
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission from the copyright holder(s)
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. Jul. 2025

The Pragmatic Annotation of Academic Lectures: An Exploratory Study Applied to Engineering Disciplines

SIÂN ALSOP

COVENTRY UNIVERSITY

NOVEMBER 2015

ACKNOWLEDGMENTS

I would like to thank the Department of English and Languages and the Academic Registry at Coventry University, whose financial support by way of a doctoral scholarship enabled this research. Additionally, travel scholarships from Coventry University's Centre for Global Engagement have allowed me to attend academic events around the world, through which my work has been refined.

The recordings and transcriptions used in this study come from the Engineering Lecture Corpus (ELC), which was developed at Coventry University under the directorship of Professor Hilary Nesi with contributions from ELC partner institutions. I am thankful for the work of the ELC project team, especially Dr. Ummul Khair Ahmad and Dr. Lynne Grant.

I have benefited enormously from the expert guidance of colleagues at home and abroad, from conferences to the staffroom. I want to acknowledge the hospitality offered by all who hosted me at the University of Tübingen.

Generous and skilled feedback also contributed to the development of the data visualisation tool (*ELVis*) which was created to support this research. I am particularly grateful to Professor Chris Culy who shared his time and knowledge with me, and to Joel Priestley for his patient programming advice.

My academic experience has been in the company of some most remarkable people, whose influence and support cannot be overstated. I am extremely lucky to count both mentors and friends amongst those who have enabled my research, from undergraduate to doctoral level.

My interest in research began with Professor Elizabeth Clarke, to whom I will always be grateful. I owe a special thanks to Professor Sheena Gardner for her unwavering support and encouragement. I am most indebted to Professor Hilary Nesi for matchless supervision, and for her kindness.

ABSTRACT

This thesis explores ways in which a corpus of engineering lectures can be annotated to identify, categorise and analyse pragmatic features. It also looks at the distribution of these features across individual lectures. The data on which the annotation system is tested comes from The Engineering Lecture Corpus (ELC), a growing corpus of English-medium lectures from across the world, currently including transcripts from Malaysia, New Zealand and the UK. Unusually, the ELC encodes functions that recur across large numbers of transcripts, which is referred to as *pragmatic annotation*.

The annotation allows features that are typical of the discourse to be identified and described. The *ELVis* data visualisation tool and corpus linguistic techniques are used to communicate and explore patterns at the macro-level, which guide finer analysis of the authentic language data. Comparison of the styles of English-medium engineering lecturers in different parts of the world is made, and the current role of English-medium instruction (EMI) in the discipline of engineering is also explored.

Recurrent functions in ELC transcripts have been found to include *storytelling*, *summarising* and *humour*. Sub-categories have been assigned to these functions; for example, storytelling is marked as an *anecdote*, *exemplum*, *narrative* or *recount*, and two types of *preview* and two types of *review* have been attributed to summarising. The purpose of this middle ground annotation system between video data and textual transcription is to provide a layer of description that allows more accurate conclusions concerning discourse features to be drawn.

The main purpose of the thesis is to refine systems of annotation so that they can be applied to the ELC and other corpora, and also to discover the features of engineering lectures that could be of value to engineering lecturers and students around the world. In so doing, this thesis provides a novel framework for discourse annotation and challenges some established views about the nature of lecture discourse.

Although engineering lecturers around the world may use a common language to deliver the same kind of syllabus for the same broad purpose, engineering lectures are likely to remain both context- and culture-specific. Lectures of all kinds often include pragmatic elements that serve to entertain, instruct, and make key information more memorable. The way in which these features are presented varies from place to place, however, and cultural differences may represent a challenge both to those who attend lectures and to those who deliver them. Such variation is important to take into account when designing ESP and staff development programmes.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

ACE	Asian Corpus of English
BASE	British Academic Spoken English
BAWE	British Academic Written English
BNC	British National Corpus
CA	correspondence analysis
CES	Corpus Encoding Standard
CLiC	Corpus Linguistics in Cheshire
DAMSL	Dialogue Act Markup in Several Layers
DART	Dialogue Annotation & Research Tool
DCMI	Dublin Core Metadata Initiative
DM	discourse marker
DOM	Document Object Model
DRI	Dialogue Resource Initiative
DTD	Document Type Definition
EAC	Engineering Accreditation Council
EAGLES	Expert Advisory Group on Language Engineering Standards
EAP	English for Academic Purposes
EC	Engineering Council
ELC	Engineering Lecture Corpus
ELF	English as a lingua franca
EMI	English-medium instruction
ESP	English for Specific Purposes
GeWiss	Gesprochene Wissenschaftssprache kontrastiv (Spoken Academic Discourse in Contrast)
HE	Higher Education
HEA	Higher Education Authority
HESA	Higher Education Statistical Agency
HTML	Hyper-Text Markup Language
IAR	inter-annotator reliability
ICEF	International Consultants for Education and Fairs
IMDI	ISLE Metadata Initiative
IPENZ	Institute of Professional Engineers New Zealand
ISLE	International Standard for Language Engineering
ISO	International Standards Organization
JSON	JavaScript Object Notation
KWIC	key word in context
L1	first language
L2	second language
LL	London-Lund
LOB	Lancaster-Oslo/Bergen
LR	laughter response
MAE	Multi-purpose Annotation Environment
MATE	Multilevel Annotation, Tools Engineering
MICASE	Michigan Corpus of Academic Spoken English
MICUSP	Michigan Corpus of Upper-Level Student Papers
MS	Malaysia
NLP	natural language processing
NLR	no laughter response
NNS	non-native speaker
NS	native speaker
NZ	New Zealand
OECD	Organisation for Economic Cooperation and Development
OLAC	Open Language Archive Community
OTA	Oxford Text Archive
PDF	Portable Document Format
PEI	professional engineering institution
PMI	Prime Minister's Initiative
PMW	per million words
Pr.A.T.I.D	Pragmatic Annotation Tool for Italian Dialogues
PSRB	professional, statutory and regulatory body
PY	Python
QAA	Quality Assurance Agency
RB	Ruby
RDF	Resource Description Framework
RGB	red, green, blue
SGML	Standard Generalised Markup Language
SHS	Self-denigrating Humor Schema
SPAAC	Speech-Act Annotated Corpus of Dialogues
STTR	standardised type-token ratio
T2K-SWAL	TOEFL 2000 Spoken and Written Academic Language Corpus
TEI	Text Encoding Initiative
TIS	Teaching International Students
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
TTR	type-token ratio
TUSNELDA	TUebinger Sammlung Nutzbarer Empirischer Linguistischer DAtenstrukturen (Tübingen collection of reusable, empirical, linguistic data structures)
UCAS	Universities and Colleges Admissions Service
UCREL	University Centre for Computer Corpus Research on Language
UK	United Kingdom
UNESCO	United Nations Educational, Scientific and Cultural Organization
W3C	World Wide Web Consortium
XCES	XML Corpus Encoding Standard
XML	eXtensible Markup Language

CHAPTER 1. INTRODUCTION

1.1. Engineering as a discipline

Engineering can be categorised as a hard, applied discipline (Biglan 1973: 198) that emphasises the solution of real-world problems. One UK government report equates engineering with "turning ideas into reality", explaining that engineers "solve problems, and the end, not the means, is the motivating factor" (House of Commons 2009). This overarching purpose appears to span the broad remit of engineering, which encompasses fields across mathematics, science and technology. Divisions such as civil, chemical, electrical and mechanical engineering are commonly made, each of which branch into multiple sub-divisions.

The matrix of possible pathways within engineering is reflected by the study options open to students. For example, UK students who wish to complete a full-time, single honours undergraduate degree in some form of engineering can choose from 132 providers in the UK offering over 500 courses. Coventry University alone offers 16 such courses (or 29 including a placement year) in areas as diverse as aerospace systems, manufacturing, and environmental engineering (UCAS 2016). It might therefore be expected that many different types of teaching methods exist within engineering, some with a theoretical focus, and others more oriented towards professional practice.

In reality, it seems that professional practice is the dominant driving force behind teaching practice. Professional bodies certainly influence the context and content of engineering degrees. Nationally, degree-level engineering awards are commonly overseen by professional bodies. In the UK, the Engineering Council (EC) is the umbrella organisation – or, professional, statutory and regulatory body (PSRB) – for 36 licensed professional engineering institutions (PEIs) who accredit academic programmes across the discipline, largely for undergraduate or postgraduate degrees (HEBRG 2011). Accreditation is based on whether a programme meets standards defined by the EC (Engineering Council 2016). The

Quality Assurance Agency (QAA) has adopted these standards in its Engineering Subject Benchmark Statement, which lists the desired learning outcomes for engineering graduates in the UK (QAA 2016).

Bodies with similar structures and functions exist worldwide. The Board of Engineers which oversees engineering education standards in Malaysia, for example, confers degrees through the Malaysian Engineering Accreditation Council (EAC) (EAC 2016). The Institute of Professional Engineers in New Zealand (IPENZ) similarly stewards national award programmes (IPENZ 2016). Substantial equivalence across these nationally accredited programmes is recognised. In formal acknowledgment of this equivalence, all three countries have signed the *Washington Accord* (IEA 2016), an international agreement which specifies the academic standard accredited graduates should meet in order to practice in signatory countries. Degree content is a highly regulated concern based on professional requirements (and experience) in disciplines like engineering, both nationally and internationally.

1.2. English-medium instruction

On university campuses worldwide there is increasing provision of English-medium instruction (EMI), especially in disciplines related to business and technology where global language skills are emphasised. It is probably the case that around the world more non-native than native speakers use English as a sole, partial or primary means of instruction (Jenkins 2014: 5). English is the most common academic lingua franca because so many academic materials are published in the medium of English, and because this is the first, second or foreign language that students and lecturers are most likely to have learnt at school.

The increase in the number of tertiary level students studying abroad is apparent from the figures in Table 1.1; based on UNESCO (2015) statistics, the rise is over one hundred per cent between 1999 and 2013.

	1999	**2013**	**increase** (%)
world	1746946	3546552	103
Africa	218599	373303	71
Asia	770835	1966513	155
Europe	554543	858713	55
North America	121586	187776	54
South America	59912	127567	113
Oceania	21471	32680	52

Table 1.1: Number of outbound students studying abroad (adapted from UNESCO 2015)

The number of internationally mobile students reportedly more than tripled between 1990 and 2011 from 1.3 million to 4.5 million (OECD 2013). Student mobility in Higher Education (HE) is still gaining momentum, and was expected to exceed five million in 2014 (ICEF Monitor 2014). The growing preference for spending only part of a degree course studying abroad, which is not factored into these statistics, further amplifies the pattern. The observed mobility in the student population is now regarded as a "mass movement" (ICEF Monitor 2014).

The direction of the geographical traffic tends to be towards countries where English is used as a first language (British Council 2015), or towards Western countries in general (de Wit and Jones 2012, Maringe 2010: 31). Where instruction in English was previously considered to give universities an "edge in the world market" (Wilkinson 2005), failure to do so is now framed as a "linguistic disadvantage" (OECD 2013) or "linguistic handicap" (Wächter 2008: 3). As a result there has been a rapid growth in EMI in areas of Europe where English is not used as a first language (L1), as reported in large-scale projects such as Brenn-White and Faethe (2013), Wächter and Maiworm (2008), and Ammon and McConnell (2002). Accelerants for this growth are commonly linked to the reforms in HE surrounding the Bologna Declaration (The European Higher Education Area 1998, 1999) and the impact on research and HE surrounding the Lisbon Strategy (European Council 2000).

In terms of institutional visibility, offering EMI expands the catchment area for potential home and overseas students. University prospectuses continue to be shaped by the

internationalisation agenda; branding straplines and vision documents declare the status of, or intention to become, a "leading global university" (for example, Brown University 2011, National University of Singapore 2012, University College London 2012, University of Western Australia 2014), or contain similar rhetoric. Students are promised membership into a world of *global citizens* in which opportunities for career and personal development are optimal. Under these circumstances, lecturers are often expected to impart the skills required to gain entrance into this world, and to do so in English.

There are tensions, however, arising from the process of "Englishization" (cf. Kachru 1994). It has been claimed that there is a lack of understanding of the process of implementation of EMI, and its setting-specific cultural and economic challenges (Doiz, Lasagabaster and Sierra 2013, Li 2013). National-level concerns have been expressed regarding English monolingualism versus pluralinguilism (Gotti 2015, Plo Alastrué and Pérez-Llantada 2015).

Moreover very little is known about the nature of EMI at university level, or the nature of lecture discourse generally. In most institutions and across most disciplines, the lecture remains the main method of communication between staff and students (Deroey and Taverniers 2012, Flowerdew 1994, Lee 2009, Thompson 1994), but the form and even the function of lectures may vary from context to context. For example, according to the Higher Education Authority (HEA 2014):

> One experienced teacher who has delivered a familiar lecture in many different contexts reports that she covers about 10-15% less material when the audience is listening to English as a second or third language.

Crawford Camiciottoli (2005: 189) records the same type of adjustment to the delivery of an EMI economics lecture in the UK and in Italy. Overall the lecturer increased adjustment for the Italian audience – for whom English was a second language (L2) – in: rate of speech, redundancies (reformulations and glosses), interpersonal features (questions and asides), and references to local cultures. Less adjustment was made in discipline-related lexis and the use of metaphor.

In naturally occurring lecture speech, where the intention behind an utterance cannot always be identified from its form alone, there is potential for misinterpretation – especially if the language of instruction is not native to the lecturer or students. There may also be differences in understanding the role of lectures in different settings, such as whether their primary aim is to instruct, to encourage, or to apply textbook knowledge.

One of the major practical concerns regarding EMI is the lack of English for Academic/Specific Purposes (EAP/ESP) support for students and particularly for lecturers. There is longstanding uncertainty about the best methods of teaching in English as an academic lingua franca (Wilkinson 2005), and about the quality of the infrastructure for enabling EMI provision. Many lecturers who have transferred from teaching in their native language to teaching in EMI have not had training; anecdotal evidence suggests that the shift from language medium delivery in some cases happens almost overnight. As Dearden (2014: 2) points out:

> […] there is a shortage of linguistically qualified teachers; there are no stated expectations of English language proficiency; there appear to be few organisational or pedagogical guidelines which might lead to effective EMI teaching and learning; there is little or no EMI content in initial teacher education (teacher preparation) programmes and continuing professional development (in-service) courses.

It seems clear, however, that before we can support either staff or students engaging with EMI lectures, we need a fuller description of the discourse of lectures, in order to understand their purpose in different cultural contexts, and how they are constructed and delivered.

1.3. Thesis purpose and design

Although corpus linguistics is a fast-developing field, on the whole corpora are not annotated for interpretative features. After creation, corpora tend to be analysed at and below sentence level. Therefore, once corpora have been created, it is difficult for people referring to them to identify the pragmatic function of any part of the text. If corpora are

developed for pedagogical as well as research purposes, it makes sense to identify some pragmatic features at the corpus creation stage. Unfortunately, very few corpora are annotated in this way. The identification of pragmatic features adds a layer of interpretation between video footage and text transcription that researchers can incorporate in analysis if they wish. This kind of interpretation is of value to those interested in pedagogic issues, such as the best way to deliver lectures to aid student absorption and retention of material.

Engineering as a discipline is notably concerned with good academic practice. For example, the *Journal of Engineering Education* and the *International Journal of Engineering Education* examine pedagogical issues such as the teaching of engineering and student attitudes to learning engineering. However, there has been little investigation of the discourse of engineering, and so little or no use has been made of corpus linguistic methods.

Generally speaking, engineering lecturers and students are not interested in linguistic features per se; they are interested in the meanings that are made in lectures. Pragmatic annotation facilitates access to these meanings. Even using the same medium of instruction, the way in which meaning-making is achieved in engineering lectures will vary from institution to institution, and particularly from country to country, due to differing educational cultures.

This thesis approaches the description of engineering lecture discourse from both pragmatic and corpus linguistic perspectives. In order to do so, it offers a novel framework for the pragmatic annotation of discourse features. Analysis is based on the Engineering Lecture Corpus (ELC), a dataset of authentic EMI lectures from Malaysia (MS), New Zealand (NZ) and the United Kingdom (UK). After evaluating the prior research into the linguistic structure of lecture discourse in Chapter 2, an explanation of the ELC system of *pragmatic annotation* is given in Chapter 3, followed by an account of the processes of data preparation, testing and analysis. Chapter 3 also includes a description of the visualisation tool designed to both interrogate and communicate data patterns in this thesis: *ELVis*.

Chapter 4 overviews patterns across all annotated data, and Chapter 5, Chapter 6, and Chapter 7 discuss in detail the three discourse functions that have been annotated: *summary*, *humour,* and *story*. Quantitative findings guide qualitative analysis. Macro-patterns are identified, showing where and for how long each function and its attributes occur, and in turn direct examination of the linguistic character of the encoded text. Each chapter also compares usage of the discourse function in lectures in the three subcorpora. A summary of major findings is given in Appendix VI.

After consideration of the findings, the conclusion of this thesis in Chapter 8 looks at possible future improvements to the corpus and the system of annotation and analysis employed, and returns to the question of pedagogical implications.

CHAPTER 2. LITERATURE REVIEW

2.1. Introduction

This thesis takes a pragmatic approach to monologic lecture discourse in academic settings. It looks at some of the functions that lectures realise, and also the linguistic features that characterise these functions. Very little previous work addresses all of the areas in focus, namely the genre of academic lectures, the discipline of engineering, and corpus linguistic and data visualisation approaches. This chapter starts by reviewing the more general literature of relevance and in later sections focuses on spoken discourse and pragmatics in relation to corpus studies. It opens with a brief consideration of the role of pragmatic meaning in language (2.2), and then surveys general features in academic text (2.3). This is followed by a more detailed examination of those features that apply to spoken academic discourse, including a review of the case for pragmatic corpus annotation (2.4). The final section outlines the research aims of this thesis (2.5).

2.2. Pragmatic meaning in language

The pragmatic study of language occupies a *logico-philosophic* position (Eggins and Slade 1997: 40) rooted in maxims from early enquiry in the field of semiotics (Carnap 1942, Morris 1938, Peirce 1934). It addresses the Wittgensteinian (1922) problem of symbolism in language use: unpicking the psychology of intention combined with the epistemological relationship between words and meaning.

A pragmatic approach places at the forefront the situational and interactional aspects of understanding language in use. Its remit is concerned with "the inter-relation of language structure and principles of language usage" (Levinson 1983: 9), which is distinguished from the relationship between linguistic forms alone (syntax) or linguistic forms and world

entities (semantics) (Aijmer and Rühlemann 2015: 1, Levinson 1983: 5, Morris 1938: 21-22, Yule 1996: 4). The pragmatic stance is that background assumptions are often irretrievable from semantic structure; in other words, the meaning of an utterance often equates to more than its literal meaning (Searle 1978: 210, Straker Cook 1975: 29). Emphasis is placed on "*how* utterances have meaning in situations" (Leech 1983: 1), and *how* intention is delivered then reconstructed.

In spoken discourse, Speech Act Theory identifies these units (or *speech acts*) based on intended purpose: the direct or indirect "illocutionary force" of what is said (Austin 1962, Searle 1969, Searle 1976). Searle (1976: vii-viii) divides the "the full blown illocutionary act" (including illocutionary force and propositional content) into five categories: *assertives*, *directives*, *commissives*, *expressives*, and *declaratives*. Multiple attributions are allowed. Searle offers the example of the utterance "Sir, you are standing on my foot", which functions as an *assertive* (the non-literal, *direct* primary *speech act*) through which a *directive* is performed (the implied, literal, *indirect speech act*). The scope of the execution of illocutions through performative utterances is laid out:

> We tell people how things are, we try to get them to do things, we commit ourselves to doing things, we express our feelings and attitudes, and we bring about changes through our utterances (Searle 1976: 22-23)

Decoding meaning, or meanings, requires contextual knowledge.

Inherent to understanding the pragmatic meaning of utterances is recognition of what is not said, or "how more gets communicated than is said" (Yule 1996: 3). Grice (1975) identified the gap between what is said and what is meant as *conversational implicature*, an inferred and predictable process (and result) based on adherence to the principles of co-operation in conversation. Alongside indexicality/deixis (versus anaphora) and presuppositions, implicature (versus entailment) is one of the major areas of study in the field of linguistic pragmatics (Bublitz and Norrick 2011: 4).

The production and reception of pragmatic meaning through language is fundamentally interactional, operating between producer and receiver through both verbal and non-verbal mechanisms. Communicative interactivity results from a common language *and* shared world knowledge; an overlap in the mental models of producer and receiver. Both a *situation* model based on semantic understanding and a *context* model based on pragmatic understanding is required to decode the present discourse in relation to its present environment (see, for example, van Dijk 2012).

Most pragmatic studies of discourse have looked at non-academic discourse, and dialogues are commonly analysed. In all cases, the analysis of pragmatic meaning is premised on first identifying meaning that may be linguistically invisible. However, there is also a substantial body of research that has looked at pragmatics in spoken academic discourse. Many of these studies are concerned with the interpretation of utterances or conversational turns, thus delimiting the units of analysis.

2.3. Pragmatics and academic discourse

The pragmatic relationship between signalling language and information processing has more often been discussed with reference to written academic texts, although some of this discussion can be applied to academic speech. Analysis has been premised on the idea that particular linguistic features are used to make the discourse structure more transparent, thus facilitating comprehension. This section looks at general pragmatic features of academic text, including prediction and summation (2.3.1), lexicogrammatical features (2.3.2), and larger structural patterns (2.3.3).

2.3.1. Prediction and summation in academic text

The categorisation of pragmatic functions in texts can be based on the location of their occurrence. Tadros (1985: 38), for example, distinguishes *recapitulation* from *summary,* the latter being a function of text that terminates a chapter or section. Both recapitulation and

summary require information to be recalled from elsewhere in the text, but recapitulation is characterised as an anchor for new information and is therefore predictive; if previously given information is recalled at the end of a text it has no predictive potential and is called summary (1985: 35). A similar structural distinction is made when information is recalled at the end of – in the case of writing – a paragraph. Tadros (1985: 38) identifies the function of recalled information as *comment* (a reminder of relevance), again because it does not predict new information.

Tadros models such predictive structures using set units: pair, member, and sentence. The pair is comprised of two members that function in a one-way relationship: the predictive (V) always precedes the predicted (D), which may be made up of one or more sentences. The internal structure of the members is composed of a (pre-head), head, and (post-head).. The V head is the predictive signal carrier, and the D head in turn carries out this prediction. Pre-heads provide the contextual environment for the head, and post-heads function as comments or expansion of the head (Tadros 1985: 10-11). Prediction is given particular emphasis because of its pedagogical value; it enables students to benefit from predictive signals when reading, or to fulfil them when writing (Tadros 1985: 64). Identifying and decoding predictive signals is a key skill when understanding pragmatic meaning within all types of academic discourse.

One of the most common predictive patterns is *advance labelling*, which occurs when the author commits to a certain discourse act by first labelling that act. For Tadros (1985: 34), there are four V membership criteria, all of which must be met: 1. the sentence must contain a labelling of a discourse act, 2. the labelling of the act must be prospective, 3. the writer is the actor, and 4. the sentence labelling the act must not include its performance.

The other common predictive structure identified in text is *enumeration*, which tends to rely on a sequential naming of items within a set resulting in some form of ordered list. Tadros (1985: 15-20) outlines three criteria for V membership classification in enumeration, at least

one of which must be met. The first two, types A and B, require the presence of a colon and are not applicable to the spoken mode. Type C contains a syntactically complete V member, a numerable (exact or inexact), an enumerable, and new information.

Type C can be applied to speech. Based on a statistical analysis of enumeration in casual conversation in the 1984 Montréal corpus, Dubois and Sankoff (2001: 286) state that:

> Enumeration represents a cumulative discursive procedure made up of at least two different components that belong to the same or equivalent morphological and functional categories.

The "two different components" referenced are the D heads, the prediction of which is committed to in the enumerative V head signal. The speech data used by Dubois and Sankoff is largely dialogic and elicited, so the focus is on the effects of interaction (2001: 289). The variable linguistic features (or factors) identified include: 1. number of components, 2. complexity of components, 3. reduction, 4. expansion, 5. repetition, and 6. the presence of explicit markers.

Contrary to their initial hypothesis, Dubois and Sankoff (2001: 290) found that number and complexity of components does not increase processing difficulty, perhaps due to the associated increase in expressivity gained. Discourse markers were also identified as a means of compensating for increased processing difficulty in the case of reduction and expansion, where parallelism is lost. Repetition, on the other hand, increases parallelism, which has a positive effect on processing. It was found that interactional level, however, had little influence on the structure and complexity of the enumeration.

As well as predictive structures, summative devices also function pragmatically at the level of individual lexis. Francis (1986) identified an anaphoric X-A discourse relationship in text in which the X-member describes the text preceding a clause containing an anaphoric noun (A-noun). Three criteria guide the categorisation of A-nouns, which must: 1. function

metadiscursively, 2. operate as an anaphoric pro-form, and 3. be presented as the given element within a clause containing new information (1986: 10).

The term *shell nouns* (which is hereafter adopted) is also applied to this class of nouns which "require lexicalization in their immediate context" (Hunston and Francis 2000: 185, see also Nesi and Moreton 2012). Following large-scale corpus investigation, Schmid suggests that this label should be considered as shorthand for "use-as-shell noun", because:

> […] shell nouns make up an open-ended functionally-defined class of abstract nouns that have, to varying degrees, the potential for being used as conceptual shells for complex, proposition-like pieces of information. (2000: 4)

One of the most common lexical signals of summary in oral discourse is the presence of abstract nouns, the meaning of which is only supplied by preceding or proceeding co-text (in anaphoric or cataphoric relation). Common examples include *case, chance, fact, idea, news, point, problem, position, reason* and *thing* (Schmid 2000: 3).

Schmid emphasises that the status of shell nouns is defined by use, rather than any inherent linguistic property. As *stand-ins* for longer, more complex notions that are expressed more lengthily elsewhere, shell nouns function at three levels: 1. semantic characterisation and perspectivisation, 2. cognitive temporary concept-formation, and 3. textual linking of the nominal concepts encapsulated (Schmid 2000: 14). The effect is the creation and maintenance of textual continuity, either by "allowing utterances to be condensed into nouns so that complex meanings can be easily carried forward in the discourse" or by "providing a convenient label ahead of time that indicates something about the nature of an idea that will be unpacked and explained in the upcoming text" (Flowerdew and Forest 2015: 2).

2.3.2. Lexicogrammatical features in academic text

Related to the idea of stand-ins in language, a function that is specific to the spoken mode is deixis. Like A-nouns, deictic devices only make sense in the context of the information to

which they refer (whether anaphoric or cataphoric) and require pragmatic contextualisation to be understood. The scope of this device is sketched out by Lyons (1977: 637):

> […] deixis is […] the location and identification of persons, objects, events, processes and activities being talked about, or referred to, in relation to the spatiotemporal context created and sustained by the act of utterance and the participation in it, typically, of a single speaker and at least one addressee.

Bamford (2004) further interrogates the spatiotemporal context, using the example of *here* to distinguish between: 1. the verbal companion to a physical gesture, such as pointing at a board (spatial), and 2. to indicate a specific point in a lecture (temporal), as in the example "I think I will stop here and we will continue our discussion tomorrow" (2004: 120).

Transitions and specific linguistic devices also function pragmatically to indicate thematic change in both writing and speech. For example, Collins (1991: 2) defines *pseudo-clefts* as constructions in which content is divided into distinct clauses: the *highlighted element* (the item in focus) is accompanied by a *relative clause* (a presupposition). The highlighted element and relative clause are reversible; a *simple pseudo-cleft* can become a *reversed pseudo-cleft* (1991: 3). The highlighted element in simple pseudo-clefts conveys a new message (the *news*), whilst in reverse pseudo-clefts the highlighted element tends to refer to background material given in the relative clause, providing a type of internal reference (Collins 2004[1987]: 93). The content of reverse pseudo-clefts is "informationally low", including forms such as generalisations and explicit repetitions; because no new information is given, structurally, reverse pseudo-clefts are considered to be suited to endings (2004[1987]: 87). Like Tadros' (1985) *comment*, certain language features are associated with certain parts of a text.

Collins (2004[1987]) examined the occurrence of clefts and pseudo-clefts across writing and speech in the Lancaster-Oslo/Bergen (LOB) and the London-Lund (LL) corpora. In this data, pseudo-clefts are far more common in speech than in writing (at a ratio of 3.3:1) (2004[1987]: 91). By comparing public and private (that is, not recorded before an

audience) speech, Collins concluded that pseudo-clefts were most frequent in private speech, and that reverse pseudo-clefts are most commonly found in face-to-face private speech where participants are intimate or equal (2004[1987]: 92). Although pseudo-clefts are less likely to occur in written texts (at a ratio of 1:4.1), the exception was imaginative discourse (especially passages of dialogue) in which the author/speaker is describing a context or constructing a landscape rather than presenting facts objectively. Clefts, however, are most common in the purely descriptive written genres (2004[1987]: 92).

Deroey (2012) argues that these '*wh*-clefts' are useful grammatical devices for marking important points in lectures. She identified 1221 basic *wh*-clefts by conducting concordance searches for 'what' within 160 lectures from the British Academic Spoken English (BASE) corpus (2012: 114). Deroey categorised the immediate clause complex according to five main discourse functions: informing (67.3%), organising discourse (15.3%), evaluating (8.4%), elaborating (5.3%), and managing the classroom (3.8%) (2012: 118). Like Collins (2004), Deroey considers these clefts to be "quite low in communicative content" in lectures, suggesting that they "serve to signal to the audience that an important elucidation follows" (2012: 122). Pronominal subjects were found in around 70% of *wh*-clefts, particularly *we* (24.2%), *I* (18%) and *you* (13%) (2012: 115). Within the discourse categories, procedural descriptions drawing on the inclusive pronouns *we* and *you* were found to be prevalent in the physical sciences, leading to the conclusion that "basic wh-clefts lend themselves particularly well to structuring such descriptions by allowing the highlighting of a new step, causal relationship or solution" (2012: 120). Deroey's work demonstrates the potential insights that can be gained by using corpus linguistic methods to analyse spoken academic discourse and disciplinary differences.

Predictive and summative features can be identified at the level of individual or small groups of lexical items. Textual prediction of this kind falls within Sinclair's category of *prospection*, "where the phrasing of a sentence leads the addressee to expect something specific in the next sentence" (2004: 88), or the inverse, *retrospective encapsulation* (2004:

88). These features help to signal the way information is organised in small segments of discourse, but it is also possible to examine prediction and summary across larger stretches of text.

2.3.3. Larger structural patterns in academic text

Some studies have investigated larger structural patterns within academic texts. There have, for example, been many analyses of research article introductions, starting with (and heavily influenced by) the work of Swales (1981, 1990, 2004). Swales (1990: 137-165) identified the presence (not bound by order) of three rhetorical *moves* (1. establishing a research territory, 2. establishing a niche, and 3. occupying the niche) and their subordinate *steps*. The unit of moves follows Coulthard and Sinclair's (1975) model of classroom discourse in which the lesson is composed of various dynamic transactions, which are subdivided hierarchically into *exchanges*, which are comprised of *moves*, which are in turn comprised of the smallest unit, *acts*.

The conventions for structuring research articles are relatively stable and are well-understood by members of the relevant research communities; introductions are often demarcated by section headings, for example, and typically consist of a series of moves (Swales 1990) aiming to create a *research space* for the article to occupy.

Swales' oft-cited work has prompted closer analysis - and annotation within corpora - of dialogue moves. Corpora of professional texts have been manually encoded in the Swalesean tradition in order to identify rhetorical moves. For example, Connor and Upton (2004) identify and define ten moves within grant proposals, and Kanoksilapatham (2005) identifies fifteen distinct moves within biochemistry research articles. This approach has been successful in the analysis of highly structured examples of written language, and has facilitated the identification of distinct genres with distinct characteristics. For example, in an analysis of moves within 48 political science academic research abstracts, Živković (2010: 85-86) outlines common linguistic features and their functions within certain moves, such as

the anaphoric use of the third person pronoun *it* after the introduction of move 2 and move 3. Such features can be construed as characteristic of summaries because abstracts have a primarily summarising function.

In many other academic genres, however, there is greater variation in the purpose and structure of introductory sections. Even if they all function to "introduce the academic work" (Bhatia 1997), practitioners do not necessarily agree about their generic features. Nesi and Gardner (2012: 98) found considerable variation in the role of introductions in student essays, for example, and Bhatia's (1997) informants disagreed about the distinctions between *introductions, prefaces* and *forewords* to academic books.

Analysis of the component parts of written texts should be easier than that of spoken discourse because the component structures of written texts can be pre-planned and are distinguished at the highest level through heading hierarchies. The demarcation of sections in some way does seem to enable description of linguistic features, which have also been shown to function pragmatically at a more micro-level of language. Yet the analyses of sections that might be expected to lend themselves to a formulaic structure, such as introductions or abstracts, point to variation across written genres.

Although these studies were largely concerned with written discourse, it is possible that many of the features identified are also present in lecture discourse. However, lectures do not contain the same amount of pure information content as highly informational written texts, so different - or different degrees of - functions can also be expected.

2.4. Pragmatics and spoken discourse, particularly lectures

Section 2.3 looked at attempts to identify and describe pragmatic meaning in academic text, which occurs at various levels of language structure. This section looks at levels of analysis of pragmatic meaning specifically within spoken discourse, primarily transcribed lectures.

The highest structural unit into which discourse can be divided is commonly described as the macro-level, or macrostructure, which can be understood as "the particular global content of a particular discourse" (van Dijk 1977: 22). It follows that, hierarchically, the synopsis provided by the macrostructure is composed of smaller, microstructures of information, which exist at various levels.

This section opens with a brief consideration of why it is important to identify lecture structure (2.4.1). Discussion of various analyses of macrostructural (2.4.2) then microstructural (2.4.3) units of lectures follows. Findings from studies that have taken a corpus linguistic approach to identifying units of pragmatic meaning are then reviewed (2.4.4).

2.4.1. Lecture structure and lecture comprehension

Much of the discussion relating to lecture macrostructure draws on principles of cognitive science. The concept of the knowledge *frame* has been used to characterise experiential knowledge that enables cognitive acts (for example, perception and language comprehension) (van Dijk 1977: 19). Fillmore (1976: 29) explains that the frame is "a kind of outline feature with not necessarily all of the details filled in"; comprehension occurs during the active process of filling in these details, as in bottom-up processing. van Dijk refers to these frames – or "higher level organising principles" – as *episodes* (1977: 21). When this cognitive information processing approach is applied to authentic data, macrostructures - units such as episodes – emerge at the level of rhetorical function.

Linguistic features and functions at all levels of lecture structure affect, to some extent, students' ability to digest, to recall, and to take effective notes. The studies into lecture structure which are reviewed in this section, however, concur that the biggest obstacle to lecture comprehension is understanding the function of larger discourse patterns, rather than utterance-level meaning or individual lexical items.

For Olsen and Huckin (1990: 40), successful comprehension by students who are native and non-native speakers (NS and NNS) can be attributed to their ability to understand "how things fit together", rather than their sentence-level linguistic competence. Straker Cook similarly observed of his student participants (for whom English was L2) that:

> [c]ontent presented no difficulties […] the problem is one of scale rather than of structural patterning *per se*: students simply failed to recognise the patterning of extended spoken discourse, still less manipulate such patterning as a productive skill. On their own admission they could follow most of the individual parts of a lecture, yet not grasp the whole. (1975: 27-28)

Chaudron and Richards concur that, in addition to content knowledge, listeners:

> […] may benefit from knowledge of the macro-structure and discourse organization of lectures. Prior knowledge of this sort helps top-down processing by initiating expectations and predictions about the lecture. These expectations are then confirmed and supported by the speaker's use of discourse signals of the relationship between successive episodes and moves within the lecture. (1986: 116)

Understanding the organisation of discourse, it is proposed, enables the main points of lectures to be identified and processed.

As part of their recent Teaching International Students (TIS) project, the Higher Education Academy (HEA) (2014) has developed suggestions on how to present lectures to L2 speakers of English to make discourse structure more transparent. Significant focus is placed on lecture organisation:

- use an introduction and a summary and repeat key ideas, for example: "this is an introduction/summary…"
- state links to previous/future lectures and topics
- signal moves between sections as in "now I am going to talk about how you apply this idea with an example…"
- try pointing to where you are: "on the outline, I am now here…", or name the section: "in summary, this lecture has covered…" (HEA 2014)

Although they appear to be based on intuition, rather than research evidence, the outlined suggestions emphasise the need for clear signalling language.

The issue of comprehension of academic discourse is not limited to L2 speakers, however. Allison and Tauroza (1995), for example, found that L1 and L2 science students had similar difficulties in comprehending the main points of the lecture when its discourse structure had a complex organisation. For pedagogic purposes, they recommend that focus should therefore be on the identification and analysis of structures within lectures that require "higher-level", pragmatic, contextualisation.

Discussions of text comprehension and how it can be achieved have been received from schema theory in psychology/cognitive science where processing involves bottom-up fleshing out of schemata in combination with top-down, conceptually-driven assimilation based on previous knowledge. Two types of schematic knowledge are distinguished: 1. *formal* knowledge of the rhetorical organisation and function of text types or genres, and 2. *content* knowledge of the subject area (Carrell and Eisterhold 1983: 560). In his work on second language lecture comprehension, Flowerdew (1994: 9) explains a two-stage process in which problems arise in the "higher-level" stage of contextualisation based on world knowledge, rather than in the "lower-level" stage of cognitive language decoding.

Young (1994: 173-174) highlights the emphasis that schema theory puts on understanding form and content in order to process information, suggesting that "[s]tudents need […] a schema for expository spoken discourse; without it they cannot accurately predict, which hampers their ability to understand". To improve note-taking skills, Young (1994: 174) suggests that students need to know that "information is imparted in several ways, through theoretical discussion, through exemplification, and through summarization", and that the same information is commonly revisited through each of these means. She then holds up the presentation of an accurate macrostructure of the lecture as a means of filling gaps in understanding, particularly for non-Western foreign students, who may have significantly different schemata for processing information in the academic context.

An identifiable relationship between macrostructure and microstructure is therefore predicted, and commonly regarded as useful. Various studies have sought to identify these structures within lectures, with various degrees of success.

2.4.2. Phasal analysis of lectures

Following the Swalesean genre-based approach discussed in 2.3, several attempts have been made to identify rhetorical move structures in academic lecture introductions (for example Lee 2009, Shamsudin and Ebrahimi 2013, Thompson 1994, Yaakob 2011, Živkovi 2014). Most studies of lecture openings have been undertaken using models similar to those used to analyse written academic genres, although clearly there are differences between lectures and for example, research articles. Building on Thompson's Lecture Introduction Framework (1994), introductions are treated as a subgenre of the academic lecture, and two or three main introductory stages involving warming up (housekeeping and previewing), setting up (in terms of topic, scope and aims) and putting the topic into context (in terms of its importance, and the students' prior knowledge) are identified (Lee 2009, Thompson 1994, Yaakob 2013).

Yet, the structural conventions of spoken academic genres are particularly difficult to identify. Moves are not usually labelled in a manner analogous to titles or section headings, and speech events unfolding in real time are of necessity more disorganised and idiosyncratic than texts carefully drafted for publication or coursework submission. It may be that body language and other visual clues (Rowley-Jolivet and Carter-Thomas 2005, Yaakob 2013, Yeo and Ting 2014) or *phonological paragraphs* marked by changes in pitch and intonation (Thompson 2003) signal transitions between stages in lectures. However, only small samples of lectures have been analysed with this in mind because the major spoken academic corpora are not annotated for visual or prosodic features.

Upcoming lecture content may be signalled in some way in the early part of lectures, to aid students' cognitive organisation of the speech event they are about to hear. Beyond this, it

is not clear why focus should be on the first part of lectures rather than any other part. The research into lecture introductions does not establish if something significantly different happens in this opening part because there is no analysis of the discursive strategies operating in non-introduction (that is, the rest of the lecture).

Some researchers identify the macrostructure of lectures from a holistic point of view that tries to divide *all* discourse into such top-level component parts (for example, Straker Cook 1975, Young 1994). This approach accounts for every token, so that all content is categorised. It is these more comprehensive approaches to functional categorisation that come closest to offering a structural breakdown of the academic lecture at the macro-level.

Straker Cook's (1975) analysis of extended monologue constitutes an early attempt to describe the discursive building blocks of a lecture in terms of the whole. His investigation into discourse structure was based on a 47 minute recorded and transcribed Soil Science lecture, which was delivered to 17 students (1975: 21). Grounded in a detailed manual analysis of the data, the lecture structure is depicted at the primary and secondary levels as "a hierarchical arrangement of units identified in terms of rhetorical functions" (emphasis original. 1975: 64), as depicted in Figure 2.1.

What emerges from this organisational analysis is that although various units of the lecture can be identified and divided hierarchically at the structural level, there is no indication of a clear or determining pattern of occurrence.

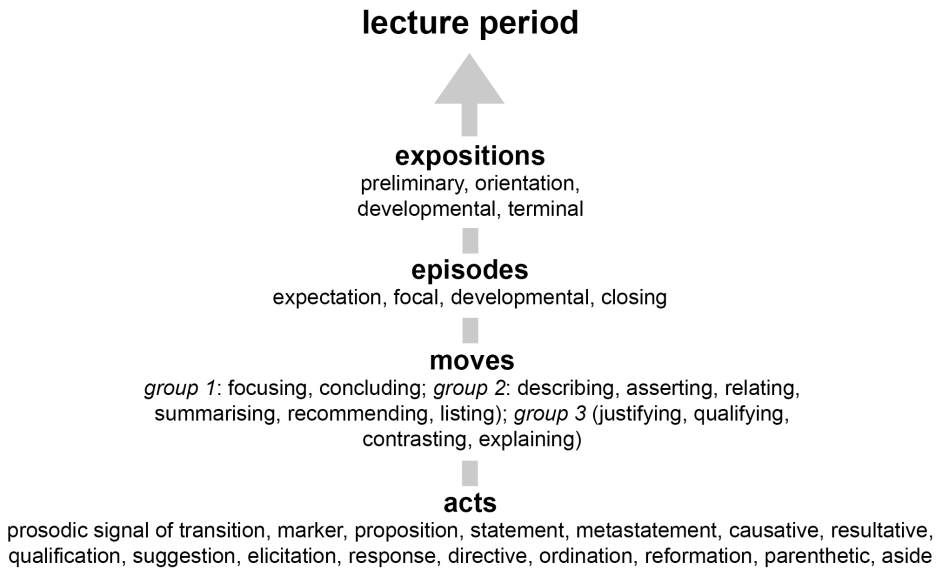


Figure 2.1: Arrangement of lecture discourse units (adapted from Straker Cook 1975)

These units are ranked from the highest to lowest level of discourse organisation, as shown in Table 2.1.

Regarding lecture *episodes*, Straker Cook observes that:

> It seems likely that some ordering should operate at this juncture, but it may only become evident through work with a larger corpus of recordings. (emphasis original. 1975: 67)

The description of the function of moves is cautious because "the limited data make it a little difficult to identify structural constraints" (1975: 68). Some estimations are proposed as to the likely position and occurrence of bound moves (1975: 74), but these estimations are also immediately qualified with reference to the size of the dataset. Despite problems of generalisation of the data, for Straker Cook an identifiable and hierarchical – if not prescriptive – relationship exists between the different structural parts of the lecture he analysed.

discourse unit	description	structure
Lecture	The *lecture-period* (excluding reference to external arrangements)	An unordered series of expositions. No differentiation is made in the structure (for example, initial, medial and terminal) of these expositions
Exposition	Roughly equated to a pedagogical stage in the lecture, but with no discernible pattern of occurrence	Composed of four elements (*preliminary*, *orientation*, *development*, and *terminal*) and realised by four classes of episode (*expectation*, *focal*, one or more *developmental* episodes, and one or more *closing* episodes)
Episode	Length varies from 15 seconds to three and a half minutes, partly dependent on pedagogical aim: whether the lecturer is laying the ground or developing an idea	The structural function is defined as *expectation*, *focal*, *developmental* and *closing*. Realised by three classes of moves (a *focusing* move, one or more moves of classes other than focusing/concluding, and a concluding move) and may contain both free and bound moves
Move	Three groups (comprised of 12 classes) of moves are identified	Group 1 (*focusing*, *concluding*) shape and demarcate the episode. Group 2 (*describing*, *asserting*, *summarising*, *relating*, *recommending*, *listing*) largely function rhetorically – offering propositions contained in the focusing move of Group 1 - and operate freely. Group 3 consists wholly of bound moves (*justifying*, *qualifying*, *contrasting*, *explaining*) which may modify a proposition and must modify one of the free moves in Group 2
Act		Seventeen rhetorically distinct *acts*, along with various subclassifications, are identified: *prosodic signal of transition*, *marker*, *proposition*, *delimitation*, *statement*, *metastatement*, *causative*, *resultative*, *qualification*, *suggestion*, *elicitation*, *response*, *directive*, *ordination*, *reformation*, *parenthetic*, *aside*

Table 2.1: Hierarchy of lecture discourse units (adapted from Straker Cook 1975: 64-86)

Perhaps the closest a corpus linguistic approach has come to a full structural analysis is presented by Young's (1994) phasal analysis of the macro-features of monologic lectures. Young examined the way in which situational factors generate the linguistic choices behind the language events that occur during lectures. These language events are placed within a framework of *phases*, which are defined as:

> […] strands of discourse that recur discontinuously throughout a particular language event and, taken together, structure that event. These strands recur and are interspersed with others resulting in an interweaving of threads as the discourse progresses. (Young 1994: 165)

Young (1994) based her analysis on seven two-hour-long university lectures. Three were given in a Western European university by L2 speakers of English in the subject areas of soil physics, sociology, and economics. The other four lectures were given in two North American universities in the subject areas of geology, sociology, economics, and engineering. The phasal analysis procedure of the corpus entailed: 1. analysing each line for semantic and syntactic choices, 2. identifying the language choices – and thus phases – made within each line, and 3. distinguishing the phases by labelling them and listing their constituent lines (Young 1994: 165). The labels that Young applied in this last stage of analysis describe the macrostructure of the lecture, as outlined in Table 2.2.

phase (type)	**description**	**features**	**occurrence**
Discourse Structuring (metadiscourse)	An indication/ announcement that a new direction will be taken in the lecture. Content prediction facilitates comprehension	Use of: particular verbal groups, rhetorical questions to indicate upcoming content, commands to emphasise information, and modals of prediction and intention to signal upcoming topics	Roughly equivalent to how many points the lecturer makes
Conclusion (metadiscourse)	Points made throughout are summarised. Reinforces the discourse structuring phase	The process is relational as already discussed information is classified. This is a *neutral* phase: no variation in mood, no significant use of modals or evaluation. Information is presented as *factual*	Roughly equivalent to how many points the lecturer makes
Evaluation (metadiscourse)	Information given (or to be given) in the lecture is evaluated, and so the audience is instructed how to weigh content	Predominating process is attributive relations. Like Conclusion, no variation in mood, no significant use of modals or evaluation offered as information is presented as *factual*. Accompanied by evaluative language	Less frequent than the Discourse Structuring or Conclusion phases
Interaction (content)	The occurrence of contact initiated by the lecturer with audience – undertaken to ensure understanding and decrease distance	Dialogue through asking and answering questions (lecturer) – particularly through polar interrogative questions. A difference with the Conclusion phase is the use of polar interrogatives (intended to be answered by someone other than the speaker) instead of rhetorical questions (the main realisation of the interrogative in the Discourse Structuring phase)	
Theory or *Content* (content)	The transmission of theoretical information (e.g. theories, models, definitions)		Interspersed with metadiscoursal phases and Interaction
Examples (content)	Illustration of theoretical concepts through concrete and familiar examples		Often more common than theory (exemplification is an important aspect of this type of speech act)

Table 2.2: Six *phases* of discourse within university-level lectures (adapted from Young 1994)

Like Straker Cook's approach, in opposition to the simple linear discourse pattern of *introduction*, *body*, and *conclusion*, a phasal analysis recognises various segments that repeat and are entwined; Young's phases echo Straker Cook's expositions in function. Predictive or introductory – and their ensuing conclusive – strands occur throughout the lecture. Both Young (1994: 173) and Straker Cook (1975: section 3.3) argue that such phasal/expository analysis offers a more realistic description of the genre of the academic lecture than the beginning, middle, and end pattern that is typically outlined in EAP listening materials.

Although Young's research is perhaps the most influential example of an analysis of spoken corpora at the macro-level, certain questions remain concerning the methodological approach to the data on which her conclusions were based. Young (1994: 160) suggests that the identification of "a macro-structure across levels and across disciplines" is a driving force behind the research. Walsh and Crawford Camiciottoli (2001: 174), however, point out that her corpus is made up of lectures delivered largely by native speakers – with the implication that her results cannot be generalised across all EMI lecture contexts. Perhaps most importantly, the small size of the corpus on which the research is based, and the diversity of disciplines it includes, brings into question the extent to which her claims can be generalised to all lectures in these disciplines. Despite these limitations, Young's phases remain unique in the perspective they offer for analysing the structure of the academic lecture.

This section has looked at strategies for classifying lecture content at the highest level. Straker Cook (1975) attempted to divide lecture monologue into the type of distinct moves and acts that Sinclair and Coulthard (1975) were so influentially able to distinguish in classroom discourse of the same period. Young's (1994) model of free-floating phases and Straker Cook's episodes offer a useful way of categorising all lecture content.

Work on macro-markers suggests that they can indicate some key propositions, and thus discourse units, within lectures. Yet, as Flowerdew (1994: 16) points out, identifying lecture macrostructure poses a particular challenge. To date, there does not appear to be an unproblematic or predictable model for categorising all parts of the lecture at the macro-level; a description of a set of categories that provide a comprehensive functional description of lecture discourse has not been identified.

2.4.3. Micro-features of lectures

Some studies have suggested that certain micro-features have a metadiscoursal role in lecture macrostructure and perform specific functions which can be disciplinary-specific. Although there is an assumption that micro-features can indicate larger structures, there is little systematic or large-scale quantitative data that convincingly binds the two concepts.

For example, *you* and then *I* were found to be by far the most common pronouns in the small cross-disciplinary subcorpora of lectures constructed by Plaza and Álvarez (2013: 190) and Fortanet (2004: 50) from the Michigan Corpus of Academic Spoken English (MICASE) (Simpson et al. 2002). MICASE contains approximately 1.7 million words of spoken academic language gained from various, largely non-monologic speech events. Plaza and Álvarez (2013: 190) conclude that academic discourse in the sciences has a more collectivist orientation because when they looked only at lectures from biology and health sciences, contrary to the general pattern, *we* was more common than *you* or *I*. In a small corpus of mathematics lectures, Rounds (1985) similarly found that *we* occurred up to three times more often than *I* or *you*, and Dafouz, Núñez and Sancho (2007: 653) report that *we* (particularly the inclusive *we*) occurred on average three times as often as *I* and almost twice as often as *you* in their corpus of three aeronautical engineering lectures.

The size of these corpora, however, brings into question their representativeness. In Dafouz, Núñez and Sancho's three lectures, for example, significant variation occurred; in lecture 1, *we* occurred 31 times more per 1000 words than *I* (which is more than ten times as often),

but in lecture 3 *we* only occurred 2.5 times more per 1000 words (2007: 653). The generalisations concerning quantitative patterns are not supported by large datasets.

Based on more qualitative analysis of the retrieved pronouns and associated verbal clusters, there is agreement that lecturers generally use pronouns to establish their position in relation to students, other colleagues, and scientific procedures. *We*, for example, reduces speaker distance and *I* tends to refer to personal experience and knowledge, as opposed to professional standing (Dafouz, Núñez and Sancho 2007: 648, Plaza and Álvarez 2013: 189).

Micro-features have been situated within larger lecture structures and functions through qualitative analysis. Plaza and Álvarez (2013), for example, locate pronouns within a phasal analysis based on Young's (1994) categories. They assert that nominative pronouns occur most often at the beginning of clauses, often at the start of a turn, frequently in the discourse structuring phase (cf. Young 1994) in anticipation of upcoming content (Plaza and Álvarez 2013: 191). *I* is also identified as common in the concluding phase, which is attributed to an individualist attitude (2013: 192). With reference to commonly occurring clusters, *you can* is identified as frequent in the discourse structuring phase, and presented as an indicator of logical possibility as part of explanatory discourse. Following Thompson (1994: 176), *we* is reported to frequently occur at the start of the discourse structuring phase and collocate with modal verbs that have a theoretical or exemplification function (Plaza and Álvarez 2013: 194).

Dafouz, Núñez and Sancho (2007) also looked at linguistic and pragmatic features of lectures, focusing on speaker stance in the setting of an international audience. They conclude that:

> […] in addition to a solidarity strategy, *we* works as a macro-organisational principle guiding both lecturers and students throughout the speech event. (emphasis original. 2007: 658)

Fortanet (2004: 63-64) concurs that one of the main usages of *we* is in a metadiscoursal capacity. She found that *we* was used in discipline-specific ways, and was often found in utterances that functioned to clarify, generalise, indicate a joint deduction, or recount a research process.

A metadiscoursal role for particular pronouns is therefore claimed by Fortanet (2004), Dafouz, Núñez and Sancho (2007), and Plaza and Álvarez (2013), but no systematic identification of lecture macrostructure is described; the pronouns were first identified and *then* situated within a larger structure based on immediate context. Discipline-specificity and the occurrence of particular language patterns were also identified, but the number of lectures analysed in each of these studies was small. In order to establish a relationship between the micro and the macro, a more intuitive entry point might have been to first identify the phases and *then* describe their constituent micro-features, using larger datasets.

Aside from pronouns, analysis of various other micro-level lexical items that are common in the spoken mode and perform a discourse function (such as *ok*, *right*, *yeah, now*, *just*, and *so*) is regularly performed. Different taxonomies refer to these micro-features as *discourse markers* (Fraser 2009), *discourse connectives* (Blakemore 1992), *discourse particles* (Schorup 1985), *pragmatic particles* (Fried and Östman 2005) *pragmatic markers* (Fraser 1988, Fraser 2009, Schiffrin 1987), or *relevance markers* (Hunston 1994). Together they constitute:

> […] a class of lexical expressions drawn primarily from the syntactic classes of conjunctions, adverbs, and prepositional phrases. With certain exceptions, they signal a relationship between the interpretation of the segment they introduce, S2, and the prior segment, S1. They have a core meaning, which is procedural, not conceptual, and their more specific interpretation is 'negotiated' by the context, both linguistic and conceptual. (Fraser 1999: 931)

Such markers can be single lexical items or longer strings of text. Strings that function at a similar metapragmatic level to indicate temporal or presentational sequence have been referred to as *macro-organizers* (Chaudron and Richards 1986) and *cue phrases* (Knott and

Dale 1994). In this thesis, *discourse markers* (cf. Fraser 2009) will be employed as an umbrella term for such single- or multi-token units of text.

Discourse markers that function as metacommentary are explicitly linked to overall text coherence in studies of general discourse. Schriffin (1987: 24), for example, explicitly views discourse markers as "indicators of the location of utterances within the emerging structures, meanings and actions of discourse". She found that the discourse markers *oh*, *well*, *and*, *but*, *or*, *so*, *because*, *now*, *then*, *y'know*, and *I mean* function within *participation frameworks* (through which speakers and hearers relate), *ideational structures* (made up of ideas and propositions), *action structures* (situating speech acts in terms of preceding and following acts, and what was intended), and *exchange structures* (turn-taking) (1987: 24-26). Many of these structures are not applicable to lectures, however, which are largely monologic.

Studies that look specifically at discourse markers in lectures generally assign to them an identifiable structural function in the organisation of the whole text. There is a prevailing assumption that certain markers "display important signalling features in non-turn-taking events such as lectures" (Othman 2010: 678), and that these signals can reveal the overall discourse structure (Fraser 2009: 893). To varying degrees, discourse markers are considered to provide some sort of skeletal map of the main points of lectures, and the focus of much of such research is on enabling lecture comprehension, for both L1 and L2 listeners.

The association of discourse markers and comprehension was established in early studies of lecture discourse. Wijasurija (1971), for example, counted and categorised the discourse markers and inter-sentence connectives in 46 hours of university level lectures in order to improve procedures in testing and teaching listening comprehension. Morrison (1974, cited in Jordan 1997: 38) investigated the features of seminars, tutorials and lectures that he thought were most problematic for postgraduate students of science, and categorised these

features in three ways: 1. the referential system (logical connectors, reference, and predictability), 2. lexis (particularly idiom and nominalised groups), and 3. phonology (such as delivery speed, accent and pronunciation). The field of investigation has since then become what Fraser (1999: 932) describes as a "growth industry", with increasing emphasis being placed on the signalling and organisational properties of the micro-features of language in relation to macrostructure. There continues to be, however, a largely intuitive approach to the initial identification of discourse markers, and a lack of systematic large-scale investigation into their function.

Discourse markers have been categorised in various ways. Fraser (1988, 1996, emphasis original 2009: 892-894), for example, identified four types (plus subtypes) and interpreted their (speaker intended) meaning as follows:

1. *basic markers* signal illocutionary force, such as "I admit that I feel a little ill"
2. *commentary markers* (including *parallel markers*) convey an attitude, such as "[a]mazingly, John made it home before dark"
3. *discourse markers*, such as "I agree but I can't do it today"; a label that was originally employed by Fraser as a type of *pragmatic markers* (1996) and later used as an umbrella term for all identified *discourse markers* in Fraser (2009)
4. *discourse management markers* "signal a metacomment on the structure of the discourse":
    a. *discourse structure markers*, as in "[i]n summary, the economy has not flourished under the Bush administration"
    b. *topic orientation markers* (including: *return to a previous topic*; *continuation with the current topic*; and *digression from the current topic*), for example: "[t]o change the topic, when are we going to have dinner"
    c. *attention markers*, including: *ah*, *alright*, *anyway*, *anyhow*, *hey*, *in any case*, *in any event*, *now*, *now then*, *oh*, *ok*, *so*, *so good*, *well*, and *well then*

Fraser's data are drawn from the British National Corpus (BNC), MICASE, internet blogs and political press conferences, which are largely spoken, but also include constructed examples. Quantitative information about sources and occurrences is not given in these studies; Fraser expressly states that he is "not concerned with the relative frequency of these terms or the context in which they occur", and that some analysis is "intuitive" (2009: 894, 896). Although many examples given have been validated in corpora of authentic

speech, the proposed interpretations may be radically changed when the role of context and prosody is taken into account (Fraser 2009: 897). Pragmatic function, it seems, is not systematically addressed.

Although his data are not entirely authentic, nor systematically identified and extracted, and are taken from a range of speech events, Fraser's categorisations do seem to map onto functional analyses of discourse markers identified in studies of authentic lectures. The function of Fraser's (1996) *discourse markers*, for example, is to point to the relationship that was intended between adjacent discourse chunks. This category shares ground with, for example, Chaudron and Richards' (1986: 127) *segmentation discourse markers* (*well*, *ok*, *now*, *and*, *right*, *alright*) and Biber et al.'s (1999: 1046) *utterance introducers* (*well, now*). Fraser's (2009: 893) *commentary* markers also seem to fall within the remit of Chaudron and Richards' (1986: 127) *emphasis markers* (*of course*, *you can see*, *you see*, *actually*, *obviously*, *unbelievably*, *as you know*, *in fact*, *naturally*).

Fraser's first three types of discourse markers relate to functions within an utterance, whereas the fourth relates to overall discourse organisation, a distinction that Chaudron and Richards (1986: 127) refer to as *micro-markers* and *macro-markers*. The distinction seems to be based on whether the discourse marker/s have an organising/signalling function solely in their local context (for example, at the utterance level), or whether they have the potential to indicate overall discourse organisation, and so enable the specific features of discourse genres (such as lectures) to be identified. Some studies have identified such organising potential at the narrow level of the single lexical item, whilst others take a wider approach.

Schleef (2005) looks at the sociolinguistic role of discourse markers such as *okay*, *right*, *like* and *you know* and identifies four categories of usage: *transition markers*, *modal question tags* (asking for confirmation or information), *progression check question tags*, and *backchannel signals*. This categorisation suggests that discourse markers may perform very

different functions in lectures and in other spoken academic genres; progression check tags, for example, might not be a common feature of seminars.

Lindemann and Mauranen (2001) interrogate the function of the discourse marker *just* at the level of utterance. Using MICASE, their analysis of 3-grams within a random sample of 100 instances of monologic and interactive/dialogic speech events led to three observations concerning *just*:

1. it appears to occur in metadiscursive frames (for example: *let me just*, *I just wanted to*)
2. it tends to co-occur with hedges (either mitigators like *a little bit* or vagueness indicators such as *sort of*, *kind of*, or *something*, or *so*)
3. it can co-occur with both metadiscursive items and hedges

Five functional categories (and subfunctions) of the instances of *just* in cross-disciplinary subcorpora were then identified and quantified (Lindemann and Mauranen 2001: 465-468):

1. minimising (74%):
    a. limiting function, paraphrase: i. only (a neutral limiter), ii. merely (limiter with the connotation *not enough*), iii. simply (limited to a simple interpretation)
    b. mitigating (with or without limiting) function (connotation of unimportance)
2. emphasising (7%), paraphrase *really/absolutely*
3. particularising (2%), paraphrase *exactly*
4. specificatory/temporal (8%)
5. ambiguous (9%)

At 74%, *just* overwhelming functions as a minimiser. The findings of this study are based on data extracted systematically with an element of inter-annotator reliability (IAR) testing.

The findings of Lindemann and Mauranen (2001) differ from those of Aijmer (1985), who reported that *just* functioned most often as an emphasiser in the British English casual conversation component of the London-Lund Corpus. Lindemann and Mauranen (2001: 472) suggest that either the language variety (American English) or the speech mode (academic speech events) of MICASE may account for this difference. Varieties of English, it seems, may have an effect on the lexical content of academic speech.

Grant's (2011) investigation into the use of *just* in the BASE corpus, however, negates the language variety explanation. The BASE corpus is comprised of 160 lectures and 40 seminars distributed evenly across four broad disciplinary groups (Nesi and Thompson 2006). Grant examined the occurrence of just in 32 lectures and eight seminars. Lindemann and Mauranen's (2001) categories were used to distinguish the function of the 1427 instances found. The main function of just was – as Lindemann and Mauranen (2001) found – as a minimiser (84% overall and 91% in the physical sciences). Grant (2011) associates minimisers with a metadiscursive *teacher talk* frame (cf. Richards, Platt and Platt 1992) which emphasises important information through suggestions, directions or contrast. Unlike Lindemann and Mauranen, Grant also identifies frequently occurring groups of words, such as *just about*, *just in case*, *just as well*.

What none of the investigations describe, however, is the wider pragmatic context of the use of *just*; whether, for example, it tends to occur when the lecturer is delivering new content or in periods of summation. The local use of the lexical item is explained in detail, but not contextualised in the larger lecture structure.

Categorising small strings of language as performative of discourse functions in lectures is clearly a common research focus. There is, however, inevitable disagreement over which discourse markers to examine and which functions to associate them with, as function depends on context. Brinton (1996), for example, details 13 possible functions for *well,* referenced in the work of at least as many separate scholars (1996: 36-37). Without larger context, function is extremely difficult to establish, as is its relation to the macrostructure of lectures. Fraser (2009: 898) acknowledges that considerably more work on discourse markers would be necessary to establish the details of the "big picture".

Fraser (2009) does make some attempt to situate micro-markers within this big picture. His fourth group of discourse markers, *discourse management markers*, is related to a larger organising function: they "convey the contribution of the following discourse segment

within the overall structure of the discourse" (2009: 893). Examples include (emphasis original. 2009: 893):

> We have three topics today. **First**, we will discuss the ethics of cheating. **Then** […].
>
> **In summary**, the economy has not flourished under the Bush administration.

Not only does Fraser suggest that discourse markers can situate a given segment within the overall discourse structure, but he also stresses the "critical role" played by such markers in the interpretation of utterances (2009: 892).

Some studies have examined the presence of discourse markers from a cross-cultural perspective, to ascertain whether speakers naturally adapt their speech based on the L1 of the audience. Crawford Camiciottoli (2004), for example, identified lexicogrammatical patterns, or *audience-oriented relevance markers*, in a small corpus of L2 guest lectures on business studies delivered in an economics faculty in Italy. Comparison was made with 14 L1 guest lectures from MICASE. Focus was placed on the relevance markers that are used to comment on the organisation of the lecture, or signal upcoming content (such as "[w]hat I'm going to talk about today", "[f]irst let's take a look at", "[w]e'll come back to that later", "[y]ou'll see that in just a minute"). The function of these markers is described as "a form of interaction between lecturer and audience that interrupts the flow of informational content" (2004: 40). Differences in the usage of interactive discourse structuring between L1 and L2 speaker lecturers was shown, with a tendency for greater usage by the latter. The posited hypothesis is that the more frequent use of such structures reflects a (conscious or unconscious) attempt to aid student comprehension, perhaps resulting from the lecturer's awareness of their own language needs (2004: 49). The context of delivery and language background of both speaker and receiver, then, may affect not only lexical choice, but also how structures in lectures are signalled, and possibly how the lecture itself is structured.

In perhaps the most comprehensive study of micro-linguistic features in lectures to date, Deroey and Taverniers (2012) examined the entire lecture component of the BASE corpus

using a semi-bottom-up procedure. They define *relevance markers* (cf. Hunston 1994) as metadiscursive "lexicogrammatical devices that overtly mark the relative importance or relevance of points which are presented verbally or visually" (2012: 222). A manual search for relevance markers was performed on 40 (out of 160) lectures and the identified forms were then retrieved from the whole corpus, in conjunction with concordance-checking.

In total 782 examples of various patterns were found. The most commonly occurring were verb patterns such as "remember slavery had already been legally abolished", followed by noun patterns such as "the point is" (Deroey and Taverniers 2012: 224). Through close and systematic analysis of linguistic patterns, the mechanism by which importance is evaluated, and relevance identified, is demonstrated at the level of utterance – in the local context. The leap between identifying how important points are signalled to mapping the overall structure of informational content (from the micro to the macro), however, is not made clear.

Several studies have attempted to test in some way the special role performed by discourse markers in overall meaning comprehension. Chaudron and Richards (1986), for example, presented four versions of a lecture on American history to two groups of students for whom English was a second language. In one version of the lecture they tested the effect on comprehension of the inclusion of discourse markers that signal major transitions, measured by: 1. a recall cloze measure of sample lecture content, 2. a multiple-choice test covering all lecture content, and 3. a true/false test using ten items which covered the entire lecture content. The lecture that included discourse markers was more successful in aiding content recall than the version that included only markers of segmentation and intersentential connection (1986: 122). Chaudron and Richards (1986: 117) refer to their data as the product of a "natural lecture". As the lecture was scripted to accommodate test variables, it can be argued that this presents a different discursive mode to more spontaneous, reactive lectures, which are delivered without this level of scripting.

Olsen and Huckin (1990: 34) similarly examined a 16 minute "realistic sample lecture" in mechanical engineering, which led to a distinction between *point-driven* and *information-driven* lectures. Signalling devices within the lecture transcripts that were identified as important include "[t]he real problem is that […]", "[t]he whole idea here […]", and "I'll indicate one in just a minute" (1990: 37). Key information points were marked in the text by the signalling devices and the immediate recall of these points by L1 and L2 speakers of English was tested (1990: 36). The sample lecture was deemed to be point-driven because successful comprehension required students to understand the organising rhetorical framework of the lecture in combination with the role of theory within this framework. Students who missed the Problem-Solution (cf. Hoey 1983) discourse structure and the main points therein reportedly failed to grasp the overall gist of the lecture.

Information recall dependent on whether rhetorical signalling cues were present in the lecture discourse was also the focus of investigation for Dunkel and Davis (1994). Two lectures were delivered in two forms to L1 and L2 speakers of English. One form contained explicit rhetorical cues pertaining to lecture organisation and the other did not. The two groups took notes in their native language during the lecture delivery and were then asked to recall everything they remembered about the lecture. Three main research findings are discussed in Dunkel and Davis' (1994) paper: 1. the L2 group took more lecture notes than the L1 group, 2. where rhetorical cues were given, more lecture notes were taken, and 3. the post-lecture recall of the L1 group was almost twice as great in terms of words written and information units present. However, the key finding in this study is that the presence of rhetorical cues did not impact significantly on the quantity of information units or words recalled, which contradicts the findings of Chaudron and Richards (1986) and Olsen and Huckin (1990).

Such tests of recall based on the presence or absence of such markers assume that the sole intention of lectures is to convey a series of facts to students. Focus on the variable of information retrieval omits other functions particular to the oral delivery mode, such as

stimulating ideas or modelling critical thinking. Given the reported pedagogic motivation of such research, measuring the *success* of lecture comprehension based on the recall effect of rhetorical cues may result in unwarranted, perhaps even distracting, emphasis on this narrow feature of discourse in teaching and teacher training materials.

The organising/signalling potential of microstructures at the macrostructural level of lectures has not been systematically established. Certainly other functional interpretations are possible. For example, particular micro-markers may sometimes simply allow the listener more time to process concepts. Disfluencies (hesitations, fillers and false starts) are common to the oral mode (Chafe 1985), and Chaudron and Richards (1986: 116) acknowledge that signals such as *well*, *so*, and *now* may also sometimes:

> […] serve as filled pauses giving listeners more time to process individual segments of a piece of discourse; they hence provide more opportunities for bottom-up processing.

Brinton (1996: 37-38) also suggests that discourse markers may serve to: sustain discourse (as a filler or delaying tactic) whilst the speaker holds the floor, initiate or close discourse (claiming listener attention), or repair discourse. The primary function of some discourse markers in some contexts, then, may bear little relation to discourse structuring which aims to highlight important concepts. It would be necessary to explore examples in context in a large dataset to draw any quantitative conclusions regarding function.

Researchers who look at the role of discourse markers have not agreed on a standard taxonomy. They all claim, however, that discourse markers are functionally important and not limited to general conversation. The discourse markers identified perform functions of some use within lectures, such as conveying speaker attitude or signalling illocutionary force at the level of utterance. The case is repeatedly made for the metacommentary function of some discourse markers on overall lecture structure, which indicates an assumption that discourse markers have the potential to reveal a structural map of lecture discourse.

The logic that lecture structure can be identified without quantitative, macro-level analysis implies that qualitative analysis of decontextualised examples of micro-level features is sufficient to inform broad claims regarding structure. What is not clear from analyses at this level is how these discourse markers contribute to the creation of a coherent whole. Although analysis of micro-markers could be useful for identifying the linguistic features of larger structural units, there is a tendency for micro-markers to be identified alone and not placed within any larger unit.

It is probably not very helpful to consider micro-features in isolation without considering what the lecturer was doing at the point of their usage: the wider context. At different moments, lecturers are doing different things, such as drawing attention to a particular piece of information, making a joke, telling a story, or summarising what they have said and done to that point. Inevitably, different parts of lectures function differently. The consideration of micro-linguistic features, such as discourse markers, may be meaningless without consideration of the larger component part of the particular lecture in which they occur.

2.4.4. The case for pragmatic corpus annotation

The analysis of the macro- and micro-features of lecture discourse in the literature discussed so far points to the need to identify pragmatic functions in the various structural phases of the lecture. Progress with making generalisable conclusions, however, has been hampered by a lack of large-scale quantitative data. Pragmatic annotation is a slow and labour-intensive task, and automated solutions have not yet addressed pragmatic features beyond micro-features or speech acts, or dealt with the problems of segmentation associated with spoken monologue.

This section will discuss some studies that have attempted to encode pragmatic aspects of text in larger datasets (corpora). This approach requires computational assistance, which in turn means adding data about structure (markup) and linguistic features (annotation) to the

raw texts. The associated technical terms will be discussed in detail in 3.1.1. The first part of this section provides an overview of the theoretical pros and cons of the process of adding encoding information to texts. Findings from software options are then discussed. Finally, some examples are given of large bodies of lecture data in which, in some way, pragmatic information has been encoded in academic monologue.

In the field of corpus linguistics, the annotation of data can be viewed as an adulterating, as well as an enriching, process. Although the camp is fragmented, the majority view is that corpus annotation (especially when computer-aided) has a place in language analysis on the proviso that clear and agreed standards are followed. Smith et al. (2008: 163) define adding interpretive encoding information as "the fundamental need of the linguist" as this "allows the linguist to produce more rigorous descriptions — and theories — about language in use". The notion of language in use is fundamental to pragmatic analysis.

The addition of inline metadata enables specific contextual information to be indexed. Rühlemann (2010: 290) depicts the corpus linguist as impoverished compared to the original producer/s of a text, as "merely a kind of eavesdropper cut off from the wealth of background knowledge ratified participants share". Normally, the corpus-using eavesdropper cannot access the participants (to confirm or clarify meaning) or the social/non-verbal context of the situation. Rühlemann (2010: 290) explains that "corpora have long been seen by some researchers as unfit for use in pragmatic research" because they lack the contextualisation upon which an understanding of pragmatic phenomena depends.

Baker (2006: 18) advocates reintroducing context (provenance, authorship, motivation) to decontextualised language examples, but stresses that: "[o]ur findings are interpretations, which is why we can only talk about restricting bias, not removing it completely". This acknowledgement, Baker points out (citing Hunston 2002: 123), can be positively construed

as a methodological advantage because it forces the inevitably subjective ground between observation and interpretation to be laid bare.

Culpeper et al. (2008) discuss a variety of ways in which spoken corpora have comprehensively and consistently captured a range of information, including the contextual. Pragmatics is envisaged as a triadic relationship between linguistic forms, world entities and the user, with the potential meaning of an utterance described in terms of five components: formal, illocutionary, implied/inferred, interactional and contextual (2008: 615).

The debate concerning the value of encoding pragmatic meaning is part of a wider discussion. The view of annotation as enrichment is widely, rather than universally, held, and there have been a few dissenting voices, most notably Sinclair's (2004) adhering to the idea of what Garside et al. (2007: 4) refer to as the "raw" or "pure" corpus. Sinclair (2004: 191) describes the "interspersing of tags in a language text" as a "perilous activity" due to the ensuing loss of integrity of the text, not least because the metadata (tags) provide an unchallenged and potentially skewed lens through which data are viewed. He warns researchers against studying metadata and not language. In untagged text, on the other hand, "uncontaminated" patterns are observable (2004: 191).

Hunston (2002: 93) also comments on the potential dangers of viewing encoding as *value added*, describing it as a "double-edged sword" because:

> […] the categories used to annotate a corpus are typically determined before any corpus analysis is carried out, which in turn tends to limit, not the kind of question that can be asked, but the kind of question that usually is asked.

These cautions flag up an ever-present consideration in the bottom-up versus top-down corpus linguistic methodology: the choice between encoding metatextual linguistic features as they are encountered (perhaps with a theoretical framework in mind) versus encoding a predetermined list of such features.

Existing in parallel to the adulteration versus enrichment debate is the question of interpretation, and the argument that any automated identification of language patterns alone is void of meaning. Fowler (1991: 90) suggests that "there is no constant relationship between linguistic structure and its semiotic significance". Additional metatextual layers of interpretation can index certain language phenomena, but these will always require contextualisation.

The subjectivity of the process of segmentation – or division of a text into constituent parts – is a significant concern.

> Segmentation requires us not only to state what unit we will be analysing, but also to define it in a way that will enable us to measure one unit against another, and, by so doing, ensure a level of consistency. (Culpeper, Archer and Davies 2008: 632)

Moreover, there is no purely objective, mechanistic way of deciding what label or labels should be applied to a given linguistic phenomenon (Garside and Rayson 1997: 2). In order to divide the text, a clearly defined taxonomy must be established. This does not necessarily mean a top-down implementation of particular categories, but that texts can be studied and relevant categories gathered and revised.

As work on retrieving relevance markers shows (for example Deroey 2013, Deroey and Taverniers 2011), one approach to identifying speech that serves a pragmatic function is based on predetermined categories. Cheng (2010), for example, looked at the speech act of *thanking* in data from MICASE speech events and the spoken part of the BNC, including dialogue in unscripted informal conversations and formal meetings. The findings indicate that in addition to the predicted (and dominant) function of expressing gratitude, thanking can also mark an ending or the rejection of an offer (2010: 268). Cheng (2010: 265) states that "[t]hanking is easy to recognise because the speaker almost always uses an explicit expression". Some features of language may be relatively easy to retrieve, but many are not realised through such explicit and predictable forms.

Other researchers have put more emphasis on illocutionary force when retrieving examples of speech acts. McAllister (2015: 32), for example, used an "identification in context" approach to defining and extracting *directive* speech acts in the academic context. The categories identified were not predefined, but rather based on close reading/listening to the transcription and audio versions of conversations between two speakers taken from the T2K-SWAL (Test of English as a Foreign Language (TOEFL) 2000 Spoken and Written Academic Language Corpus) Corpus. As well as confirming (with empirical quantitative evidence) the presence of speech acts that were predicted in service encounters (such as *information requests* and *suggestions*), this purely corpus-driven approach crucially enables the identification of previously neglected categories of speech act (such as *warnings* and *corrections*) (McAllister 2015: 45).

Various forms of automated extraction have also been used to identify pragmatic functions in text. These attempts usually look at the speech between two speakers, because dialogue offers a particular type of "contextual embeddedness", where interpretation is partially informed by what was said in the previous utterance (Aijmer and Rühlemann 2015: 2).

Specific tools are being developed to help encode pragmatic aspects of transcribed speech. Historically, focus was on lexical or syntactic units, such as the Cast3LB project's focus on coreference and anaphora (Navarro et al. 2003). Pragmatic annotation schemes (both hand-coded and semi-automated) are, however, moving towards encoding information about speech acts, discourse moves and the contexts in which they occur (Culpeper, Archer and Davies 2008: 614). Pr.A.T.I.D (the Pragmatic Annotation Tool for Italian Dialogues) (Savy 2010), for example, offers a multilevel structure in which the status of the dialogue act is encoded differently at each of its three (embedded) levels. By utilising this multi-level construction, pragmatic statements about various aspects of a text can be made and the influence of context more thoroughly addressed.

In addition to extracting and encoding pragmatic functions after they have been identified by humans, attempts have also been made at perhaps the more challenging work of automatically identifying these functions. Even in written texts at the utterance level however, there is inevitable disagreement between annotators concerning the pragmatic interpretation of speech acts, which heralds a particular difficulty for attempts to automate the process. For example, De Felice and Deane (2012) developed a computational model for automating the identification of speech acts in emails written by L2 students as part of the Test Of English for International Communication (TOEIC) writing test. Utterances were encoded with two sets of information: linguistic form and speech act category (which are roughly equated with locutionary and illocutionary acts). Overall, the automated system classified speech acts with 79% accuracy; unsurprisingly, the classification of indirect speech acts caused a large proportion of the 21% of errors (2012: 29-30). A noteworthy observation, however, is that almost half of the misclassified speech acts also caused dispute amongst the three human annotators on whose classifications the model was trained (2012: 30).

A recent step forward in identifying pragmatic categories within transcribed spoken data is Weisser's (2014) Dialogue Annotation & Research Tool (DART), which automatically annotates speech acts, among other pragmatic-related linguistic phenomena. This project is a development of the Speech-Act Annotated Corpus of Dialogues (SPAAC) project (Leech and Weisser 2003). DART was trained using four speech corpora, all of which contain dialogues between two speakers, either in task-driven scenarios or spontaneous telephone exchanges. Conversation topics relate to everyday life, such as goods transportation or music. Weisser (2015: 85) reports that the automated process can identify only high-level or generic speech acts. He gives the example of the utterance "[w]e'll be there at five o'clock", which can be automatically identified as a predictive speech act, but which cannot be classified at a more nuanced level of pragmatic meaning (for example as a *promise*).

DART enables the identification of a range of speech acts. For example, in a telephone conversation between a customer and agent from the Trainline corpus, high numbers of statements, requests for information and directives were identified through DART. In total, within the 35 dialogues analysed, 13 types of speech act occurred with a minimum frequency of 40 (Weisser 2015: 105-106).

In Weisser's opinion, although analysis within DART is based only on dialogue:

> of course we should assume that similar syntactic/ functional units to the ones described here also occur in monologues, as well as to some extent even in written language. (2015: 86)

However, as Weisser also acknowledges, dialogue is easily segmented because it involves clear division (speaker turns). Like punctuation in writing, these turns demarcate a text (in this case, a transcription) into more easily analysable segments. It can be argued that monologues contain no such intuitive segments. Weisser states that identifying a consistent form of units through which to segment monologue is possible, if controversial (2015: 88-89). There do not appear, however, to be any findings from DART analyses which prove this to be the case.

In dialogue, as in written text, the range of speech acts that can be identified is extremely wide, depending on the level of delicacy and the nature of the communication. Discourse from a single event that is largely monologic – such as a lecture – does not involve the same range of relationships and contexts, and its overarching purpose is unchanging. The range of potential speech acts is therefore more limited; a lecturer is unlikely to give many *compliments*, *apologies*, *invitations* or *refusals*, for example, and the speech acts commonly used (such as *directives* or *statements* that inform) are unlikely to vary.

Speech act identification has proved problematic for both human and automated annotators. De Felice and Deane (2012: 5) argue that even in written text it is difficult to analyse speech acts at any level higher than the utterance (such as the level of message)

because longer stretches of text commonly fulfil more than one function. Broader categories of communicative function, encompassing longer stretches of text, may be less contentious. For example, there is less likely to be disagreement over whether to characterise a passage as a *story* than whether a smaller section of that passage is functioning as a *description* or a *request*.

Some attempts at identifying pragmatic functions in lecture monologue beyond the scope of speech acts have been made. MICASE (Maynard and Leicher 2007), for example, has been pragmatically annotated in part. The MICASE taxonomy is based on an "inventory" of what was considered to be the "pedagogically interesting pragmatic content of each speech event" (2007: 108), rather than specific speech acts. To identify and implement the pragmatic tags, a three-pronged approach was taken. Firstly, an abstract was compiled for each of the transcripts in the corpus. The inventory was based on three criteria, which require that the features are: 1. "not easily searchable", 2. "prevalent in the data" to ensure a significant occurrence for researchers to use, and 3. "relatively unambiguous" - the aim was to create an accurate, not exhaustive list (2007: 112). Any unclear instances were excluded. A researcher listened to each speech event whilst reading its transcript. Using the checklist of 25 pragmatic features, the frequency of the features was noted, using the categories of *none, few, or numerous* (2007: 109). This information was included in the abstract for each transcript, along with an overview of the event and its content.

Secondly, the frequency of the 25 pragmatic features in each separate speech event was recorded in the header of its transcript. Thirdly, a subcorpus of 50 transcripts was pragmatically tagged for 12 (out of the 25) features, because "pragmatic tagging identifies specific instances or examples of language clearly performing any of a set of various pre-determined pragmatic features" (Maynard and Leicher 2007: 111). Tagging was carried out by two researchers: the first encoded the transcript and the second checked the annotation and entered the tags into a database.

This three-pronged approach was designed to enable:

> […] three different entry points into the corpus, thus accommodating different research approaches or styles (e.g. top-down vs bottom-up) and allowing access to different groupings of information or vantage points from which to view a single event or the entire corpus. (Maynard and Leicher 2007: 108)

The motivation behind pragmatically annotating the subcorpus was to "expose interesting linguistic phenomena" occurring in MICASE for pedagogic use, rather than to provide a platform on which to base broader generalisations (2007: 108).

However, only a subcorpus of 50 out of 152 transcripts was manually tagged for pragmatic features, bringing into question the representativeness of any conclusions drawn. These transcripts were selected for richness of features rather than at random, although all speech events and disciplines were equally represented (2007: 111). There are 17 speech event types in MICASE, 10 of which are categorised as *non-classroom events* (2007: 82); the choice of academic speech events covers a wide range.

The MICASE transcriptions were created from audio recordings only, which constrains the accuracy of the representation of pragmatic content. Visual clues are helpful in interpreting pragmatic meaning; speaker body language, for example, may enable the identification of humour (such as irony) that may not be retrievable from audio only. Similarly, it is not clear how the inventory of 25 pragmatic features was initially devised, or how this was reduced to the final tagset of 12 features that were noted in transcription headers.

The second of three criteria that determined the compilation of the long list was to include features that are "prevalent in the data"; an aim that is, however, hedged by the need "to strike a balance between features that were prevalent and features that were ubiquitous" (Maynard and Leicher 2007: 112). In some cases, this criterion led to instances of exclusion. For example, humour was removed from the final tagset on the grounds of too many occurrences (2007: 112). This is particularly interesting as the decision not to include humour, based on difficulty of identification and recording, flags it up as an important

lecture function. Its non-inclusion in the MICASE taxonomy seems to reflect the largely intuitive nature of the development of the initial long list of features and resultant final tagset, which does not result from a systematic or corpus-driven methodology.

Perhaps one of the most curious design aspects of the pragmatic annotation of MICASE is the rationale behind including frequency information in the header only rather than annotating specific occurrences within the speech event. The information about the frequency of pragmatic features included in the header is intended to guide users to the transcripts that are most rich in the feature/s of interest. However, although this strategy would allow potentially relevant transcripts to be identified, it crucially does not allow comparison or contextualisation of pragmatic categories within the corpus, or across similarly annotated external corpora. Additionally, no attempt was made to examine the linguistic features of the text identified as having a pragmatic function. Inline annotation (that is, indexing the boundaries of categories identified) would have enabled such analysis.

The MICASE approach does, however, demonstrate the particular value of pragmatic annotation at two levels: 1. in the compilation of the list of categories annotated as a record of types of language that are actually used in spoken academic data, and 2. in the potential for systematically identifying, contextualising and comparing specific language functions.

As far as academic lectures are concerned, progress with pragmatic mark-up has been very slow. Early attempts by Young (1994) and Straker Cook (1975) describe a sort of generic move structure in a limited number of academic lectures, which identified various phases, each with a different communicative function, and each with particular boundaries of fixedness (2.4.2). However Maynard and Leicher (2007: 108) observe that “specific examples of the language that is actually used to accomplish things in the academic community (e.g. explaining, defining) are still not readily accessible”. Even one of the largest fully documented and publicly available corpus of lectures, the BASE corpus, for example, is only encoded for part of speech, pausing, and contextual information.

The identification and assessment of formal characteristics of lecture discourse has been hampered by the very limited quantity of available authentic spoken data, and a lack of information about possible variation across cultural/educational contexts. Aside from Maynard and Leicher's (2007) experimental tagging of a small subcorpus of MICASE transcripts, there does not seem to have been any other attempt to annotate a corpus of lectures to reflect their structure or purpose.

This situation is changing, however, and the pragmatic annotation of MICASE should lead to further such work on the increasing amounts of data becoming available within new corpora of academic speech.

2.5. Research aims

It would perhaps be most helpful to examine larger structural units in terms of their linguistic functions *and* their micro units across a range of cultural/educational contexts, which is the intention of this thesis. My aim for this thesis is to identify certain functions that are specific to engineering lecture discourse and the way in which these functions are dispersed across individual lectures. I also aim to identify the linguistic features that characterise these functions, including features that are specific to given cultural/educational contexts. In order to achieve these aims, I will develop a framework for the pragmatic annotation of academic lectures, and I will test a range of tools and techniques to see how these can contribute to the framework and lead to the development of reusable processes for data extraction and analysis.

CHAPTER 3. METHODS

3.1. Introduction

Data analysis within this thesis is primarily empirical, within the positivist tradition of seeking objective proof through data. The quantitative, partially bottom-up approach enables finer qualitative analysis. A marriage of natural language processing (NLP), corpus linguistic and data visualisation methods determine the first-wave identification of data patterns, which in turn direct qualitative analysis.

After a brief account of the terminology used in this thesis (3.1.1), the groundwork for adopting a largely quantitative, corpus-driven approach is laid out in 3.2. An overview of the data, the Engineering Lecture Corpus (ELC), follows in 3.3, and the stages in data preparation (collection, transcription, markup and annotation, and reliability testing) are described in 3.4. Attention then turns to inter-annotator reliability (IAR) testing in 3.5 and methods of analysis (including simple data mining, corpus linguistic and data visualisation techniques) in 3.6.

3.1.1. A note on terminology

As terms are often used seemingly interchangeably in corpus linguistics, and variation over time occurs, this section clarifies the use of terminology in this thesis.

A *text* is widely understood to describe spoken or written language data (Garside and Rayson 1997: 2), in which linguistic interaction occurs in an operational rather than citational context (Halliday 1993: 23). At the narrowest level, the *text* can be understood as only the sentence that is being processed (Sinclair 2004: 14). More encompassing definitions refer to any piece of information (spoken, written, visual) that can be encoded and stored, usually in an electronic database (Keats 2009: 181). In this thesis *text* refers to plain (raw) transcriptions of spoken lecture data and the language items (clauses and utterances) of which they are composed.

Databases in which texts are stored, or *corpora*, are distinguished from other collections of information (such as archives) by function: they represent a language variety or genre, providing a standard point of reference (Baker, Hardie and McEnery 2006: 48), and are normally carefully sampled from some population data. In addition, the ELC adheres to the definition of a corpus as a collection of language data that has been processed to make it accessible for research purposes (Wallis 2014: 1) and as "a large set of texts for studying language as it is used in real life" (Kilgarriff n.d).

Originally associated with general linguistics, *metalanguage* is a broad term for what is most commonly understood to mean language about language (Berry 2005: 3, Culpeper 2012: 66); it is part of the *text*. Drawing on the same epistemological prefix, the term *metadata* is commonly applied to data about data (Petrillo and Baycroft 2010: 2, Wittenburg, Broeder and Sloman 2000: 2); these are the machine-readable language/symbols used to describe text, either technically or conceptually. *Metadata* encompasses header fields and can be regarded as descriptive of the specific attributes of a resource (Garside and Rayson 1997: 3, Kilgarriff n.d, Taylor 2003). The term describes two separate concepts: the structural (data about how data structures are designed and contained) and descriptive (data about individual instances of content, usually language structure). Petrillo and Baycroft (2010) explain that metadata, unlike *annotation*, is not anchored to a specific point in a text. Although this distinction is technically correct, few projects distinguish metadata from annotation data in this way (Broeder and Wittenburg 2001: 79, ft.1). Although it can be argued that metadata is one kind of annotation, by convention (and in this thesis), the term is used as an umbrella to refer to all description external to the text.

Metadata less commonly refers only to the *structural markup* of data (Meyer 2002: 81). In this thesis, structural markup (or *markup*) describes the procedure for and result of formatting, processing and classifying data structures *(text* and *metadata).* It can be regarded as a subtype of annotation that shows the actual linguistic structure of the text (for example, at the level of utterances). In the ELC, however, resultant markup is

distinguished from the processes of tagging or annotation (and the *tagset* utilised), which refer to the identification of linguistic items specifically in the body text. Both markup and annotation are part of the metadata.

The terms annotation and tagging are frequently used synonymously, seemingly with three slightly different implications:

1. in its widest usage, data *tagging* is used as shorthand for the general process of all *annotation* (Baker, Hardie and McEnery 2006: 154, Meyer 2002: 81)
2. more specifically, *tagging* refers to the identification of parts of speech within a text (Meyer 2002: 86). The Sketch Engine website, for example, defines a tagset summary as "the list of part of speech (POS) classes used for annotation (tagging) of the corpus" (Kilgarriff n.d). A *tag* here refers to the word category label that is assigned to each word (e.g. Atwell et al. 2000: 8, Hunston 2002: 18)
3. an intermediary understanding aligns with the use of corpus *annotation* in reference to the linguistic information encoded (Dickinson and Lee 2009, Leech 2005)

In the sense of the third implication, a synonymous usage of tagset refers to all the elements and attributes used in any annotation scheme. This is the definition employed in this thesis. Tagging, then, refers to the process of adding linguistic interpretation to the text (resulting in *tags*). It is used interchangeably with annotation, which can also be a process nominal or a result nominal.

The term *string* is used in this thesis to denote an uninterrupted sequence of characters that form the tokens (individual linguistic units, typically words) that comprise the raw ELC texts. This definition corresponds to the entity referred to as a *text node* in the Document Object Model (DOM) for eXtensible Markup Language (XML). Annotated *strings* are the continuous stretches of text that are enclosed (or, indexed) by annotation categories – the opening and closing tags that identify a distinct instance of a particular pragmatic function within the text. These categories are also referred to as *elements*, and any assigned subcategories/category *types* are described as *attributes* (examples are given in 3.4.8, Table 3.5).

3.2. The overarching approach: quantitative and qualitative data analysis

This section reviews methodological approaches that inform the discussion of how the ELC data were collected and prepared for the purposes of answering the research questions (2.5).

A dataset containing multiple texts is required to address questions of comparability and to make generalised conclusions. To avoid information overload, large or complex data are most efficiently stored and processed electronically. As Norman (1993: 43) notes, "[t]he power of the unaided mind is highly overrated". Computational analysis can identify patterns that may not be detected by the naked eye (Flowerdew 2013: 161). The identification and analysis of the characteristics of engineering lectures (in general and across educational contexts) in this thesis is thus based on an electronic corpus of transcribed lectures (as described in 3.4.1).

The metadata that guides the implementation of computational techniques used to create the ELC is based on a series of methodological choices that, somewhat paradoxically, require subjective decisions in the pursuit of objectivity. Fully top-down approaches to forming research hypotheses have long been rejected as too subjective. Mouly sums this up in his alternative and more objective inductive-deductive approach:

> […] a back-and-forth movement in which the investigator first operates inductively from observations to hypotheses, and then deductively from these hypotheses to their implications, in order to check their validity from the standpoint of compatibility with accepted knowledge. After revision, where necessary, these hypotheses are submitted to further test through the collection of data specifically designed to test their validity at the empirical level. (Mouly 1978, cited in Cohen, Manion and Morrison 2000: 4-5, and Nunan 2013: 49-50)

Echoing Mouly's approach specifically in the field of corpus linguistics, Wallis (2014: 2) argues that traditional top-down corpus-based and bottom-up corpus-driven approaches to the analysis of speech corpora are "one-sided" and so "usefully subsumed into an exploratory cyclic approach to research". Wallis (2014: 4) proposes the *3A perspective*:

annotation, abstraction, and analysis – a recursive approach in which knowledge addition and critical reflection occur at each level (see Figure 3.1).

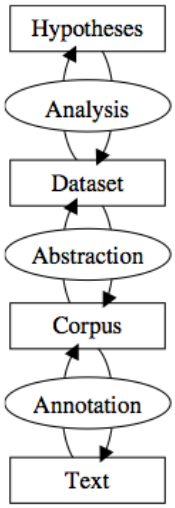


Figure 3.1: The 3A perspective in corpus linguistics (Wallis 2014: 4)

Multiple iterations of – and constant adjustments to – the ELC annotation system have been undertaken, as detailed in the workflow model in Figure 3.2 and list of adjustments in Table 3.3.

The dangers of studying metalanguage over language is a concern, as is the potential for using predefined taxonomies to confirm patterns that were either known to linguists prior to computerised assistance or that meet pre-determined criteria. As Culpeper et al. (2008: 632-633) suggest, when analysing language:

> […] dividing the whole is very often the best way of making sense of it, and the debate amongst corpus linguists tends to surround not whether this should be done, but how it should best be achieved.

The approach taken in this thesis is the type of continuously revised cyclical procedure outlined by Wallis (2014). As metadata applied to the ELC is towards the non-automated

end of the annotation-type spectrum, information processing stages feed heavily into the determination of annotation categories.

3.3. Data collection

3.3.1. Data collection in Phase 1 (2007-2009) and current holdings (2014)

The starting point for this thesis was a set of 76 videos and partial transcriptions of English-medium lectures from the UK, New Zealand, and Malaysia. Data collection was funded by the British Council (PMI 2 Connect Research Cooperation, British Council (RC 90)) and a research grant from Auckland University of Technology. Collection took place between 2007-2009 (see Phase 1 in the workflow model, Figure 3.2), before this thesis began. The ELC project is led by Professor Hilary Nesi (Coventry University). Dr. Lynn Grant (Auckland University of Technology) and Dr. Ummul Khair Ahmad (Universiti Teknologi Malaysia) collaborated in Phase 1 data collection.

Most of the lectures collected are in civil, mechanical and electrical engineering, and similar topics are often covered in the different cultural/educational contexts. The lectures were recorded using video and audio equipment and vary in duration between 41-104 minutes. The lectures were delivered across undergraduate degree programmes (from years 1-3), largely as part of courses that are mandatory to the programmes. A range of lecturers were filmed from each institution. A breakdown of the composition of the (current) 2014 version of the ELC that informs this thesis is given in Table 3.1.

		Coventry University (United Kingdom)	**Universiti Teknologi** Malaysia (Malaysia)	**Auckland University of Technology** (New Zealand)
abbreviation		UK	MS	NZ
identifier series		1000-1030	2000-2018	3000-3028
token size		156838	120211	251108
engineering type	civil	27	6	0
	electrical	0	0	17
	graphics	0	0	3
	mechanical	3	12	2
	fluid mechanics	0	0	3
	solid mechanics	0	0	3
total lectures		30	18	28
total lecturers		4	9	4
average lecture length (tokens)		5228	6678	8968

Table 3.1: Summary of ELC holdings (2014)

A single video camera was trained on the lecturer from the back of the lecture theatre to preserve student anonymity. The video files were then roughly transcribed as plain text, including instances of code-switching in the Malaysian lectures. A summary of transcription protocols is given in Appendix I.

Like many spoken and even written corpora, Table 3.1 shows that the ELC is not balanced across all of its variables, particularly engineering type. As discussed in 1.1, engineering is not one discipline. The multi-disciplinary nature of engineering has a direct impact on studies that take a corpus linguistic approach to understanding the nature of its discourse. For example, in the Michigan Corpus of Upper-Level Student Papers (MICUSP), out of 829 papers, 105 relate to engineering disciplines: 31 civil and environmental, 42 industrial and operations, and 32 mechanical (Römer and Brook O'Donnell 2011: 164). The British Academic Written English (BAWE) corpus also regards engineering as belonging to the physical sciences, one of four domains (along with arts and humanities, social sciences and life sciences), each of which contain approximately 32 assignments at each level from 6 or 7 subject areas. Due to its size and diversity, engineering is double-weighted; the final corpus contains 238 engineering assignments

split across four years of study (Alsop and Nesi 2009: 74, 80). Even at the disciplinary level, the range of engineering branches makes developing a balanced corpus difficult.

Disciplinary balance is also an issue In corpora of spoken academic discourse. In MICASE, for example, there are 13 speech events (including a dissertation defence, lectures and office hours) representing eight types of engineering (Simpson-Vlach and Leicher 2006: 26-28). Of the 160 lectures and 40 seminars that comprise the BASE corpus, six lectures and three seminars are from engineering, across diverse areas such as renewable energy, tension structures, and writing issues for engineers (Coventry University 2016). Despite issues of composition, it does seem that engineering is treated as an homogenous discipline in some respects from the perspective of corpus analysis. It follows that there is value in identifying language patterns across the various branches.

As I was not involved in Phase 1 (the initial data and metadata collection stage), the metadata for the corpus described in this thesis is not complete, although it does include information about the lecturers, their topics, and some of the context in which the lectures take place.

3.3.2. Data collection: ethics

All the base ELC data (video files) were collected prior to the start of this thesis, and ethical clearance for the project was gained from Coventry University and Auckland University of Technology. All lecturers who provided data signed an informed consent form, the wording of which is given in Appendix II. Potential future uses of the corpus were fully explained to the lecturers and audiences. Additional ethical *low risk* clearance was granted for this thesis by Coventry University (reference 1179. Appendix III).

3.4. Data preparation

3.4.1. Workflow

In Phases 2-4 of the project (2010-2014, see Figure 3.2), I made a searchable corpus from the initial holdings for the purposes of responding to the research questions in this thesis (2.5). Data preparation involved transcription completion/verification, the addition of consistent metadata and structural markup, and the annotation of certain functional categories along with IAR testing.

.

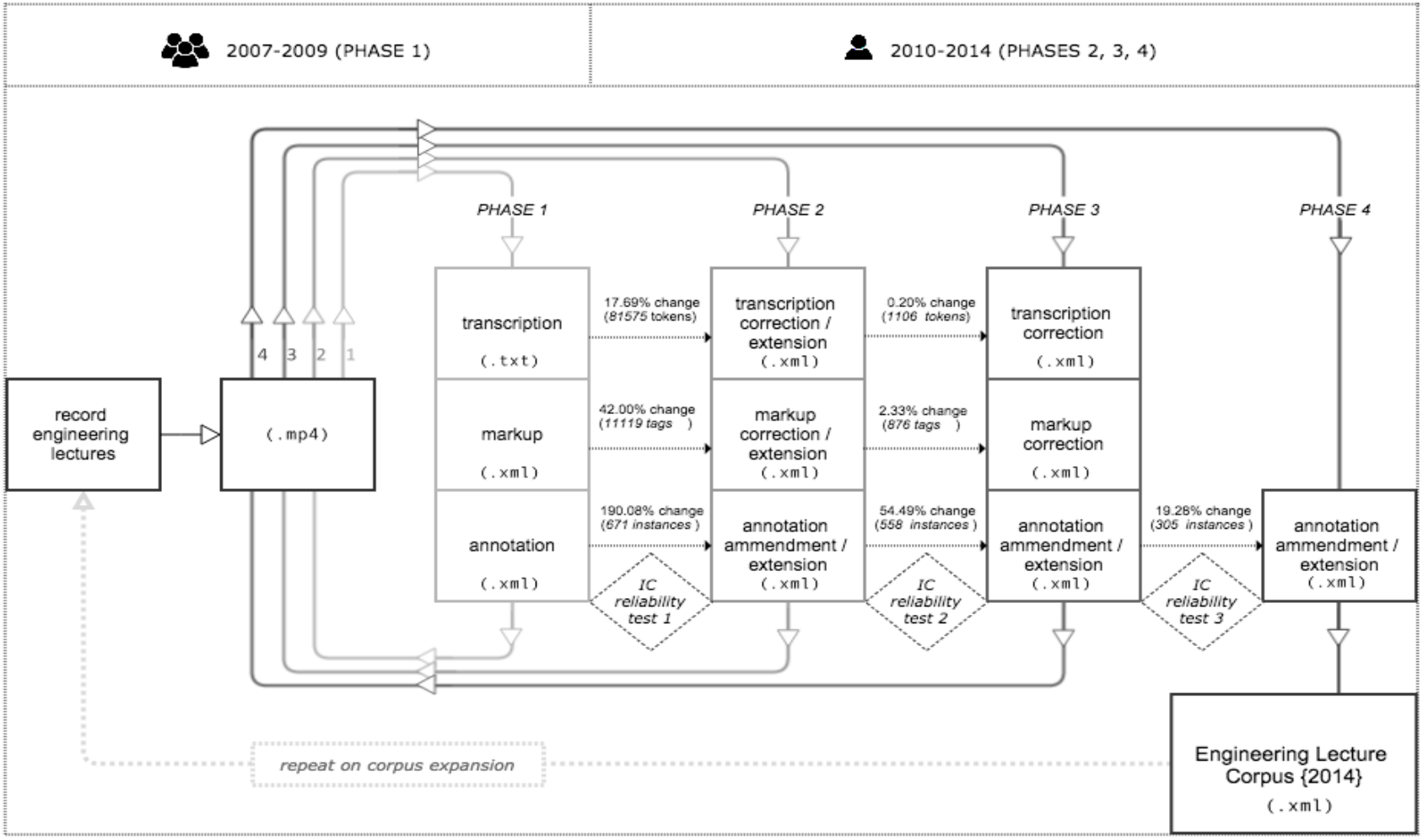


Figure 3.2: ELC workflow model

Processes in Phase 1 (the original project work) and Phases 2-4 (work carried out for this thesis) are visually rendered in the workflow model in Figure 3.2, along with the percentage changes to transcription, markup and annotation following each Phase and accompanying iteration of IAR. The percentage changes were calculated using Python scripts I wrote to compare the difference in the number of tokens, structural tags, and instances of annotation at each juncture between Phases 1-4

3.4.2. Transcription

Each lecture was assigned a unique identifier (*idno*) based on subcorpus membership; the 30 lectures from the UK are numbered 1001, 1002, …, 1030, the 18 lectures from Malaysia 2001-2018, and the 28 lectures from New Zealand 3001-3028. The raw text transcriptions constitute the body of the ELC files. The corrections and additions I made in Phase 2 to Phase 1 draft transcriptions resulted in an 18% change to the raw text (see Figure 3.2), not including corrections to existing tokens.

3.4.3. Structural markup and encoding standards

The workflow model (Figure 3.2) also shows that I made a 42% addition to the markup (counted in number of new tags) between Phases 1-2, and a further 2% between Phases 2-3. This addition improved the description of the structural data and enabled full searchability. All existing markup was also standardised.

The need for encoding standards for corpora has received significant attention from the NLP community (Widlöcher and Mathet 2012), especially regarding long-term, clear systems of annotation that are reusable across multiple corpora (Leech 2005, Petrillo and Baycroft 2010: 2). Many researchers tailor systems or create their own (Smith, Hoffmann and Rayson 2008: 164). Elliott and Elliott (2003: 201-202) liken the progression of corpus linguistics to the history of the railways where various designs were developed based on the priorities of individual enterprises:

> Finally, when the individual developments […] had to be integrated into a single network for the system to truly serve the users' needs, it was suddenly apparent that many lines were incompatible as many had chosen different gauge tracks.

The ELC is intended to be a publically available resource and so decisions regarding structural metadata were guided by principles of reuse, merging and comparison across systems.

The purpose of adding markup and annotation is to aid recovery of original context and to allow extra-linguistic features to be encoded. A variety of encoding languages are available, including HTML (Hyper-Text Markup Language), SGML (Standard Generalised Markup Language), and RDF (Resource Description Framework) (CLARIN 2009, Taylor 2003). In corpus-based work in NLP, the current standard language used is XML. The main standardisation bodies that develop guidelines for corpus creation and curation are the International Standards Organization TC C37/SC 4 language resource management (ISO 2011) and the W3C Consortium (W3C 2011).

There have been some initiatives that have tried to provide annotation standards for encoding various predefined levels of pragmatic content. As in DART (Weisser 2015), utterance-level speech acts are the focus. The Discourse Resource Initiative (DRI) developed the Dialogue Act Markup in Several Layers (DAMSL), which identifies dimensions (or levels of pragmatic content) using utterance tags (Allen and Core 1997: 4). The Multilevel Annotation, Tools Engineering (MATE) (Klein 1999) annotation guidelines also contain recommendations for representing descriptive annotation of transcribed spoken dialogue.

Various general standards for XML encoding have also been developed for a range of user communities. Generic metadata categories for digital language resources include the Dublin Core Metadata Initiative (DCMI 2010) and the Open Language Archive Community (OLAC) (Wynne 2004). Standards to describe multi-media and multi-modal language resources include the IMDI (ISLE (International Standard for Language Engineering) Metadata Initiative) (Broeder and Wittenburg 2001). A strong option for work in NLP is XCES, the XML version of the Corpus Encoding Standard (CES) (XCES 2008), developed by the Expert Advisory Group on Language Engineering Standards (EAGLES) to support simple to highly annotated corpora. Spin-off standards exist, such as TUSNELDA (Kallmeyer, Meyer and Wagner 2009). TUSNELDA also incorporates conventions from the other major

contender in the digital humanities: the Text Encoding Initiative (TEI) (TEI Consortium 2011).

The TEI delivers guidelines and supporting resources for encoding digital texts. It offers both generic and text-type specific *modules*, which draw on a pool of around 500 descriptive *elements*. Documents that adhere to this standard of markup are represented by XML notation, specifically TEI XML. The complexity of the guidelines and human-reader unfriendly nature of the markup fuels some arguments against using the TEI (for example, Grönqvist 2003: 12, Meyer 2002: 98). The difficulties inherent to recording spoken data in written form (such as consistency of transcription and quality of recorded material) amplify the tension between compliance with the standard and customisation to meet user needs. Although the TEI provides tailored modules for various language events, difficulties in use are inevitable. For example, when marking up the BASE corpus, Creer and Thompson (2005: 163-164) report ambiguity surrounding the correct way to encode a speech event and note the added linearity imposed on speech events by the TEI model (for example, that there is no way of representing overlapping speakers within utterance tags).

TEI standards have not traditionally been used to signal the function of larger stretches of discourse, such as pragmatic annotation, and appropriate coding strategies are still under development. However, TEI P5 XML standards have proved to offer a stable document encoding format and are flexible and comprehensive enough to allow tailored schema. Capacity for machine readability is also high, as is the level of customised resources and support offered to *TEI-ers* (christened in the dedicated Wiki (TEIWiki 2010)). As a result, TEI is the preferred standard for encoding electronic texts in the humanities and social sciences (Meyer 2002: 84).

Following the BNC (2007) and the BASE corpus (Nesi and Thompson 2006), the ELC adheres to TEI-compliant structural markup standards. Encoding in the ELC files is in XML, which was written using the XML editor <oXygen/> 15.2 (SyncRO Soft SRL 2014). Structural elements include container elements, empty elements, and context-specific event descriptions. The metadata requirements enable the ELC to be fully computationally searchable.

The general markup design of the ELC follows choices made within the BASE corpus (cf. Nesi and Thompson 2006). Container elements include speaker utterance tags that identify sex and academic status. For example: <u who="sm"> is a male student, <u who="sf"> is a female student, and <u who="ss"> is a group of students. The lecturer is identified by a two letter code for sex (nf/nm = female non-student/male non-student) followed by a four digit unique identifier, such as <u who="nm1003">, which signifies a male lecturer from the UK component who is identified as 1003. The close of each utterance is marked by the end tag </u>.

Six empty elements are also commonly used: <gap reason="inaudible"/>, <gap reason="pause"/>, <vocal desc="laughter"/>, <vocal desc="voice from video"/>, <event desc="writes on board"/>, and <event desc="draws on board"/>. Other context-specific descriptions record unusual occurrences, for example <event desc="drops pen"> has been inserted to make sense of the proceeding “oops” (3028). All pauses of perceivable length are recorded as <gap reason="pause">. Significant gaps for breaks in recordings and inaudible speech are identified in the same way with the addition of length in time data, for example <gap reason="break in recording" dur="00:01:12"/>.

In addition to the structural markup of body text in the ELC, header metadata was added to each file: a file description (including title and citation information along with a source description of recording and transcription information), a description of encoding, and a profile description of non-bibliographic information (such as the number of participants, the meaning of unique identifiers, level and module). The full header tagset is given in the example from lecture 1001 in Appendix IV. The requirements of the element and attribute structures are declared in the associated document type definition (DTD), which states the constraints to which the metadata must adhere to be valid. The ELC DTD was adapted from the BASE DTD (cf. Nesi and Thompson 2006) and is TEI-compliant.

The header metadata was compiled in a tab delimited spreadsheet that was outputted using a script written in Python to XML format to create skeletal files, which include empty body tags for later population (as indicated by the comment <!-- marked up and annotated transcript here--> in the sample in Appendix IV). This method ensures the consistency of

layout and information across all files and allows changes to be easily overwritten without editing individual files. Each header file, like the marked up body transcripts, was labelled according to the lecture identification number. Two sets of adjacent directories (headers and bodies) were then merged using another Python script to create a searchable corpus of XML files. A separate directory of raw (that is, not marked up or annotated) files was maintained.

For the purposes of readability in this thesis, metadata is only included in reproduced ELC text where informative. Further, utterance markup has been simplified, for example the markup enclosing an utterance by lecturer 2001 has been changed from <u who="nm2001">[…]</u> to <lecturer>[…]</lecturer>. Where only plain text is reproduced, it can be assumed that all metadata has been stripped. In some cases, markup that does not form part of the ELC taxonomy has been added to reproduced text for communication purposes. Units within stories in Chapter 7 are identified, for example, by markup such as <orientation>[…]</orientation>. Such non-taxonomical encoding is clearly identified.

3.4.4. Pragmatic annotation

The starting point for the pragmatic annotation of the ELC transcriptions was a list of 15 pragmatic categories (outlined in Table 3.3) compiled by Nesi and Ahmed (2009) based on a sample of ELC recordings. The list does not attempt to cover all pragmatic possibilities, but was composed in accordance with four selection criteria. According to Nesi and Ahmed (2009), the categories must:

1. not be realised by a single predictable form
2. shed light on the specific nature of lecture discourse
3. identify features which were not easily recoverable from context
4. occur more than once in the corpus

These rules continue to underpin the current 2014 working list of ELC categories, the broad definitions of which are given in Table 3.2. Full details of the evolution of both elements and attributes are given in Table 3.3.

explaining	where lecturers define, demonstrate or translate concepts or terms
housekeeping	where lecturers talk about academic commitments and events external to the lecture
humour	where lecturers use irony, mock-threats, teasing, sarcasm, self-denigration, wordplay, or bawdy, black or playful language for comic effect
prayer	self-explanatory (only occurs in the Malaysian component of the corpus)
story	where lecturers discuss personal or work-related matters in the form of anecdotes, exempla, narratives or recounts
summary	where lecturers preview the content of current and future lectures, or review the content of current and past lectures

Table 3.2: Working list of pragmatic categories in the ELC (2014)

3.4.5. Situating the ELC annotation categories

The first, and (as far as I know) only other example of annotating a corpus for such pragmatic features is the MICASE system of pragmatic annotation (outlined in 2.4.4), which includes speech events from multiple disciplines and genres, ranging from advisory sessions to lectures. The focus on a single discipline in the ELC means that a more focused and accurately weighted taxonomical hierarchy was identified, which better responds to the research questions of this thesis (2.5).

A comparison of the ELC and MICASE taxonomies including definitions and any assigned attributes is given in Appendix V. The ELC tagset is compared with the original full MICASE pragmatic inventory (Simpson-Vlach and Leicher 2006: 68-69), which contains a long list of possible pragmatic categories and their definitions. The final refined tagset (Maynard and Leicher 2007: 112-114) that was used to note pragmatic features in the header information of a selected subcorpus of MICASE is also given, along with further definitions that were not present in the original full inventory. The comparison of the ELC tagset to two iterations of the MICASE tagset reveals overlaps and differences in both the categories deemed most relevant to the speech events considered and in the hierarchical weighting of these categories and the subcategories assigned to them.

The MICASE element *narrative* is roughly synonymous with the ELC element *story*, which functions as an umbrella category for four genre types. The *narrative* attribute within ELC can, in this case, be understood as a specific genre type (cf. Martin 2008), and is therefore hierarchically inferior to both the ELC story element and the MICASE narrative element, which is defined as "a story of two or more sequential clauses using the past tense or the

historical present" (Simpson-Vlach and Leicher 2006: 68-69). The MICASE element *introductory roadmap* partly overlaps with the *preview content of current lecture* attribute of the ELC element *summary*. Shared ground may also exist, to a lesser extent, between MICASE's *speaker introductions* and the ELC attribute *preview content of current lecture*.

In terms of overlaps between the ELC and MICASE, only *defining* – which broadly refers to examples of the definition or glossing of terms in both corpora – can be found in all descriptions. Defining is an attribute of *explaining* in the ELC, which was considered to be outside the remit of this thesis for reasons of space, because it is extremely widespread (which is supported by its presence in the MICASE tagset). Explaining can be considered to be one of the main purposes of lectures alongside the presentation of new information. MICASE's final tagset element *assigning homework* and long inventory elements *assigning homework*, *logistics / announcements*, *returning or going over homework or an exam*, and *reviewing for an exam* can also be located within the ELC's broader description of *housekeeping*. The non-inclusion of housekeeping in this thesis is also primarily for reasons of space.

There are clear differences between the two final tagsets, not least in the number of categories identified: the ELC identifies five elements and 17 attributes, MICASE's final tagset is comprised of 11 elements, and the longer MICASE inventory identifies twice as many elements again. *Humo[u]r* is present in the original MICASE inventory and the ELC tagset, but not in the MICASE final tagset because, as discussed in 2.4.4, it was considered that there were too many occurrences to record (Maynard and Leicher 2007: 112). The ELC tagset includes humour, to which nine attributes are assigned.

Given the variety of genres and disciplines in MICASE, it is unsurprising that some of the categories identified are not found in the ELC lectures. The ELC taxonomy aims to identify the functions specific to lectures in a single discipline and the linguistic features of which they are composed.

3.4.6. Refining the annotation elements and attributes

As noted, two ELC elements, explaining and housekeeping, are not discussed at length in this thesis but remain part of the overall project tagset. Analysis focuses on elements which are not commonly regarded as central to the purpose of the lecture, but which nevertheless appear to be important because they occur frequently across the subcorpora, namely: *summary*, *humour* and *story*.

The current 2014 ELC working list emerged gradually during the process of annotation (as illustrated in the workflow model in Figure 3.2). Initially, the process involved identifying features in a selection of files, checking the resulting long list of features against the four selection criteria (3.4.4), and collapsing the list to remove instances of inefficient and overlapping description. Throughout, the transcripts were annotated whilst watching the corresponding videos. Facial expressions and phonological features could be accessed in the video component, and sometimes helped in the construal of pragmatic meaning. Where it was felt that a feature was important and interesting but not sufficiently frequent to warrant a distinct category (or element), sub-categories (or attributes) were created. The four-stage process of tagset adjustment is outlined in Table 3.3.

A liberal approach was taken to the annotation. As far as possible, text was annotated according to the following three principles:

1. sufficient contextual data should be captured so that the annotated text makes sense as a standalone string
2. introductory and evaluative sections that enclose the core text should be included
3. when in doubt, more rather than less of the transcript should be included within the annotation

In terms of the first principle, category boundaries had to take meaning into account in order to facilitate later micro-level analysis when examples of the pragmatic annotation were extracted from the corpus.

The first pass at annotation is considered to be the notes on possible functions made by local experts from the UK, Malaysia and New Zealand involved in the initial project funded by the British Council and Auckland University of Technology (Phase 1 in Figure 3.2,

resulting in Adjustment 1 in Table 3.3). Local experts worked through samples from each subcorpora looking at the original working list and some of the functions that actually occur in the corpus. This corpus-driven approach resulted in the first adjustment to the working list.

At this stage it became clear that some of the elements identified in the original working list needed to be expanded to include attributes, and some should be hierarchically demoted and subsumed under a more general umbrella element. As a result, *review lecture content* and *preview lecture content* were made attributes of the umbrella element summary, *personal narrative* was made an attribute of story, and the humour element was expanded to include five more attributes and *wordplay* (which was formerly an element). Two other elements from the original working list (*reference to students' future profession* and *greetings*) and part of one element (*register*, from *register and wordplay*) were not evident in sufficient quantity to justify their inclusion in the adjusted list when considered in terms of the original criteria of identifying and describing typical engineering lecture discourse features (2.5).

I made a second pass at refining the working list with the aim of ensuring consistency across all identified features (see Phase 2 in Figure 3.2, resulting in Adjustment 2 in Table 3.3). This was the first time that all lectures had been overviewed. The summary element was augmented to distinguish two types of *review* (of previous lecture content and of current lecture content) and two types of *preview* (of current lecture content and of future lecture content). The story element was also expanded to include *narratives of professional experience* (where lecturers tell second-hand narratives) in addition to *narratives of personal experience* (where lecturers tell narratives about situations in which they were involved). The manual annotation of elements and attributes across the corpus occurred during the second pass. This was the most time-consuming aspect of the workflow process, involving the draft annotation of more than 80 hours of recorded speech.

At the third pass (Phase 3 in Figure 3.2, resulting in Adjustment 3 in Table 3.3), I again reassessed and expanded the story element. The distinction of four genres - *anecdote*,

exemplum, *narrative* and *recount* (Martin 2009) - was considered to be more useful than the former limited description of narrative type (further details are given in 7.2). In the humour category four attributes were condensed into two (*irony/sarcasm* and *teasing/mock-threat*) because the boundaries between types was judged to be too blurred for distinction. The fourth pass at annotation largely involved refining boundaries. At each pass, I returned to the original videos to ensure maximum accuracy, as indicated by the spiralling loop that traverses the MP4 rectangle that connects the Phases in Figure 3.2.

Adjustments were made between each pass to reflect the findings of inter-annotator reliability (IAR) testing as indicated numerically between Phases in the workflow model (Figure 3.2). Explanation of the procedure is given in the sections on IAR testing (3.5.2-3.5.4). The four adjustments to the working list are outlined in Table 3.3.

(Nesi and Ahmad 2009)	**Adjustment 1** (2010)		**Adjustment 2** (2012)		**Adjustment 3** (2013)		**Adjustment 4** (*current 2014 tagset*)	
	element	*attribute*	*element*	*attribute*	*element*	*attribute*	*element*	*attribute*
prayer	prayer		prayer		prayer		prayer	
housekeeping	housekeeping		housekeeping		housekeeping		housekeeping	
defining term	defining		defining		explaining	defining reasoning translating	explaining	defining reasoning translating equating
review lecture content preview lecture content	summary	review lecture content preview lecture content	summary	review lecture content preview lecture content	summary	review previous lecture content review current lecture content preview current lecture content preview future lecture content	summary	review previous lecture content review current lecture content preview current lecture content preview future lecture content
personal narratives	story	personal narrative	story	personal narratives professional narratives	story	anecdote exemplum narrative recount	story	anecdote exemplum narrative recount
teasing self-recovery self-denigration black humour disparagement of out-group member mock threat register and wordplay	humour	bawdy black disparagement irony jokes mock threat playful sarcasm self-denigration teasing wordplay	humour	bawdy black disparagement irony jokes mock threat playful sarcasm self-denigration teasing wordplay	humour	bawdy black disparagement irony/sarcasm jokes mock threat/playful self-denigration teasing wordplay	humour	bawdy black disparagement irony/sarcasm jokes mock threat/playful self-denigration teasing wordplay
greetings								
reference to students' future profession								

Table 3.3: Four adjustments to the ELC elements and attributes (2009-2014)

3.4.7. Refining the annotation boundaries

Umbrella categories in the ELC describe discourse function, and the attributed types refer to content. For example, text annotated as “humour type=bawdy” will contain content that is bawdy. Where more than one category fits a particular string of text, the dominant function is annotated. In the case of summaries, for example, “right so let's go and have a look at that cylinder that we did last week” (3003) is identified as a preview of current content because the review aspect performs a supporting function in the prediction of upcoming content. Further examples are given in Table 3.4.

example	dominant function	secondary function
what I want you to do in lecture twenty one is um impress upon you the fact that that work we did on charging a capacitor with voltage and charging the coil with current is just two examples of a very generally theory that applies to just about everything (3015) before we move on to the actual tests we need to understand something that we've already looked at with stress (3025)	preview content of current lecture	review content of previous lecture
but I just want to er review the aspect of safety and health at the workplace (2011) so back to our definition of efficiency (2017)	preview content of current lecture	review content of current lecture
right this week and next week then we're going to look at beams (1006) today we are going into the cycle yeah the refrigeration cycle and we are only going to take a look at the vapour compression there are many types of refrigeration cycle we will cover them in term of two but the one that we gonna look at in our class refers to the vapour compression (2017) then next I want to go on to talk about combinations of capacitors today rather than w- we'll do some more about the derivation of C tomorrow for different shapes but for the rest of today I want to talk about combinations of capacitors (3004)	preview content of current lecture	preview content of future lecture

Table 3.4: Examples of previews of current content that also function to review previous, review current, or preview future content

In some cases, however, doubt about categorisation remained. For example, one lecturer advises that “now rating is a word I'm going to be talking about a lot” (3006). Following content within that lecture confirms that the lecturer does indeed talk about the concept of rating, so the example is a preview of current content. What is not - and cannot - be known, however, is whether it is also a preview of future content, because the corpus does not

contain all the lectures delivered on the course by that lecturer. Similarly, one lecturer states that "I said it many times now that water is one of the most important er element yah in the piece of concrete yah […]" (2003). Again, the reiteration of the importance of water is recoverable from the given lecture, but whether this topic also occurs in previous lectures (in this module or on the course) is unknown. In such cases, the confirmed attribution was given, which is a preview of current content in the example from 3006 and review of current content in the example from 2003.

Hesitation markers and fillers, which frequently co-occur, are included within the boundaries of pragmatic categories. Common examples include *so, ok, now, right, er* and *um,* as in "*now um* in the second semester in about three months time we will be going over electric motors" (emphasis added. 3011). As Fraser (2009) recognises, these markers commonly precede – and have a semantic relationship with – topic orientation markers, as in the preview "*right so um* what I'm going to do is first of all define the temperature coefficient" (emphasis added, 3006). All markers and fillers are indexed if they occur at the beginning of annotated text.

Markers of disfluency (such as false starts) (cf. Chafe 1985) are common and have been included within the boundaries of annotation. For example, "wh- what I've done over here is I've shown you […]" (3006) and "I'll ki- I'll die if I eat them this many Mars bars" (1005).

Markers of topic shift are also included within the boundaries of pragmatic annotation, as in the following example where *but* divides two stories:

> <story type="narrative"> […] they tried to lift too much the crane topped over slightly embarrassing that would happen</story><story type="narrative"> *but* it's not as embarrassing as the one I saw on YouTube […]</story> (emphasis added. 1001)

This example somewhat ties in with Norrick's (2001: 49) finding that *but* has the special discourse function at the conclusion of narrative action in oral storytelling. However, in the ELC system markers at the end of pragmatic strings are not indexed because they largely mark the start of a new type of pragmatic string, or a return to non-pragmatic lecture content.

Most commonly, for example, the use of *but* follows a section of pragmatic language and indicates a shift *back* to new content, as in the following example:

> <summary type="review content of current lecture">[…] that in summary is the method of joints cutting through each joint in turn applying the rule two equations of equilibrium and finding the values of the unknown forces</summary> *but* the general principle is the whole structure's in equilibrium because we know it's got the reaction forces (emphasis added, 1004)

The marker in this case lies outside the pragmatic string and is not indexed.

3.4.8. Examples of ELC pragmatic categories

Some of the ELC categories are fairly self-explanatory, such as summary, or most usefully clarified by the subcategories attributed to them, such as humour or story. Given the inevitably subjective nature of the annotation process, and the ongoing bottom-up adjustments to the taxonomy as the corpus expands, the categories are described fairly loosely without specifying linguistic features. In Table 3.5, the current (2014) ELC tagset is given alongside typical examples of each element and attribute discussed in this thesis.

element	attribute	description	example of discourse
humour	bawdy	a lewd or vulgar reference (direct or implied) usually related to sex	what I'm going to ask you to do is trust me so I want you to all strip naked and fall back like this no not that amount of trust is needed (1006)
	black	satirical treatment of taboo or dark topics	ok this this slide show how children go to school in India yah and then again from here we can see that er this is quite a hazardous way to cross the river if say er if this is in Malaysia then Sarawak for instance then there'll be some crocs waiting underneath here so if the any one of the children fall then the croc will surely have a very heavy meal (2010)
	disparaging	an utterance that belittles something or someone - often an out of group member	I was waiting for someone to ask if I would kindly derive these two equations and um my answer to that was going to be er A we can't spare the time B you wouldn't understand it so there really isn't any point (1029)
	irony/ sarcasm	an utterance that means the opposite of what is said	there's delta and there's rectangular and obviously those words are quite similar so it's easy to er be confused when you look at the words (1029)
	joke	a time out short story with a set-up and a punchline	he said ah this reminds me of th- the time that ah the students went to the lecture and um there was no lecturer there was just a tape recorder and it said when you've all assembled just switch this on and take notes and the lecturer came in for the next lecture the following week and what did he find there were twenty tape recorders scattered about the room with a little note saying when you come switch this one and just start talking (1030)
	playful	untargeted uncritical humour intended solely to amuse	you can't go wrong if you follow the method we'll say fun-sized Mars bar if you hav- if you write something down but get the wrong answer family pack Mars bar double whammy if you don't even try it and because no one complained this is a formal wager (1005)
	self-deprecating	negative reference to the self for comic effect – commonly linked to physical attributes and cognitive abilities	if I would have to machine this I would pull my left hair out my few I have left (3019)
	teasing/mock-threat	gentle mockery of an audience member or the audience / exaggerated threats made in jest to underline a requirement or concept	I will open it up again for another two weeks except for the person whose phone's going off cause they're not gonna be able to sit down for about a month (1004)
	wordplay	a display of wit for amusement where meaning centres of word choice	so remember I talked about relationships concentric coincident yeah we've got the same here um now there's an opportunity to laugh again it's called mate yeah no laughers today ok so we are mating now parts together (3020)
story	anecdote	where events are problematised and remain unresolved, accompanied by an emotional reaction	if you put a a lump of concrete in a microwave oven just take a uh a little well it works best wi- with actually grout y- you don't do it with the aggregate um good strong mix um fully saturated put it in a microwave put it on full heat and you'll probably break the plate because it explodes um certainly I did once much to the disgust of the owner of the microwave at Leeds University (1014)

	exemplum	where events are problematised and remain unresolved, accompanied by scientific or moral judgment	so from the video you can see that the the girl was hit by the forklift because because of very very simple reason she did not hear anything because of her I-tunes er normally when you use I-tune you listen the music very very loud so it will cut you off anything from outside so even though the forklift driver he er use the horn or whatever so the the the girl in this video yeah even though it's acting she did not hear anything and hence she was hit by the forklift this type of accident actually occur sometimes (2010)
	narrative	where events are problematised and resolved	I hate to admit to this one but one site I was on we had cube failures and the reason was that when I'd been sending the cubes off I'd been having to break the ice on the top of the tank before I could get them out and um the tank had a heater in we just hadn't bothered to get the spark to wire it in and ah fairly obviously by the time the area manager appeared to ah come and have a look and see what had gone wrong it was all wired in and working fine and we said oh no no problem with that would we do a thing like that and ah but ok sort of nevertheless it caused endless hassle the fact that we'd had these cube failures if you keep them too cold they'll go down a low strength (1012)
	recount	an unproblematised retelling of events	yeah rim on a steel wheel you know th- the good old horse carts yeah that's how they put the rims on there they heated up the rims and hammer them on and then le- just let them cool down and you could never get them off the wooden yeah wheels yeah that's how that was done just basically shrunk on there (3019)
summary	review previous lecture content	a backward reference to information already given on the module or course prior to the current lecture	let's just review back what we did yesterday we talked about the refrigerator yeah we talked about the refrigerator and you were introduced to refrigerators and the heat pump (2017)
	review current lecture content	a backward reference to information already given within the current lecture	main three things that have come out of here though out of these tests is yield stress ultimate stress and modulus of elasticity (3026)
	preview current lecture content	a forward reference to information that is upcoming in the current lecture	so what are we going to do today is we are going to wrap up chapter five the second law of thermodynamics yeah so today we should be able to determine finally the thermo efficiencies and the coefficient of performance for our ideal our reversible or our Carnot cycle (2019)
	preview future lecture content	a forward reference to information that is upcoming in future lectures on the module or course	in the next two lectures we're actually going to delve a little bit into material properties and then we're going to get back into the solid mechanics (3024)

Table 3.5: Examples of the ELC pragmatic categories and attributed types

The process of identifying pragmatic categories (and their attributes) raises an issue common to any kind of pragmatic analysis: the intention of the speaker (illocutionary force) and how it was interpreted / acted upon by recipients (perlocutionary force) (Austin 1962, Searle 1969, 1976) cannot always be understood with certainty. Even if participants are interviewed, they may not be able to make explicit their communicative intentions and understandings. This issue applies to all kinds of functional analyses of text. In the case of this study, no attempt was made to identify the function of every utterance. Instead, the broader communicative purposes of longer strings of text were sought. Although the same difficulties in decoding intentions and understandings may still arise in this case, it might be that they do so to a lesser extent.

The annotation categories were identified in dialogue with all those involved in the original development of the ELC, as shown by the progression of category adjustments in Table 3.3. Once these categories were established, I made every effort to ensure that fellow researchers would generally agree with my categorisation decisions through a series of reliability tests (described in 3.5).

3.5. Testing the annotation: IAR

3.5.1. Statistical measures

This section outlines the inter-annotator reliability (IAR) testing undertaken at different stages of the development of the ELC (see Figure 3.2 and 3.5.2-3.5.4), and also the statistical measurements used and their application.

An appropriate statistical measurement of inter-annotator agreement is necessary to confirm the validity of the annotation scheme, and the guidelines that underpin it. It is widely agreed that content analysis without some measure of agreement is not an effective endeavour (Artstein and Poesio 2008: 557, Lombard, Snyder-Duch and Campanella Bracken 2010). Limiting the subjectivity of manual discourse annotation is particularly important in

spoken corpora. If annotators demonstrate an acceptable level of agreement, the scheme is shown to be internally consistent, which gives more credence to findings. Reliability testing also identifies insufficient levels of calculated agreement, which is useful for demonstrating weaknesses in an annotation scheme.

The ELC is manually annotated using a fairly experimental taxonomy, and this presents specific challenges in terms of defining (as well as measuring) agreement. Leech (2005: Chapter 2) notes that the lack of accord over linguistic terminology and phenomena means that even seemingly objective labels are open to interpretation, making it difficult to reach consensus; "there is no absolute 'God's truth' view of language or 'gold standard' annotation" against which even POS tags can be measured.

Pragmatic annotation is, of necessity, particularly subjective, and for this reason the quality of the ELC annotations are measured in terms of IAR, rather than Artstein and Poesio's (2008) notion of *accuracy* (defined as measured against a gold-standard), or intra-annotator reliability (defined typically as the results from the same annotator over time). The focus is on the reproducibility of annotation decisions, both to ensure consistency and to provide a foundation for future expansion and comparability if more subcorpora are added to the ELC.

Two measures of IAR for nominal data are commonly used to calculate paired (that is, between two annotators) agreement. Both Scott's pi (1955) and Cohen's kappa (1960) are classification measurements of agreement based on mutually exclusive categories. Both measurements calculate observed agreement, expected agreement, and both take into account the element of chance. The same formula can be used for both at the highest level because Pr(*a*) is the relative observed agreement:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

Figure 3.3: Formula for calculating inter-annotator agreement

The difference between the two formulae lies in the calculation of expected agreement, Pr(*e*). Scott's pi assumes that the distribution of responses from annotators is equal, whereas Cohen's kappa uses the actual number of responses from each annotator to calculate chance probability. These measures, however, only calculate agreement in terms of an exact match.

The ELC annotation does not pose a classical classification problem because it indexes strings of text rather than single lexical items. The categories annotated are more subjective than other types of NLP annotation, such as POS tags, both in terms of definition and in terms of the identification of start and end boundaries. The calculation of an exact match between annotators is, therefore, of less value than the calculation of a percentage match (the overlap in tokens annotated as a certain category). Exact matches are not expected to be the norm, so calculating the *fuzziness* of the boundaries that different annotators identify (that is, the intersection in agreement) was deemed to be a more valuable test of the reliability of ELC annotations.

A common occurrence in the ELC is slightly offset annotation indices when two annotators deal with the same text. This is illustrated by Figure 3.4 in which the text highlighted in grey marks the overlap (or intersection) of agreement between annotator 1 (ANN1) and annotator2 (ANN2).

Intersection

<story type="exemplum"_ANN1>ok the other one if we look <story type="exemplum"_ANN2>remember last time when I show you the er Sampoong er which is the one which happen in Korea this is what happen when you don't properly consider design yeah ok so they put up a new equipment on the roof top of the building and this is very heavy they push along the slab ok which is not practical ok so I push forward a bit ok this is the the events yeah that lead to the failure ok ok see what happen alright so that is a more serious case related to failures yeah hopefully none of our students in the future will be what we call it er will be relat- will be what we call it er link with such event yeah if you look at the video just now er many people were found guilty and some of them were sentence to several years in jail er including the the C E O I believe some of the engineers as well negligence yeah greedy</story_ANN1> ok those are the thing that happen at the at the what we call it at the projects</story_ANN2>

Results (in tokens, not including annotation/markup)

String length = 194
Ann1 = 176
Ann2 = 187
Intersection = 169

Figure 3.4: Intersection in agreement in two annotated examples (non-ELC annotation added)

Agreement intersection is calculated using the formula:

$$agreement(A,B) = \frac{(\text{indices\_A} \cap \text{indices\_B})}{(\text{indices\_A} \cup \text{indices\_B})}$$

Figure 3.5: Formula to calculate the intersection of annotation boundaries

The ELC IAR testing calculates percentage agreement, instead of simple observed agreement (that is, agree or not agree). In the equation in Figure 3.5, ∩ refers to an intersection and ∪ to a union. So, the level of agreement between ANN1 and ANN2 (A,B) is equal to the intersection between A,B (the overlap shown in grey: 169 tokens) divided by the union of A,B (the total string: 194 tokens). Perfect agreement - an exact overlap in the boundaries of the text annotated - would equal an agreement of 1, or 100%. The extent to

which the annotators agree in Figure 3.4 is 169/194, in other words: 0.87 probability, or 87% agreement.

3.5.2. IAR test 1: establishing overall agreement

As shown in the workflow model (see Figure 3.2), IAR testing was performed at three points: between Phases 1-2, between Phases 2-3, and between Phase 3 and the output of the current (2014) version of the corpus (Phase 4). The three steps of IAR testing correspond to various needs at each point between the four Phases. They also represent an increasingly coarse approach to agreement at each pass as confidence in both the tagset and indexing grew.

Inline annotation of the ELC transcripts was carried out independently by four people: myself as lead annotator (rf1002) who marked up and annotated all transcripts, two local experts (rm1001 and rf3001) who did partial markup and partial annotation of transcripts in two subcorpora (one each), and an independent language expert (rm1003) who annotated a sample from all subcorpora (see Table 3.6). The project leader, who is identified as rf1004, examined the annotations in the final stage of the process (see 3.5.4).

annotator identifier	**rf1002**	**rm1001**	**rf3001**	**rm1003**
UK transcripts (total = 30)	30	26	0	1
MS transcripts (total = 18)	18	0	0	5
NZ transcripts (total = 28)	28	0	20	2
total transcripts annotated	76	26	20	8

Table 3.6: Breakdown of annotator workload

In the first round of IAR testing, the four annotation sets were split into four directories. I wrote a Python script to compare my annotations as lead annotator (rf1002) to those of each of the other three annotators. To compare annotations, the script firstly matches files of the same name (1001, 3010, 2009 et cetera) within the two directories. It then traverses each pair of files, logging: 1. the umbrella annotation element (humour, story, or summarising), 2. the start index of each annotation (that is, position of first token following

the opening tag), and 3. the number of tokens within the annotation tags (which is used to generate the end index). These three layers of information are used to build an array (an ordered list).

For each pair of files, the array – containing the start and end indices of each annotation element – is used to match pairs of the same element annotations and calculate the overlap in tokens. This overlap is based on the intersection divided by the union (see Figure 3.5). Where an overlap occurs, it is thus reported as an agreement value (with 1 as absolute overlap). Where no overlap occurs, the indices of potential matches are printed in parentheses. It is this step which allows for the manual post-correction of annotations which have been offset due to alterations to the static text. Where parentheses were reported, I checked each null match manually.

Table 3.7 summarises the results by giving an overall agreement value for each of the elements according to annotator pairs. A total number of indexed strings is given for each annotator within the pair (that is, how many instances of that language function each identified). Overlaps are instances where two annotators (at least partially) identified the same string as performing the same function. The agreement value is the intersection divided by the union of the tokens within those matching indexed strings. The *misses* are the number of strings indexed that had no overlap between the annotator pairs. The results are broken down according to subcorpus, and an overall probability value is given in the final row of Table 3.7.

annotator pairs		humour	story	summary
rf1002 vs rm1001	overlap probability	0.46	0.50	0.62
	rf1002 total	262	64	433
	rm1001 total	59	10	34
	overlaps	29	7	30
	misses	231	60	407
rf1002 vs rf3001	overlap probability	0.49	0.71	0.69
	rf1002 total	59	45	327
	rf3001 total	29	21	117
	overlaps	0.46	15	88
	misses	231	36	268
rf1002 vs rm1003	overlap probability	0.25	0.40	0.53
	rf1002 total	18	18	93
	rm1003 total	30	67	99
	overlaps	6	13	50
	misses	36	59	92
rf1002 vs all	overlap probability	**0.40**	**0.54**	**0.61**

Table 3.7: IAR test 1: agreement probability based on annotator pairs per pragmatic element

In non-exploratory coding systems, coefficients between 0.80-0.90 and in some cases as low as 0.70 are widely considered to be acceptable (Lombard, Snyder-Duch and Bracken 2002: 593, Neuendorf 2002: 145). At the other extreme, however, a measure of probability can be considered meaningless because any deviation from absolute agreement renders the results uninterpretable (Krippendorff 2004: 413). This extreme approach is not maximally useful given the expected levels of fuzziness in agreement in the ELC annotation. More flexible agreement scales rate over 0.75 as "excellent beyond chance", 0.40-0.75 as "fair to good beyond chance", and less than 0.40 as "poor beyond chance" (Capozzoli, McSweeney and Sinha 1999: 6).

This first IAR test resulted in an average overlap agreement of 0.52 (52%) between the lead annotator and three other annotators, with some variation (0.40-0.61) (see Table 3.7). The lowness of the agreement is linked to two factors: the subjectivity of the categories, and, from a procedural point of view, the evolution of category definitions. IAR test 1 occurred relatively early in the process of category definition and refinement. Although no major changes were made at the element or attribute level, annotation guidelines (such as the inclusion of hesitation markers and false starts) were established after rf3001 and rm1001

completed their partial annotations. An additional confounding factor was that the transcripts were corrected by rf1003 in Phase 2 (resulting in an 18% change, see Figure 3.2), which caused minor contraction or extension of some of the text within annotated boundaries (although the Python script used to compare texts allowed for some variation in matching pairs).

Most helpful were the complete misses (instances of no overlap), as these signalled examples for further consideration and helped to refine the category descriptions in Phase 2, as did instances where overlap was present but small. Another possible method of dealing with the issue of non-static base texts in the calculation of these early IAR figures would be to visualise the arrays, but the approach adopted provided the necessary statistical data.

Overall, 52% agreement was considered to be an encouraging start at this stage. It falls within the "fair to good beyond chance" range (cf. Capozzoli, McSweeney and Sinha 1999), and reflects an in-progress taxonomy of subjective categories. In an overview of the available approaches, Joyce (2013) suggests that a percentage agreement measure can yield meaning if used cautiously and not as the only indicator of reliability. Although misses were identified, one concern with the first test was that averaging out results might obscure particular disagreements. A variant of the first test was therefore undertaken between Phase 2 and Phase 3 to check specific agreement.

3.5.3. IAR test 2: checking specific agreement

In two separate 90-minute sittings at Coventry University on 01/12/2012 and 17/09/2014, 14 different participants (two groups of 8 and 6) were asked to manually encode a hardcopy sample of ELC transcripts. All participants were language experts with no prior connection to the project. Four participants were from Asia and 10 from Europe (including four from the UK). Three 500 word strings were extracted from the corpus (one per subcorpus) for the sample. The only selection criteria were that at least two sets of annotation indices (start

and end) as identified by the lead annotator were present in the sample. All participants were given a description of the categories and annotation principles (outlined in 3.4.6), the sample texts were stripped of all annotation, and clips of video data aligned to the samples were shown.

The annotations of the 14 participants (hereafter p1-p14) were then transferred to electronic form and the formula to calculate intersection boundaries between participants and lead annotator (Figure 3.5) was applied. The results are shown in Table 3.8.

	sample 1 (1010)			**sample 2** (2017)			**sample 3** (3024)			*average*
	humour	*story*	*summary*	*humour*	*story*	*summary*	*humour*	*story*	*summary*	
p1	0.59	0.68	0.92	0.41	0.69	1.00	0.85	0.71	0.79	0.74
p2	0.67	0.78	0.85	0.67	0.77	0.95	0.52	1.00	1.00	0.80
p3	0.82	0.21	1.00	0.60	0.65	1.00	0.60	0.67	1.00	0.73
p4	0.76	0.69	1.00	0.76	0.70	0.81	0.64	0.90	0.85	0.79
p5	1.00	0.72	0.76	0.87	0.85	0.62	0.59	0.82	0.42	0.74
p6	0.43	0.80	0.72	0.56	0.78	0.90	0.89	0.95	0.80	0.76
p7	0.87	0.56	0.95	0.89	0.81	0.82	0.45	0.84	0.51	0.74
p8	1.00	0.70	0.81	1.00	0.90	0.57	0.53	1.00	0.93	0.83
p9	0.87	1.00	0.68	1.00	1.00	0.87	0.75	0.80	0.95	0.88
p10	0.65	0.65	0.92	0.74	0.81	0.98	0.67	0.80	0.90	0.79
p11	1.00	0.94	0.67	0.65	0.83	0.62	0.66	0.59	1.00	0.77
p12	0.82	0.79	1.00	0.89	0.33	0.85	0.76	1.00	0.74	0.80
p13	0.65	0.23	0.83	0.85	1.00	0.73	0.74	0.96	0.94	0.77
p14	0.90	0.80	0.82	0.59	0.96	0.79	0.55	0.70	0.81	0.77
average	0.79	0.68	0.85	0.75	0.79	0.82	0.66	0.84	0.83	0.78

Table 3.8: IAR test 2: intersection agreement for p1-p14

The average intersections at this stage of testing on the smaller sample were both higher and more consistent. Average overall agreement across categories was 0.78 (with a variation range of 0.66-0.85). Some low overlap was recorded, for example 0.21 (p3, sample1, story). There were no complete misses and 19 instances of exact matches in annotation indices occurred. Instances of low overlap were discussed. Although the sample was smaller, this round of testing pushed the reliability of the annotation just into the "excellent beyond chance" category (Capozzoli, McSweeney and Sinha 1999), or at least towards acceptable reliability (Lombard, Snyder-Duch and Bracken 2002).

3.5.4. IAR test 3: hit or miss

The final IAR test was intended to eliminate remaining examples of misses to ensure the highest possible overall reliability. In this case, all instances of pragmatically annotated text at the finest level of element and attribute type were extracted for review by the project lead (rf1004). Instead of identifying boundaries, the primary purpose of this final test was to accept or reject annotations. Some examples were queried for further discussion with the lead annotator (rf1001). The results are summarised in Table 3.9.

	humour	**story**	**summary**	**total**	**average**
a. initial number of annotated strings	695	161	1306	2162	
b. miscategorisation identified by rf1004 resulting in the removal of annotation	65	9	60	134	
c. partial miscategorisation identified by rf1004 resulting in either adjustment to attribute type, or contraction/expansion of annotation boundaries	92	12	86	190	
d. query raised by rf1004 resulting in no change to annotation	12	11	87	110	
e. final number of annotated strings (a-b)	630	153	1246	2029	
f. % boundary adjustment = (c/a)*100	14.60	7.89	6.90		9.80
g. % rejection = (100-(e/a*100))	9.35	5.59	4.59		6.51

Table 3.9: IAR test 3: results of complete annotation review

The difference in the number of categories originally identified and those rejected by rf1004 (no overlap) is less than 10% in each category and less than 7% on average (see Table 3.9, row g), and the percentage of instances where boundaries were adjusted is between 7-15% and less than 10% overall (see Table 3.9, row f). In both cases, humour was the most complex category to identify, as shown by its low intersection overlap score in the first IAR test (Table 3.7).

As a final failsafe to ensure consistency across the corpus, the third IAR test delivered high agreement on the identification of categories, including some fuzziness (93%), and 84% of these agreed (non-rejected) instances returned a perfect match on annotation indices. Overall agreement at this stage had a probability value of 0.84-0.93, which is in the upper range of general acceptability.

This section has discussed the three types of IAR testing conducted on the ELC and the way in which they were calculated. Overall, acceptable agreement concerning both the indices of boundaries and overall category types was reached, and the process was useful for refining descriptions and identifying problems between Phases 1-4 of corpus creation.

3.6. Analysing the annotated data

3.6.1. Simple data mining

The main advantage of the ELC inline annotation is that text is indexed by tokens, allowing for extraction, comparison and statistical measure at the level of individual lexical items or longer strings. Analysis of lexis both within the pragmatically annotated string and in comparison to strings of different types or non-annotated data is thus possible.

The indices allow quantitative data to be extracted and then normalised, a step which is particularly significant in the ELC because lecture length (tokens) and subcorpus size (lectures) are not equal (as described in Table 4.1). The statistics underpinning all analyses in discussions in Chapter 5, Chapter 6, and Chapter 7 are therefore normalised, with reference to the raw data where informative.

3.6.2. Corpus linguistic techniques

The normalised data are extremely useful for looking at decontextualised quantitative patterns concerning how often and for how long strings of pragmatically annotated text occur. They cannot, however, give any information about the nature of those strings at the level of lexis. This section gives an overview of the statistical corpus linguistic methods that, along with the techniques discussed in sections 3.6.1 and 3.6.3, enable the qualitative analysis.

All statistical analysis related to lexis is based on the initial construction of a list of tokens and the calculation of their frequency. Raw frequencies, however, give little information

without context. As McEnery and Wilson (2001: 81) point out, "[t]he use of quantification in corpus linguistics typically goes well beyond simple counting". This observation applies particularly to the type of quantification needed to make the initial corpus-driven broad-brush discoveries required in this thesis.

The importance of individual lexical items in a corpus can be measured statistically in two ways: 1. normalised frequency difference, or 2. statistical significance (normally measured through a log-likelihood or chi-squared test). The aim of such measurements is to identify words that are most salient, or *key*, to the analysed text. Scott (1997: 236) explains that:

> [a] key word may be defined as a word which occurs with unusual frequency in a given text. This does not mean high frequency but unusual frequency, by comparison with a reference corpus of some kind.

The value of *keyness* is calculated as the statistical probability (p value) that a lexical item will occur with greater (non-accidental) frequency in a text compared to a reference corpus (Scott 2012). The maximum threshold for the p value is commonly set at p=0.01 in corpus linguistic investigation, which means that the probability that any reported positive statistical difference in frequency is due to chance is less than 1% (Gabrielatos and Marchi 2011). The salience of a word is therefore indicated by high keyness and low p value.

The p value calculated by chi-squared is the probability that frequency is based on chance alone through the measurement of the difference between observed frequencies (actual data) and expected frequencies (if the only factor at work is chance), based on an assumed normal distribution. The p value calculated by a log-likelihood test does not depend on an even data distribution. In both cases, the use of aggregate data in the statistical analysis of frequency can be criticised for yielding falsely positive results (such as inflated log-likelihood values) because potentially significant variation within groups of data is occluded by the calculation of inter-group difference. This is a concern in the analysis of a relatively small sample of data such as the ELC, which is susceptible to skewing due to the idiosyncrasies of

individual lecturers. Manual analysis of concordance lines and interactive visualisation techniques are employed to guard against the influence of outlying results.

In comparison to chi-squared, Dunning (1993: 65-66) explains that, as a frequency measure, log-likelihood operates more effectively "with very much smaller volumes of text than is necessary for conventional tests based on assumed normal distributions". He also regards log-likelihood as more useful for comparing the significance of both rare and common phenomena. Gabrielatos and Marchi (2011) also found that although log-likelihood is sensitive to word frequencies and corpus sizes, in calculations of statistical significance the absolute and relative size of compared corpora does not make a difference within the same genre. In the ELC dataset, the keyness of individual lexical items was thus calculated using a log-likelihood test. The threshold of greater than or equal to a probability of 1% false positive ($p \geq 0.01$) corresponds to a log-likelihood return of 6.63, which will be the cut-off point for items considered as key.

The analysed text is contained within the annotation indices at the level of element or attribute, which can be filtered by subcorpus membership. A Python script was used to extract a series of raw text files based on the possible combinations of variables of elements (total 4), attributes (total 17), and subcorpora (total 3), resulting in 80 files. In addition, a mirror set of 80 inverse files (comprised of all text *except* the element/attribute/subcorpora variable specified) was outputted to provide a set of reference corpora. For example, to identify the keywords in all humour in the New Zealand subcorpus, the text annotated within the humour element in this subcorpus was extracted and compared against a reference corpus of all non-humour (that is, all text not indexed within humour tags) in the New Zealand subcorpus.

Although *keyness* testing could be achieved using a simple script or specific application (such as UCREL's online calculator (Rayson n.d.)), this thesis uses the *keyword list* function within *AntConc* (Anthony 2011). *AntConc* is freely available, lightweight (in terms of

processing capacity required), and well-maintained. The results can therefore be easily checked by other researchers once the ELC has been made public. Given the size of the ELC, the other corpus linguistic techniques that inform the qualitative analysis can also be applied using *AntConc*.

For reasons of space, discussions of keyword results will be limited to the presentation of 40-50 items (split between positively key and negatively key), assuming that their statistical value is above the acceptable probability threshold of log-likelihood 6.63. Grey areas in tables of keywords denote either an absence of items, or items that did not achieve this required p value.

In addition to identifying highly salient, or key, individual lexical items, light is also shed on the character of the strings of discourse within elements and/or attributes by calculating which items are most frequent in contiguous sequence, namely n-grams (where n refers to number of tokens). This process of automatically identifying strings of tokens that commonly co-occur bypasses researcher hunches (based on domain, syntactic and lexical knowledge) and returns the results within a single window, allowing comparison of strings across the variables of: 1. subcorpora, and 2. text annotated with a pragmatic function versus non-annotated text.

In Chapter 4 - Chapter 7, n-grams are calculated in *AntConc*, then exported and normalised according to subcorpus/corpus size (tokens) (based on Table 4.1). N-grams (most commonly 4-grams) tend to be more indicative of genre than topic (Nesi 2012b: 418). As the ELC is not a large corpus, and the text for investigation against the various *inverse* reference corpora can be as small as hundreds of tokens, 3-grams are calculated where: 1. the most common 4-grams do not retrieve a minimum frequency of five in any element or attribute across the corpus or subcorpus, or 2. the 3-grams augment the picture rendered by the 4-grams. The concordance and collocate functions within *AntConc* are then consulted to gain more

nuanced, qualitative understanding of the language patterns that emerge from the bottom-up keyword and n-gram approaches to pattern identification.

Lexical variety is also measured through calculation of the ratio of unique tokens (*types*) to total number of tokens, which is expressed as a percentage: (number of types/number of tokens)*100. As the ELC segments analysed are of different lengths, Scott's (2014) standardised type/token ratio (STTR) approach is followed. STTR calculates type/token ration (TTR) every *n* words and then calculates the average. A Python script was written to calculate the STTR where *n* is 1000, the results of which are given in Table 4.4 and Figure 4.2.

The broad NLP quantitative methods discussed, along with the approaches from data visualisation laid out in section 3.6.3, guide the qualitative analysis of individual examples. The aim is to first gain a good understanding of overarching patterns, and then to pick out typical examples for finer analysis and comparison.

3.6.3. Data visualisation

3.6.3.1. VISUALISATION TECHNIQUES

After the pragmatic categories were indexed and reliability was checked, data visualisation techniques were applied alongside the exploration of the context and co-occurrence of the lexis.

The inline pragmatic annotation gives a set of indices that can be extracted to provide a picture of the occurrence and duration of elements and attributes across the corpus. Numerical indicators of similarity/difference across subcorpora such as frequency of discourse feature and average token count were extracted, and statistical tests on individual lexical items were calculated to reveal patterns at the level of language. What this simple data mining does not show, however, is the bigger picture, and, crucially, comparability across variables such as subcorpora and pragmatic feature.

Answering the research questions is heavily reliant on being able to clearly process patterns related to where and for how long the stretches of text that perform particular pragmatic functions occur. Overall comprehension of the occurrence and duration of data patterns is arguably inaccessible (or at least less accessible) without abstracting macro patterns in the metadata through visualisation techniques.

For example, basic frequency information about the annotation of the summary element and its four attributed types within an ELC lecture might show that previews of the current lecture are common. Unmediated visual scanning of the XML might show that this category occurs quite often at the beginning of lectures, perhaps more often in one subcorpus than in another. Keyness tests and the calculation of n-grams might give an indication of salient lexis. Familiarity with the data following multiple cycles of annotation would certainly have afforded the lead annotator intuitions for pattern spotting. As Baker (2006: 25) observes:

> The process of finding and selecting texts, obtaining permissions, transferring to electronic format, checking and annotating files will result in the researcher gaining a much better "feel" for the data and its idiosyncrasies. This process may also provide the researcher with hypotheses as certain patterns are noticed – and such hypotheses could form the basis for the first stages of corpus research.

The types of analysis hypothesised above, however, represent only such *first stage* investigation.

A bird's eye view of the annotation data in combination with the extracted statistics was taken to help establish a more informed starting point for qualitative analysis. This view was achieved through data visualisation, or *Infovis*, techniques. As Yi et al. (2007: 1226) explain, "[o]ne of the essential purposes of Infovis is to reveal hidden characteristics of data and the relationships between them". Visualisation techniques rely on "computer-supported, interactive, visual representations of data to amplify cognition" (Card, Mackinlay and Shneiderman 1999: 6). Visualising corpus data constitutes a second wave of abstraction, yet as Rayson and Mariani point out, despite the wealth of data that existing corpora provide, "visualisation techniques have not been widely explored within corpus linguistics" (2009:

article 426). The justification for visualising data, then, lies in the gap between these statements. Corpus linguists have the data and are already looking for patterns through other computational means, such as keywords, concordances and n-grams; visualisation offers a powerful and customisable tool which adds value to this process.

3.6.3.2. REQUIRED VISUALISATION TYPE

The dataset genre (engineering lectures) and the inquiry question (comparison of where and for how long pragmatic features occur across the corpus and comparatively across subcorpora) determine the annotation, and thereafter the appropriate type of visualisation.

Like corpus linguistic methods, at the general cognitive level the visualisation of data functions as an external aid to thinking. When data are visualised, the burden on the working memory is relieved, which makes processing easier. An important consideration in this process is accurate representation, as "[t]here are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not" (Tufte 1997: 45, cited in Card, Mackinlay and Shneiderman 1999: 5).

The ELC data contains three variables which require accurate comparison: subcorpora, elements, and attributes. Multiple repetitions of the same graphic showing different data points can be used to interrogate combinations of variables. As Tufte (1990: 67) notes:

> At the heart of quantitative reasoning is a single question: Compared to what? Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives.

A common approach draws on two common data visualisation techniques: small multiples (repeating graphical elements) and timelines (a graphical representation of a time period on which events are marked), as exemplified in the sporting visualisations in Figure 3.6 and Figure 3.7.

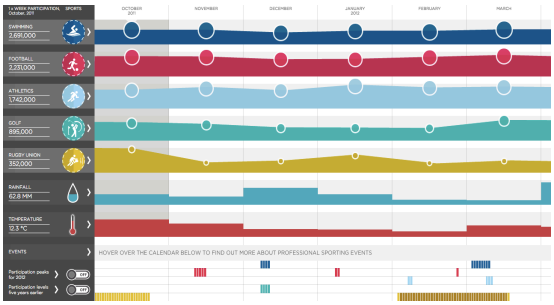


Figure 3.6: User-filtered visualisation of participation in five sports, by month (Sport England n.d)

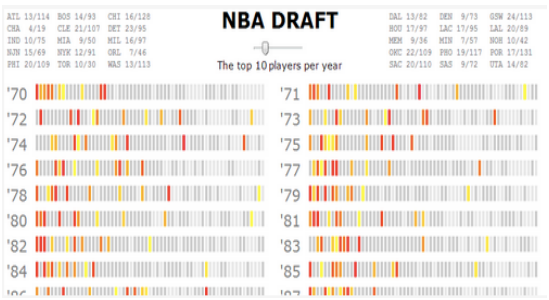


Figure 3.7: Non-user-filtered visualisation of NBA draft top players, by year (visual.ly 2013)

Stacking timelines according to the principles of small multiples in this way enables focus to be placed on the data presented, rather than how the data are presented.

3.6.3.3. EXISTING VISUALISATION SOLUTIONS

The timeline format is familiar and so easily processed cognitively. Many construction options exist, from tailored software like *Dipity* (2011), *Preceden* (2014), or *Timeglider* (2008), to graphing libraries like *Google Developers'* (2013) annotated chart. In the field of commercial business intelligence, user-friendly options such as *Spotfire* (2014) and *Tableau* (2014) allow data to be uploaded and suggest suitable visualisation types. These packages, however, allow limited or no customisation. The *details on demand* function in *Tableau*, for example, cannot display the original source text.

Interaction with original source text is naturally more of a concern for field-specific software where data visualisation is directly linked to other types of data view, such as concordances. Some corpus analysis packages, such as *WordSmith Tools* (Scott 2012, Version 6) or *AntConc* (Anthony 2011, Version 3.2.2), will plot the occurrence of linguistic features, but cannot display multi-token durations (as shown in Figure 3.8).

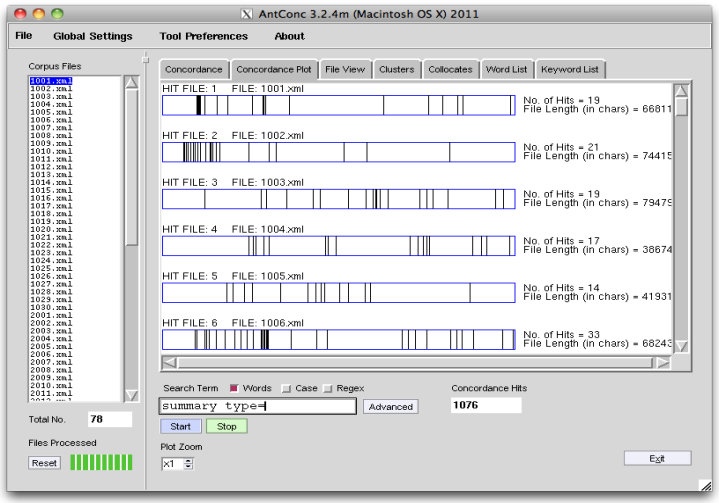


Figure 3.8: Concordance plot of pragmatic annotation data using *AntConc*

The Corpus Linguistics in Cheshire (*CLiC*) interface – used by Mahlberg and Smith (2012) in their work on literary analysis – displays the occurrence of searched term/feature in a similar way. Figure 3.9 shows *CLiC*'s plot view of a simple search for "lawyer" in long suspensions in a corpus of 15 novels by Charles Dickens.

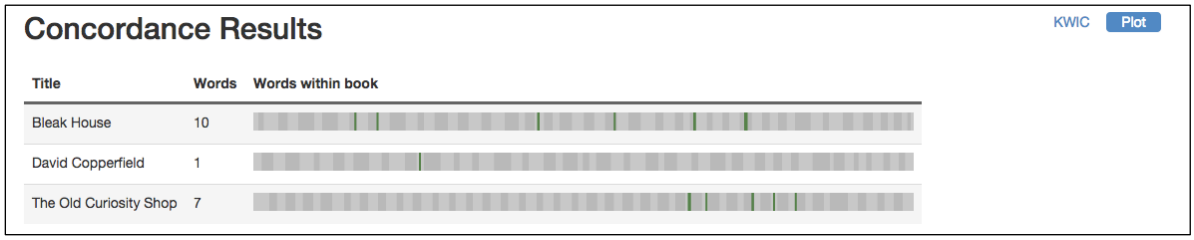


Figure 3.9: Example of a plotted concordance view in *CLiC*

Like *AntConc*, *CLiC* visualises occurrence through stacking – presumably normalised – results per unit (in this case, novels).

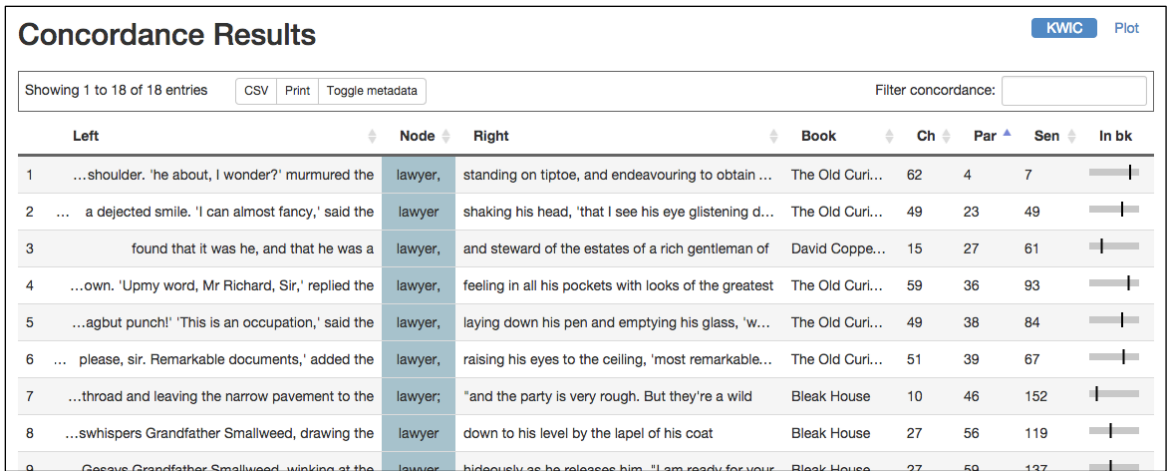


Figure 3.10: Example of a KWIC concordance view in *CLiC*

In addition to the textual key word in context (KWIC) display in Figure 3.10, *CLiC* renders a mini visualisation of the relative occurrence of the keyword in the far right column (labelled *In bk*). The stacking of the visual element effectively mirrors the stacking of keyword terms, offering the dual possibility of comparing concording text and relative occurrence within that text. This approach offers an alternative to a click-through interaction with the textual source by weighting the two aspects (textual and visual) equally and presenting them simultaneously.

The need for direct interaction between the visual overview and the source data, and the ability to compare multiple linguistic features and sort by different metadata variables, is, however, not a priority in available software packages. Existing tools from both business intelligence and corpus analysis/presentation contain powerful and extensive features, but do not meet the ELC need to display the relative duration of annotated strings or offer direct interaction between plotted view and source text.

3.6.3.4. The ELC solution: ELVis

Three distinct aims for the visualisation of data apply to this thesis: exploration, comprehension/analysis, and communication/presentation (cf. Card, Mackinlay and

Shneiderman 1999, Culy 2013). A small-scale, customised piece of software was created for these purposes, which will be referred to as *ELVis*.

ELVis is weighted towards exploration and analysis, following Card's (1999: 4) definition of a visual knowledge tool as a "[s]ubstrate into which data is poured and/or tool for manipulating it". Design is biased towards pattern-detection and knowledge crystallisation. The communication of patterns is a useful outcome to illustrate discussion in Chapter 4 – Chapter 7, but the primary purpose is to create a manageable way of processing all data through a filter-controlled visualisation that identifies otherwise unobservable patterns. *ELVis* visualises the comparative occurrence of indexed strings across cultural subcorpora in an interactive form.

The *ELVis* display is comprised of three visual components: the *core visualisation* is most "expressive" (Card, Mackinlay and Shneiderman 1999: 23), the *secondary visualisation* is a dependent summary of the data and filter choices of the core visualisation, and the tertiary *source text view* is also dependent on core choices but displays limited, untransformed information (that is, the original text). The webpage layout and final rendering of components is shown in Figure 3.14 and Figure 3.15.

The core visualisation uses the stacked timeline technique combined with an aggregated bar-chart to provide a dashboard view of the occurrence of annotated features, especially at the level of cultural subcorpora. The three subcorpora are distinguished by three blocks of background colour, which contain the 76 lecture identifiers on the y-axis. As shown in Figure 3.11, block 1 is comprised of the 30 UK lectures, block 2 the 18 lectures from Malaysia, and block 3 of the 28 lectures from New Zealand.

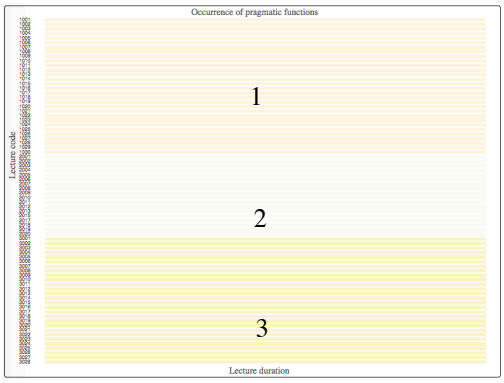


Figure 3.11: *ELVis* colour partitions

Following the principles of small multiples, in the core visualisation each stacked timeline (one row per lecture) directly compares the occurrence and duration of annotated pragmatic features on a base layer of subcorpora identification, as in the example of annotated instances of the humour element given in Figure 3.12.

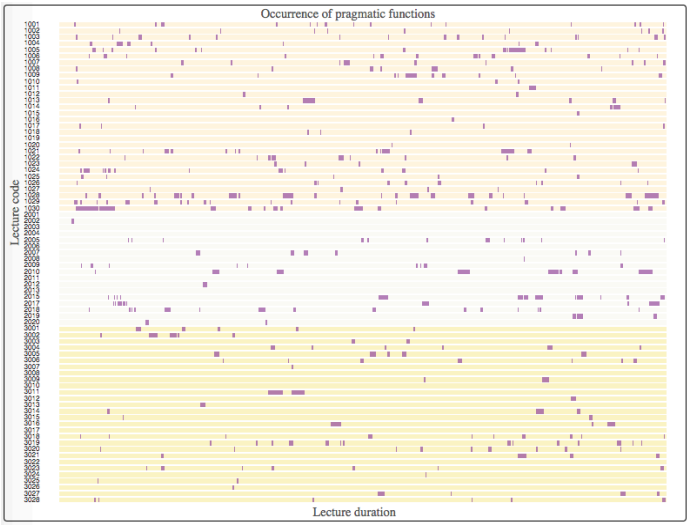


Figure 3.12: *ELVis* core visualisation: the distribution and duration of humour across subcorpora

No data aggregation was undertaken in order to avoid rendering complex or over-described single items. As bottom-up pattern identification is the goal of exploration, the loss or

occlusion of any data was not considered to be a desirable option, even at the expense of maintaining simplicity. A direct comparison of data is presented through displaying lecture identity on the y-axis and normalised duration on the x-axis in a single-columned series of small multiples.

3.6.3.5. THE ELVIS PIPELINE

Detecting patterns at the global level within annotated data does, however, involve some form of data transformation. Infovis techniques can be divided into two components working in symbiosis: 1. data representation (from mapping to rendering) and 2. user interaction with the system (Yi et al. 2007: 1230). A pipeline model can be seen as a dataflow network comprising modules, connections and execution (Moreland 2013: 368). The data representation sequence is broadly agreed: data are acquired/analysed, filtered, mapped, and then rendered, and interaction (through direct or indirect manipulation and interpretation) occurs during the transformational processes (Card, Mackinlay and Shneiderman 1999: 17, Carpendale 2003: 17, Chi 2000: 70, Yi et al. 2007: 1225). The construction of *ELVis* involved a pipeline that transforms raw text speech transcriptions into a form that shows the occurrence and duration of linguistic features in a comparative way.

The ELC data lends itself to the abstract notion of a data table as the identified indices are easily mathematically encoded. *ELVis* includes explicit transformations (data-data and data-visual structures) within the software. Transcription of the recorded lectures is the first step in the transformation of raw data to a data table (via automatic data transformation). This is an automatic transformation of value to derived value (MP4 -> TXT). Metadata is then added to plain text: body pragmatic annotation and structural markup, and header identifier and provenance information (as shown in the workflow model Figure 3.2). Again, this is an automatic transformation of value to derived value (TXT -> XML). The data table is then created from ordinal and nominal data extracted using a Python script: a flat array of file names (akin to data table metadata) and a series of hashes within an array for annotation information. This is an automatic transformation of value to derived structure

(XML -> (via PY) -> JSON). The hashes, one per annotation instance, contain the annotated text and relevant metadata, as exemplified in Figure 3.13.

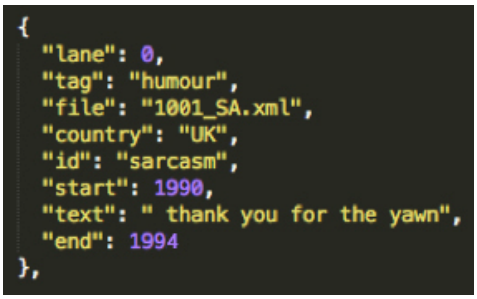


Figure 3.13: Sample annotation hash from lecture 1001

In Figure 3.13, the lecturer comment “thank you for the yawn” (*text*), which was identified as humour (*tag*) of the irony-sarcasm type (*id*), is given a (normalised, to adjust for small variations in lecture length) token start point of 1990 and end point of 1994 in the first row (*lane*) of the timeline, which is occupied by lecture 1001.

Data table information is then used to map visual structures and views (via visual mapping or view transformations). The two visualisations are mapped from data in the hashes, to which shape, class, fill colour, position and size are assigned. This is an automatic transformation of structure to derived structure (JSON -> D3 -> SVG within HTML). An on-click function assigns the text to the source text view (DOM element *view*). Filtering options for category and type are enabled via the *key* (a div element constructed from colours assigned by category and type). Based on filtering choices, data are aggregated to show the dependent secondary visualisation (JSON -> JS within HTML). Mouseover and mouseout events give an overview of the annotation metadata.

The normalised annotation indices are plotted on simple x-y axes. In the core visualisation, the x-axis is an ordinal number line showing 0-100% of the lecture. The y-axis gives the nominal lecture names, ordered by subcorpus. In the secondary visualisation the x-axis

displays the nominal data. Raw token counts for the pragmatic type (attribute/s) selected are shown on the y-axis.

3.6.3.6. ELVIS DISPLAY DECISIONS

Card et al. (1999: 23) define good, or fully "expressive" visualisations as those that preserve all and only data from the data table. The final transformation of the data table in *ELVis* contains quantitative representations of all annotation metadata retrieved from the body of the source texts (the indices of the pragmatic annotations) and header metadata necessary to describe the source text to which the indices belong. Sorting is thus enabled by lecture identity and subcorpus membership.

These two sets of information from inside and outside the data table are by default fully expressed in the unfiltered visualisation, with an option added for user-filtering. By providing a (linked) secondary visualisation of the same data in different graphical form, *ELVis* partly draws on *encode* techniques (cf. Yi et al. 2007: 1227). Specified variables are displayed through filtering, but both the secondary visualisation type (bar-chart) and its graphical properties are fixed. With emphasis on data exploration and analysis, the default display is pitched at a middle ground regarding user-filtering control: all lectures are shown on the y-axis and all categories on the x-axis, but not category types.

The inclusion of an option to display the original text from which the core visualisation indices were extracted is intended to mediate the tension between expressiveness and reconfiguration. The source text view fulfils the final part of Shneiderman's (1996: 2) experience-based Visual Information-Seeking Mantra by providing "details-on-demand", which is essential to the goal of informed qualitative analysis. By showing the source text alongside its visual abstraction, patterns and content can be simultaneously processed.

The composition of the *ELVis* webpage is based on a simple grid system arranged loosely on the rule of thirds, and adheres to a standard 1366 x 768 pixel display. Figure 3.14 shows this

page division, with proportions of the four major cells (core visualisation, source text view, secondary visualisation and key). Eye-scanning percentages for the rule of thirds (taken from Codrops (2014)) have been added. The composition represents a trade-off in ideal proportions, allowing enough space for data to be comprehended.

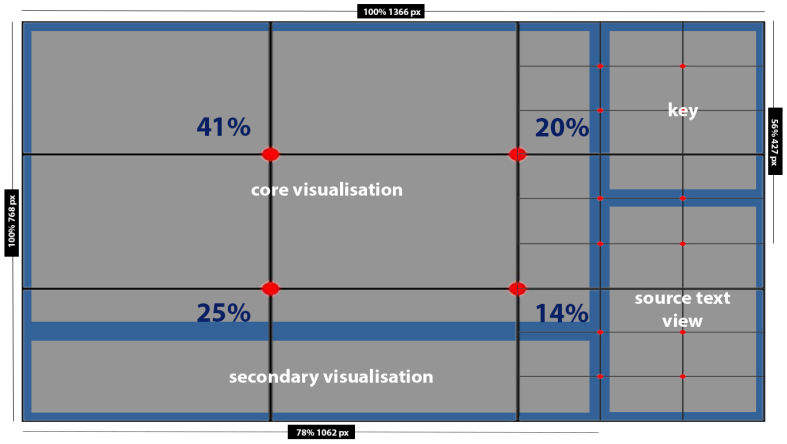


Figure 3.14: Proportions of the *ELVis* webpage with typical page scanning information

Effectively mapping data tables onto visual structures (that is, processing the data patterns) involves biological mechanisms of human perception and cognition, and types of graphical representation. At the low level of perception are the basic visual elements that are processed from a multi-element display *preattentively* (cf. Treisman 1985, Ware 2004: 152), that is, typically in less than 200 to 250 milliseconds (Healey 1996). Chipman's (1996) review shows that colour, size, width and closure are important preattentive features, and so are employed to convey the most salient aspects of the *ELVis*. Consideration of spatial substrate (cf. Card, Mackinlay and Shneiderman 1999: 26) works at two levels: the macro composition of the whole webpage and the microstructuring of the cells.

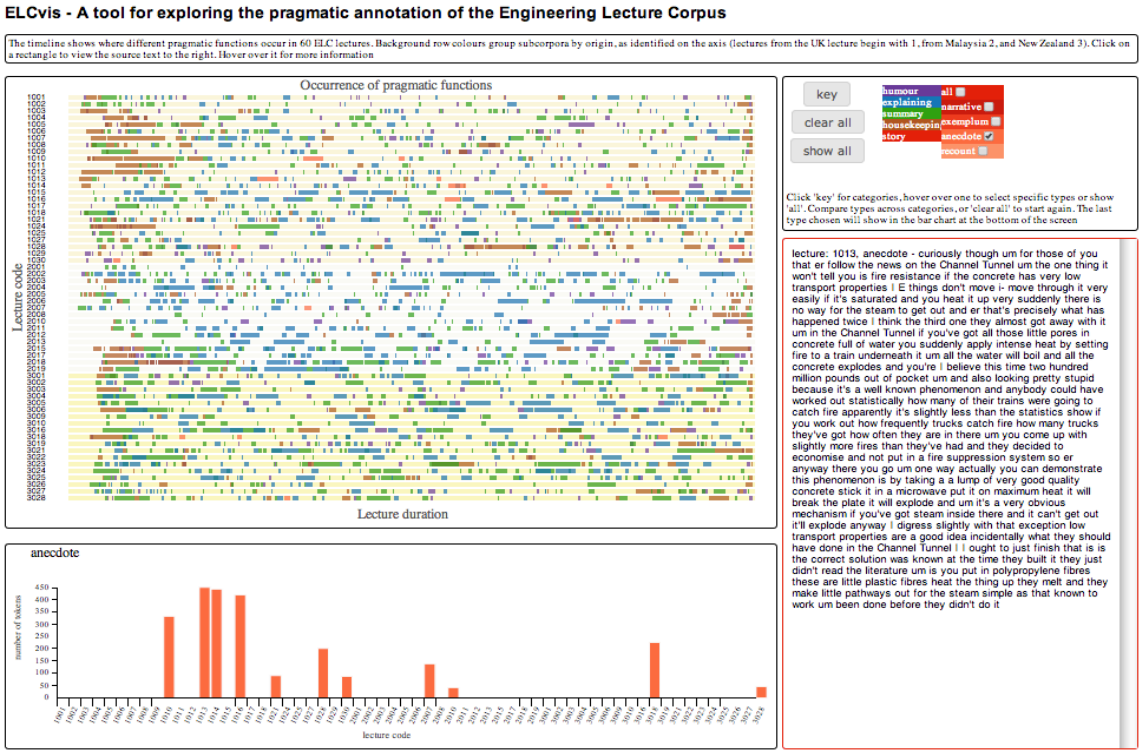


Figure 3.15: *Elvis*: all categories in the core visualisation, and the anecdote story type in the secondary visualisation and source text view

Drawing on Carpendale's (2003) discussion of Bertinian principles (adjusted for computational display), the most significant visual marks in the spatial substrate of *ELVis* include the area of the rectangles within the core visualisation, as these express the quantitative relationship between marks; height is consistent and width represents lecture duration in tokens as a normalised percentage. Position renders normalised occurrence in tokens of the lecture (on the x-axis), and lecture/subcorpus membership on the y-axis. The visualisation relies on the selective and quantitative characteristics of position.

In terms of graphical properties, also known as visual variables (Carpendale 2003), the colour fill of the rectangular marks in both the core and secondary visualisations is significant. Colour is extremely effective for nominal information encoding; it is a learnable form of labelling (Ware 2008: 123), it can speed up parsing, and it can enable meaning to be

extracted preattentively. An important role of the *ELVis* key is that visual variables (for example, colour and shape) are weighted more heavily if what is to be found is known; a type of goal-directed processing (Ware 2008: 13). Labelling storytelling as red, for example (as in Figure 3.15) facilitates top-down processing as perception is biased to find this colour. Associations are reinforced through reuse in the secondary visualisation, and in the border around the source view text.

Colour is a separable feature channel (Ware 2004: 167), but it has limited length – a measure of across how many changes distinction, or perceived difference, remains possible (Carpendale 2003). The number of lengths that allow distinction varies between up to seven (Carpendale 2003, Healey 1996: 270) and up to 12 (Ware 2008: 125-126). The choice of colours for the *ELVis* key addresses the ELC research criteria of enabling comparison of data across elements and attributes. The required distinctions can be ordered with hierarchical importance as follows: 1. across elements, 2. across same element attributes, and 3. across different element attributes.

Colour is commonly used as an umbrella term for hue, saturation and transparency, but it is distinguished from value (Carpendale 2003, Ware 2008). For the *ELVis*, a qualitative colour scheme of five data classes of hue (cf. Brewer 1999) was used to represent difference between the three pragmatic elements discussed and also those outside the remit of this thesis (*explaining and housekeeping*). As there is no distinction in the importance of categories, attention was paid to value and saturation to avoid contrast difference between categories (cf. Brewer 1999, Carpendale 2003). Allocation of the colours used is arbitrary with no intended cultural semantic meaning.

Encoding the data using colour underlines the problem of limited length as there are 17 types attributed to the three categories discussed; more than the seven to twelve distinguishable colours recommended. Colour was therefore extended to encompass the value variable, which Carpendale (2003) assigns a length of between seven to ten. As they

are based on value, sequential schemes normally represent ordinal data in the form of lightness/darkness as an indicator of low-high data values. However, multiple variables may be represented by a combination of colour schemes.

To enable data exploration, sequential schemes were assigned to enable cross-element *and* cross-attribute comparison (as exemplified in Figure 3.16). To address the second most important required distinction (within elements), five corresponding single hue sequential schemes were assigned to category types ranging from zero to nine classes; for example humour has nine subtypes, and story and summary both have four. All RGB (the *red*, *green*, *blue* system of representing colours on a computer display) codes were adapted from *ColorBrewer* (Brewer 1999).

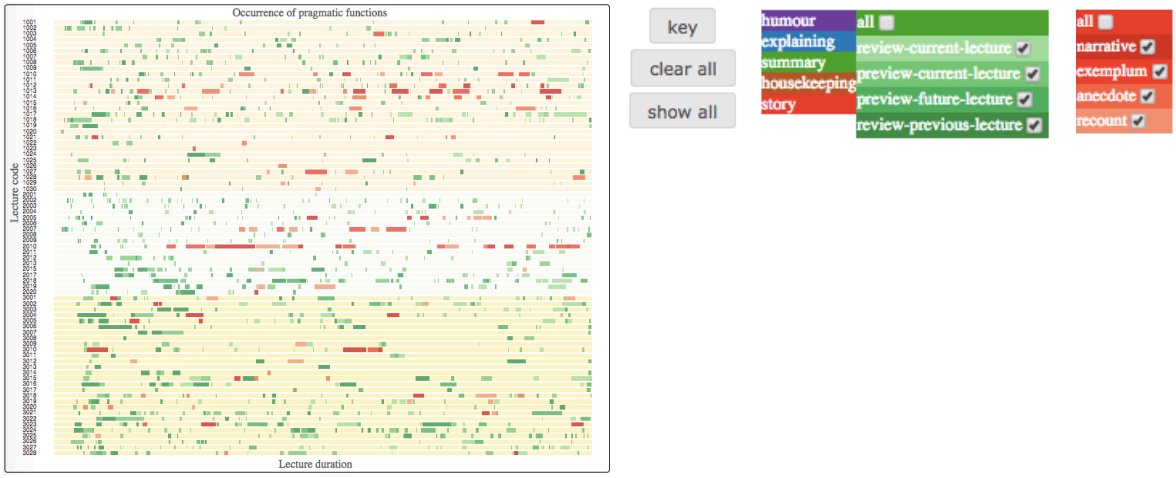


Figure 3.16: Humour and story types visualised through sequential colours

The light-to-dark value ranking of category types indicates both difference (across attributes) and similarity/association (within elements). Although distinguishing between elements was the priority, a middle ground was attempted in satisfying the other two requirements: attributes were distinguished within an element, yet, to an extent, remained associative - thus distinguishable across attribute range - by using the single hue schemes.

This is shown in the key in Figure 3.16 which represents the four summary attributes with four values of green alongside the four values of red that represent the four story attributes. Using this qualitative-sequential combination also enabled the third distinction (across attributes of different elements) to be made.

For the purposes of data presentation through screenshots in this thesis, however, if only a single annotation element is in focus, a qualitative scheme is assigned to the attributes, as shown in Figure 3.17 where the four summary types are rendered non-associatively for contrast.

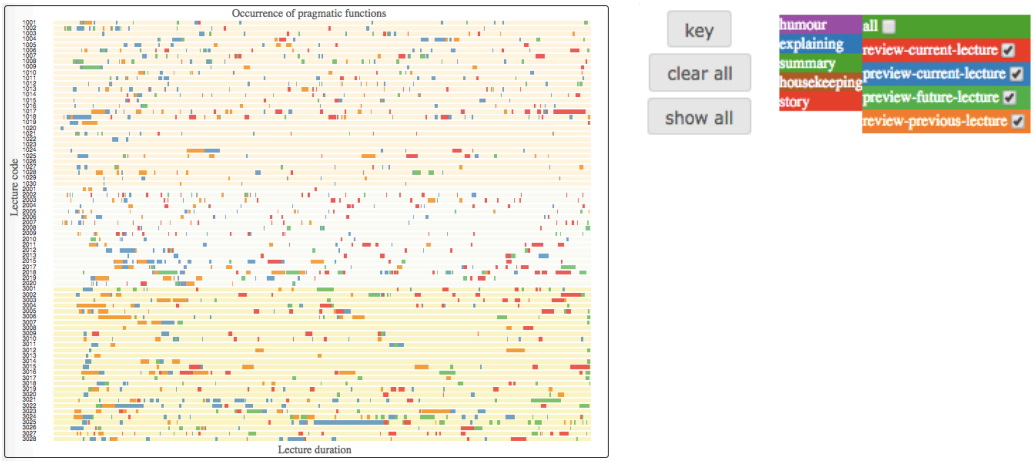


Figure 3.17: Summary types visualised through a diverging colour scheme

The first step to understanding the nature of the annotated text is to understand patterns of occurrence and duration within the 528,157 token ELC. Alongside and supporting statistical analysis at the level of lexis, the visualisation of the metadata enables language patterns to be *seen* in a way that is not possible from the raw data alone, in the sense of analysis and presentation (to researcher and to reader). Using *ELVis*, cross-subcorpora difference in occurrence and duration of the identified functional elements is made possible, along with the identification of patterns of attribute-chaining and any indications of structuring at the macro-level of the lectures.

3.7. Conclusion

Indexing the start and end of stretches of text that perform certain pragmatic functions within the ELC enables the linguistic description of typical lecture discourse features. A semi-bottom up procedure was cyclically undertaken to establish a hierarchy of pragmatic elements and attributes. Once an acceptable level of annotation reliability was achieved, patterns at the level of lexis were analysed through statistical corpus methods and macro-level structural patterns were identified through the custom-built data visualisation tool, *ELVis*. Combined, these quantitative approaches guide finer qualitative analysis of the annotated text. The triangular methodological approach informs discussion of the three ELC elements discussed in this thesis: summary (Chapter 5), humour (Chapter 6), and story (Chapter 7). An overview of the macro- (element) level data findings is first presented in Chapter 4.

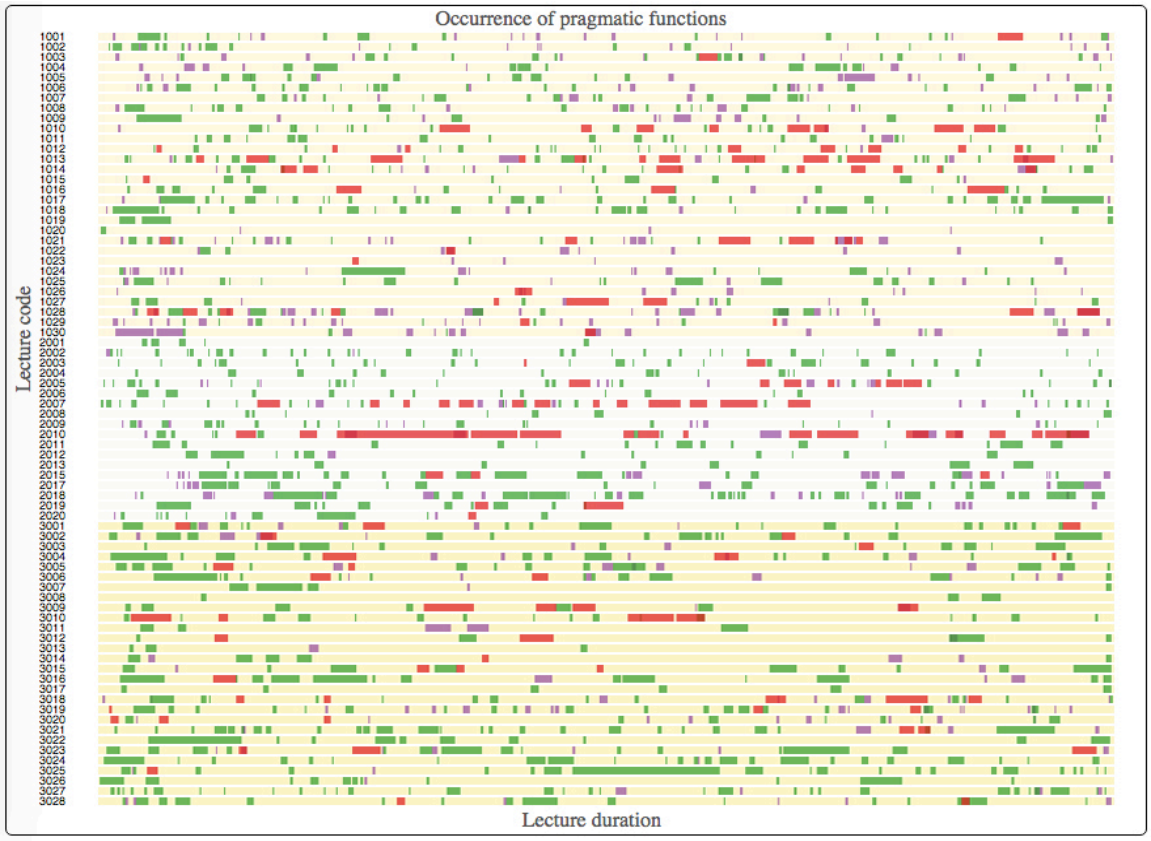
CHAPTER 4. DATA OVERVIEW

4.1. Introduction

This chapter overviews the data that informs later macro and micro analyses. A top-level visualisation of the occurrence and distribution of indexed humour, story and summary elements is first presented (4.2), followed by a more detailed breakdown of the corpus which shows raw and normalised token counts for each element and attribute across the subcorpora (4.3). Element-level analyses of lexical variation in terms of STTR (4.4), salient lexis in terms of keywords (4.5), and lexical sequences in terms of n-grams (4.6) are then presented. Features are not explored exhaustively or in depth. Rather, this overview chapter provides initial views on what can be extracted from the data, using the information as a starting point to guide investigation in Chapter 5, Chapter 6, and Chapter 7.

4.2. Visualisation overview

Figure 4.1 is an *ELVis*-generated rendering of where and for how long each instance of annotated text occurs across the corpus and subcorpora (as described in 3.6.3.4 - 3.6.3.6).



KEY: humour | story | summary

Figure 4.1: Occurrence and duration of all pragmatic elements

Based on a token count, pragmatically annotated text constitutes 16.41% of the corpus (see Table 4.1). The visualisation shows that instances of each element (humour, story and summary) punctuate each lecture with a degree of regularity; highly distinct patterns are not immediately apparent. There is some indication of tendencies within the data at the element level shown, for example summary occurs somewhat more densely and for longer duration in the first quarter of lectures. Patterns at the level of individual elements and attributes are discussed in greater detail in: 5.2.1 (summary), 6.3 (humour), and 7.3 (story).

4.3. Breakdown of the ELC

The ELC pragmatic annotation results in two variables that offer equally important analysis perspectives: 1. number of tokens, and 2. occurrence of strings. The normalisation of token counts is calculated based on the size of subcorpus (tokens) in relation to the total number of tokens of which any element (or any constituent attribute) is comprised (see Table 4.1).

		UK		MS		NZ		ELC	
element	*attribute*	*raw*	*%*	*raw*	*%*	*raw*	*%*	*raw*	*%*
humour		8872	3.53	3094	2.57	3263	2.08	15229	2.88
	bawdy	145	0.06	0	0.00	130	0.08	275	0.05
	black	241	0.10	683	0.57	91	0.06	1015	0.19
	disparaging	2373	0.95	189	0.16	504	0.32	3066	0.58
	irony/sarcasm	1859	0.74	162	0.13	314	0.20	2335	0.44
	joke	524	0.21	0	0.00	0	0.00	524	0.10
	playful	2288	0.91	1355	1.13	933	0.59	4576	0.87
	self-deprecating	874	0.35	245	0.20	880	0.56	1999	0.38
	teasing/mock-threat	310	0.12	127	0.11	199	0.13	636	0.12
	wordplay	258	0.10	333	0.28	206	0.13	797	0.15
story		8418	3.35	5937	4.94	3863	2.46	18218	3.45
	anecdote	2190	0.87	179	0.15	521	0.33	2890	0.55
	exemplum	2506	1.00	1952	1.62	193	0.12	4651	0.88
	narrative	2348	0.94	1446	1.20	1830	1.17	5624	1.06
	recount	1374	0.55	2360	1.96	1319	0.84	5053	0.96
summary		18318	7.29	11699	9.73	23212	14.80	53229	10.08
	review content of previous lecture	3807	1.52	2996	2.49	5622	3.58	12425	2.35
	review content of current lecture	4443	1.77	3777	3.14	5568	3.55	13788	2.61
	preview content of current lecture	7404	2.95	3137	2.61	7430	4.74	17971	3.40
	preview content of future lecture	2664	1.06	1746	1.45	4592	2.93	9002	1.70
all		251108	14.17	120211	17.24	156838	19.34	528157	16.41

Table 4.1: Token length (raw and %) of strings: corpus, subcorpora, elements and attributes

The number of occurrences of pragmatic strings was also normalised based on the number of lectures in each subcorpus, as shown in Table 4.2.

		UK (30 lectures)		MS (18 lectures)		NZ (28 lectures)		ELC (76 lectures)	
element	*attribute*	*raw*	*per lecture*	*raw*	*per lecture*	*raw*	*per lecture*	*raw*	*per lecture*
humour		354	11.80	148	8.22	128	4.57	630	8.29
	bawdy	11	0.37	0.00	0.00	11	0.39	22	0.29
	black	8	0.27	12	0.67	3	0.11	23	0.30
	disparaging	72	2.40	9	0.50	15	0.54	96	1.26
	irony/sarcasm	73	2.43	11	0.61	11	0.39	95	1.25
	joke	4	0.13	0	0.00	0	0.00	4	0.05
	playful	122	4.07	76	4.22	47	1.68	245	3.22
	self-deprecating	40	1.33	10	0.56	27	0.96	77	1.01
	teasing/mock-threat	15	0.50	10	0.56	5	0.18	30	0.39
	wordplay	9	0.30	20	1.11	9	0.32	38	0.50
story		65	2.17	48	2.67	40	1.43	153	2.01
	anecdote	16	0.53	2	0.11	9	0.32	27	0.36
	exemplum	15	0.50	15	0.83	2	0.07	32	0.42
	narrative	16	0.53	10	0.56	15	0.54	41	0.54
	recount	15	0.50	21	1.17	14	0.50	50	0.66
summary		487	16.23	337	18.72	422	15.07	1246	16.68
	review content of previous lecture	81	2.70	64	3.56	81	2.89	226	3.05
	review content of current lecture	110	3.67	112	6.22	106	3.79	328	4.56
	preview content of current lecture	216	7.20	117	6.50	141	5.04	474	6.25
	preview content of future lecture	80	2.67	44	2.44	94	3.36	218	2.82
all		906	30.20	533	29.61	590	21.07	2029	26.98

Table 4.2: Occurrence (raw and per lecture) of elements and attributes in the corpus and subcorpora

A Python script was written to extract both sets of data by looping through the annotated texts and calculating indices and occurrences per file based on variables of subcorpus, element and attribute (as described in section 3.6.3 on data visualisation techniques). Three sets of data were then written out to a tab delimited file: 1. raw number of tokens, 2. raw occurrence of annotated strings, 3. normalised number of tokens, and 4. normalised occurrence of annotated strings. These base data were then used to calculate a third related set of metrics: average length (tokens) per occurrence, as shown in Table 4.3.

		UK	MS	NZ	ELC
element	*attribute*	*average token length per annotation indices*			
humour		323	164	219	235
	bawdy	13	0	12	13
	black	30	57	30	39
	disparaging	33	21	34	29
	irony/sarcasm	25	15	29	23
	joke	131	0	0	131
	playful	19	18	20	19
	self-deprecating	22	25	33	26
	teasing/mock-threat	21	13	40	24
	wordplay	29	17	23	23
story		136	124	97	121
	anecdote	137	90	58	107
	exemplum	167	130	97	145
	narrative	147	145	122	137
	recount	92	112	94	101
summary		38	35	55	42
	review content of previous lecture	47	47	69	54
	review content of current lecture	40	34	53	42
	preview content of current lecture	34	27	53	38
	preview content of future lecture	33	40	49	41
all		166	108	124	132

Table 4.3: Average length (tokens) per annotation indices for the corpus and subcorpora

All sets of data for elements and attributes across subcorpora and on average across the corpus (including normalised token count, normalised occurrence per lecture, and average length per indices) inform the quantitative then qualitative analyses in Chapter 5, Chapter 6, and Chapter 7.

4.4. Lexical variation: STTR

Table 4.4 compares text annotated within each of the three pragmatic elements, their inverse texts, the three elements combined (all annotated), and all text not contained within any of the pragmatic annotation (all non-annotated) using an n value of 1000 (as described in 3.6.2).

The lower the STTR, the lower the lexical diversity of the text, which can be an indicator of simplicity or specificity. Low lexical diversity tends to result from the production of unrehearsed, spontaneous text (such as real-time speech), whereas higher STTR is

associated with more diverse forms (such as planned writing). In general, lexical diversity is lower in speech than in writing (Halliday 1985).

	STTR
humour	39.14
non-humour	29.63
story	36.91
non-story	29.66
summary	31.54
non-summary	29.86
all annotated	34.34
all non-annotated	29.90

Table 4.4: STTR of humour, story, summary, non-humour, non-story, non-summary, all annotated, and all non-annotated text

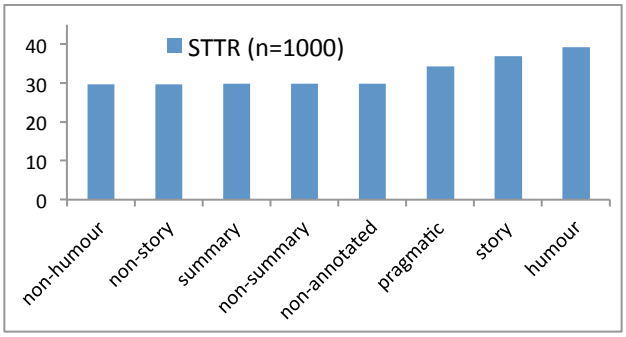


Figure 4.2: STTR of humour, story, summary, non-humour, non-story, non-summary, all annotated, and all non-annotated text. Arranged in ascending order

Perhaps unexpectedly, the greatest lexical diversity (highest STTR) occurs in the text that is identified as serving one of the pragmatic functions discussed; humour, story, and summary have a higher STTR than their inverse counterparts. The STTR of all text annotated as pragmatic is 34.34 whereas the remaining non-annotated text is 29.90. Summary and non-summary have the closest STTR (31.54 and 29.86), whereas humour and non-humour have the greatest range of lexical diversity (29.63 and 39.14) (Table 4.4).

Summary particularly was expected to display a comparatively low level of lexical diversity as formulaic language patterns were predicted. The (slightly) lower STTR of non-summary compared to summary may reflect the more technical nature of the lexis used in this discourse that primarily functions to deliver content, as indicated by the keyword analysis in Table 4.5. It may be that in the non-summative majority of the lecture, alternatives to the lexis required are not readily available, and paraphrasing might compromise accuracy. Additionally, certain concepts may need to be frequently repeated. The more extreme

reversal of expected patterns shown in the results for humour and story (compared to the majority of the rest of the lecture) may indicate that when using these types of pragmatic language, the lecturer is relying more heavily on rehearsed (perhaps scripted) – rather than spontaneous – discourse, more so than when summarising.

4.5. Salient lexis: keyword analyses

If pragmatic language performs a particular function within lectures, it is likely that it will be characterised by particular lexis. To test this, keyword analyses were run comparing all tokens within annotated strings against a reference corpus of the inverse non-annotated text, and vice versa (as described in 3.6.2).

Table 4.5 shows the 25 tokens that emerged as most key to the summative language (positive keywords) and the 25 tokens that were most key in the non-summative reference corpus (negative keywords).

Perhaps the most apparent difference between summative and non-summative text revealed by keyword analysis relates to the use of pronouns. First person pronouns are especially important because they give insights into the relationship between participants and the interactive nature of lecture discourse (Dafouz, Núñez and Sancho 2007, Fortanet 2004, Plaza and Álvarez 2013). The use of *we* in summative language has the highest keyness value of all lexical items. In non-summative language, the second and third person plurals *they* and *you* have a higher keyness measure. When summarising, however, the more inclusive first person plural *we* is employed. Noticeably absent from either list is the first person singular *I*, which indicates that lecturers do not insert themselves into the discourse with any more significance when summarising compared to any other part of the lecture.

The keyword analysis also shows that temporal deixis (in the form of references to *week*, *today*, *next*, *later*, *tomorrow*, *now*) is salient in summative language, which implies that an

important aspect of summaries is situating reviewed or previewed information chronologically.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
we	1727	847.145	it	869	119.536
going	533	283.517	hundred	18	92.541
week	163	265.846	t	182	76.075
last	152	239.999	got	129	70.317
look	276	225.824	twenty	18	65.109
today	96	182.820	they	112	62.290
lecture	96	169.543	s	742	61.245
ll	255	162.938	times	31	59.188
remember	125	156.545	five	74	50.843
at	510	133.260	if	290	48.879
next	127	123.528	thirty	5	45.992
chapter	56	104.349	ten	26	45.752
looked	50	102.252	here	160	44.806
re	456	100.025	you	1352	39.164
cycle	62	89.756	my	30	37.241
later	53	87.749	top	7	35.108
mentioned	33	81.821	metres	4	34.943
talked	32	79.885	project	5	34.840
processes	34	79.179	fifty	11	31.576
reversible	55	75.060	one	348	30.415
to	1728	74.348	forty	6	29.871
about	274	68.220	down	37	29.258
tomorrow	30	64.318	because	122	29.148
now	288	62.041	so	691	29.002
as	336	60.105	zero	43	28.937

Table 4.5: 25 most highly ranked positive and negative keywords in summary

The importance of following the syllabus is highlighted by the positive keyness of *chapter*, which is more key in summaries than references to general temporality (for example, “later”) (Table 4.5). The ELC lecturers often refer to textbooks to contextualise previously given or upcoming information. A raw count of references to *chapter* shows that it does not occur in UK summaries, occurs only twice (which equates to 86 times per million words (pmw)) in the New Zealand summaries, and is used 57 times (4872 pmw) in the Malaysian summaries. In the Malaysian lectures, textbooks form the basis of the syllabus, which is rigidly followed from week to week, chapter by chapter.

Summative speech is not strongly characterised by numerical references, unlike the non-summative speech, in which nine of the 25 most negatively key items are numbers (Table

4.5). The type of language that typifies other parts of the lecture, in formulae, calculations, and workings out is noticeably absent when lecturers are reviewing or previewing content. Even though summaries function to review and preview important lecture content, the condensed form does not privilege specific numerical detail. Discussion of the lexis of summary types takes place in 5.3-5.3.4.

Table 4.6 shows the results of a keyword analysis of the umbrella category humour compared to a reference corpus of non-humour.

positive			**negative**		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
i	483	225.43	the	615	109.72
he	66	132.11	we	129	61.92
me	70	117.04	so	134	56.80
who	52	94.66	is	225	56.66
t	192	85.53	point	23	38.49
my	68	66.07	beam	4	37.63
no	70	59.14	concrete	2	32.23
exam	23	45.57	zero	3	31.54
hands	12	45.46	steel	4	24.54
cafe	7	43.48	area	1	23.92
today	27	42.24	value	4	23.71
yes	22	39.82	two	43	22.85
you	596	39.63	section	2	22.13
don	73	39.04	voltage	1	21.53
thank	17	38.30	between	1	21.37
people	32	35.11	and	280	21.24
him	11	34.88	equal	1	21.09
m	85	33.92	x	3	20.74
was	57	33.73	different	1	20.71
not	110	33.72	by	15	20.22
waiting	7	32.05	this	143	19.33
know	88	31.30	four	11	19.17
playing	6	31.27	minus	4	18.46
hair	8	30.52	moment	8	18.27
apple	5	30.11	plus	2	17.97

Table 4.6: 25 most highly ranked positive and negative keywords in humour

The most positive and negative keyword entries in Table 4.6 also highlight the importance of pronouns in the language of humour. *I* and *he* (and to a lesser extent *you* and *him*) figure prominently, whilst in non-humorous language *we* is more salient. The use of *she/her* does not figure highly in the ELC as a whole; there are 16 references in non-humour, and only

four in humour (which occur during a description of a woman falling on a catwalk in a health and safety lecture, 2010). The language of the nine types of humour identified is examined in 6.5.

The most salient lexis in text annotated as performing a story function is given in Table 4.7.

positive keywords			**negative keywords**		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
was	182	347.436	point	12	114.670
they	281	298.447	you	442	70.059
he	101	227.073	beam	5	55.472
er	178	149.975	value	1	55.326
bridge	33	140.645	five	14	52.871
years	48	139.338	two	47	52.525
were	55	101.347	stress	2	48.025
had	61	96.992	three	17	46.175
lego	19	89.374	moment	5	45.017
workers	20	84.750	going	42	44.638
station	14	61.368	is	362	41.797
accident	14	59.504	four	9	41.662
used	42	59.060	we	226	40.054
said	41	56.542	x	2	38.785
ago	20	56.362	times	7	37.929
guy	20	54.153	force	11	36.029
malaysia	16	49.573	one	110	35.819
building	42	48.050	six	4	35.421
contractor	24	47.373	section	2	34.865
um	206	46.936	plus	1	31.844
london	10	45.280	ll	13	31.445
built	19	45.043	minus	4	30.480
aircraft	7	44.771	f	1	27.816
happen	29	46.012	zero	9	27.754
been	37	45.543	eight	3	26.403

Table 4.7: 25 most highly ranked positive and negative keywords in story

In stories, *they* and *he* are privileged over *you* and *we*. Concordance analysis shows that the favoured pronouns are largely used to describe people external to the university context. The keywords in stories also mark pastness (*was*, *were*, *ago*, *been*) and refer to time (*years*). Retold events involve people (*he*, *they*, *workers*, *guy*, *contractor*), places (*Malaysia*, *London*) and large structures (*bridge, station*). Type-specific patterns are discussed in 7.4.1-7.4.4.

Table 4.8 shows which words are positively and negatively key when all pragmatic text (a combination of humour, summary and story) is compared against a reference corpus of non-annotated text (that is, the ELC minus humour, summary and story).

positive keywords			**negative keywords**		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
we	2082	350.968	point	194	137.183
was	332	240.196	five	109	118.812
last	184	202.956	times	49	113.383
week	185	185.548	twenty	39	91.221
today	123	181.605	hundred	57	89.800
lecture	124	178.633	one	543	87.157
he	183	174.702	zero	55	85.275
going	644	129.587	so	1114	80.795
look	317	114.295	four	84	76.460
remember	149	113.821	s	1322	73.725
ll	327	109.488	got	263	71.018
year	83	103.198	is	1738	71.003
had	136	102.599	here	285	66.393
next	158	92.172	beam	88	66.044
later	68	89.688	minus	43	61.323
were	109	87.212	ten	55	58.137
er	413	86.656	two	337	55.590
ago	44	83.729	six	54	55.427
years	71	82.020	if	536	54.923
looked	54	71.540	equal	28	54.465
said	100	70.960	moment	78	53.509
i	1621	68.244	metres	9	52.870
things	168	64.059	seven	48	52.448
did	114	63.703	load	52	52.190
bridge	36	63.630	three	160	48.672

Table 4.8: 25 most highly ranked positive and negative keywords in all pragmatic text

As in the element-specific analyses, the use of pronouns emerges as an important difference. In this case, *we*, *he*, and *i* are more salient in pragmatic text, whereas no pronouns rank in the top 25 key negatively key items. Temporal deixis (*week*, *today*, *year*, and *later*), reiteration (*looked*, *said*, and *did*) and other markers of pastness (*was*, *were*, and *ago*) are also key to the pragmatic text analysed. Looking at the relevant concordance lines, reiteration of information in the form of *I*/*we*/*you* + *looked*/*said*/*did* is also a feature of non-annotated text.

Non-annotated text is highly characterised by numerical references (such as *five*, *twenty*, and *hundred*) and technical terms (such as *beam*, *load*, and *moment*). It appears that during

these strings, lecturers focus on calculations and working through engineering concepts. The higher overall lexical diversity of pragmatic text (see Table 4.4 and Figure 4.2) supports the idea that in the parts of the lecture in which content delivery is key, more technical lexis within a comparatively narrow range is used.

Analysis of the words that are most salient and formulaic sequences in the annotated strings compared to the non-annotated strings shows that a general shift in lexis occurs when lecturers tell stories, use humour, or summarise information. This broad-stroke analysis immediately highlights a number of differences in lexical choice, which shape the analyses in Chapter 5, Chapter 6, and Chapter 7.

4.6. Lexical sequences: n-grams

Table 4.9 shows the number of different types of 4-grams (pmw) calculated (as described in 3.6.2) for humour, story, summary, all annotated, and non-annotated text, and Table 4.10 shows the total instances of these 4-grams across the same categories. Figure 4.3 renders the results of Table 4.10.

	0-49	50-99	100-149	150-199	200-249	250-299	300-349	350-399	400-449	450-499	500+	total
humour	0	0	342	33	0	14	3	3	0	1	4	400
story	0	0	312	35	14	3	3	2	1	1	0	371
summary	1975	550	53	33	16	4	4	5	0	3	8	2643
all annotated	1975	550	707	101	30	21	10	10	1	5	12	3414
all non-annotated	0	0	342	33	0	14	3	3	0	1	4	400

Table 4.9: Discreet types (pmw) of 4-grams in humour, story, summary, all annotated, and all non-annotated text categorised by number of occurrences

	0-49	50-99	100-149	150-199	200-249	250-299	300-349	350-399	400-449	450-499	500+	total
humour	0	0	44914	6501	0	3677	985	1182	0	460	3480	61199
story	0	0	34252	5764	3074	823	988	768	439	494	0	46602
summary	74208	35789	6481	5411	3269	1071	1277	1897	0	1390	9938	140731
all annotated	74208	35789	85647	17676	6343	5571	3250	3847	439	2344	13418	248532
all non-annotated	207413	10211	2709	865	1171	537	0	990	0	908	1015	225819

Table 4.10: Total instances (pmw) of 4-grams in humour, story, summary, all annotated, and all non-annotated text

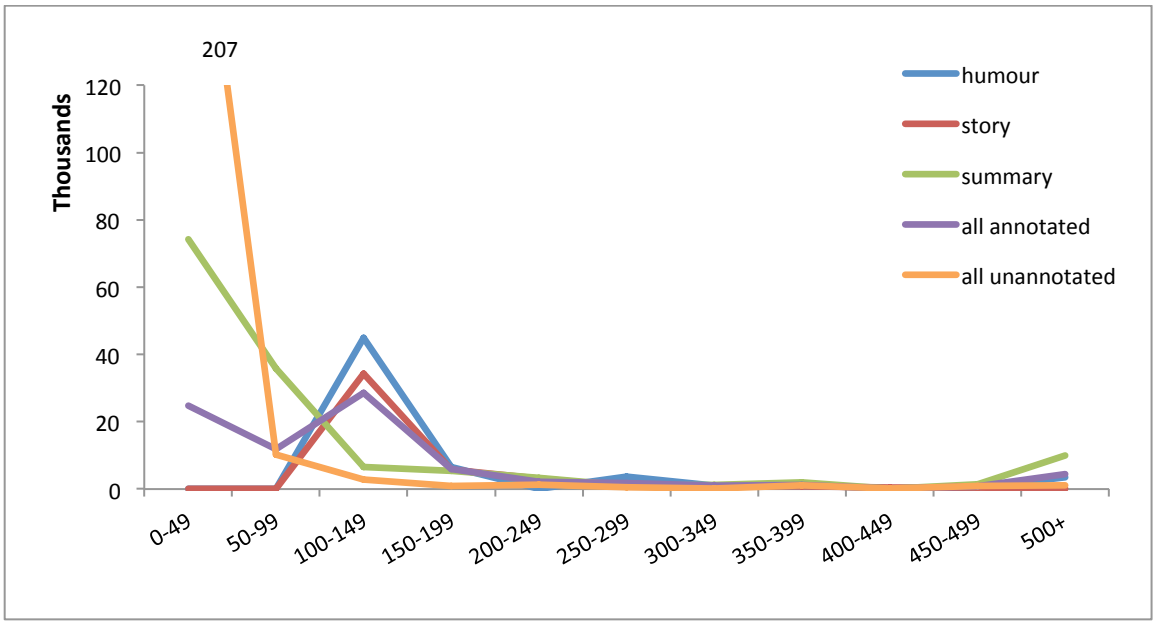


Figure 4.3: 4-grams (pmw) in humour, story, summary, all annotated, and all non-annotated text

Looking at types, there are a similar total number of types of 4-grams in humour (400), story (371), and non-annotated language, and around eight times as many in summary (2643) (Table 4.9). The highest results by far are in the 0-49, 50-99 and 100-149 brackets. Figure 4.3 clearly illustrates that the distribution of the total occurrence of the 4-grams is similarly low across all categories in the 150+ (pmw) bracket. Stark differences are evident in the lower brackets of (pmw) frequency. No 4-grams occur in humour and story with a frequency of less than 100 pmw, although both categories register very similar (and the highest of all four types) results for the 100-149 pmw bracket. A wide range of 4-gram types are relatively infrequent in both non-annotated and pragmatic text, especially in summaries.

The calculation of (pmw) frequency of each of these types (Table 4.10) augments this finding, this time showing that although pragmatic and non-annotated texts overall return roughly the same frequency (pmw) of 4-grams, there is a particular concentration of these 4-grams at the lowest end of occurrence (0-49 bracket) within non-annotated text. The combined results suggest that non-annotated - and to an extent summative - text relies

more heavily on formulaic sequences, but that the range of sequences used is extremely wide. The result tallies with the findings of the STTR tests described in 4.4.

The patterns revealed at the higher end of pmw 4-gram type occurrence also indicate clear differences that are somewhat occluded in Table 4.9. Table 4.11 shows that there is variation across categories in both the content and frequency of 4-grams that occur (pmw) more than 150 times.

humour			**story**			**summary**			**non-annotated**		
4-gram	*freq*	*pmw*	*4-gram*	*freq*	*pmw*	*4-gram*	*freq*	*pmw*	*4-gram*	*freq*	*pmw*
i m going to	19	1248	you can see here	9	494	we re going to	208	3908	i m going to	227	514
i don t know	16	1051	what we call it	8	439	i m going to	93	1747	you ve got to	221	501
those of you who	10	657	as what you can	7	384	re going to do	59	1108	we re going to	201	455
you re going to	8	525	what you can see	7	384	what we re going	39	733	you re going to	200	453
get the right answer	7	460	this is what happen	6	329	going to look at	36	676	it s going to	157	356
it s going to	6	394	we call it er	6	329	re going to look	34	639	you ve got a	142	322
m going to ask	6	394	what we can see	6	329	you re going to	33	620	s going to be	138	313
we re going to	6	394	as what we can	5	274	we ll look at	27	507	if you ve got	122	276
i m not going	5	328	can see here this	5	274	have a look at	26	488	times ten to the	115	260
if you ve got	5	328	in the middle of	5	274	we are going to	24	451	ten to the minus	108	245
m not going to	5	328	all over the place	4	220	when we look at	24	451	if you look at	104	236
gave me the four	4	263	i don t know	4	220	m going to do	21	395	if you want to	104	236
if i want to	4	263	i m going to	4	220	re going to be	21	395	is going to be	104	236
if you don t	4	263	if you ve got	4	220	going to do is	20	376	have a look at	97	220
in it at the	4	263	if you want to	4	220	what i m going	20	376	i don t know	87	197
it at the time	4	263	it s a very	4	220	that s what we	19	357	we ve got a	77	174
put the force on	4	263	on top of the	4	220	to be able to	18	338	so you ve got	76	172
re going to do	4	263	the end of the	4	220	if you want to	17	319	you ve got the	74	168
so i m going	4	263	the rest of the	4	220	it s going to	17	319	you don t have	68	154
t in it at	4	263	the weight of the	4	220	you ve got to	16	301			
thank you	4	263	this is a very	4	220	let s have a	15	282			

for playing											
we weren t in	4	263	um it s a	4	220	and we re going	14	263			
weren t in it	4	263	we can see here	4	220	let s look at	14	263			
you gave me the	4	263	you can see that	4	220	we look at the	14	263			
you ve got a	4	263	a look at the	3	165	ll come back to	13	244			
a bit of a	3	197	all sorts of problems	3	165	to look at the	13	244			
a look at the	3	197	and so on and	3	165	and that s what	12	225			
an apple a day	3	197	and there s a	3	165	is we re going	12	225			
and you re riding	3	197	and things like that	3	165	s have a look	12	225			
at the end of	3	197	as you can see	3	165	second moment of area	12	225			
do a bit of	3	197	at half past eleven	3	165	the way in which	12	225			
don t know i	3	197	at the end of	3	165	a little bit of	11	207			
don t know if	3	197	bridge ok this is	3	165	a piece of material	11	207			
go wrong uh go	3	197	can see here the	3	165	i m not going	11	207			
had a lecture at	3	197	can see that the	3	165	if you don t	11	207			
have a look at	3	197	don t don t	3	165	m not going to	11	207			
i can see that	3	197	don t have to	3	165	re going to get	11	207			
i want you to	3	197	i m not mistaken	3	165	re going to use	11	207			
if it is negative	3	197	i was on we	3	165	so that s what	11	207			
if you want to	3	197	if i m not	3	165	we ll come back	11	207			
it s not a	3	197	if you read the	3	165	be able to do	10	188			
not going to get	3	197	in the u s	3	165	going to be doing	10	188			
on your t shirt	3	197	it is it s	3	165	i mentioned just now	10	188			
quick show of hands	3	197	it s going to	3	165	need to be able	10	188			
s going to be	3	197	it s not just	3	165	so we re going	10	188			
t get the right	3	197	might be made in	3	165	that we re going	10	188			
take moment about q	3	197	not even one year	3	165	and i m going	9	169			
that s how you	3	197	ok this is the	3	165	and then we ll	9	169			
that you don t	3	197	see here this is	3	165	going to go through	9	169			
the end of the	3	197	so that s why	3	165	if you have a	9	169			
times ten to	3	197	the area of	3	165	is going to be	9	169			

the			the								
to be an interesting	3	197	the middle of the	3	165	so that s the	9	169			
u i t m	3	197	this accident occur in	3	165	that s what you	9	169			
we had a lecture	3	197	this type of scaffolding	3	165	the other one is	9	169			
wrong uh go wrong	3	197	to do with that	3	165	then we re going	9	169			
you don t know	3	197	ve got all the	3	165	we re looking at	9	169			
you know what they	3	197	weight of the bridge	3	165	we will look at	9	169			
you ve got to	3	197	yeah so that s	3	165	what we ve done	9	169			
			you can t make	3	165	a little bit more	8	150			
						as i told you	8	150			
						get you to do	8	150			
						going to do now	8	150			
						not going to go	8	150			
						now we re going	8	150			
						re going to go	8	150			
						re going to have	8	150			
						that s what the	8	150			
						the end of the	8	150			
						to do today is	8	150			
						we did last week	8	150			
						we ve looked at	8	150			
						we were looking at	8	150			
						you ve got the	8	150			

Table 4.11: 4-grams (150+ pmw) in humour, story, summary, and non-annotated text

In the 150+ pmw range, non-annotated text has the lowest number of entries (19), compared to humour (58), story (59), and summary (65) (Table 4.11). This confirms that although sequences are regularly used in non-annotated text, their content is varied. In other words, their distribution is long and thin. However, summaries, which also had a high number of sequences at the lower end of pmw occurrence, contain a high number of sequences at the higher end of pmw occurrence.

Figure 4.4 compares all 4-grams that occur over 300 times pmw across elements, all annotated text, and all non-annotated text.

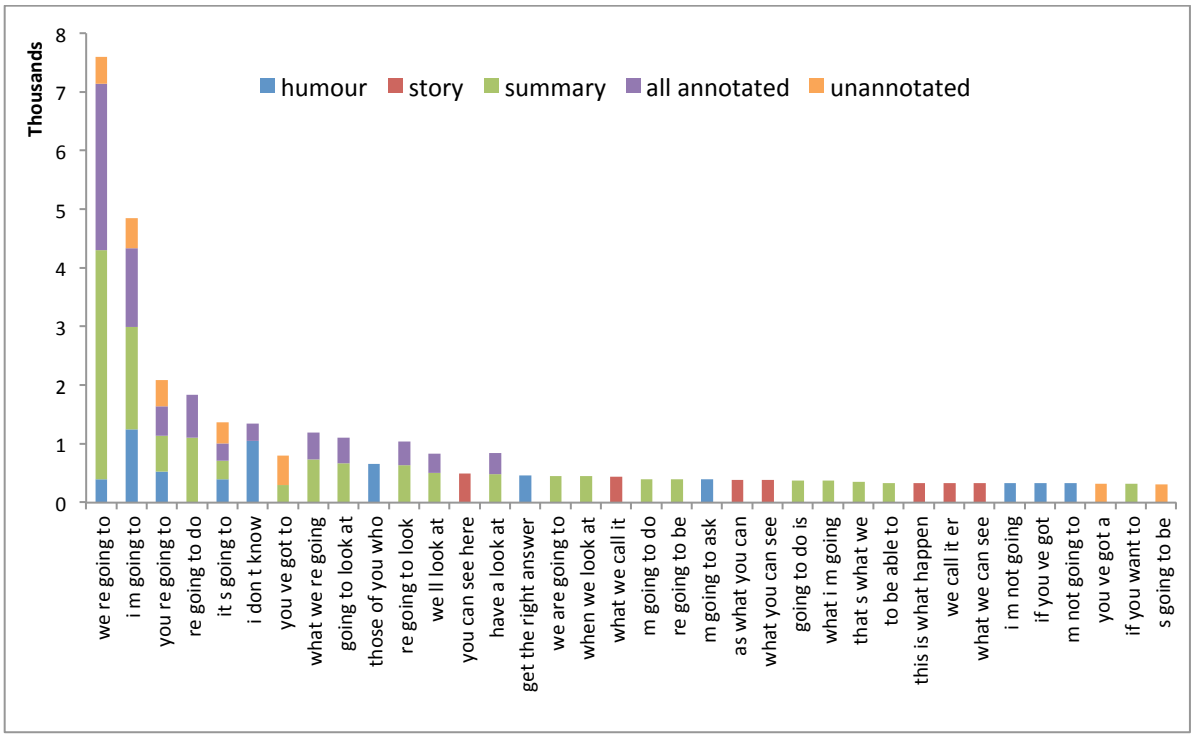


Figure 4.4: 4-grams (occurrence 300+ pmw) in humour, story, summary, all annotated, and all non-annotated text

The most frequent 4-gram in summary (*we re going to*) occurs 3908 times pmw, compared to the most frequent 4-gram in non-annotated text (*i m going to*), which occurs 514 time pmw (Table 4.11, Figure 4.4). The range of 4-grams in summaries is relatively long and thin, but becomes fat at the higher end of occurrence, relying very heavily on a few sequences. With relatively few 4-grams of low occurrence, both humour and story are shorter in range. Humour particularly becomes fatter at the higher end of pmw occurrence, as its most frequent n-gram (*i m going to*) occurs 1248 times pmw. Across the categories, the 4-gram results are dominated by the sequence: pronoun + (*be*) + going to.

4.7. Conclusion

The macro-level results show that lecturers commonly use summaries, stories and humour in the engineering lectures analysed, which suggests that these pragmatic features warrant

further investigation. Summarising, for example, certainly seems to deserve its often anecdotally presented position as recommended practice (cf. HEA 2014). In terms of EAP teaching (such as when giving note-taking instruction), the results also indicate that the features discussed do not occur in lectures in predictable positions.

The annotation of the ELC, for example, undermines the possibility of applying the type of Lecture Introduction Framework identified by Thompson (1994), which has been widely developed (for example, Lee 2009, Shamsudin and Ebrahimi 2013, Yaakob 2013, Yeo and Ting 2014). This model seems to assume a *beginning*, *middle* and *end* structure where each section contains noticeably different types of discourse function. For instance, the introduction is the "preliminary part before the lecturer embarks on a new topic or subtopic for the lecture proper" (Yeo and Ting 2014: 28), the end of which is marked by the first presentation of new information (Shamsudin and Ebrahimi 2013, Yeo and Ting 2014). ELC findings do not accord with a model in which *preview* type summaries would occur at the start and *review* type summaries at the end. Both types occur throughout the ELC lectures, with little evidence of significant patterns of clustering.

So far, the element-level findings instead support Young's (1994) phasal model in which preview, conclusion and evaluation phases are interspersed with theory, example and interaction phases discontinuously, throughout the lecture. Figure 4.5 displays the overall percentage (tokens) assigned to the three elements discussed in this thesis split between the first 10% and remaining 90% of the lectures. Figure 4.6 displays the same split, this time based on a count of instances of occurrence (strings) across the corpus.

The occurrence of pragmatic features across the whole lecture shows that although elements associated with the Lecture Introduction Framework – such as previewing upcoming information - do occur more frequently in the first part than in the remainder of the lecture, pragmatic features occurring in the first 10% also occur in other parts of the lecture.

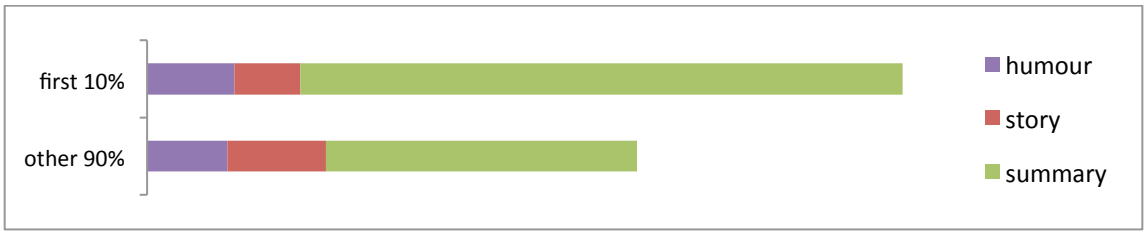


Figure 4.5: Elements (tokens) in the first 10% and following 90% of the lecture

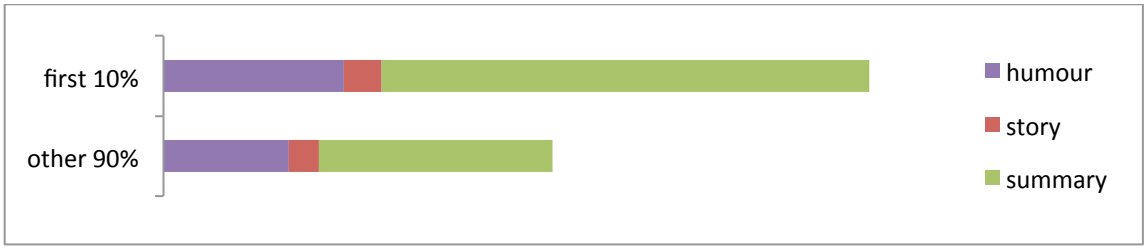


Figure 4.6: Elements (strings pmw) in the first 10% and following 90% of the lecture

The breakdowns indicate that, for example, the explanation of new concepts does not necessarily mark the end of an introduction and the beginning of the *lecture proper,* as research that identifies a distinct introductory phase suggests. Thus it may not be true that lecturers should deliver certain types of pragmatic features at particular points in the lecture, in order to optimise student comprehension. Further investigation at the attribute level is necessary to better understand the role of these pragmatic elements in lecture discourse.

This data overview has looked at macro, element-level patterns in pragmatic text in terms of the occurrence and duration of strings, lexical diversity, lexical salience, and lexical co-occurrence. Analysis included a comparison of pragmatic text with non-annotated text within the corpus. The patterns identified guide and are further interrogated at the level attributes in Chapter 5, Chapter 6, and Chapter 7.

CHAPTER 5. SUMMARY

5.1. Introduction

A summary occurs in a lecture when topic content (not including housekeeping issues) is reviewed or previewed by the lecturer. Reviews contain restated information or reminders that information has previously been given, and previews look forward to upcoming content, whether in the current lecture or in future lectures.

The purpose of this chapter is to describe and analyse the linguistic features of summarising in the ELC. The terminology used to classify functional categories related to summarising in lecture discourse is not entirely consistent with the terminology used in the prior research. However, echoes of the language and/or criteria used in this thesis to describe summative functions in lectures can be identified in various other taxonomies.

Chapter 2 charted a range of linguistic features in lecture discourse, including shell nouns (Hunston and Francis 2000, Nesi and Moreton 2012), pseudo-clefts (Deroey 2012), enumeration (Tadros 1985), and deixis (Lyons 1977), all of which can function as summative devices at the micro-level. Certain expressions were also identified as predictive of upcoming content, such as “I’ll indicate one in just a minute” (Olsen and Huckin 1990: 37), "what I'm going to talk about today" (Chaudron and Richards 1986: 177), “today we’re going to talk a little bit about” and “[f]irst let’s take a look at” (Crawford Camiciottoli 2004: 40). These microstructural features have been a main focus of research into lecture discourse but are not normally discussed in relation to lecture macrostructure.

Different types of summation have been associated with different phases of the lecture, however. Previewing is often associated with the beginning, or introductory, phase, because it serves to put the topic into context (Lee 2009, Thompson 1994, Yaakob 2013). The *introductory roadmap* in the long inventory of MICASE pragmatic features is defined as “at least two or more statements or phrases outlining or announcing the topics or course of the

class or events" (Maynard and Leicher 2007: 109-110). This use of summary is in accord with the common pedagogic advice that an outline should be delivered at the beginning of a lecture along with a metadiscursive signal, as in "[n]ow I am going to talk about [...]" (HEA 2014). Previewing upcoming content, then, can be tied, but not limited to, introductions.

Likewise Straker Cook's *focusing moves* "lay the ground for extensions […] and generally point the way forward", and his *concluding moves* "summarise and round off the episodes and are capable both of recapitulating the ground covered and of pointing the way forward" (1975: 70). Straker Cook describes summarising moves as "giving a résumé of the immediately preceding discourse" (1975: 76). However, unlike in other models, the moves identified by Straker Cook (1975) as having a summarising function can occur at any point in the lecture; focusing moves, for example, are not bound moves like the introductory roadmap identified by Maynard and Leicher (2007). Straker Cook's concluding moves serve the same purpose as summary, recapitulation, and comment in written text, as defined by Tadros (1985), but are less constrained in terms of their position in the text.

The three metadiscoursal phases described by Young (1994) (as outlined in 2.4.2) resemble in part Straker Cook's (1975) concluding and focusing moves. The first of Young's phases, discourse structuring, is:

> […] marked by processes of verbalization, with participant roles realized by first and second person pronouns, by rhetorical questions alternating with statements and imperatives, and by a type of modality that indicates the purpose of the phase, to announce future directions, namely that of intentionality and prediction. (1994: 172)

In the discourse structuring phase the lecturer indicates upcoming content, letting the audience know which direction the discourse is going to take. Young (1994: 168) gives the examples of "[so] what I will do now first is to give you some description […]" and "[l]et me give an example from Belgium". There groups are assigned to the category of verbalisation.

The second metadiscoursal phase identified by Young is *conclusion*, which is characterised by relational processes signalled by the verbal group *is*/*are*, as in "[s]o there are different

ways of expressing that soil water content" (1994: 171). In conclusion phases key terms are repeated with morphological variation, forming a chain of related items to ensure that students are aware of the most important terms and concepts in the lecture (1994: 172).

In the third metadiscoursal phase, *evaluation*, lecturers emphasise concepts and approaches by passing judgment on them, ensuring that students know which to adopt and, by implication, which to reject (1994: 172-173). Young's examples include: "[t]he larger the code the more efficient the code can be" (1994: 171). Evaluation is characterised by attributive relational processes rather than attitudinal elements such as modals. As in Straker Cook's analysis, Young's phases can "recur discontinuously" throughout the lecture (1994: 165).

In largely monologic lectures, summative structures are considered to have both an information-shaping function and an information-management function. Summary offers huge potential for condensing and organising information, but its various linguistic structures can pose challenges in processing and delivery.

5.2. Macro-level patterns in summary

Four types of summative discourse units were identified within the ELC and attributed to the umbrella element summary: 1. review of previous lecture content, 2. review of current lecture content, 3. preview of current lecture content, and 4. preview of future lecture content.

Lecturers commonly look backwards to material covered both previously on the course, and in the current lecture, as in:

> now if you were here last week I was saying that delta rosettes actually do come in three forms (1029)
>
> I just covered the two things one is the bond the other one is th- the work insurance which has to be purchased by the contractor (2007)

Some reviews restate information, as in:

> so remember I told you yesterday that if you have a current flow you always get magnetic fields there's no exception (3011)

Other reviews function as reminders that information has been given, as in:

> we've done axial stress both in materials and we did a little bit of it last term (1008)

All discourse that reiterates previously given information (whether in full or in grammatically encapsulated form) is classed as informative, and thus a review of lecture content.

Lecturers also look forward within the lecture they are giving and forwards to future lectures, as in:

> next week I intend to move on to looking at columns new work completely (1019)

> what we're going to do today is figure out how strong the field is at a point some distance away from various objects (3002)

There is no upper or lower limit on the amount of information needed to constitute review or preview summary types in the ELC taxonomy, provided that the string of text makes sense as a standalone unit (see 3.4.6).

A general discussion of summary within the ELC lectures from a quantitative and qualitative perspective will be followed by more detailed discussion of the four attributed types and variation/similarity within lectures in the three subcorpora.

5.2.1. Occurrence and duration of summary types

When the data are overviewed, one of the most remarkable aspects of summarising in the ELC is the frequency and duration of its occurrence. The visualisation in Figure 5.1 shows that summaries occur multiple times in all lectures, and their occurrence is not limited to a particular part of the lecture.

Figure 5.1: Occurrence and duration of all types of summary

The breakdown of tokens given as a normalised percentage in Table 5.1 shows that summarising constitutes just over ten per cent of all lecture discourse, with some variation across types and subcorpora.

The classification of the occurrence of summary by duration (token count) in Table 5.1 shows that previewing current lecture content is the most common form across the corpus (3.4%) and previewing future lecture content occurs for the least overall duration (1.70%), with reviews of previous content (2.35%) and reviews of current content (2.61%) lying in between. The same pattern emerges by a measure of instances of summary per lecture, as show in Table 5.2. Whether looking at the frequency or duration of summaries, previews of

current lecture content are the most common form of summarising in the ELC, and the most common form in each of the three subcorpora, except by a measure of token duration. The lecturers from MS dedicate more tokens, but not more individual instances, to reviewing the content of the current lecture.

	UK (251108 tokens)		**MS** (120211 tokens)		**NZ** (15683 tokens)		**all** (528157 tokens)	
	tokens	%	tokens	%	token	%	tokens	%
review content of previous lecture	3807	1.52	2996	2.49	5622	3.58	12425	2.35
review content of current lecture	4443	1.77	3777	3.14	5568	3.55	13788	2.61
preview content of current lecture	7404	2.95	3137	2.61	7430	4.74	17971	3.40
preview content of future lecture	2664	1.06	1746	1.45	4592	2.93	9002	1.70
all summary types	18318	7.29	11699	9.73	23212	14.80	53229	10.08

Table 5.1: Breakdown (raw tokens and %) of summary types across subcorpora

	UK (30 lectures)		**MS** (18 lectures)		**NZ** (28 lectures)		**all** (76 lectures)
	raw	*per lecture*	*raw*	*per lecture*	*raw*	*per lecture*	*per lecture*
review content of previous lecture	81	2.70	64	3.56	81	2.89	3.05
review content of current lecture	110	3.67	112	6.22	106	3.79	4.56
preview content of current lecture	216	7.20	117	6.50	141	5.04	6.25
preview content of future lecture	80	2.67	44	2.44	94	3.36	2.82
all summary types	487	16.23	337	18.72	422	15.07	16.68

Table 5.2: Instances (raw and per lecture) of summary across types and subcorpora

At the level of subcorpora, the token duration of summarising is largest in the NZ lectures (14.8%), smallest in the UK lectures (7.29%) and midway in the MS lectures (9.73%) (Table 5.1). When broken down by average instances per lecture, however, the pattern is reordered. There are on average 18.72 instances of summarising per lecture in the MS subcorpus, 16.23 in the UK subcorpus, and 15.07 in the NZ subcorpus (Table 5.2). The quantitative data show that on average instances of summary occur most frequently but with middling token duration in the MS lectures. Instances of summary in the NZ lectures take up the most token space, yet occur with the least frequency; lecturers give summaries less often, but when they do each summary is on average much longer than in the other

subcorpora. In the UK subcorpus, lecturers do not dedicate as many tokens to summarising, and they use this function less than their MS, but more than their NZ, counterparts.

Calculation of the average length (tokens) of each instance of summary gives a slightly different picture than looking either at subcorpora in terms of percentage duration (tokens) or average instances per lecture (as shown in Table 5.1 and Table 5.2 respectively). As Table 5.3 shows, reviewing the content of previous lectures emerges as the longest form of summarising in terms of tokens in each of the subcorpora, followed by reviewing the content of the current lecture. Overall, previews of future content are ranked third and previews of current content fourth, with some variation in position across the subcorpora.

	UK	**MS**	**NZ**	**all**
review content of previous lecture	47	47	69	54
review content of current lecture	40	34	53	42
preview content of current lecture	34	27	53	38
preview content of future lecture	33	40	49	41
all summary	38	35	55	42

Table 5.3: Average token counts per instance of summarising

Therefore reviews of previous and previews of future lecture content in general take up less space in the corpus overall (see Table 5.1) and occur with the least frequency per lecture (see Table 5.2), but when they do occur, particularly reviews of previous content, their duration is on average longer than reviews and previews of current lecture content (see Table 5.3). Previews of current lecture content take up the most space in the corpus (see Table 5.1) and occur by far the most frequently per lecture (see Table 5.2). Their average duration is, however, the shortest of all the types (see Table 5.3), indicating that in general lecturers preview current content often but not for very long. Reviews of current content follow almost the same pattern. They take up the second most amount of space in lectures and occur with the second most frequency per lecture. Their average duration, however, is second longest (not second shortest).

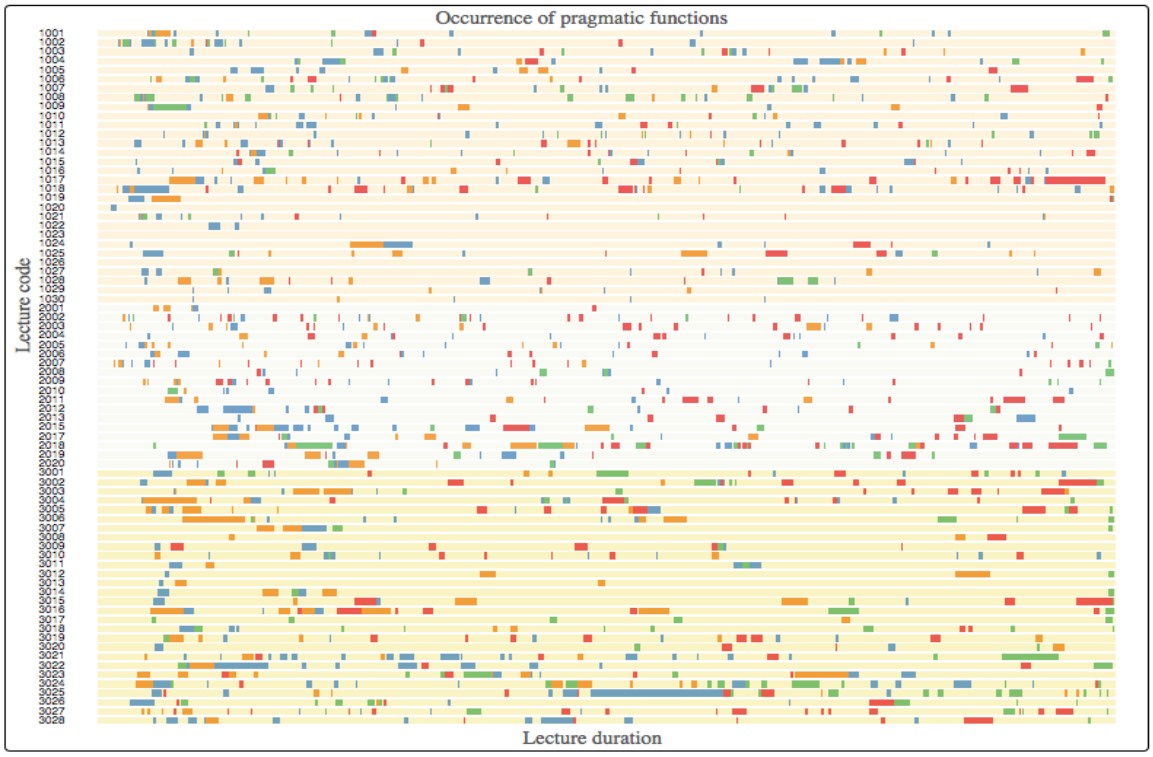
So far the qualitative analysis has shown three main patterns. Firstly, summaries have a substantive role to play in lecture discourse, of which they constitute about a tenth and occur roughly 17 times per lecture. Secondly, different types of summary occur with different frequency and duration within the corpus. Thirdly, variation in the occurrence and duration of summary exists at the level of the subcorpora. More detailed discussion of the occurrence and duration of the four types of summary across the corpus and with reference to subcorpora takes place in sections 5.3-5.3.4.

5.2.2. Co-occurrence of summary types

A visualisation of the occurrence and duration of the umbrella category summary in all ELC lectures was given in Figure 5.1. Figure 5.2 further distinguishes between the occurrence and duration of the four types attributed to the umbrella category.

The overview in Figure 5.1 points to two patterns for further discussion. Firstly, different types of summary tend to occur at different stages in the lecture. Such macrostructural patterns are analysed in sections 5.3-5.3.4. The other pattern that emerges when the data are visualised by type is a chaining phenomenon, whereby different types of (colour-encoded) summary co-occur seamlessly in a continuous stretch within the same lecture.

Figure 5.2 reveals that such chaining occurs throughout the lectures. As might be expected, there is a tendency for chains of reviews of previous content and previews of current content (orange and blue) to cluster around the beginning of the lectures. Reviews of current content and previews of future content (red and green) also predictably tend to cluster in approximately the final fifth of lectures, although to a lesser extent than introductory chaining.



KEY: review previous | review current | preview current | preview future

Figure 5.2: Occurrence and duration of summary types in the ELC

The occurrence of summary type chains indicates a relationship between old and new information that spans previous, current and future lectures. Many of the chains begin with a review, which can be conceptually likened to Tadros' (1985) understanding that recapitulation occurs when the V head predicts that there will be new information (but not its nature) by recalling information from elsewhere in the text. The text in this case is the current lecture and also preceding and proceeding lectures on the course.

This more complex chaining effect is evident at the beginning of lecture 3024, for example, where three summary types form five continuous links (review previous – preview current – preview future – preview current – preview future):

<summary type="review content of previous lecture">where we were last time was we had a look at an auxiliary loaded piece of material with a very simple load situation but what we found was that when we originally started looking at this auxiliary loaded piece of material it had a very simple set of stresses on it had a set of normal stresses acting on a plane cutting across the piece of material however after the previous lecture we now understand that if we look inside a piece of material at some arbitrary angle we are likely to find more than just normal stresses and in this case even though it's a very simple auxiliary loaded piece of material it has both the combination of normal stresses and shear stresses on a particular surface</summary><summary type="preview content of current lecture"> now what I'm going to do with you today is for the first ten minutes of this lecture we're going to look at general loading conditions we're going to look at the absolute worst case you're going to get some groundwork as to what the worst case looks like</summary><summary type="preview content of future lecture"> and over the next twelve weeks we're going to be working towards filling in the gaps as to how we understand how we actually work out the values of all these stresses</summary><summary type="preview content of current lecture"> so we're going to start with that we're then going to jump to another topic we're going to look at factors of safety and I'm going to do an example towards the end of the lecture ok it's bits and pieces at this stage </summary><summary type="preview content of future lecture"> next lecture on Wednesday we're actually going to get into a big piece of work but we're laying groundwork at the moment</summary> (3024)

In the context of summary chains, previews of future lectures seem to function somewhat like a D member in predictive structures. In the example above from lecture 3024, the already given information is contained in the review of a simple auxiliary load (previous content) and the review of a combined stress load situation (current lecture). These reviews anchor the new, upcoming information – the preview of worst-case general loading conditions (current content) and working out stresses (future content) and factors of safety (current content) – which all link to a related future piece of work (future content). In this case, the recapitulated information informs the new information.

The process by which recapitulated information predicts and cognitively anchors new (previewed) information is explicitly acknowledged by the lecturer in 3024. The reviewed information is characterised as “what we had a look at” and “what we now understand”, which enables “filling in the gaps as to how we understand how we actually work out” the reviewed theme of stress calculations. The culmination is the “big piece of work” for which

“we're laying groundwork at the moment”: the work that has been predicted by the recapitulated information.

The tendency for more chaining to occur at the beginning of lectures may indicate that the effect of summarising on cognitive processing is considered to be greater early in the lecture. However, barring the illogical (reviewing current lecture content at the very beginning, or previewing current content at the very end of lectures), concrete patterns of occurrence of summary types are not identifiable. Despite some clustering according to type, the most pertinent macro pattern shown by the visualisation is that summaries punctuate lectures throughout, which supports the idea of discursive cycles – or recurring phases (cf. Young 1994) – within lectures.

5.2.3. Macro-level language patterns in summary types

The overview of instances of summary at the macro-level in 5.2.2 revealed patterns for further investigation at the attribute level. This section takes the same macro approach, applying it to the language within the instances of summarising identified.

The keyword analysis in 4.5 (Table 4.5) showed that the inclusive *we* is more salient in summative language whereas *you* and *they* have a higher keyness value in non-summative language. In terms of occurrence (pmw) in specific and general summary compared to non-summary, the picture is a little different, as Figure 5.3 shows.

We is relatively more frequent in all four summary types (and so all summary) compared to non-summary. At the level of type, *you* is relatively more frequent in reviews and *we* is relatively more frequent in previews. The standardised account of frequency mirrors the absence of *I* in the keyness ratings, as this shows the least variation in usage across summary and summary types. *They* is relatively more frequent in non-summary, but its range ranks significantly lower in relative frequency than all other categories (0-3753 pmw, compared to *I* = 13053-18750 pmw, *you* = 18750-29779 pmw, *we* = 12279-39959 pmw).

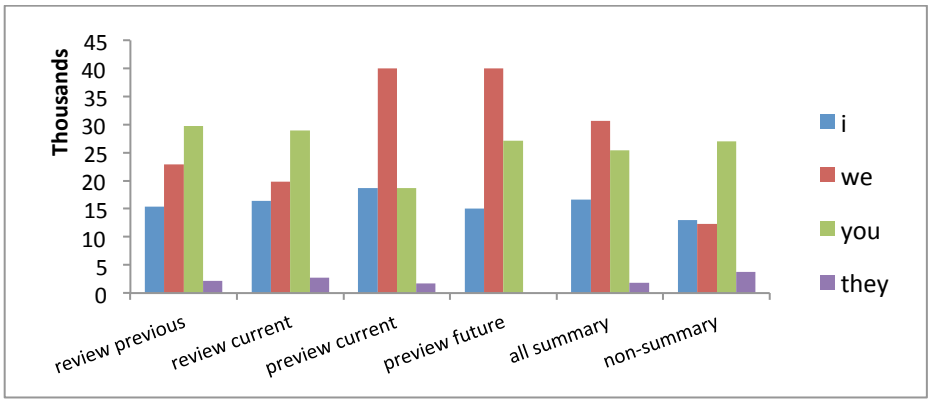


Figure 5.3: Occurrence (pmw) of *I*, *we*, *you* and *they* in summary types, summary, and non-summary

The quantitative data show that lecturers use *you* slightly less in general when they are summarising than when they are not summarising, although reviews employ *you* slightly more. The picture is of no highly remarkable difference in the usage of *you*, *I*, or *they* in summaries and non-summaries overall. *We*, however, is relatively more frequent in general summarising than non-summarising, and particularly so in previews. This difference will be discussed in more detail in the sections on previewing current and future lecture content (5.3.3-5.3.4).

Analysis of multiword lexical items at the level of summary type was undertaken to further interrogate the findings of Table 4.11 (visualised in Figure 4.4) and build on the keyness analysis of single lexical items (Table 4.5). The ten most common 4-grams per individual type were calculated for all summary, summary types and non-summary, as shown in Table 5.4. The pmw data from Table 5.4 are rendered in Figure 5.4 to give a sense of differences in the relative frequency of 4-grams and to show overlaps between attributed summary types.

review previous			review current			preview current		
4-gram	*freq*	*pmw*	*4-gram*	*freq*	*pmw*	*4-gram*	*freq*	*pmw*
if you want to	8	644	i mentioned just now	10	725	we re going to	153	8514
we did last week	8	644	so that s the	8	580	i m going to	69	3840
we were looking at	7	563	you ve got to	8	580	re going to do	43	2393
as i told you	6	483	as i mentioned just	7	508	what we re going	34	1892
if you have a	6	483	if you want to	7	508	going to look at	33	1836
last week we looked	5	402	so that s what	5	363	re going to look	29	1614
second moment of area	5	402	that s what the	5	363	have a look at	24	1335
we did on tuesday	5	402	a one one oh	4	290	going to do is	19	1057
we talked about the	5	402	mentioned just now you	4	290	m going to do	18	1002
we ve looked at	5	402	now as i say	4	290	what i m going	18	1002
preview future			**all summary**			**non-summary**		
4-gram	*freq*	*pmw*	*4-gram*	*freq*	*pmw*	*4-gram*	*freq*	*pmw*
we re going to	48	5332	we re going to	207	3889	i m going to	227	478
i m going to	21	2333	i m going to	93	1747	you ve got to	221	465
when we look at	19	2111	re going to do	59	1108	we re going to	201	423
re going to do	14	1555	what we re going	38	714	you re going to	200	421
to be able to	13	1444	going to look at	36	676	it s going to	157	331
we ll look at	12	1333	re going to look	33	620	you ve got a	142	299
you re going to	11	1222	you re going to	33	620	s going to be	138	291
re going to be	10	1111	we ll look at	27	507	if you ve got	122	257
need to be able	9	1000	have a look at	26	488	times ten to the	115	242
be able to do	7	778	we are going to	24	451	ten to the minus	108	227

Table 5.4: 10 most common 4-grams (raw frequency and pmw) in summary types, all summary and non-summary

The thrust of the 4-grams is towards predictive language using pronouns. The lecturers commonly talk about what *we* or *I* are *going to do* or *look at* or *be able* to do. As previewing current lecture content is the most common form of summarising in the ELC, this forward-looking emphasis is unsurprising. As the non-summary category shows, 4-grams are more commonly used when lecturers are summarising than in other parts of the lecture. When the ten most common 4-grams per type are compared (see Figure 5.4), it becomes clear that previewing types contain substantially more 4-grams - particularly *I'm going to* and *we're going to* – than in other reviewing types or in non-summative language.

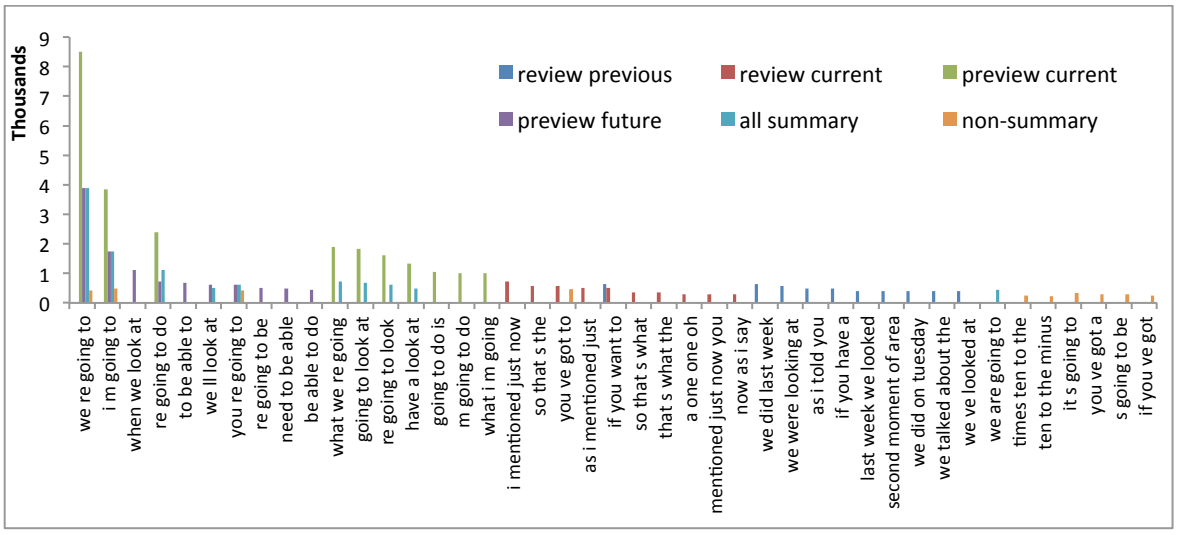


Figure 5.4: 10 most common 4-grams (pmw) in summary types, all summary and non-summary

The type of metatext that is noticeably missing in either the positive or negative keyword lists is boundary markers. In terms of specific characteristics of summative language, a comparison of the presence of boundary markers was expected to denote a shift in function between summative and non-summative language. The results show that boundary markers, especially *ok*, are in fact less likely to be contained in (or contain, in the sense of immediately surrounding) summative language (see Table 5.5).

	all summary	all non-summary
so	14248	15581
yeah	3282	3592
ok	10416	66755
right	2846	3171

Table 5.5: Occurrence (pmw) of boundary markers in summary and non-summary

The marker *well* can perform a discourse initiation function in the preview types of summary, as in:

> *well* I'm now going to be a a client regarding that need and you need to interview me ok so that's what we're going to be doing today […] (3028)

However this is rare – only three such examples were found.

Another strategy for emphasising the importance of summarised information is labelling the significance of specific concepts or processes. Young (1994: 171) found that lecturers in the evaluation phase reiterated and weighed up information that had already been given; material was appraised through the process of attributive relations, not by attitudinal elements (for example, modals). In the conclusion phase key terms were identified and classified. Those points that had already been summarised were reinforced through repetition and evaluation of key terms and theories (1994: 171). The process of weighing up and classifying already given information is also visible in the summative sections of the ELC. For example:

> as I say at the introduction to the questions if you're within point one or point two of water cement ratio that's as accurate as this method is (1011)

The anaphoric shell noun in "this method" serves to repeat and summarise a previously given extensive explanation about how to graph various aspects of concrete mixes, and the evaluative use of "as accurate" draws further attention to its significance within the greater lecture discourse. As in Young's *conclusion phase*, a participant chain of information is accompanied by evaluative language.

In previews, the evaluation commonly precedes the delivery of the information, as in:

> a *really important* point for what follows in the in the coming lectures starting on Thursday is that the energy and the force and the whole effect of the field is in the air it's not here (emphasis added. 3001)

In reviews, evaluation commonly follows the delivery of information, as in:

> so that's the last bit about bracing then it's *important* (emphasis added. 1003)

> what I want to do today is look at the um approach to serviceability of limited state design […] so we're going to look today at the three um three limit states some of this is a bit factual so *just drawing attention* to things um others when we get onto the cracking

> bit we start to get into analytical work again and we'll finish the day with a few sums and calculations on crack recalculation (emphasis added. 1018)

Most commonly, this act of evaluating information occurs in reviews of previous or current content. Examples include:

> I said *very important* the stress strain graph is *very important* (emphasis added. 2003)

> if you remember ok the *all important* ten-fifty-seven yah the standard generic form of the second order ordinary differential equation (emphasis added. 2020)

The evaluative language within summaries thus reinforces content and indicates how information should be interpreted.

As well as emphasising the importance of certain concepts/processes, lecturers also use summaries to indicate hierarchies in significance of the information reviewed, as in:

> all that theoretical stuff is really *less important* than the practical ideas (emphasis added. 3005)

Labelling significance is also used to contextualise summarised material in the overall syllabus. One lecturer explains that:

> this whole course the whole year's work you're going to do with me on electrical principles if there was one board that I could you know peel off the board and stick it in your minds it will be this one because this is the one you use the most the ideas you see here are *really sort of central* to everything else we do pretty much (emphasis added. 3016)

In the same way as labelling the summative discourse act, labelling the significance of summarised information gives it a contextual place within and beyond the current lecture.

Anaphora is another commonly used linguistic device in summaries. For example, amongst the most common 4-grams (Figure 5.4), the anaphoric use of *so that's the*, *so that's what* and *that's what the* emerged as typical to summative language, especially in previews of current content. Particularly common are shell nouns which anaphorically and cataphorically function as micro-level summaries within summative strings (Nesi and

Moreton 2012). Lecturers employ such nouns to refer back in a compressed way to immediately preceding information, as in:

> so let's have a look at th- this how we're going to deal with this particular *problem* then (emphasis added. 1005)

> that's the physical *issue* we're looking at we're now going to use this to find out things primarily where is the meta centre our objective here is (emphasis added. 3021)

Details of the referent (the *problem* or *issue*) are delivered in the immediately preceding clauses.

The anaphoric reference is usually given in combination with a contextualisation of its relevance to upcoming content. Understanding the problem or the physical issue is the platform for knowing *how* to deal with it or *how* to find related information. Understanding the formula allows it to be rearranged, which functions to enable application to a new problem. The anaphora not only summarises a more complex concept or process, but also often heralds the introduction of related, expanded content.

Shell nouns also realise an important predictive function, particularly when the lecturer is highlighting key engineering topics, as in:

> we'll see next week um when you have two two plates separated by a distance that's a capacitor and we can charge up the capacitor by putting a voltage on it […] the *point* is when you get between those two plates you will feel a force (emphasis added. 3002)

Many of these shell nouns are nominalised verbal processes. For example, in the following, the shell noun *discussion* is a nominalised form of discuss:

> if the contractor fails to rectify the defect the client has the right to engage third party to do the work I already mention this one earlier in the in the *discussion* today (emphasis added. 2007)

Lecturers commonly talk about process/es, method/s, problem/s, equation/s, solution/s, question/s, objective/s, and thing/s. For example:

> we're gonna look at particular *method* called the method of joints which is the sim- simplest application of the pri- rules of equilibrium to these structures erm but it's quite a long-winded *process* (emphasis added. 1004)
>
> I've said that we are um running out of aggregates the obvious *solution* to that is to make aggregate out of waste and that is being done (emphasis added. 1010)
>
> with mix design we had this *problem* that we needed to add water to get workability if you remember (emphasis added. 1013)

Perhaps there is a particular need for this kind of cohesive device in engineering and other disciplines where complex processes and cause and effect relations are described.

In terms of the summary element, shell nouns (and anaphoric and cataphoric references in general) are common to, rather than necessarily indicative of, the four types of summary. Not all strings containing endophoric devices were counted as summary. Moreover, as the summative strings that were extracted had to make sense as standalone units, this might in some cases have resulted in the exclusion of text containing anaphoric devices, especially at the beginning of strings. To illustrate this distinction, at the borderline is the anaphoric use of *example*. Where no other indication of summative language was present and it was not contained within a complete clause, *example* was not annotated. The use of *for example* most frequently precedes the introduction of new information, as in:

> plastics are very very bad news in a fire so *for example* if you look at the light fittings in here you'll find that despite the fact that metals are not a very good material for making light fittings because it conducts electricity um they're all made out of metal (emphasis added. 1014)

Only in complete clauses is the use of *example* annotated as summative, such as its cataphoric use in the previews “I give you an example” (3010) and “let's deal with an example here” (3027).

The lexical item *key* is frequently used evaluatively, followed by a shell noun. Examples include:

> we're going to start off by doing key points (1006)
>
> let's look at some key points (3027)
>
> what I'm going to do today is just summarise some of the key principles (1001)

Countable shell nouns are not always used enumeratively in previews, but the term *key* is commonly used with enumerative language, in both previews and reviews. When summing up, lecturers often specify the number of important ideas covered, as in:

> so two key reasons why one w- w- w- wants to control deflections (1018)
>
> so all of these factors um how many are ther- there's five of them altogether are the key factors that govern the durability of the structure (1018)

The same model is predominantly used when upcoming content is previewed:

> I've got the three well four key bits of information (1006)
>
> I'll just take you through that and see what the code says about how to design serviceability limit states and essentially the three ke- key conditions of three serviceability limit states that are important (1018)

By specifying the number of significant concepts/processes, in combination with evaluative language, lecturers provide an explicit framework for processing given or forthcoming information within and beyond the context of the current lecture.

Signals of enumeration (cf. Tadros 1985: 20-21) in the ELC summaries include inexact numerals, as in the example:

> so *several type* of insurances that should be provided by the contractor one is we call it the performance bond in short we call it P-B ok the other one is related to what we call it insurance of works yeah we call it work insurance this is to prot- to protect the work especially against event such as fire things like that we will cover in short while yeah (emphasis added. 2007)

Most commonly, summaries include exact numerals, as in the previews:

> so we have t- *two reaction forces* at the cross section one is a compressive load fifteen kilonewtons the other one is the bending moment (emphasis added. 2013)
>
> this is a numerical example to show you *two things* firstly how this GG2 equation works not too difficult you simply take the I value of that free surface through its own centroid around the axis around which the whole shifting system is rotating the second part is to actually give you a sense of what this does (emphasis added. 3022)

And in the reviews:

> main *three things* that have come out of here though out of these tests is yield stress ultimate stress and modulus of elasticity (emphasis added. 3025)
>
> so those are the *four main processes* expansion compression and heat transfer (emphasis added. 2018)

In the 54 instances of exact enumeration found in the summaries, 31 occurred in previews of current content, 18 in reviews of current content, and five in reviews of previous content.

Most of the examples of non-predictive enumerative signals occur in reviews. Their function is recapitulation, as in:

> now if you were here last week I was saying that delta rosettes actually do come in *three forms* (emphasis added.1029)
>
> so these are *three type* you can divide the aggregate into three types (emphasis added. 2003)

In all cases the lecturer is simply summing up information that is given to context – it can be recovered from the previous text rather than following the enumerative signal.

There are, however, many examples of enumerative recapitulation where the signalled information is (re)given, as in:

> remember yeah there are *two types* of of tension beam one is post-tension the other one is pre-tension (emphasis added. 2005)
>
> I just covered the *two things* one is the bond the other one is the the work insurance which has to be purchased by the contractor (emphasis added. 2007)

> I've gone over *three different kinds* of charge surface the sphere the cylinder the plates (emphasis added. 3002)

The lectures also contain collaborative enumeration, as in:

> <lecturer> the reason why I'm showing this to you is because the *four main components* of the steam power plant as I have demonstrated yea- on the white board there are four one is the</lecturer>
> <students>boiler</students>
> <lecturer>boiler next is the</lecturer>
> <students>turbine</students>
> <lecturer>turbine nex- let's do this all together I like to hear your voice boiler</lecturer>
> <students>boiler turbine condenser pump</students> (emphasis added. 2015)

Enumerative recapitulation occurs as the lecturer proposes the V head and students and lecturer fulfil the D heads in co-operation. The example is like classroom discourse, but unlike written text and most forms of spoken discourse.

Information is flagged up as important through the relatively high repetition of certain lexis within previews and reviews compared to non-summative text, as shown by the comparison of positive and negative keywords (Table 4.5). A type-token analysis shows that the STTR of summary is 31.54, and the STTR of non-summary is 29.86 (Table 4.4 and Figure 4.2). The proximity of the lexical density somewhat masks the stronger tendency in summaries for certain lexis to commonly co-occur. The most frequent n-grams have significantly higher occurrence in summative compared to non-summative text (see Table 5.4 and Figure 4.4).

At the attribute level, repetition within reviews provides a type of double recapitulation of information. For example:

> so what we've done so far is E so there's E now remember what that is it's the field strength it's the force you feel when you charge something up so if I do this if I get a battery that's a battery symbol and I gave it two plates of metal and they're separated by an insulator if I do that if I set that real system up then in the space between the two plates we get a real mechanical force and th- the strength of that force is E E is electric field stress (3002)

The imperative “remember what E is” (a prompt to recall previously given new information) is shored up by the immediately following recapitulated definition “E is electric field stress”. In this example, the significance of E that is signalled in this summary is made explicit in a non-summative comment that occurs in the following lecture:

> if you forget about the integration and you forget about all the other stuff what I really want you to know is a concept in your mind of what E is and a concept in your mind of what V is (3003)

Reviews can provide multiple layers of reiterated information to signal special importance.

The same approach to repeating and reformulating especially important concepts also occurs in previews. For example:

> right now let's get into the big bits now we know round how it turns and how it twists now we need to go along and develop an equation that shows us how to find numerically the distance between the centre of buoyancy and the meta centre we want to peg the meta centre in space yeah we're going to develop an equation to find the height of the meta centre above the centre of b- buoyancy (3021)

The key message of the previewed information is repeated, in slightly different forms: the twice referenced “equation” will measure “the distance between the centre of buoyancy and the meta centre”, which is in other words “the height of the meta centre above the centre of b- buoyancy”. Glosses occur frequently in non-summative language within the corpus, especially during explanations. In summaries they reinforce the significance of the contained information and allow students another opportunity to absorb the central concept.

In the qualitative analysis, one of the linguistic patterns that emerged as particularly common to previews of current lecture content is the use of simple and reverse pseudo-clefts, as defined in 2.3 as constructions that express a relationship of identity between the highlighted element and relative clause (Collins 1991: 2).

In terms of their discourse organisation function, Deroey (2012: 121) identifies three main functions of clefts: 1. orientation to topic or aims, 2. delineation/ordering of discourse parts, 3. marking relevance through previewing or reviewing information from the same or other lectures. The main discourse function within clefts found in summative strings within the ELC certainly performs the relevance marking function identified by Deroey, particularly in terms of previewing content.

In all summary types, 98 examples of simple pseudo-clefts were found; 59 of these belong to the category of previews of current lecture content, as laid out in Table 5.6.

	UK		**MS**		**NZ**		**all**	
	raw	*per lecture*	*raw*	*per lecture*	*raw*	*per lecture*	*raw*	*per lecture*
review previous	1	0.03	2	0.11	2	0.07	5	0.07
review current	7	0.23	6	0.33	10	0.36	23	0.30
preview current	32	1.07	6	0.33	21	0.75	59	0.78
preview future	3	0.10	5	0.28	3	0.11	11	0.14
total	43	1.43	19	1.06	36	1.29	98	1.29

Table 5.6: Simple pseudo-clefts (raw and per lecture) in summary types across subcorpora

The fronting of *what* follows a fairly predictable pattern across all examples:

> what I want to do um get you to do in a minute is to just produce a um an interaction curve (1017)
>
> what I want to highlight here is the safety aspect of the bending process (2010)
>
> what are we going to do today is w- we are going to wrap up chapter five or the second law of thermodynamics (2019)
>
> what we're going to do today is figure out how strong the field is at a point some distance away from various objects (3002)
>
> what I'm going to do for you now is simply give you a sense of what kinds of things go into a factor of safety (3024)

The same device is also used in reviews of current content, as in:

> what we've covered so far is we looked at the analysis analysis and design of the sections (1019)
>
> what I've tried to show you in this introduction to AC is first of all that AC is really important (3015)

It occurs to a much lesser extent in previews of future content, as in:

> what we'll be doing next week and the week after is actually looking at f- figuring out how the forces get transferred through those frameworks (1003)
>
> what I'm going to emphasise on I'm not going to get you to make very complicated parts I'm going to get you for the next five weeks to make good parts" (3019)

Most infrequent is the use of simple pseudo-clefts in reviews of previous content, as in:

> what I said in week one the cost is a criteria the the more work you put in the higher the value of the component (3019)

Fronting does not always rely on the use of *what*, as shown in the following examples:

> to calculate the Z effective is the the one that I mentioned just now (2002)
>
> we've shown it graphically that the resultant has the same overall effect on the body so its goes from the same start point to the same end point as when we draw the forces nose to tail (1001)

Simple pseudo-clefts are particularly used in previews of upcoming content, to discuss *what we're going to do/have done*, *what I'm going to* say/*have said* or *going to do*/*have done*, *what* is about to happen/has happened. By fronting *what*, attention is drawn to the importance of the message that is being delivered. It would be less linguistically complex to dispense with the *what*. For example: "what we're going to do now is some calculations on ultrasound" (1015) could be expressed as *we're going to do some calculations on ultrasound*. Instead the simple pseudo-cleft functions as a concept emphasiser, signalling that attention should be paid to the information that follows.

Less commonly found in the ELC are reverse pseudo-clefts. Four were identified in reviews of current lecture content, as in:

> plasticisers and superplasticisers are what I've just done in the demonstration (1013)

Five were found in previews of current content, as in:

> so how do you actually find resultants in components and that's what today's class is really about (1001)

Collins (2004[1987]) argues that simple pseudo-clefts give specific contextual background knowledge in the relative clause before delivering the *message*, whilst the highlighted element in reverse pseudo-clefts refers to material given in the relative clause (distinguishing contextual and cotextual reference). It follows that the simple type occurs significantly more frequently in previews of current content (see Table 5.6), as this is when lecturers highlight information that has immediate relevance (or much lower distance to delivery, compared to future previews). Simple pseudo-clefts are also associated with descriptive genres (Collins 2004[1987]: 93), which suggests that summaries share commonalities with this text type. Relatively more reverse type pseudo-clefts would be expected to occur in reviews, but the total occurrence (9 reverse, compared to 98 simple) is too small to inform any useful conclusions.

Also characteristic of previews of current content are features functioning to minimise and hedge. The softening function of *just*, for example, is noticeable in ELC summaries. As discussed in 2.4.3, *just* is a commonly occurring lexical item within corpora of academic speech events, and functions overwhelmingly as a metadiscoursal hedge or minimiser (Grant 2011, Lindemann and Mauranen 2001). Lindemann and Mauranen (2001: 268) identified the "mitigating just" as a device that was "used to soften (sometimes implicit) requests, challenges, or other potentially face-threatening acts".

In non-summative language, *just* occurs 2303 times in 528157 tokens (4360 pmw). Within summaries, *just* occurs 277 times in 52708 total tokens (5255 pmw), which means that *just* is more likely on average to occur when lecturers are summarising content than when they

are delivering other parts of the lecture. The picture of average frequency is more revealing when calculated according to summary types, as Table 5.7 shows:

	review previous	review current	preview current	preview future
total tokens	12425	13788	17493	9002
instances of summary	226	351	474	218
instances of *just* within summary	50	106	104	17
just pmw	4024	7688	5945	1888

Table 5.7: Occurrence of *just* within ELC summary types

The breakdown by type reveals that *just* is particularly more common in summaries of current lecture content (reviews=7688 and previews = 5945, Table 5.7).

Using Lindemann and Mauranen's categories, Table 5.8 overviews the functions that were found for *just* in each of the four types of summary.

	minimiser		emphasiser		particulariser		temporal		ambiguous		total	
	raw	*pmw*	*raw*	*pmw*	*raw*	*pmw*	*raw*	*pmw*	*raw*	*pmw*	*raw*	*pmw*
review previous	38	3058	3	241	1	0	4	0	4	0	50	4024
review current	56	4062	4	290	3	218	42	3046	1	73	106	7688
preview current	89	5088	1	57	2	114	7	400	5	286	104	5945
preview future	16	1777	1	111	0	0	0	0	0	0	17	1888
total	199	72	9	700	6	332	53	3446	10	358	277	5255

Table 5.8: Functions of *just* in summary types (cf. Lindemann and Mauranen 2001)

In line with the findings of Lindemann and Mauranen (2001) and Grant (2011) in academic spoken English, the use of *just* in ELC summaries predominantly functions as a minimiser; 72% (199 of 277) of all instances fall into this category. For example:

> what we're going to do is just work through finding the resultant of those three forces (1001)
>
> this semester you just learn how to determine yeah the current or actual efficiency and the maximum efficiency (2015)
>
> what I'm going to do today I am just going to um simplify simplify this er derivation yeah (2019)

Further, 45% (89 of 199) of all minimisers occur within previews of current lecture content. This is interesting in the context of summative language because the temporal function was expected to rank more highly, particularly in reviews of current content. Unlike when they discuss what they will do *later*, however, even in reviews lecturers do not commonly refer to what they have *just* done.

An additional (secondary) attribute type was noticed alongside the preview/review current content types of summary. Multiple examples were found where during previews lecturers explicitly limited the scope of upcoming information by specifying the material that would *not* be included. For example:

> we'll see in a moment especially when we look at steel structures in reality the way things are connected together aren't this simple and we'll talk about it briefly this morning but we're not going to actually be doing anything in terms of calculations with it (1002)

Alternatively, reviews of information are contextualised by limiting the scope of their importance, as in:

> I'm not going to dwell on this um it's not a usual situation it can happen with sort of stocky T shaped beam sections um but essentially you can go back to first principles as we did last week and you can work out the area of concreting compression um which now incorporates the flange and the web itself so I'm not going to dwell on that (1017)

In the majority of cases found, summative strings that have this *negative* aspect contain the pronoun *I*, which opposes the common pattern discussed of using the inclusive *we* and locates the ownership of content firmly with the lecturer, enabling a disclaimer function. By acknowledging the vastness of topics and narrowing commitment to cover a particular aspect, this negative type of summary appears to function as a means of narrowing the field, contextualising the importance of topics, and pre-empting criticism.

Following the discussion of the macro-level structures and general linguistic features of summary, the following four sections (5.3-5.3.4) look in more detail at the characteristics of the four summary types identified, across the ELC and in its three subcorpora.

5.3. Summary types

5.3.1. Summary type 1: reviews of previous lecture content

Reviews of previous lecture content rank third in relative duration and frequency out of the four types identified: lecturers dedicate 2.35% of tokens (Table 5.1) to this type, and use it on average 3.05 times per lecture (Table 5.2).

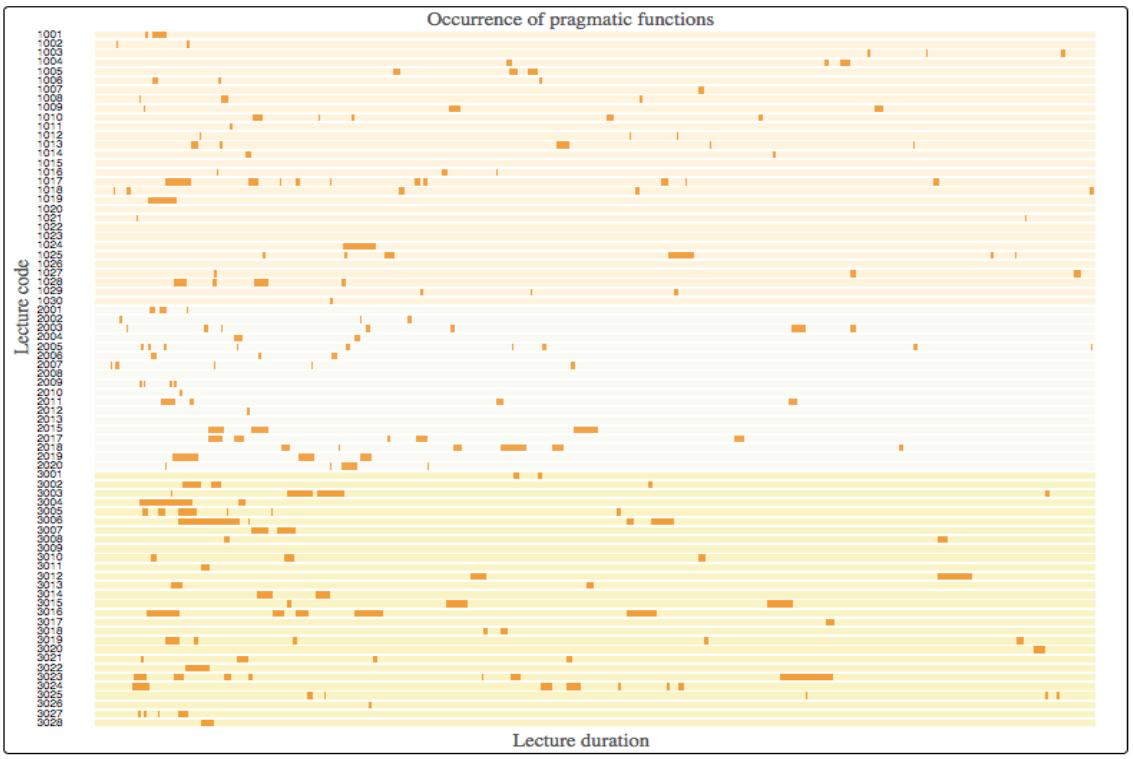


Figure 5.5: Occurrence and duration of reviews of previous lecture content

The average length of each instance of reviewing previous content is 54 tokens (Table 5.3), which is the longest of the four types. As the visualisation of occurrence in Figure 5.5 shows,

reviews of previous content tend to cluster around the first third of lectures in general, but can occur at any point.

Figure 5.6 renders for comparison selected patterns from data given in the earlier macro overview of summary type distribution (Table 5.1 and Table 5.2). The extracted data shows that reviews of previous lecture content occur for the longest duration (total tokens) in the lectures from New Zealand (3.58%), then Malaysia (2.49%), then the UK (1.52%). This sequence is re-ordered when the measure is of average occurrence (strings) per lecture: Malaysia (3.56), New Zealand (2.89), UK (2.70). The sequence arranged by average duration (tokens) of each discreet strings mirrors the sequence of overall duration: New Zealand (71), Malaysia (47) and UK (47).

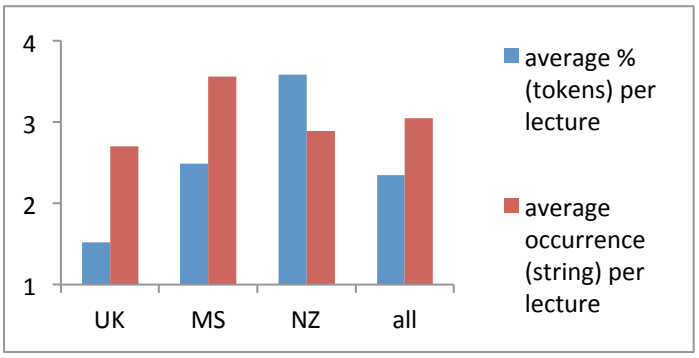


Figure 5.6: Average % (tokens) and occurrence (strings) of reviews of previous content per lecture

At the quantitative level the lecturers from New Zealand review previous lecture content a middling number of times per lecture, but each of these reviews is much longer than those found in the other subcorpora, as is the overall duration. The Malaysian lecturers review the content of previous lectures more frequently per lecture, but each instance (and the overall duration) is shorter in tokens. The UK lecturers employ this summary type by far the least in terms of overall duration and have the lowest number of discreet occurrences (the average length of which is equal to the average length of the Malaysian occurrences).

Reviews of previous lecture content function to recapitulate previously given information. Some of the concepts or processes under review are simply named, as in:

> we looked last week all about cement manufacture (1010)

Others are reiterated in detail, as in:

> I've given you a brief introduction in the last class um if you recall what we did in the first chapter we were looking at basic concepts definition terms that we use in thermodynamics yeah and then we started to analyse processes er we er apply the first law to a process yeah either a close system or an open system so we were looking at processes er if you look you've you've look at some of these processes those expansion processes those compression processes polytrophic processes you're able to apply the first law and you're able to determine how much heat is required to produce how much work you did all that yeah in the last four chapters (2015)

Reviews of previous lecture content include reviews of information in the immediately preceding lecture on the module, and also any other preceding lectures across all modules and years of study on the engineering course. Some references are made, for example, to "last year's work" (3025), or work covered "in first year" (3023), or "in the last four chapters" (2015).

Most commonly, however, lecturers review information delivered in the most recent lecture. This is highlighted by the keyword analysis, in which the two most salient tokens in previews are "last" then "week" (see Table 5.9).

Within the 226 instances of this summary type, explicit reference to "last week" occurs 78 times, which is almost every third time the lecturer reviews previous content. A collocation search within a range of three tokens to the left and three to the right shows that the five most common collocates of "last week" are: *we*, *I*, *you*, *did*, *looked*. The other most common simple past tense verbs within this range include *said, promised, talked, mentioned* and *discussed*; the last three of which rank as statistically key to reviewing previous content.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
last	112	408.863	five	9	29.441
week	82	256.514	t	37	25.851
remember	70	196.336	so	133	23.304
we	384	191.932	hundred	4	23.290
looked	36	150.897	going	21	21.353
did	58	143.927	be	40	21.201
talked	19	85.174	point	26	18.761
tuesday	19	82.787	s	167	18.446
told	20	70.978	ll	5	18.286
lecture	25	56.690	ok	22	18.175
showed	11	48.297	will	22	16.892
discussed	9	48.060	ten	4	16.565
yesterday	10	47.568	to	249	16.410
year	19	38.098	ok	12	16.189
processes	11	37.039	because	21	16.089
cycle	18	36.472	not	25	15.315
sigma	22	35.319	got	32	14.840
mentioned	10	34.777	they	27	14.150
were	25	33.855	three	16	14.098
covered	10	32.802	don	10	13.851
at	113	28.123	it	229	13.829
was	47	27.617	need	8	13.481
done	31	25.752	re	35	13.176
had	25	24.733	down	6	12.207
law	18	23.598	put	8	11.517

Table 5.9: 25 most highly ranked positive and negative keywords in reviews of previous content

Reviews are strongly characterised by the following linguistic pattern: temporal deixis (for example, *last week*) + pronoun + simple past tense verb + topic reference. Examples include:

last week we looked at resolving forces into components (1002)

last week we looked at shear design (1018)

last week we looked at er pressure and ultrasound (1021)

Or variations of this pattern, as in:

the other one we looked at briefly last week (1002)

we mentioned them briefly last week (1013)

that's the way the Xerox system works as I told you last week (3001)

Within the Malaysian data, there is some variety in linguistic structure:

> last week we have discussed equilibrium of a particle for two D problem (2001)
>
> last week we discuss on the setting and hardening (2003)
>
> last week I've gone through this (2005)

The verb is often not marked as simple past tense.

The dominant bigram *last week* is also substituted by other time references in the same pattern across subcorpora. For example, *yesterday* occurs 42 times, as in:

> we noted yesterday there are two functions or two purpose of this reverse heat engine (2017)

Specific days of the week are also named, for example: "on Tuesday" (3007), "last Thursday" (1023), and "on Friday" (3024). Fixing the point at which information was delivered is an important strategy when summarising content delivered prior to the current lecture.

Lecturers are likely to locate prior content chronologically using the simple past tense rather than the present perfect tense. Sometimes references to information delivered at a non-fixed point in time precede a chronologically-identified reference, as in:

> we've done axial stress both in materials and we did a little bit of it *last term* (emphasis added. 1008)

Alternatively temporal deixis is tacked on, as in:

> we've done it briefly an intro to it when we looked at masonry as a material ah *last term* (emphasis added. 1027)

Lecturers put effort into pinpointing information. Where chronological specifics are not recoverable, broad references are made to oft-repeated or widely known information, as in:

> recycled aggregate also one of the things that *we discussed so far* in the field of civil engineering (emphasis added. 2003)

> if I were to draw a schematic of a power plant like I've done *so many times* before say this is the boiler and then you have a turbine and then you have a condenser (emphasis added. 2015)

If the information cannot be associated with a particular point in time, it is normally affixed to another concrete point in the progression of the module or course, as in:

> we've seen this picture before *when we've looked at bracing* (emphasis added. 1025)
>
> you've done this *elsewhere* in S-H-N (emphasis added. 3023)

The thematic location of reviewed information appears to act as a substitute cognitive anchor, giving students a reference point for recall.

In reviews, lecturers frequently explicitly name, or give an advance label (Tadros 1985) to, the discourse function with which they are about to engage. Reviews often begin with a metastatement concerning what is about to occur. The use of *recap* is common in reviews of previous content, as in:

> a little bit of recap with shear (1007)
>
> to recap what you learned during your first year (2005)
>
> just to recap what we learned last week (2005)
>
> to recap on what we did yesterday (2019)

Synonyms are also used:

> so just to *summarise* what we've been covering (emphasis added. 1017)
>
> to just *go over* just *pull together* and *summarise* some of the work we've done so far (emphasis added. 1019)
>
> *summarising* what we've got what this theorem is about (emphasis added. 1024)
>
> let's just *sum up* the principles once more (emphasis added. 3002)

By flagging up the discourse type, it seems that lecturers are alerting students to the importance of the summary. As one lecturer explains:

> let's just recap and make sure that you have th- the concepts here (3005)

The significance of grasping key information is also highlighted by another lecturer:

> I'll just say this once more right because everything else is built upon tha- that first principle (3004)

The act of reviewing information from previous lectures seems to have two functions: to stress the importance of key concepts, and to ensure that the building blocks for future learning are in place.

The significance of reviewed information is also highlighted through the use of the verb *remember*, which ranks as the third most key lexical item (Table 5.9). *Remember* functions as a relevance marker (Deroey and Taverniers 2012). In terms of occurrence (pmw), *remember* is significantly more common in reviews of previous content compared with all other summary types and non-summative language (review previous=5714, review current=3191, preview current=229, preview future=889, non-summative=471).

Lecturers commonly use two 3-grams that contain prompts. The first is *if you remember*, as in:

> in mix design with mix design we had this problem that we needed to add water to get workability if you remember there was the table that showed how much water you need to add to get a given slump and then we found that if we added all that water we had a problem because that increased the water to cement ratio and we lost our strength so we couldn't get the workability and the strength at the same time (1013)

The second is *do you remember*, as in:

> now do you remember when we did the heat engine we say efficiency the thermal efficiency of the heat engine is what you want which is the work net of what you have to pay Q H isn't it (2017)

In both cases, the interrogative acts as a prompt for a recapitulated explanation (as annotated in the examples from 1018 and 1013 below). Synonyms are also used, and the tone of this usage is always reassuring and gentle, as in:

> <prompt>now if I can *cast your mind back* to first year maybe you did bending composite materials do you remember doing that</prompt><explanation> with beams made of two different materials</explanation><prompt> a long time back can you remember it I'm sure you did it</prompt> (emphasis and non-ELC annotation added. 1018)

The most detailed summative strings employ this prompt-explanation pattern recursively within single summaries, as in:

> <prompt>last week I was showing you some shear rate versus shear stress</prompt><explanation> so this ones effectively the other way up but what it shows is that you can track the exact effect of um superplasticisers um by measuring it on one of these reometers machines</explanation><prompt> if you remember I described it</prompt><explanation> it's effectively a food mixture that stirs up the mix and you measure the rate at which it's stirring it and you measure the amount of energy required to stir it right this is just showing that um generally when you add a superplasticiser um you can't actually measure the slump quite often you have to do it on the flow table<explanation><prompt> which again I described last week</prompt><explanation> ok and if you add your superplasticisers but remember you've also got to add your viscosity modifier you get self-compacting concrete and there's some pictures of it</explanation><prompt> um I did describe it last week</prompt> (non-ELC annotation added.1013)

When *remember* foreshadows a detailed overview of the referenced material, the review acts either as a reminder or as a means of filling in gaps to ensure that important information has not been missed.

The other use of *remember* is as a simple alert or memory jog. For example:

> remember how I showed you on Tuesday that la- that lamp glowed I held that lamp here and it glowed a bit one of these tubes well that would happen here too (3002)

> when you have your refrigeration you have a valve remember that is an isentropic constant entropy but this is isentropic constant (2018)

The focus of this usage is to make links between current concepts or topics without fleshing out the reiterated information.

Whilst all uses of *remember* function to highlight important information, their focus is different. When accompanied by detailed explanation, *remember* tends to occur at some distance from the presentation of the information that is about to be reiterated and its function is to emphasise that information. When *remember* is used as an alert, the lecturer is drawing attention to links between previous and current content through contextualisation.

As in the analysis of all types of summarising, *we* also has a high keyness ranking within reviews of previous content (see Table 5.9). Lecturers choose to use the inclusive pronoun when discussing work that has already been done, as in:

> you'll see that looks an awful lot like what *we've* done over the last well about two or three weeks ago when *we* looked at forces resolution of forces meeting at a point and finding resultants (emphasis added. 1004)

The formulaic pattern of: specific time frame + object/concept of review (in either order) identified earlier is repeated twice in this summary. It is also noteworthy that an inclusive pronoun accompanies a negative evaluation. The anaphoric "that" in the example above from 1004 refers to a preceding discussion of member forces. Earlier in the lecture during non-summative discourse, students were forewarned about the complexity of upcoming content:

> I will tell you now based on my previous years' experience of this topic this is the hardest topic that you will do in structures and structures we believe is the hardest module you'll take on your degree not sure about the H-N-D but it's definitely up there with er with one of the tough ones (1004)

In this context, the principles of forces resolution (a topic that is known to be complex) is compared to principles of member forces, a topic that we can assume is also complex due to the sense of trepidation implied by the comparison "an awful lot like". The use of "what

we've done" and "what we've looked at" (as opposed to, for example *what I showed you*) presents a united front in dealing with "one of the tough ones" in the hierarchy of topics.

Negative evaluations are also present in previews, as in:

> next week we'll look at a more sophisticated way of finding out forces in members like in pin jointed frames like this but it's a little bit conceptually *a little bit more difficult* but the principles again are the same (emphasis added.1004)

> we're going to move on to ranking these *it's going to get quite complicated* (emphasis added. 3027)

Reviews of previous information, however, contain more negative evaluations than the preview types of summary. The particular complexity of the information contained in these reviews is often acknowledged. For example:

> this is insurance of work ok take a look at the cost ok and as I told you er as mention earlier *it's not that easy to understand* from the write up here because this is based on the language of the lawyer (emphasis added. 2007)

> now as I told you before in lecture one *it is not easy it's not easy* to give a clear definition of the word energy what energy actually is I always find it *hard to explain* what energy is (emphasis added. 3006)

In reviews, the requirement for attention to be paid is doubled: the information is deemed to be worth reviewing (an indicator of complexity) and explicitly marked as complex through accompanying negative evaluations. Lecturers add emphasis to the importance of summarised information by comparing it to other equally complex concepts or processes, or recognising problems even in the delivery (let alone reception) of the information.

5.3.2. Summary type 2: reviews of current lecture content

Reviews of current lecture content occur second most commonly and for the second longest duration of the four types identified: lecturers dedicate 2.61% of tokens (Table 5.1) to this type, and use it on average 4.56 times per lecture (Table 5.2). The average length of each instance is 42 tokens (Table 5.3).

As expected, the density of reviews of current content is sparse towards the beginning of lectures. As Figure 5.7 shows, they occur with regular uniformity in the central 60%, and cluster with more density towards the end. Lecturers review the content they deliver in the current lecture throughout that lecture in quite short bursts.

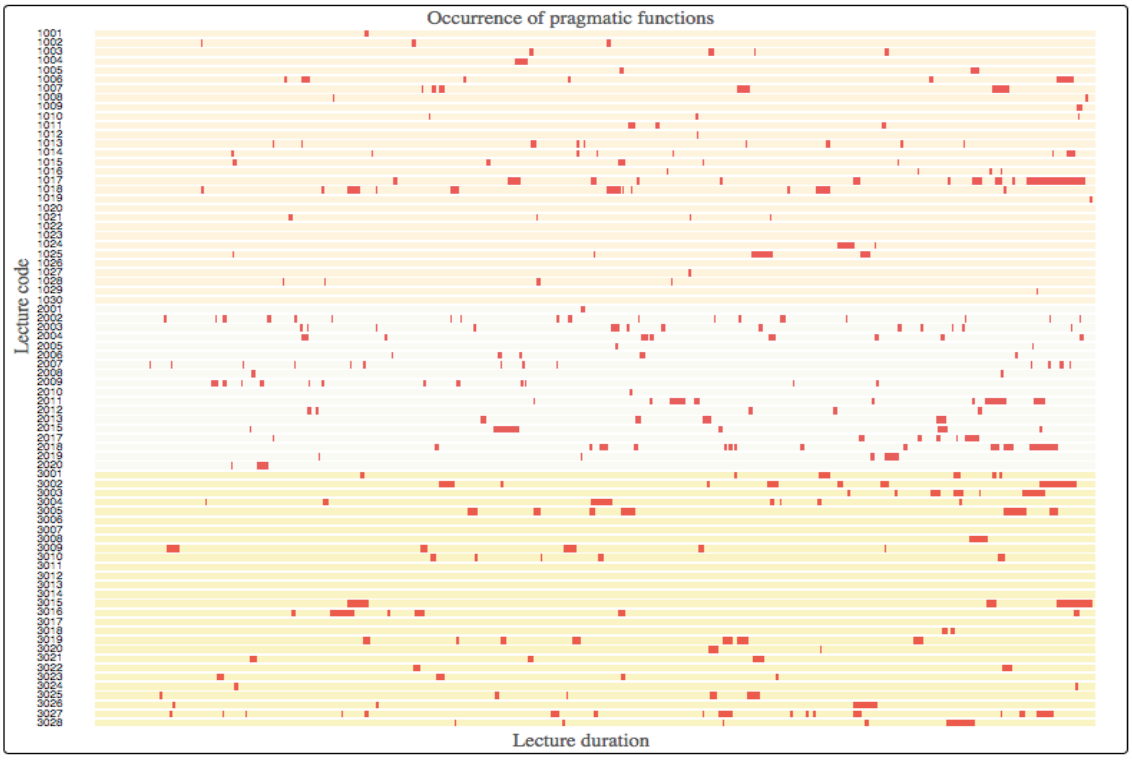


Figure 5.7: Occurrence and duration of reviews of current lecture content

Figure 5.8 shows select pattern data from Table 5.1 and Table 5.2, which illustrates that reviews of current lecture content occur for the longest duration (total tokens) in the lectures from New Zealand (3.55%), then Malaysia (3.14%), then the UK (1.77%). This sequence changes when the measure is of average occurrence (strings) per lecture: MS (6.22), NZ (3.79), UK (3.67). Both patterns follow the sequences identified in the reviews of previous content (5.3.1). When measured by the average duration (tokens) of each discreet string, the order of the last two subcorpora is rearranged: NZ (53), UK (40), MS (34).

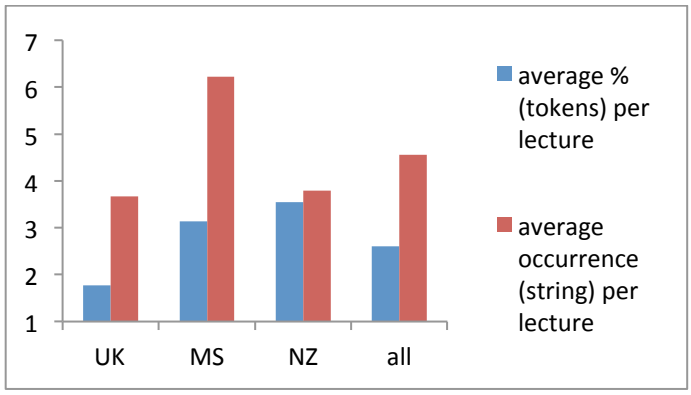


Figure 5.8: Average % (tokens) and occurrence (strings) of reviews of current content per lecture

The pattern across both types of review is that the lecturers from New Zealand dedicate the most tokens to reviewing and that the Malaysian lecturers deliver reviews more often but for less time. The UK lecturers review current content less in terms of tokens and with less occurrence, but the average length of each review is not the shortest.

Overall, the reviews of current content function as reminders and highlight the important parts of what has been said. There are some instances that follow the pattern that dominated reviews of previous content, as in:

> <discourse act>I've spoke</discourse act><topic/process of review> about the ah leadership struggles</topic or process of review> (non-ELC annotation added. 3028)

> <discourse act>so in summary</discourse act><topic/process of review> conflict is bad don't go there</topic or process of review> (non-ELC annotation added. 3028)

This pattern was expected to be common in reviews of current content following its identification in reviews of previous content. However, identifying a specific time-period is not characteristic of this type. Instead, non-specific references place the reviewed information within the current lecture. The most common mark of distance from delivery is *earlier*, which is the most key token in this review type (Table 5.10). If information that did not immediately precede the summary is being reviewed, lecturers specify that "I spoke earlier" (3027, 3028) or "I said earlier" (2015, 3001, 3028), or refer to what "we did earlier"

(2013, 3003) or what "we had earlier" (2012). Other references to current content, as in "we've worked out today" (3003) and "as I say this lecture" (3002) function like *earlier* to remind students that the process or concept under review is one that has already been discussed in that lecture.

This formulation is most commonly fronted in reviews of current content. Other formulations occur with much less frequency, for example:

> there is way of calculating it there is but you need a stress strain graph for the wood to do that so *we've already gone through* that materials process (emphasis added.1008)

Going "through that materials process" is a reference to what has already been done, with no prediction of future or new content. Using Tadros' terminology, this example would be classified as a reminder because it occurs after the information has been given, and chronologically two thirds into the lecture (that is, not at the end).

The keyword analysis in Table 5.10 also points to the use of *mention/ed* as particularly salient to these reviews. The collocates in the range of two left and two right of *mention/ed* are *i, just, now,* and *as*. Lecturers commonly draw on formulaic language comprised of these tokens, for example:

> as I mentioned just now you need to classify your your section (2002)
>
> I I mentioned just now you know to use these these formula for table in the for the deflection (2002)
>
> so OSHA as I mentioned just now it covers all sectors except the arm forces and work on board of ships (2010)
>
> so the Carnot cycle is composed of four reversible processes yeah as I mentioned just now (2018)

Variations on the theme of proximity include:

> as I said a minute ago (1017)

going back to what we said a few moments ago (1018)

Naming the discourse act and specifying its occurrence comes with a sense of immediacy. It signals a review of information that has very recently been delivered, which implies that notice should be taken because the content is worth immediate or near repetition.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
earlier	22	92.242	m	12	22.628
mentioned	20	85.823	it	242	21.673
remember	43	80.204	ok	24	20.832
reversible	29	77.455	t	49	18.587
as	115	51.580	put	7	16.930
mention	13	37.563	but	33	16.741
charge	26	33.282	question	2	16.715
curve	20	32.798	hundred	8	15.887
we	273	32.244	be	52	14.979
shoes	6	31.522	because	26	14.079
spoke	6	31.522	going	30	13.983
said	25	30.950	ten	7	11.868
cop	10	29.712	twenty	7	11.831
constraints	6	29.577	who	1	11.821
discussed	6	27.936	me	3	11.757
refrigerator	12	27.555	got	40	11.585
conflict	5	25.260	next	1	11.160
isentropic	4	23.592	any	7	10.605
law	19	23.415	thirty	2	9.875
the	959	23.160	could	4	9.657
told	10	22.838	mean	1	9.598
families	3	21.404	should	4	9.495
subsets	4	21.015	will	32	9.376
just	106	19.687	answer	1	9.100
today	14	19.500	sixty	1	8.901

Table 5.10: 25 most highly ranked positive and negative keywords in reviews of current content

There is less tense and aspect marking in the reviews in the Malaysian subcorpus. For example, the verb *mention* is used 31 times in current reviews in a context normally associated with the past tense or the present perfect. In 18 of these instances, the verb remains uninflected, as in:

I I mention just now simple construction (2002)

so the impurities that I mention or the one that we discuss is just like is like the clay (2003)

Tense and aspect marking does not appear to be crucial in signalling reviews of information in this subcorpus. In the Asian Corpus of English (ACE) – an English as a lingua franca (ELF) general corpus of general spoken – Kirkpatrick and Sophiaan (2014) found that tense in Malaysian English is only unmarked in informal situations. When explaining new concepts (in non-summative text), the same ELC lecturer tends to mark tense, as in:

> aggregate is a rock rock-like material of various sizes and shapes so there are many types of aggregate either crushed aggregate or natural aggregate yah so it is *used* in the manufacture of Portland cement (emphasis added. 2003)

Periods of summarisation may mark a shift in lecturer tone compared to periods when new information is being worked through.

A 3-gram analysis points to the place of minimising language in summaries in general (see Table 5.11 and Figure 5.9). It shows that lecturers use *a little bit* with much higher relative frequency in summary – and especially reviews and previews of current lecture content – compared to non-summary. In the discussion of the use of *recap* and its synonyms (5.3), one of the noticeable features was a tendency for collocation with minimising language, as in:

> *a little bit* of recap with shear (emphasis added. 1007)
>
> so *just* to summarise what we've just been covering (emphasis added. 1017)
>
> to *just* go over *just* pull together and summarise some of the work we've done so far (emphasis added. 1019)
>
> let's *just* sum up the principles once more (emphasis added. 3002)

It seems that minimising language, and especially the use of *just*, has a particular role to play in summaries in terms of reducing the imposition when the concept is perceived as being difficult. This is especially the case in reviews of current lecture content.

	review previous		review current		preview current		preview future		all summary types		non-summary	
	freq	*pmw*	*freq*	*pmw*	*freq*	*pmw*	*freq*	*pmw*	*freq*	*pmw*	*freq*	*pmw*
're going to	5	402	7	508	170	9718	59	6554	241	4572	410	776
we're going	4	322	3	218	154	8804	49	5443	210	3984	207	392
going to do	2	161	2	145	67	3830	22	2444	93	1764	122	231
i'm going	3	241	3	218	66	3773	21	2333	93	1764	231	437
m going to	3	241	3	218	66	3773	21	2333	93	1764	228	432
look at the	12	966	6	435	36	2058	8	889	62	1176	202	382
going to be	0	0	7	508	33	1886	21	2333	61	1157	378	716
that's what	10	805	16	1160	15	857	12	1333	53	1006	182	345
what we're	0	0	1	73	44	2515	6	667	51	968	106	201
to look at	2	161	1	73	41	2344	6	667	50	949	56	106
we look at	8	644	1	73	20	1143	21	2333	50	949	61	115
going to look	1	80	0	0	43	2458	4	444	48	911	21	40
so that's	6	483	27	1958	7	400	4	444	44	835	272	515
you've got	13	1046	11	798	14	800	6	667	44	835	825	1562
that's the	11	885	22	1596	7	400	2	222	42	797	360	682
we've got	3	241	18	1305	11	629	7	778	39	740	425	805
a little bit	5	402	2	145	21	1200	10	1111	38	721	130	246
and then we	2	161	3	218	28	1601	5	555	38	721	71	134
are going to	1	80	1	73	22	1258	14	1555	38	721	88	167
be able to	3	241	1	73	12	686	22	2444	38	721	102	193

Table 5.11: Occurrence (raw frequency and pmw) of the most common 20 3-grams in summary, summary types and non-summary

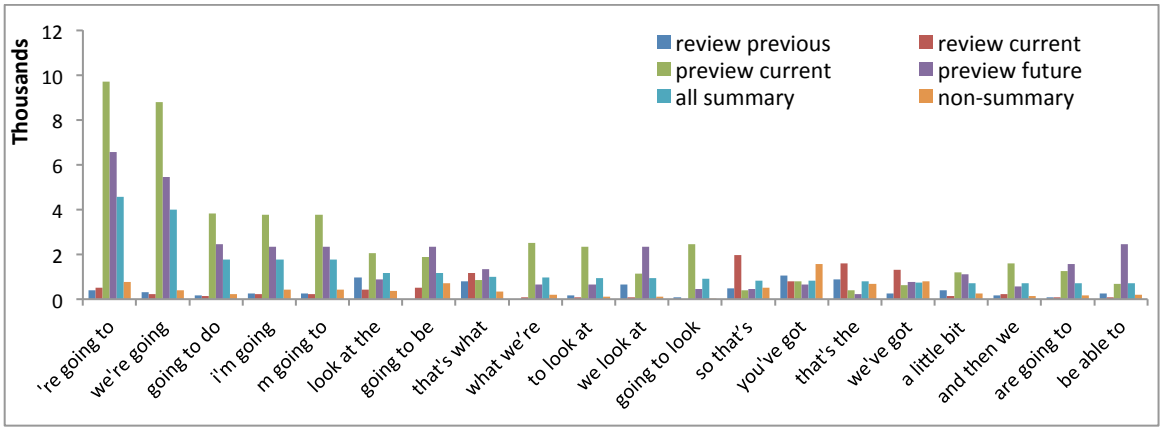


Figure 5.9: Occurrence (pmw) of 3-grams in summary, summary types and non-summary

Largely, however, lecturers do not specify that the information was delivered *just now* or at some point *earlier*. When current content is reviewed, the focus tends to be on reiterating specifics, jumping straight in with facts. For example:

so there is in an arbitrary load situation at least two kinds of forces that act on a piece of surface one normal and one along the surface (3024)

so that is a basic fundamental mechanism that can move fluids around in solids (1016)

The dominant function of this review type is to sum up information that has just been given. A commonly occurring feature is the presence of some kind of logical connector, normally *so*. Within the 351 instances of reviewing current information identified (see Table 5.7), *so* occurs 260 times. The level to which the connected information is fleshed out tends to be minimal. A common pattern is: conjunction (acting as causal logical connector) + anaphoric demonstrative reference to topic (or vice-versa). For example:

so that's what the the aggregates do (1010)

so that is how we measure things on site (1012)

so that's our general formula for calculating second moment of areas of shapes (1024)

so this is how the development of the thermodynamic temperature scale is done (2019)

Although lecturers do not explicitly name the act of repeating information, the act of review (especially when it is immediate) functions as both reminder and signal of importance.

5.3.3. Summary type 3: previews of current lecture content

Previews of current content are by far the most common summarising type in terms of both duration and occurrence: lecturers use on average 3.46% of tokens summarising upcoming content (Table 5.1) and on average 6.40 instances of this type occur per lecture (Table 5.2). The average length of each string is 38 tokens (Table 5.3), which is the shortest of the four types. This type of summarising happens most often, but for the shortest periods.

Previews of current content perform the explicit metadiscursive signalling function commonly discussed in previous research (for example, Crawford Camiciottoli 2004: 40), and have a strong discourse structuring function (cf. Young 1994). Like the *introductory roadmap* category identified in MICASE, previews of current content outline or announce

the topics or course of the upcoming class. ELC previews, however, do not have to occur in a particular position (for example, at the beginning of lectures), unlike MICASE roadmaps.

As expected, previews of current content, especially those of longer durations, tend to cluster around the beginning of lectures (see Figure 5.10). They do, however, occur throughout the lectures, often in shorter strings.

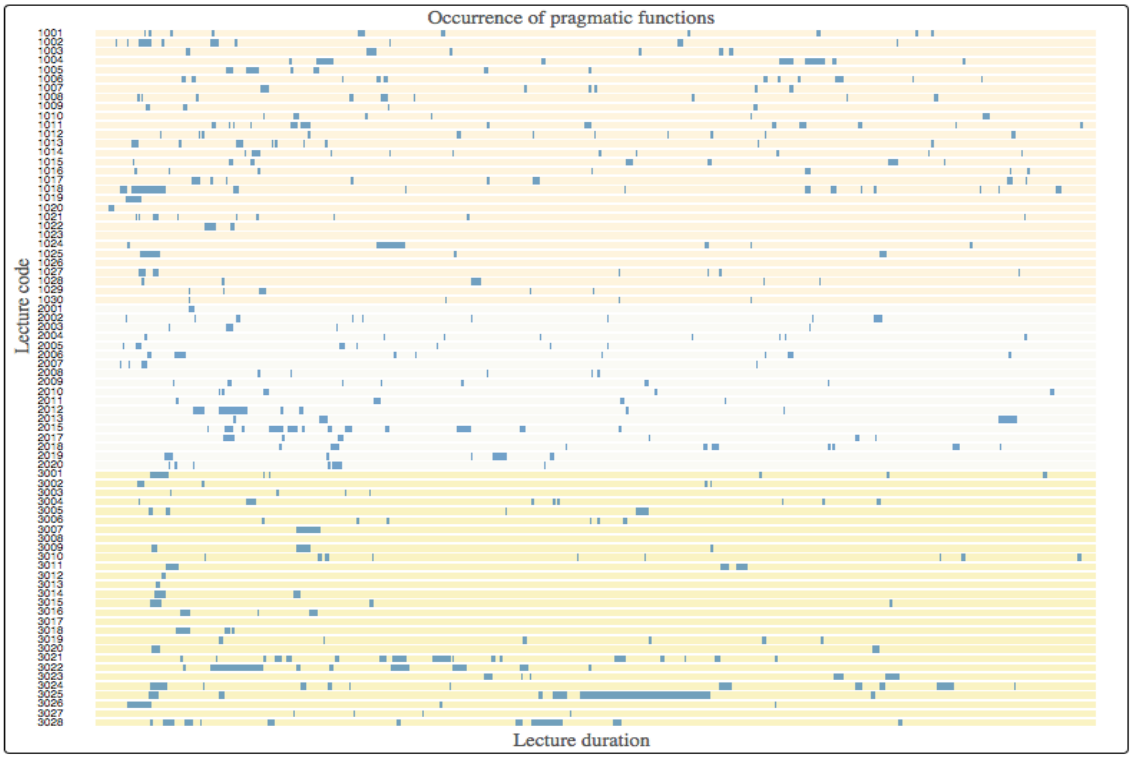


Figure 5.10: Occurrence and duration of previews of current lecture content

Figure 5.11 shows that previews of current lecture content occur for the longest duration (total tokens) in the lectures from New Zealand (4.74%), then the UK (2.95%), and then Malaysia (2.61%). This sequence is re-ordered when the measure is of average occurrence (strings) per lecture: UK (7.2), MS (6.5), and NZ (5.04). Measurement by average duration (tokens) again gives a different picture: NZ (42), UK (38), MS (35).

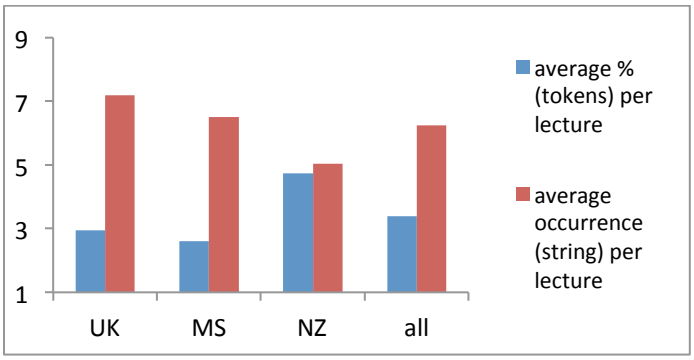


Figure 5.11: Average % (tokens) and occurrence (strings) of previews of current content per lecture

The most noticeable pattern in this data is that in the two subcorpora that dedicate almost equally small amounts of tokens to previewing upcoming content in the current lecture (UK and MS), the average number of instances of summary is higher than the norm, and a lot higher than in the New Zealand subcorpus. Lecturers from the UK and Malaysia use this type of preview frequently for very short durations, whereas lecturers from New Zealand use it less often for much longer.

The more extended previews that occur at the beginning of lectures are of the roadmap type, which plot the course of the upcoming class in some detail. For example:

> this week's class then looking at stability in structures and we'll just do a little intro to pin jointed frames the intro is really going to be about calculating reactions again because you've got some of these questions in your CAL tutorial which you should be able to do if you stick to the principles but hopefully these examples and the question that I'm going to get you to do after the break ah they'll help you see that you can actually just apply the same principles to any type of structure if you want to find out it's reactions but before we do that we're going to look at stability of structures (1003)

These roadmaps generally include references to more than one topic, and often provide a rationale for the upcoming content.

Later in the class, lecturers largely preview content in two ways. Firstly, immediately following topics are briefly outlined prior to delivery, as in:

what we're going to move onto now is looking at bending stresses inside beams (1008)

what we're talking about now is um making concrete (1011)

next off we look at air entrainers (1013)

ok we start off by defining temperature (1014)

right now we go onto a bit of thermal (1021)

ok now we'll go to the new topic on machine (2008)

now let's have a look at combinations of capacitors (3004)

Common discourse markers include *right*, *now,* and *ok* as lecturers transition between topics. These short alerts provide both a mini introduction to new topics and mark the boundary of the end of the previous topic. They occur cyclically throughout the lectures and contain an explicit authorial commitment to perform a discourse act – or *advance label* (Tadros 1985).

The second way of previewing that occurs in the current lecture is the signalling of topics that are upcoming but do not occur immediately following the preview. Some examples include a chronological reference to when the content will be delivered, as in:

these two K values these these are parameters um I'll come back to these in a second what these K values are

I took some cubes we'll talk about that in a minute (1012)

These examples function much as the previews of immediately following content, as quick alerts embedded within longer stretches of content delivery.

Other previews of non-immediately occurring current content acknowledge more distance to delivery time. The use of *later* ranks as key in this type of summarising (Table 5.12), which reflects the extent to which lecturers outline (future) current content, as in:

> a bit later on today u- um as we finish today I'm going to give you the coursework exercise and part of the coursework exercise is determining the you know particularly i- i- is a car box structure you've got to determine um what type of structure is it and what do you do with that information when you've got it (1018)
>
> how do you work out the stress we'll come back to that later when we see how to work out the neutral access position (1018)
>
> just like to the last five or ten minutes or so I'd like to show you how assemblies for assemblies I need parts (3020)

This more distanced type of previewing tends to be more fleshed out, functioning as a guide for overall structure that is delivered part-way through the lecture.

The analyses of 4-grams and 3-grams in summarising language (Table 5.4/Figure 5.4 and Table 5.11/Figure 5.9) show that approximately half of the most common n-grams occur with highest frequency (pmw) within previews of the current lecture. A condensed version of the 3-gram information is given in Figure 5.12. The calculations of n-grams highlight the importance of *we're going to*. The analysis of keywords (Table 5.12) also shows that *we*, *going*, *look*, *today* are significantly more salient to this type of summarising compared to other parts of the lecture discourse. Their collocations confirm that these common tokens co-occur in some order in a string.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
we	699	539.810	you	328	88.018
going	338	482.836	it	256	67.890
look	166	289.330	if	65	52.331
today	73	251.378	s	205	50.997
re	251	191.309	one	75	49.253
ll	135	178.319	times	4	39.768
at	251	174.369	hundred	4	39.320
to	770	172.357	t	52	37.546
lecture	37	83.943	minus	1	36.527
chapter	27	73.642	got	36	34.613
now	130	68.618	they	30	32.711
do	180	63.378	y	1	31.926
later	25	60.876	plus	1	25.152
let	60	58.200	can	62	24.813
go	110	56.171	not	34	23.900
example	53	54.016	point	39	23.728
m	107	51.954	there	71	23.454
through	66	47.446	zero	8	22.422
material	41	44.519	here	47	22.264
about	103	39.838	or	38	21.461
talk	24	38.977	twenty	7	19.726
define	15	37.706	c	5	19.589
move	30	36.723	say	14	19.318
some	75	36.601	thirty	1	18.881
how	88	33.275	ve	57	18.344

Table 5.12: 25 most highly ranked positive and negative keywords in previews of current content

Lecturers most often front previews with we are/we're going + verb, as in:

> we're going to talk today about column design (1020)
>
> today we're going to look at V (3003)

According to Peters (2004: 494), the difference between *will* and (*be*) *going to* indicates the speaker's orientation to the future event. In previews of the current lecture, pronoun + semi-modal (that is, (*be*) *going to*) is privileged over the pronoun + modal auxiliary (that is, *will*), as Table 5.13 shows.

	will/'ll		are/'re/am/'m going	
	raw freq	*pmw*	*raw freq*	*pmw*
we	48	2671	172	9571
you	28	1558	18	1002
i	49	2727	70	3895

Table 5.13: Pronoun + modal auxiliary/semi-modal (pmw) in previews of current content

When talking about upcoming information in the current lecture, lecturers most frequently discuss what *we are/'re going to* do – there is a sense of certainty about the outcome.

Previews favour the inclusive *we* much more strongly than either reviewing types or non-summative language (see Figure 5.3). Other pronouns are only rarely substituted for *we*, as in:

> I'm going to do a demonstration (1013)
>
> you're going to do an experiment (3023)

Noticeably, *you* is the most significant negative keyword (Table 5.12). *You* occurs with the least pmw frequency of all summary types: review previous=29779, review current=28938, preview current=18750, preview future=27105. On occasion, usage of *we* almost feels strained, as in:

> right what *we're* going to do now then is show *you* the short cut way to do it right (emphasis added. 1006)
>
> now what *I'm* going to do with *you* today is for the first ten minutes of this lecture *we're* going to look at general loading conditions (emphasis added. 3024)

When talking about what is about to happen in class, lecturers privilege an inclusive tone.

Another noticeable feature is that lecturers sometimes frame upcoming content as a puzzle that the class will solve together. For example:

> so *we will look* at the Carnot principles yeah and *we will examine* the Carnot cycle and then *we should be able to determine* the thermal efficiency of the Carnot cycle yeah for heat engine (emphasis added. 2015)
>
> *what we're going to do today is figure out* how strong the field is at a point some distance away from various objects (emphasis added. 3002)

Whereas in reviews facts and problems were concretely explained, in previews of current content lecturers raise questions with a sense of future projection, as if the solution is not already determined. The significantly frequent inclusive *we* contributes to the sense of

togetherness. Reaching the *how* (which ranked as a keyword in this summary type, Table 5.12) of the engineering problem is framed as a process that students and lecturer will *figure out*, *examine* then *determine* together. Lecture participants often embark on this journey by means of an *example* (another keyword, Table 5.12).

A device that is evident in both the examples of puzzle-solving language and throughout previews of current content is the use of softening, or minimising, language (as highlighted in 5.2.3). For example, following lexis related to prediction, *a little bit* is the eighth most frequent 3-gram in this type of summarising (Figure 5.12).

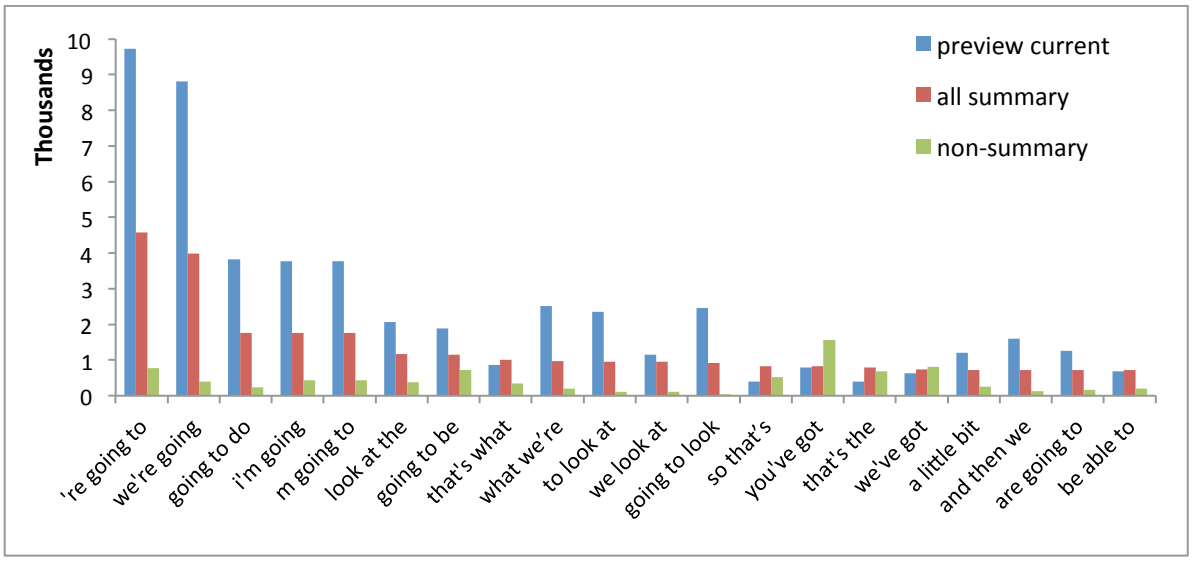


Figure 5.12: 20 most frequent 3-grams (pmw) in all summary, previews of current content and non-summary

Further analysis shows that such formulations are particularly prevalent in previews of current content. For example:

> so in the book it spends *a fair bit* of time looking at defamations of beams so I'm just going to explain *a little bit* about that and why we would want to worry about that (1006)

> now I'm going to discuss *a little bit* about th- the sort of theory of this and what we're actually measuring (emphasis added. 1012)

> we'll look at *a little bit* more mathematical approach to how we would calculate it for not so straight forward sections (emphasis added. 1024)
>
> so *a little bit* about filling in the background about retaining walls what they are how they're built and why (emphasis added. 1027)
>
> we're going to look *a little bit* carefully at some of the bits and pieces that go on over here (emphasis added. 3022)

The function of the minimiser is to offset the presentation of complex concepts and processes; things that should be looked at *carefully*, that should cause *worry*, background knowledge that needs *filling in*, the crux of what is *actually* being measured. The diminution of scale is a recurring device. For example, students are given the nod that "a little problem comes next" (3010), that they are going to do "just a bit of a bit of practice" (1017). Every effort is made to lessen the imposition of potentially daunting upcoming content.

A closer look at the qualitative data shows that *just* is predominantly used to soften previews of concepts and processes within the current lecture, functioning much like the common 3-gram *a little bit*. For example:

> what I'm going to do abou- now is *just* take a step back a bit go back into a bit more of the theory behind forces and moments (emphasis added. 1001)
>
> what we're going to do is *just* use the chance with the introduction of principles *just* to do some more calculation of reactions work (emphasis added. 1003)
>
> what we are going to move on is look at a more powerful technique called the method of sections which is *just* a continuation really of what we've done last week um but it's a little conceptually a little more difficult to understand (emphasis added. 1005)
>
> so what I want to do um get you to do in a minute is to *just* produce a um an interaction curve it's called an interaction curve axial load against moment for a given column and just see how the mathematics pans out (emphasis added. 1017)

The minimisation of particularly tricky upcoming content is realised by various language forms. There appears to be a correlation between level of difficulty and level of informality. Students are informed, for example, that:

> we're *just* going to *nail down* this idea of factors of safety (emphasis added. 3025)
>
> I'm going to *trot through* these overheads (emphasis added. 1011)
>
> you can *have a go* at designing th- the steel frame for a building (emphasis added. 1025)
>
> ok we're now going to *hop into* two side issues that we require to move forward (emphasis added. 3024)

Formulating a complex equation to calculate the height of the meta centre above the centre of buoyancy is presented as one of the "big bits":

> right now let's get into the big bits now we know round how it turns and how it twists now we need to go along and develop an equation that shows us how to find numerically the distance (3021)

A traditional "telling or transmission" (cf. Ramsden 1992: 111) model of lecturing is held up as "prattle" when students are required to engage more actively in working through problems:

> what I'm going to do now is move on to an example that illustrates some of these points but also involves some calculations as well what I intend doing instead of me stood here just and you listening to me prattle on er it's not gonna be a traditional example it's gonna be a case of explain the principles you have a go in rough if you need to in pencil and then we'll put the right answer up (1004)

The imposition of presenting complex concepts is commonly mitigated through choice of lexis and language structure.

5.3.4. Summary type 4: previews of future lecture content

Previews of future lecture content are the least common form of summarising in the ELC, constituting only 1.70% (see Table 5.1) of the total tokens in lectures, and occurring on average only 2.82 times per lecture (see Table 5.2). Based on average token count they are the second shortest of all types at 41 tokens (Table 5.3). Lecturers have a slight tendency to conclude with this type, but, like all summary types, future previews can occur in any part of the lecture (see Figure 5.13).

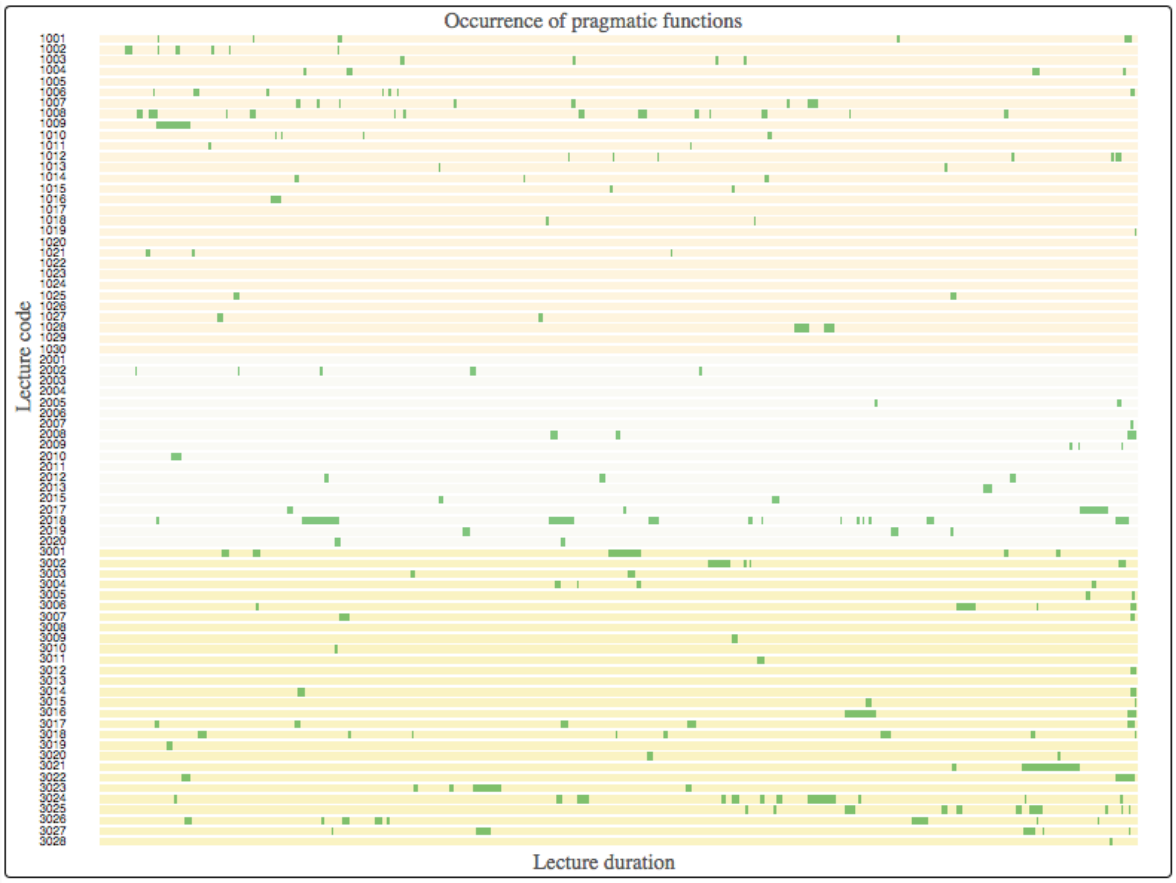


Figure 5.13: Occurrence and duration of previews of future lecture content

In future previews, *we* is in general the pronoun of choice – it is much more common than any other pronoun (Figure 5.3). The occurrence of *we* is almost equally frequent in future previews (39991 pmw) and current previews (39959 pmw), which is much higher than in non-summative discourse (12279 pmw).

Figure 5.14 shows that, as in all summary types discussed so far, previews of future lecture content occur for the longest duration (total tokens) in the lectures from New Zealand (2.93%) (followed by the Malaysian lectures (1.45%), then those from the UK (1.06%)). When the measure is of average occurrence (strings) per lecture, the sequence is: NZ (3.36), MS (2.44), and UK (2.67).

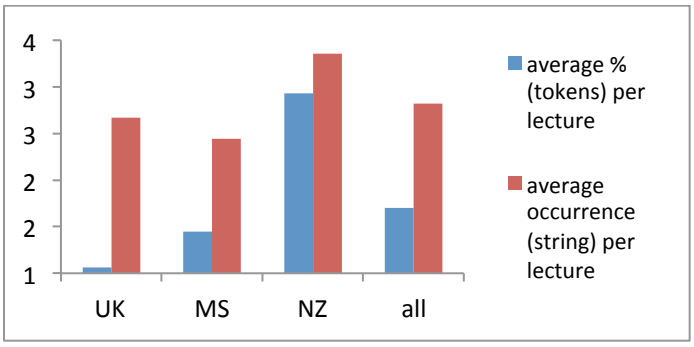


Figure 5.14: Average % (tokens) and occurrence (string) of previews of future content per lecture

The lecturers from New Zealand dedicate a lot more token space to previewing future lecture content and employ this type of summary more often. The Malaysian lecturers again preview in short bursts fairly often. A noteworthy result is the large frequency of future previews in UK lectures, but the very small amount of tokens of which they are comprised.

The practice of rigidly following a syllabus is indicated by the salience of *chapter* in this summary type (see Table 5.14). Reference to *chapter* largely occur in the lectures from Malaysia and New Zealand, for example:

> to design or to calculate the value of the M-C-X value for unrestrained beam it will be in chapter four ok I will explain later on in chapter four (2002)

Qualitative analysis also shows that the New Zealand lecturers particularly are keen to build topics incrementally week by week in quite strict order. Regular reference is made to numbered lectures and their content, as in:

> this idea of voltage divider is really important I'll be banging that over and over again throughout the semester *lecture nine* in two weeks time we'll see voltage divider's a really big deal (emphasis added. 3004)

The New Zealand lecturers are also concerned to show how content fits with the current topic, as in:

> we need to do *lecture five* we need to come again on Tuesday because the way it stands at the moment we still have this Q and that is a problem [...] now we can get round that problem and I'll show you how to do it on on next Tuesday (emphasis added. 3002)

The depth of detail of previewed information is also greater in the lectures from New Zealand and Malaysia, as an average token count comparison shows: NZ: 41, MS: 40, UK: 33. The UK lecturers tend to engage in less expanded forms of future previews, as in:

> we'll talk a bit more about these axes and the X and the Y significance when we look at bending stresses (1007)

> right but we'll deal more with that when we actually do some um design in the structures sessions (1008)

Links between topics tend to be less explicit, and the previews are not as specifically situated chronologically in terms of upcoming lectures as the future preview in the lectures from New Zealand and Malaysia. The shorter UK future previews put more emphasis on mentioning then sidelining information for another time, rather than signalling its place within the larger course.

Markers of temporality are characteristic of this type of summary. The keyword analysis (Table 5.14) shows that *week*, *tomorrow*, *year*, *Christmas*, and *Friday* are significantly more salient in previews of future content. Lecturers emphasise specificity and anchor this to the delivery of concrete and concise learning outcomes.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
next	92	312.932	if	24	44.924
we	360	286.125	one	30	39.961
ll	103	202.577	the	398	39.320
week	61	196.462	it	138	29.708
tomorrow	30	160.599	times	1	25.754
going	133	144.88	so	88	25.448
later	26	95.984	is	145	22.924
semester	17	92.960	here	17	22.818
lecture	28	83.904	just	17	20.834
look	61	79.274	hundred	2	20.679
chapter	20	69.175	down	1	20.610
christmas	12	67.033	was	1	20.575
will	94	66.305	ten	1	19.766
year	23	65.521	s	117	17.172
do	111	60.908	b	1	16.204
re	105	56.474	they	16	16.051
thermo	9	55.634	got	20	15.651
learn	14	49.882	five	10	13.714
able	22	48.769	from	11	12.921
when	66	47.882	would	4	12.511
structures	18	47.878	four	6	12.216
after	23	47.830	minus	2	12.193
friday	12	47.568	x	2	11.661
assignment	12	46.258	beam	6	11.416
cycle	18	46.063	ve	28	10.990

Table 5.14: 25 most highly ranked positive and negative keywords in previews of future content

The formula in which these references occur follows the format: *temporal deixis* + *pronoun* + *auxiliary verb* + *lexical verb*, as in:

> *next week we'll look at* a more sophisticated way of finding out forces in members (emphasis added. 1004)
>
> on *Friday we will continue* as well with the second law of thermodynamics (emphasis added. 2017)
>
> *in your third year you're going to start dealing* with fatigue (emphasis added. 3024)

The temporal deixis can occur at any point:

> we're going to then be designing beams and columns *after Christmas* (emphasis added. 1003)
>
> you'll be hearing a lot more about electric motors *in the second semester* (emphasis added. 3010)

As in reviews, situating the chronological occurrence of information for students is important, but previews of future lecture content typically contain less information than reviews. Students are alerted to upcoming topics, but the level of detail is minimal. Usually lecturers name the topic and when it will be delivered, often with an anaphoric reference to the current topic. For example:

> next week we'll extend *it* and look at uniformly distributed loads (emphasis added.1006)
>
> there is another one related to *this* we'll be talking about admixtures so I think is it next week or the week after (emphasis added. 1012)
>
> *it's* been very specifically manufactured so that these curves here and here rotate around that point and we'll discuss on Friday why that is so (emphasis added. 3023)

In terms of processing, these previews have a higher-level information structuring function; they put topics into the context of a framework for future work. This function is relatively short and infrequent in occurrence – lecturers do not prioritise forward-scaffolding in the same way as recapitulating information or predicting detail of content.

In the Malaysian data, tense is not always marked, as in:

> next week ah basically your syllabus finish for this semester so next week ah we will do some revision (2005)

In all of the types of summary discussed, marking tense does not seem to be crucial to conveying meaning in the MS subcorpus.

As topics are normally simply named and not elucidated, the down-toning strategies identified to minimise imposition in previews of current content are not present. Lecturers are brief and direct in delivering information in this summary type, as in:

> w- we'll do some more about the derivation of C tomorrow for different shapes (3004)
>
> later you'll find out in the last chapter a reversible adiabatic process is also known as an isentropic process (2018)

Few hedges or softeners occur, as processes of knowledge building are presented as simple and incremental.

Later is a keyword in previews of both current and future content (see Table 5.12 and Table 5.14). The distance to delivery time expressed, however, has nuanced implications. Concordance analysis shows that in previews of current content *later* is used to refer to upcoming content that expands on the current work, as in:

> how do you work out the stress we'll come back to that *later* when we see how to work out the neutral access position (emphasis added. 1018)
>
> you ca- as we'll do *later* on this morning we'll calculate these crack widths (emphasis added. 1018)
>
> a little bit more mathematical approach to how we would calculate it for not so straight forward sections so as an example we're going to calculate it for that shape for *later* on (emphasis added. 1024)

The expansion element of the predicted content often contains more complex calculations.

In previews of future lectures, however, *later* is used to indicate information that is relevant but not instrumental to understanding the *how* of the current content, as in:

> the main instance in which I actually use um electrical theory to do calculations doesn't come 'til a lot later in the course when I look at corrosion theory (1021)
>
> so later when you do y- when you go to your second year or third year when you study the gas turbine in detail you actually have to idealize some of the cycle and make it a complete cycle (2015)

Lecturers recognise the upcoming occurrence of topics, but frame them as information that can be put on the backburner until greater detail is required.

The lecturers' orientation is different when previewing future events compared to when previewing events in the current lecture. In the discussion of previews of the current lecture (see 5.3.3), it was noted that pronoun + semi-modal ((*be*) *going to*) is privileged over

the pronoun + modal auxiliary (*will*) (Table 5.13). In previews of future lectures, however, pronoun + modal auxiliary is privileged over pronoun + semi-modal, as Table 5.15 shows.

	will/'ll		**are/'re/am/'m going**	
	raw freq	*pmw*	*raw freq*	*pmw*
we	102	11331	56	6221
you	30	3333	15	1666
i	33	3666	21	2333

Table 5.15: Occurrence (pmw) of pronoun + modal auxiliary/semi-modal in previews of future content

In (*be*) *going to* forms:

> […] the speaker sees something which pre-dates Now as a reason for the future event. In future forms involving a modal, the emphasis is on the speaker's judgment at the moment of speaking (Lewis 1986: 145)

This distinction is made clear in the ELC examples:

> on Monday I'm going to do an example with this which will really put things into place what you're going to need for for this section is you're going to need to be able to deal with second moments of area (3021)

> we're going to find that a large number of mechanical um analyses are going to be in a range with a relationship between load and stress is linear we will discuss this it's known as the elastic range (3024)

Both lectures 3021 and 3024 are predictive, but the use of the semi-modal is based on what the lecturer already knows and the modal auxiliary is based on what is predicted to happen from this point in time. 3021 could be paraphrased, for example, along the lines of: *from what I know up to this point in time, I'm going to do an example, and I predict that this will really put things into place*. Likewise, 3024 could be paraphrased as: *from what I know up to this point in time, we're going to find a particular relationship, and I predict that we will then discuss these findings*.

In previews of current content, *we* and *going* were the two most highly ranked keywords (Table 5.12). In the previews of future content, *we* and *ll* ranked as the second and third

most common keywords (Table 5.14). The only two keywords that are the same in previewing current content and previewing future lecture content are *'ll* and *going*, the order of which is reversed in their respective rankings (Table 5.12 and Table 5.14). Of the other three words that the two lists have in common, *chapter* and *lecture* occur with almost equal frequency, but *look* is much more common in previews of the current lecture. These patterns indicate a tendency for the lecturer to build on prior evidence when previewing the current lecture and predict based on *now* when previewing future content.

The future delivery of content is not subject to any of the linguistic devices used to soften immediately upcoming content; the distance in time to delivery allows the predicted outcome to be more concrete. Students are reassured that they do not have to worry about grappling with complex problems, but simply need to understand the way in which what they are doing now fits into the bigger picture of definite learning outcomes.

5.4. Conclusion

Summarising constitutes just over a tenth of all lecture discourse, with some variation across types and subcorpora. Overall, summary is most common in the lectures from New Zealand in terms of total token count (Table 5.1) but occurs most frequently per lecture in the Malaysian subcorpus (Table 5.2). Previews of the current lecture content are the most common type in terms of word count and occurrence (Table 5.1 and Table 5.2). There is some variation in the type of summary that occurs at different stages in the lecture, and some indication of the chaining of certain summary types (such as reviews of previous content and previews of current content) (see Figure 5.2).

There is some variation in the salience of lexis across summary types, particularly in the use of pronouns (see Figure 5.3), but overall the inclusive *we* is significantly more key in summative text compared to non-summative text (see Table 4.5). In comparison to non-summative text, a much stronger tendency towards the co-occurrence of particular lexis was also noted in summaries, with some variation across types (Figure 5.9 and Figure 5.4).

Linguistic strategies such as advance labelling, the use of anaphoric shell nouns, simple pseudo-clefts, softening language, enumeration and evaluation were also identified as characterising summary types. Lecturers use various strategies for marking the significance of reviewed information. Formulaic language and chronological or thematic anchors situate the information both in terms of how it fits into the progression of knowledge acquisition at the module and course level, and also its hierarchical importance within this. Prompts such as *remember* are used as simple alerts to jog memory, or as signals of an upcoming detailed review. The complexity of the reviewed information is often emphasised.

In reviews of current content, lecturers either specify the discourse act and the distance of the summary from its original delivery, or fix it to directly preceding new information through a logical connector. The review of information functions as a means of checking that all students are up to date. As one lecturer explains:

> ok run through it again so that those who haven't thought it would be a good idea to make a note of this can do so (1025)

Less fleshing out of reviewed information occurs in this type compared to reviews of previous lectures. The reviews of current content are short and occur frequently throughout the lecture, ensuring that key concepts and connections have not been missed. The emphasis is on discourse structuring.

A characteristic feature of previews of current content is the diminution of the concepts previewed. Lecturers minimise the imposition of complex upcoming content through their choice of lexis and grammatical structures. Lecturers also reassuringly present problem-solving as a group journey. This sort of gentle priming may be particular to reviews of current content due to the lower chronological distance to the reception of concepts referenced.

As the previews of current lectures show, when the projected information becomes more immediate, the way in which it is characterised changes. Previews of current and previews

of future content establish listener expectations differently, through the use of different linguistic features. Previews of future content, the least commonly occurring summary type, seem to perform a specific high level information structuring function. They are concise, and intend to contextualise current information in the bigger picture of the module or course. The language is direct and less attention is paid to strategies of inclusiveness or the mitigation of imposition.

Although some slight patterns in occurrence were noted, all summary types discussed appear throughout the ELC lectures. Lecturers regularly punctuate their speech with summaries to support learning through either the reformulation of previously given information or the prediction of new information.

CHAPTER 6. HUMOUR

6.1. Introduction

6.1.1. Discourse and humour

Humour is one of the most difficult pragmatic devices for researchers to identify systematically. It can also cause particular problems of miscommunication in the delivery and reception of lectures.

This chapter demonstrates the use of pragmatic annotation for mapping the distribution, duration and specific function of humour and its nine attributed types.

6.1.2. Theories of humour

Discussions of humour and laughter in the academic context commonly draw on and extend theoretical models based on classical philosophy (superiority/hostility), cognitive science (incongruity) and physiological or psychoanalytic studies (relief). From a linguistic point of view, the theories are not necessarily exclusive; incongruity can often explain *how* humour occurs, while theories of superiority or hostility can explain *why* it occurs and the notion of relief can account for its *effect*.

Superiority/hostility theories consider humour or resulting laughter to be indicative of triumph. The teller gains success through the identification of weakness or failings in an *other*. The roots of such thinking can be traced back to Plato and Aristotle (Attardo 1994: 50, Glenn 2003: 19, Meyer 2000: 312, Partington 2006: 232). Hobbes treated laughter as a sign of a person's sense of social superiority, describing it as:

> […] caused either by some sudden act of their own, that pleaseth them; or by the apprehension of some deformed thing in another, by comparison whereof they suddenly applaud themselves. (Hobbes 1996[1651]: 43, also cited in: Morreall 1983: 5, Raskin 1985[1944]: 36).

Listeners as well as tellers can gain a feeling of superiority through recognising humour.

Superiority/hostility can also result in an abrasive form of humour, causing conflict not consensus (Martineau 1972: 103). In this case, elements of malice, disparagement or derision are central to boosting the teller's sense of power through attacks on individuals, groups, or on the self.

In terms of the mechanism of humour, incongruity theories identify particular cognitive demands. Incompatible *frames* (cf. Fillmore 1976, van Dijk 1977) of reference (or *planes* (Koestler 1989[1964]) or *schema/scripts* (Norrick 1986: 229)) are received, semantic processing stumbles over competing meanings, commonality is sought, and, if the contradiction can be resolved, new meaning arises. The absurdity of competing frames causes discord in expectations and humorous results (Attardo 1994: 1994, Ross 1998: 7).

The process is distinguished from simple habitual associative thinking. Incongruity lies in the "sudden bisociation of an idea or event with two habitually incompatible matrices" (Koestler 1989[1964]: 51), or a shift of mode, narrative, role, or register (Partington 2006: 25). Incongruity occurs at different levels of language in use (Ross 1998). At the pragmatic level in terms of Speech Act Theory, incongruous humour marks a gap in the sense and the force of an utterance (Austin 1962, Searle 1969) and flouts the co-operative principle (Grice 1975). It can also occur at the level of register by mixing styles or references (Lee 2006, Ross 1998). The subconscious process of combining incongruous planes creates an emotional reaction.

Relief theory emphasises this response, as the moment of realisation/resolution results in release. Laughter is most simply considered to be the expression of "suddenly perceived incongruity" (Schopenhauer 1966[1959]: 59). Built up tension "gushes out in laughter" (Koestler 1989[1964]: 51), which acts as a vent through which constraint is discharged (Morreall 1983: 20), or through which the forbidden or suppressed is released (Freud 1976[1905]). Successful humour and a laughter response are commonly linked (for example, Raskin 1985[1944]: 4). Physical pleasure is gained from laughter, which Kant describes as

"an effect arising from a strained expectation being suddenly reduced to nothing" (2007[1790]: 161, also cited in: Glenn 2003: 20, Morreall 1983: 16). After the resolution of incongruity, a physical response is expected.

At this level of effect, there is some agreement across frameworks that humour and laughter can modify behaviour and regulate the unacceptable. Control is explicitly identified as a function of humour and laughter (Hay 2000: 717, Martineau 1972, Stebbins 2012[1980]). Bergson describes laughter as a "social corrective" through which humiliation is intended (1899 cited in Raskin 1985[1944]: 17). Morreall (1983: 5) notes that Plato and Aristotle also thought that laughter serves as a "social corrective to get wrongdoers back into line". Hostile joking serves the purpose of satire or defence (Freud 1976[1905]: 97), and certain types of humour and laughter function as a form of aggression towards those of lower status, indicating social hierarchy. The use of humour in rhetoric also has a persuasive function (Meyer 2000: 310, Partington 2006: 226).

Within institutional discourse, particular humour types perform specific, often very different, communicative functions. They can enable rapport-building, construct in-group cohesion, mitigate conflict or model identities (Kotthoff 2007, Lee 2006, Martineau 1972, Nesi 2012a, Norrick and Spitz 2008, Partington 2006, Reershemius 2012, Stebbins 2012[1980]). By identifying where and to what extent functions occur, we can begin to better understand the dynamic of the academic lecture across cultural settings.

6.1.3. Humour across cultures

Understanding humour is considered to be one of the most complex cultural adjustments. We learn what to laugh about from our family, peers and culture (Norrick 1986: 228), so although humour as a field of investigation is regarded as central to understanding communication (Attardo 2003: 1292), it is often specific to certain cultures and attitudes and may not travel well across these boundaries (Chafe 2007: 127, Lee 2006: 49, Ross 1998: 2, Zhang 2005). In his interview data, Straker Cook (1975: 28) reported that one lecturer:

> […] complained not only of a lack of overall comprehension but a lack of awareness of interpersonal cues. Deprecation and praise, veiled criticism or irony, "off-the-cuff" and "off-the-record" comments, were all given equal weight and were likely to be quoted back at the lecturer in all seriousness.

Miscomprehension appears to be a global issue. Studies of Chinese students in New Zealand, for example, report that lecturers' humour style can be difficult to understand (Andrade 2006: 139, Holmes 2004: 299), and Chinese students attending lectures in a British University missed completely some attempts at humour, which resulted in alienation (Wang 2014). Yusoff found that misunderstanding humour contributed to problems in the socio-cultural adjustment process for international students in a Malaysian public university (2010: 38). In a report on the needs of International students in New Zealand, Butcher and McGrath (2004: 544) identify humour as an area of particular concern.

Humour is also mentioned in the support materials for international students visiting British universities. Alongside queuing and politeness, dry humour is often flagged up as a particularly British trait – one which students visiting from overseas are explicitly alerted may be baffling to them. Warnings are issued that immediate comprehension should not be expected, especially of irony and sarcasm, because "[t]he British have a unique sense of humour. It […] is definitely not cross-culturally funny at first" (University of Sheffield International Student Support 2013). Unfamiliarity is expected:

> It may take a while for you to get used to British humour. […] It also involves teasing and can take the form of picking on aspects of an individual's personality and exaggerating them in fun. Sarcasm and plays on words are also common. (University of Northampton Student Services 2012: 20)

Careful listening to this discourse feature is advised because:

> British humour is witty and self-deprecating […]. Much use is made of irony so do listen carefully for cues and do not take all comments literally. (Cardiff University Careers Service 2009: 13)

In addition to acknowledging a range of culturally-specific humour types, the advice for visiting students agrees that the discourse feature is ever-present at all levels of society, including in the academic arena.

As discussed in Chapter 1, current internationalisation drives in universities result in increased staff and student mobility. In turn, in different cultural settings, the need to understand the nature of potentially tricky pragmatic devices (such as humour) grows.

Although lexical issues related to limited vocabulary have been cited as a significant problems in lecture discourse comprehension (Flowerdew 1994: 19), studies of the reception of academic language confirm that it is not always the technical jargon that students struggle with, but more often than not getting used to the pragmatic applications of everyday language is challenging (Ehlich 1999 cited in Reershemius 2012: 856). Students may not struggle with discipline-specific technical terminology, which they tend to be familiar with from prior learning or close equivalents in their L1 (Straker Cook 1975: 27-28).

Lee (2006) notes that it is particularly the use of incongruous registers or references in humour that international students find difficult to comprehend, or completely fail to decode. Self-effacing humour is also identified as problematic for students with different cultural norms for public speaking (2006: 60, 57). Recognising and contextualising pragmatic language such as humour seems to be a particular problem for students receiving lectures that are not in their L1.

6.1.4. Humour in spoken academic discourse

Linguistic research in spoken academic discourse shows that humour is used as a pragmatic device in academic lectures and presentations. The importance of humour in lecture discourse was reflected in the pragmatic annotation of MICASE. *Humor* was originally identified as one of 25 relevant linguistic/pragmatic functions and discourse features, but

excluded from the final set of features for pragmatic annotation only because there were too many instances to encode (Maynard and Leicher 2007: 112).

Also using MICASE, Lee (2006) discussed the discoursal and rhetorical functions of humour and their impact on pedagogy in higher education. Taking a broad brush approach to the definition of humour, Lee (2006: 52) looks at the occurrence of laughter by producer or receiver in academic speech events. He ranks the humour-density of speech events by calculating the normalised number of laughs per minute / laughs per 1000 words, to create a *laughter index* (2006: 53). To determine humour categories, a sample of the more humour-dense speech events was qualitatively analysed more finely.

Nesi (2012a) looked specifically at *laughter episodes* (which describe the occurrence of laughter), largely from the cross-disciplinary BASE corpus. These episodes, which can indicate the presence of humour, were retrieved from the structural markup of laughter. They were analysed in light of notions of Politeness Theory (Brown and Levinson 1987), *face* (Goffman 1967) and social management and anxiety management. Six humour types performing a variety of different functions were identified: *lecturer-student teasing*, *lecturer error*, *lecturer self-deprecation*, *black humour*, *disparagement of out-group members*, and *register and wordplay*.

Reershemius (2012) studied humour in research presentations, drawing on a subcorpus from the Gesprochene Wissenschaftssprache kontrastiv (GeWiss) corpus of spoken academic discourse. The subcorpus comprises 1800 minutes of presentation speech divided equally between speakers from England, Germany and Poland. In the first instance of quantitative analysis, humour instances were identified based on the presence of markup for laughter (2012: 868). All data were then revisited for qualitative analysis, which looked for occurrences of humour based on instances of laughter in context. The findings revealed cultural differences in the distribution and function of humour; the British presenters

regarded humour as a means of challenging genre boundaries and applied it more frequently than their German counterparts.

These studies of humour and laughter in lectures and seminars suggest that different research environments may well breed different approaches to the use of humour.

6.2. Identifying humour

Common practice in spoken corpora is to recover humour via structural laughter markup. A laughter response is expected in general theoretical discussions of humour (Koestler 1989[1964], Raskin 1985[1944], Schopenhauer 1966[1959]). The corpus analyses of Lee (2006), Nesi (2012a) and Reershemius (2012) heavily rely on this vocal indicator. However, the conflation of laughter and humour is potentially erroneous (Glenn 2003: 18-19, Partington 2006: 231) because laughter and humour "are by no means coextensive" (Attardo 2003: 1288). Laughter may indicate other states such as embarrassment, anxiety, relief or repair (Meyer 2000: 311, Ross 1998: 1, Swales 2004: 165) and laughter tags do not reliably record instances of humour (Simpson-Vlach and Leicher 2006: 68-69).

Humour in the ELC was tested based on whether the text was *funny* in some way (whether the incongruity/relief caused amusement) *and* whether there was some sort of implication in the utterance. In the current 2014 ELC working list there are nine attributes to the humour element: bawdy, black, disparaging, ironic/sarcastic, joke, playful, self-deprecating, teasing/mock-threatening, and wordplay (as outlined in Table 3.3).

As in all pragmatic categories indexed, the annotation excludes humour episodes that make explicit reference to the artificiality of the setting, such as the camera or operator, as in:

<lecturer>there's a very old saying about lectures it is there is only one thing more boring than a lecture and it's</lecturer>
<student>a lecturer</student>
<students><vocal desc="laughter"/></student>
<lecturer>no it's not a lecturer<vocal desc="laughter"/> it's a video of a lecture </lecturer> (1029)

Such instances are a direct consequence of the data collection process, and tend to occur at the beginning or end of lectures.

The annotation also excludes humour that arises external to language, such as an unexpectedly ringing phone. In some studies unintentional humour is identified as a separate category, for example *howlers* in student writing (Ross 1998: 12). Humour in the ELC is restricted to the deliberate and the linguistic – the "humour-creating manoeuvres" (Fillmore 1994) that the lecturer undertakes through language choice (see Table 3.5).

6.3. Macro-level patterns in humour

Based on a token count, humour constitutes 2.88% of the ELC. Humour of some type is more likely to occur in the UK lectures than in the lectures from Malaysia or New Zealand; it makes up 3.53% of the UK subcorpus, 2.57% of the Malaysian subcorpus and 2.08% of the subcorpus from New Zealand. A further breakdown into types is given in Table 6.1, and normalised percentages are visualised in Figure 6.1 and Figure 6.2.

	UK		**MS**		**NZ**		**all**	
	raw tokens	*% tokens*	*raw tokens*	*% tokens*	*raw tokens*	*% tokens*	*raw tokens*	*% tokens*
bawdy	145	0.06	0	0.00	130	0.08	275	0.05
black	241	0.10	683	0.57	91	0.06	1015	0.19
disparaging	2373	0.95	189	0.16	504	0.32	3066	0.58
irony/sarcasm	1859	0.74	162	0.13	314	0.20	2335	0.44
joke	524	0.21	0	0.00	0	0.00	524	0.10
playful	2288	0.91	1355	1.13	933	0.59	4576	0.87
self-deprecating	874	0.35	245	0.20	880	0.56	1999	0.38
teasing/mock-threat	310	0.12	127	0.11	199	0.13	636	0.12
wordplay	258	0.10	333	0.28	206	0.13	797	0.15
all humour	8872	3.53	3094	2.57	3263	2.08	15229	2.88

Table 6.1: Humour types (raw token and %) across subcorpora

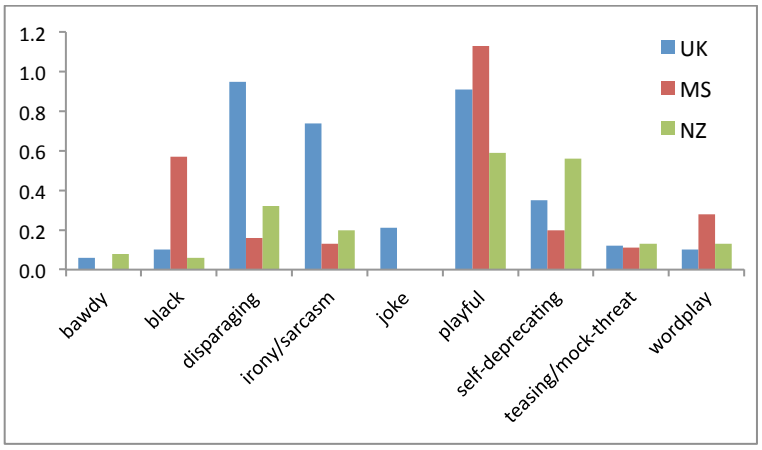


Figure 6.1: Humour types (tokens) as a % of subcorpora

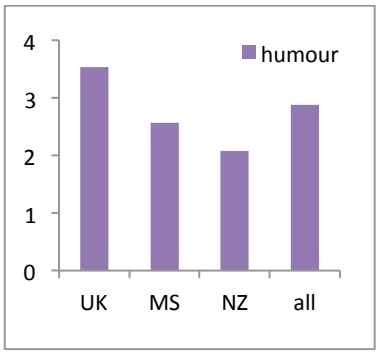


Figure 6.2: Humour (tokens) as a % of the corpus (all) and subcorpora

There are an average of 12 occurrences of some type of humour per lecture in the UK subcorpus, compared to eight in the lectures from Malaysia and four from New Zealand (see Table 6.2 and Figure 6.3). The final two positions are reversed in the sequence of the average length of each occurrence: UK: 36 tokens, NZ: 24 tokens, and MS: 18 tokens (see Table 6.2).

	UK		**MS**		**NZ**		**average**	
	instance per lecture	*length (tokens)*	*instance per lecture*	*length (tokens)*	*instance per lecture*	*length (tokens)*	*instance per lecture*	*length (tokens)*
bawdy	0.37	13	0.00	0	0.39	12	0.25	13
black	0.27	30	0.67	57	0.11	30	0.35	39
disparaging	2.40	33	0.50	21	0.54	34	1.15	29
irony/sarcasm	2.43	25	0.61	15	0.39	29	1.14	23
joke	0.13	131	0.00	0	0.00	0	0.04	131
playful	4.07	19	4.22	18	1.68	20	3.32	19
self-deprecating	1.33	22	0.56	25	0.96	33	0.95	26
teasing/mock-threat	0.50	21	0.56	13	0.18	40	0.41	24
wordplay	0.30	29	1.11	17	0.32	23	0.58	23
all humour	11.80	36	8.22	18	4.57	24	8.20	36

Table 6.2: Average occurrence (per lecture) of humour instance and average length of occurrence (tokens) by type and subcorpora

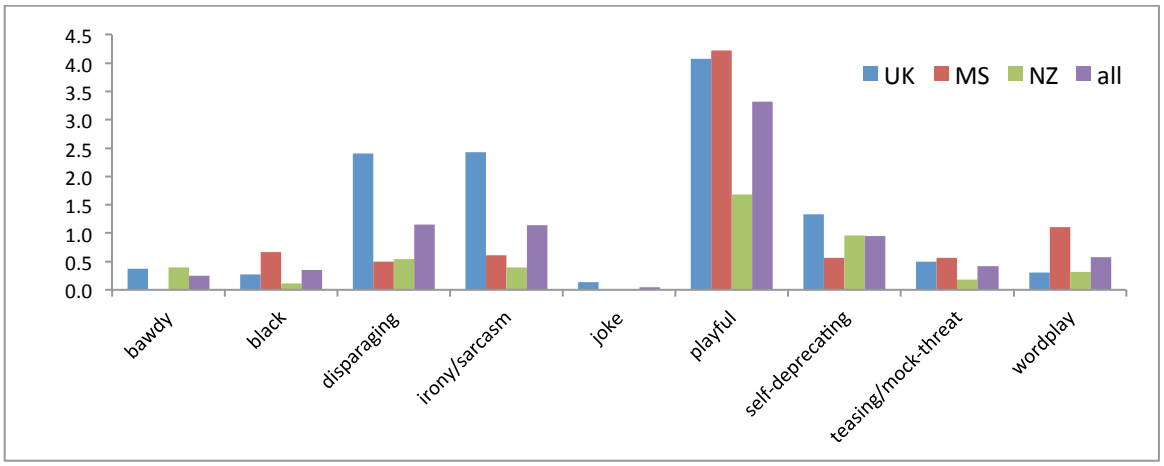


Figure 6.3: Average occurrence (per lecture) of humour types across subcorpora

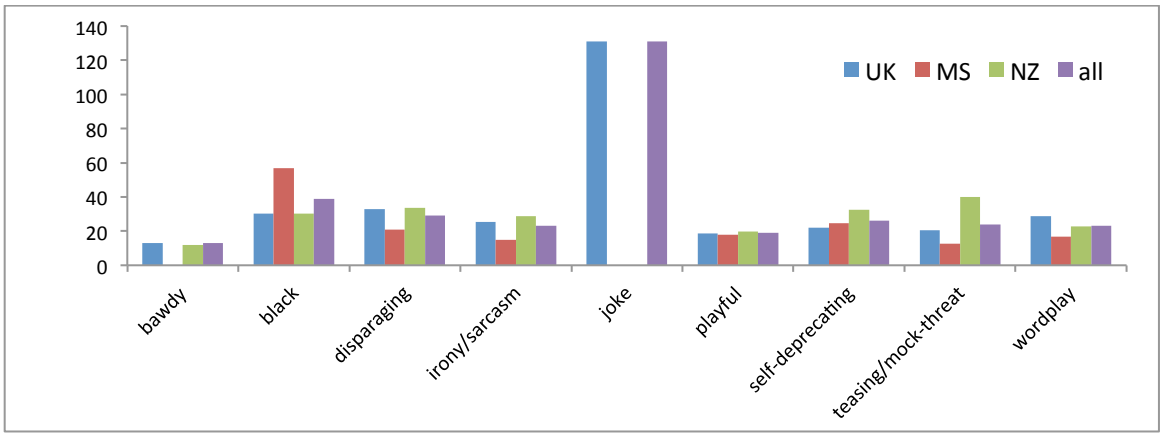
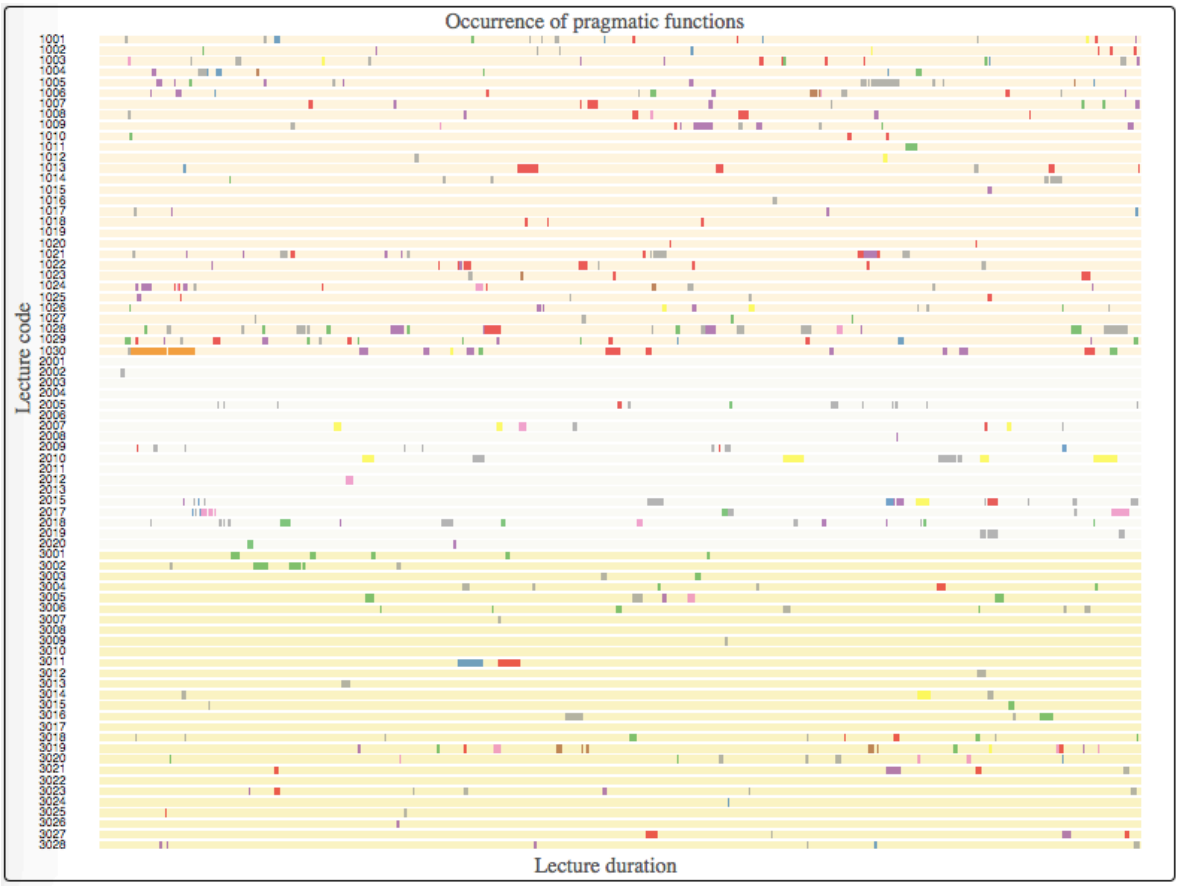


Figure 6.4: Average length of occurrences (tokens) of humour types across subcorpora

Humour is most frequent in the lectures from the UK (then Malaysia, then New Zealand) when measured by a normalised token count and normalised number of occurrences per lecture. Based on the average length per occurrence, it is most common in the lectures from the UK, then New Zealand and then Malaysia.

Based on a normalised token count, playful humour is the most popular type across subcorpora, and occurs most often in the Malaysian lectures. Black humour and wordplay

are also most frequent in the Malaysian subcorpus, where no bawdy humour was identified. Disparaging and ironic/sarcastic types are significantly more common in the UK lectures, and jokes are only found in the UK lectures. Teasing/mock-threat has almost equally low occurrence across subcorpora. Self-deprecation is most frequent in the New Zealand lectures. Where and for how long each humour episode occurs is plotted in Figure 6.5.



KEY: **bawdy** | **black** | **disparaging** | **irony/sarcasm** | **joke** | **playful** | **self-deprecating** | **teasing/mock-threat** | **wordplay**

Figure 6.5: Occurrence and duration of humour types

Table 6.3 further breaks down the raw frequency and (pmw) occurrence of select pronouns across humour types and non-humour, and comparative (pmw) data are rendered in Figure 6.6. This enables comparison of usage at the more nuanced level of attributes.

		i	**he**	**him**	**we**	**you**	**they**
bawdy	*raw freq*	13	0	0	6	9	1
	pmw	47273	0	0	21818	32727	3636
black	*raw freq*	31	1	0	9	13	6
	pmw	30542	985	0	8867	12808	5911
disparaging	*raw freq*	87	9	5	21	117	21
	pmw	28376	2935	1631	6849	38160	6849
irony/sarcasm	*raw freq*	60	6	0	34	90	6
	pmw	25696	2570	0	14561	38544	2570
joke	*raw freq*	1	14	3	6	7	0
	pmw	1908	26718	5725	11450	13359	0
playful	*raw freq*	143	13	2	33	201	19
	pmw	31250	2841	437	7212	43925	4152
self-deprecating	*raw freq*	108	3	0	12	67	10
	pmw	54027	1501	0	6003	33517	5003
teasing/mock-threat	*raw freq*	30	1	0	37	2	1
	pmw	47170	1572	0	58176	3145	1572
wordplay	*raw freq*	23	1	1	6	35	10
	pmw	28858	1255	1255	7528	43915	12547
all humour	*raw freq*	483	48	11	158	532	73
	pmw	31716	3152	722	10375	34933	4793
non-humour	*raw freq*	7292	216	25	8027	15045	2022
	pmw	13807	409	47	15198	28486	3828

Table 6.3: Pronouns (raw frequency and normalised pmw) *i*, *he*, *him*, *we*, *you*, and *they* across humour types

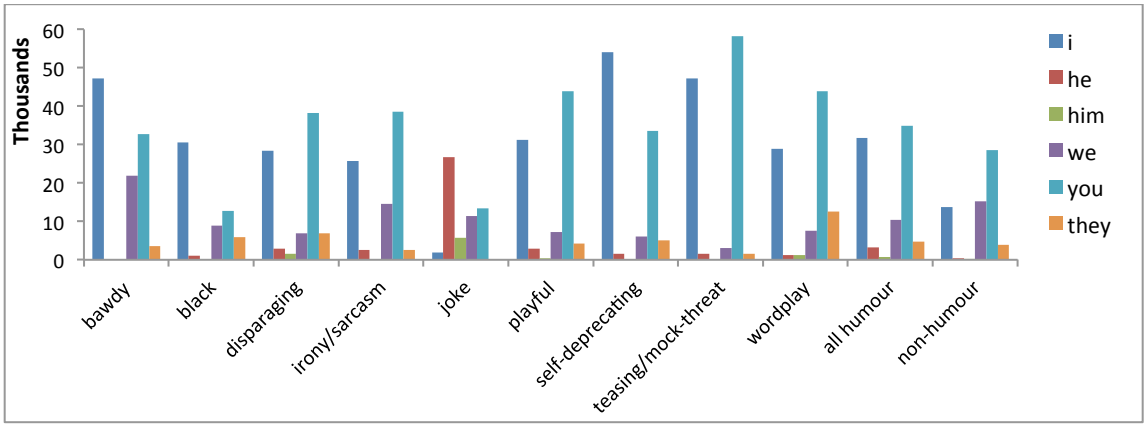


Figure 6.6: Pronouns (pmw) *i*, *he*, *him*, *we*, *you*, and *they* across humour types

Overall, *you* remains the most common pronoun in most humour types (especially in teasing/mock-threat, irony/sarcasm, disparaging, playful and wordplay) and also in non-humour when calculated by relative frequency. The use of *I* is second most frequent across all humour, and is relatively more frequent in bawdy, self-deprecating, and mock-threat types. The relative frequency of *he* is noticeable in jokes. There is a marked difference in the occurrence of *I* in all humour (31716 pmw) compared to non-humour (13807 pmw). As the 3-gram analysis (Table 6.4 and Figure 6.7) shows, *I'm going* [*to*] and *I don't* [*know*] co-occur relatively often in humour. Overall, however, the language of humour is not highly formulaic like summary; the STTR results point rather to its lexical density (see Table 4.4).

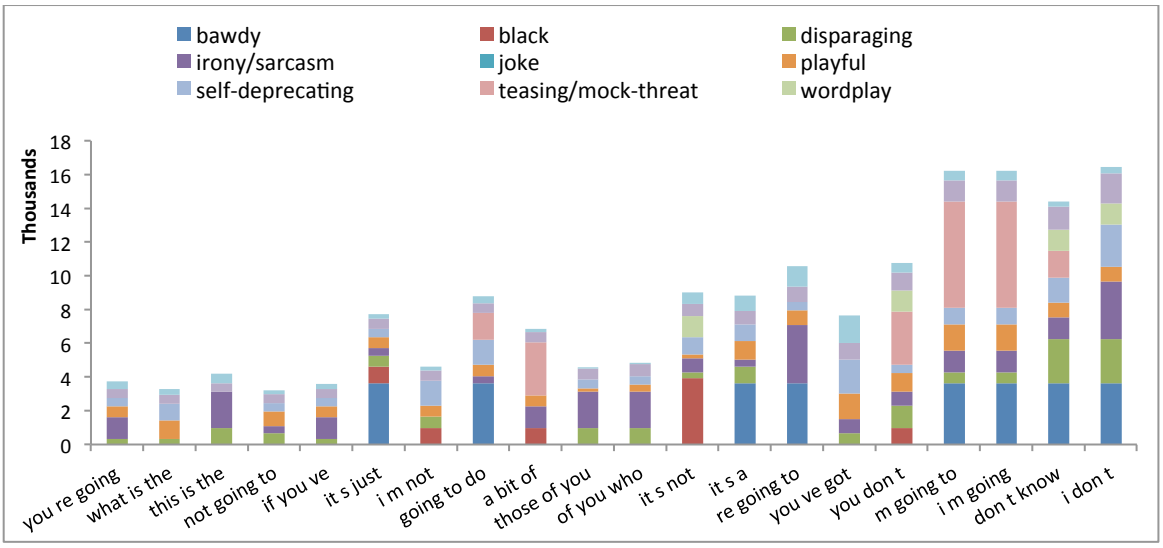


Figure 6.7: 20 most common 3-grams (pmw) in humour, humour types and non-humour

		bawdy	black	disparaging	Irony/ sarcasm	joke	playful	self-deprecating	teasing/mock-threat	wordplay	all humour	non-humour
i don t	*raw freq*	1	0	8	8	0	4	5	0	1	27	205
	pmw	3636	0	261	3426	0	874	2501	0	1021	1773	388
don t know	*raw freq*	1	0	8	3	0	4	3	1	1	21	150
	pmw	3636	0	261	1285	0	874	1501	1572	1021	1379	284
i m going	*raw freq*	1	0	2	3	0	7	2	4	0	19	306
	pmw	3636	0	65	1285	0	1530	1001	6289	0	1248	579
m going to	*raw freq*	1	0	2	3	0	7	2	4	0	19	300
	pmw	3636	0	65	1285	0	1530	1001	6289	0	1248	568
you don t	*raw freq*	0	1	4	2	0	5	1	2	1	16	306
	pmw	0	985	130	857	0	1093	500	3145	1021	1051	579
you ve got	*raw freq*	0	0	2	2	0	7	4	0	0	15	854
	pmw	0	0	65	857	0	1530	2001	0	0	985	1617
re going to	*raw freq*	1	0	0	8	0	4	1	0	0	14	642
	pmw	3636	0	0	3426	0	874	500	0	0	919	1216
it s a	*raw freq*	1	0	3	1	0	5	2	0	0	12	475
	pmw	3636	0	98	428	0	1093	1001	0	0	788	899
it s not	*raw freq*	0	4	1	2	0	1	2	0	1	11	373
	pmw	0	3941	33	857	0	219	1001	0	1021	722	706
of you who	*raw freq*	0	0	3	5	0	2	1	0	0	11	35
	pmw	0	0	98	2141	0	437	500	0	0	722	66
those of you	*raw freq*	0	0	3	5	0	1	1	0	0	10	47
	pmw	0	0	98	2141	0	219	500	0	0	657	89
a bit of	*raw freq*	0	1	0	3	0	3	0	2	0	9	106
	pmw	0	985	0	1285	0	656	0	3145	0	591	201
going to do	*raw freq*	1	0	0	1	0	3	3	1	0	9	209
	pmw	3636	0	0	428	0	656	1501	1572	0	591	396
i m not	*raw freq*	0	1	2	0	0	3	3	0	0	9	122
	pmw	0	985	65	0	0	656	1501	0	0	591	231
it s just	*raw freq*	1	1	2	1	0	3	1	0	0	9	150
	pmw	3636	985	65	428	0	656	500	0	0	591	284
if you ve	*raw freq*	0	0	1	3	0	3	1	0	0	8	156
	pmw	0	0	33	1285	0	656	500	0	0	525	295
not going to	*raw freq*	0	0	2	1	0	4	1	0	0	8	114
	pmw	0	0	65	428	0	874	500	0	0	525	216
this is the	*raw freq*	0	0	3	5	0	0	0	0	0	8	290
	pmw	0	0	98	2141	0	0	0	0	0	525	549
what is the	*raw freq*	0	0	1	0	0	5	2	0	0	8	172
	pmw	0	0	33	0	0	1093	1001	0	0	525	326
you re going	*raw freq*	0	0	1	3	0	3	1	0	0	8	237
	pmw	0	0	33	1285	0	656	500	0	0	525	449

Table 6.4: Top 20 3-grams (raw frequency and pmw) in humour, humour types and non-humour

6.4. Humour and laughter

The ELC data supports the hypothesis that not all humour elicits laughter and not all laughter indicates humour; both phenomena can occur separately or in conjunction, but the relationship is not dependent.

Non-humour related laughter was identified in 78 instances in the ELC. Lecturer error is a common cause, but attempts to save face through humour are not always made, and so laughter functions only as tension release. Recurring themes include calculation errors (1006, 2005, 3007) and mistakes with acronyms (2007). Student discomfort also triggers laughter as a means of tension release rather than as a response to humour: poor attendance (1014), failure to grasp concepts (1008, 3014), surprise (1021) and embarrassment at incorrect coursework (1030) all lead to laughter. The failure of technical equipment causes laughter outside the remit of the ELC definition of humour (1012, 3013), and there were a number of non-linguistic humorous occurrences such as ringing phones (1005, 3005), dropped objects (1024, 2008) and, for one unfortunate lecturer, electric shocks (3001 – multiple).

There was some unexplained laughter where the cause could not be recovered from the video data. Some events happened off camera (1010, 1014, 1015, 3028) or were inaudible (1020, 1025, 2005, 3005), and I could not ascertain whether the resulting laughter was triggered by humour. In other cases the motivation for laughter was clear, but was not categorised as humour because it was not deliberate. The use of certain terminology, such as “wide flanges” (1026) and “lubrication” (3019), for example, provoked a juvenile response, as did reference to the use of urine as first aid (2010). A notable laughter response was received in one lecture from New Zealand:

<lecturer>[…] I would get my hammer ready because I must press fit the bush in the housing so I I whack<event desc="mimics hammering action"/> that in there and then I do expect when I put<event desc="points to the board"/> my shaft in the bush it should just</lecturer>
<students><vocal desc="laughter"/></students>
<lecturer>slightly slide in<vocal desc="sigh"/>oh man</lecturer>
<students><vocal desc="laughter"/></students>
<lecturer>[*name removed*] say something</lecturer>
<students><vocal desc="laughter"/></students>
<researcher>it's the language</researcher>
<lecturer>it's the language yeah exactly yeah ambiguity of language</lecturer> (3019)

This example contains both laughter and two competing planes: engineering equipment/sexual innuendo (bush, housing, shaft). However there is no deliberate incongruity; the long-suffering lecturer repeatedly tries to use the terminology seriously. With increasing frustration he eventually admonishes the students:

> you guys can't go on laughing forever four years when we're talking about shafts it's just what they're called (3019)

This elicits even more laughter.

The laughter response (LR) to occurrences of humour in the ELC neatly equates to 50% (317/630). In Table 6.5, LR is defined as the identification of at least one vocal description of laughter from either speaker or audience during the episode. NLR means that there is no vocal description of laughter (no laughter response).

		all		UK		MS		NZ	
		LR	NLR	LR	NLR	LR	NLR	LR	NLR
bawdy	raw	19	3	8	3	0	0	11	0
	%	86	14	73	27	0	0	100	0
black	raw	12	11	2	6	7	5	3	0
	%	52	48	25	75	58	42	100	0
disparaging	raw	37	59	25	47	7	2	5	10
	%	39	61	35	65	78	22	33	67
irony/sarcasm	raw	20	75	10	63	7	4	3	8
	%	21	79	14	86	64	36	27	73
joke	raw	0	4	0	4	0	0	0	0
	%	0	100	0	100	0	0	0	0
playful	raw	169	76	77	45	58	18	34	13
	%	69	31	63	37	76	24	72	28
self-deprecating	raw	32	45	14	26	6	4	12	15
	%	42	58	35	65	60	40	44	56
teasing/mock-threat	raw	14	16	6	9	6	4	2	3
	%	47	53	40	60	60	40	40	60
wordplay	raw	14	24	2	7	8	12	4	5
	%	37	63	22	78	40	60	44	56
all humour	raw	317	313	144	210	99	49	74	54
	%	50	50	41	59	67	33	58	42

Table 6.5: Occurrence (raw frequency and %) of humour episodes accompanied by speaker/audience laughter (LR) or not (NLR)

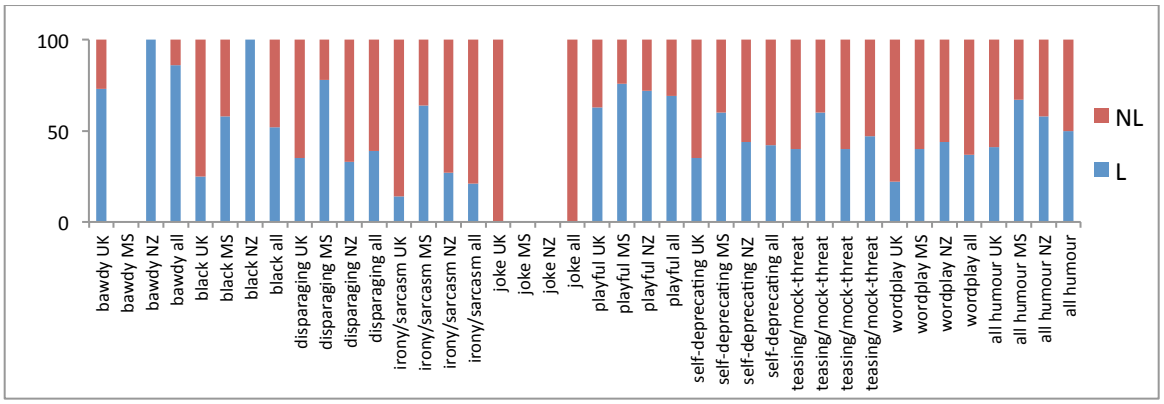


Figure 6.8: Normalised % (per occurrence) LR and NLR across humour types, humour, and subcorpora

As rendered for comparison in Figure 6.8, different types of humour elicit different LR rates, which again differ across cultural components. Self-deprecating humour, for example, does not elicit laughter in three quarters of occurrences overall, particularly in lectures from the UK (NLR=65%). Irony/sarcasm is generally met with even less laughter: 79% NLR overall and

up to 86% NLR in the UK lectures. The Malaysian lectures, however, show an LR in two thirds of occurrences of irony/sarcasm.

The same pattern emerges in disparaging humour: overall there is NLR in two thirds of cases, but the Malaysian lectures reverse this and show a 78% LR. Bawdy, joke, teasing/mock-threatening and wordplay types all show a general pattern of NLR in over half of occurrences. Significant variation exists, however, in the case of playful humour, which has an overall LR of 69%. The lowest LR rate was recorded in the UK lectures (37%) and the highest response is again in the Malaysian subcorpus where laughter accompanied humour in around three quarters of all occurrences (76%), as it does in the New Zealand lectures (72%). Laughter and humour, then, are not co-extensive in the ELC data, and humour type and subcorpus membership affects the LR rate.

6.5. Humour types

6.5.1. Playful humour

In the ELC the most common humour type identified is playful, which also elicits the highest LR. Conflation of humour and joking is common, even by the ELC lecturers who refer to "a little bit of a joke" (1001) or add "just joking" (2001). Both instances are actually identified as playful humour, which I define as good-natured banter intended only to amuse, entertain, and establish good rapport.

An overview of the tokens that are most salient to this humour type is given in Table 6.10. Following the keyword trend in humour (see Table 4.6), the pronouns *I* and *he* are particularly salient to playful humour. Because this humour type is not targeted, lecturers tend to draw on their own experience, as in:

> ah things haven't changed that much I mean if I told you some of the things that we got up to but I wouldn't want to give you ideas you see not these days (1030)

Or make (usually neutral) reference to external figures in setting-up the humour, as in:

> I shouldn't wind people up [*name removed*]'s a very nice guy and I'm sure he won't set you a nasty exam (1014)

People (Princess Diana), places (Warwick, Windsor, a castle and a highway), and food stuffs (Mars bars, steak) feature as key (see Table 6.6). The function of these seemingly non-discipline-specific references is to give colour to moments of uncritical entertainment.

positive			**negative**		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
i	143	67.41	the	158	57.30
cafe	6	50.93	we	33	26.48
he	22	50.17	um	7	20.73
windsor	6	47.69	and	65	20.72
mars	4	37.77	is	67	17.35
warwick	4	37.77	of	58	16.34
shirt	5	33.44	point	5	16.21
eat	4	32.78	so	42	14.97
me	19	30.28	three	2	14.79
camera	4	30.16	which	3	14.32
diana	3	28.33	four	2	8.67
steak	3	28.33	concrete	1	7.99
tails	3	28.33	value	1	7.88
highway	4	28.26			
playing	4	28.26			
you	201	27.51			
name	8	26.45			
castle	4	24.44			

Table 6.6: 20 most highly ranked positive and negative keywords in playful humour

Playful humour is entertaining and establishes a sense of fellowship within the lecture hall. The following example of playful humour is taken from a civil engineering lecture from the UK in which the lecturer challenges students to successfully apply a recently discussed procedure:

> <lecturer>you can't go wrong if you follow the method we'll say fun-sized Mars bar if you hav- if you write something down but get the wrong answer family pack Mars bar double whammy if you don't even try it and because no one complained this is a formal wager</lecturer>
> <students><vocal desc="laughter"/></students>(1005)

The element of play is introduced as the lecture theatre becomes a gambling arena. Elements of mock-threat and self-deprecation are detectable but play is the primary

intention. This type of humour is constructive and cohesive, and builds the positive face of the group. Although the play is mildly coercive, there is no explicit threat. A similar pattern exists in most instances of playful humour, for example:

> anyone give me a common problem with timber floors they squeak you know why they squeak is it the little rodents inside them (1009)

Incongruity exists in unexpected script (gambling/small mammals), but there is no punchline, and no resolution.

6.5.2. Joke

Unlike playful humour, jokes in the ELC are narrowly identified as a structured oral genre with set-up and punchline (Kotthoff 2007: 268, Raskin 1985[1944]: 99) and function as a time-out from the serious business of academic content. Incongruity accords to the disjunctor model in which one coherent script is developed, and a second, disturbing, script is then activated by a disjunctor. Coherence is restored by a connector (Attardo 1994: 95). For example:

> at ten o'clock the second lecturer came and he said you know seeing these TV sets reminds me of that um story about ah how we're heading towards an automatic world you know the people are in the plane and the voice comes over the tannoy ah this is a fully automated flight ah we took off ten minutes ago and there is not a single crew member on board this plane everything is automatic but all the systems have been designed to be sophisticated and um fail safe so be assured that nothing can go wrong uh go wrong uh go wrong uh go wrong (1030)

The two incompatible scripts are the achievement/danger (of automation). The setup conjures a sophisticated and failsafe automatic world. The punchline begins on the second repetition of "go wrong", which activates a script for faulty machinery stuck on a loop and the potential disaster heralded by onboard announcements. The abrupt switch forces reconsideration of the verity of information taken at face value. Bisociatively the listener must process competing scripts and arrive at the new meaning that fully automated environments cannot be trusted.

Jokes are much longer in average token length than any other category, which reflects the complexity of their required structural elements. More investment is required in terms of scripting jokes, for example compared to throwaway sarcastic comments. Telling jokes takes time and requires the attention of the audience. In the lecture theatre their operation is inclusive. The payoff is in the punchline that results from extended structural development.

Keyword data are not given for jokes as quantitative analysis is not overly revealing: all jokes occur early in a single lecture (1030) almost in an unbroken chain (as shown in the visualisation of occurrence and duration in Figure 6.5). Most telling, perhaps, is the lack of jokes in the corpus. Partly this flags up the role of lecturer idiosyncrasies. The lecturer responsible for all identified examples appears to deliver a chain of scripted jokes as a warm-up to the main event, the lecture. Normally – as shown in 5.2.2 – this opening part of the lecture would contain some type of reviewing or previewing discourse to situate upcoming content. In this case, humour is substituted.

6.5.3. Irony/sarcasm

The main functions of sarcasm/irony are behaviour modification and self-aggrandisement. Irony flouts Grice's first maxim: "[d]o not say what you know to be false" (1975: 46). In doing so its purpose is to deliver an implicit message. Saying the opposite of what is meant masks the intention to criticise or self-promote. The presence of at least one of the conditions of failed expectation, pragmatic insincerity, negative tension, and presence of a victim (Campbell and Katz 2012: 459) are likely but not prerequisite in the ELC examples.

As the advice literature to visiting students (outlined in 6.1.3) predicts, a disproportionate amount of this humour type occurs in the UK subcorpus (UK: 0.74%, MS: 0.13%, NZ: 0.20%). The visualisations of irony/sarcasm (Figure 6.9 and Figure 6.10) illustrate this preference in the lectures from the UK.

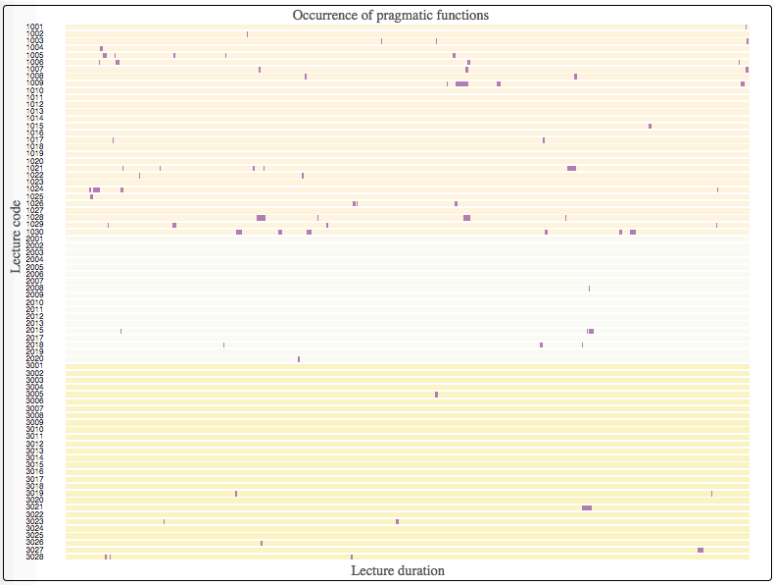


Figure 6.9: Occurrence and duration of irony/sarcasm per ELC lecture

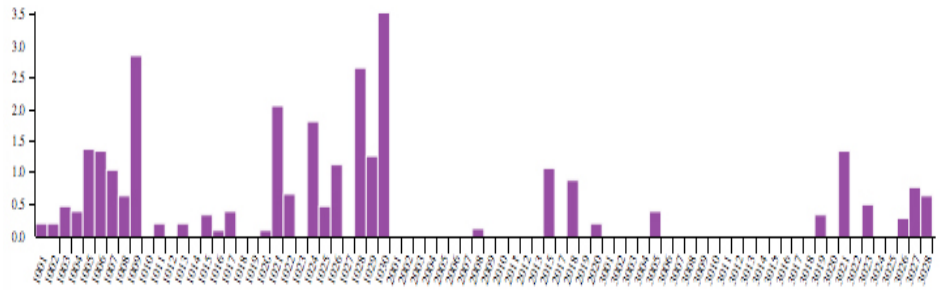


Figure 6.10: Occurrence per lecture (normalised %) of irony/sarcasm

Delivery is largely deadpan, and the LR return is the lowest of all types at 21% overall, and only 14% in the UK corpus (see Table 6.5). Detecting irony/sarcasm relies on recognition of insincerity in pragmatic meaning.

A noticeable lack of pronouns is evident in the keyword analysis compared to other types (see Table 6.7). The target of the humour is not made explicit because its operation is through evaluation of events not people. Inherently positive tropes such as “fun”, “excited”,

“yes”, “interesting” and “hurray” rank highly as key. In this context, however, they function as negatives. A “horrible mess” of an equation is presented as something that students are “going to have a lot of fun” with (3021), students are told not to get “too excited” by an upcoming experiment, one lecturer thrice dryly comments that the session will be “interesting” following repeated interruptions (1021), and “hurray” marks the end of another lecture after the lecturer has sarcastically thanked a student for yawning (1001). The sarcasm/irony functions through an emphasis by contrast: the insertion of a positive semantic context to indicate negativity.

positive			**negative**		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
who	16	50.02	is	26	18.91
fun	5	37.62	so	15	16.97
rubric	3	32.31	what	5	12.70
hands	5	29.94	one	7	11.68
megapascals	5	29.02	two	4	8.11
exam	8	28.29	then	2	8.01
twiddling	3	27.82	but	3	7.50
disappears	3	25.60	and	37	7.22
excited	3	24.02	point	3	7.18
anyone	6	22.40			
yes	7	22.21			
ellipse	2	21.54			
foyer	2	21.54			
hurray	2	21.54			
ago	5	20.99			
portfolio	4	18.04			
interesting	4	17.82			
tankers	2	17.73			
today	7	17.67			

Table 6.7: 20 most highly ranked positive and negative keywords in irony/sarcasm

Seeming affirmations also operate by contrast. The keyword “yes” is commonly inserted by lecturers to denote the absence of agreement, as in:

> I'm going to take silence as being yes agreed (1003)
>
> y- yes would be a good response (1007)
>
> I'll take that as yes (1024)

The term “yeah” is commonly used to opposite effect, as in:

> no question oh very good yeah (2008)
>
> does not make sense but hey this is out of the [*name removed*] book must be right yeah (3019)

In the example from 3019, "hey" functions as a mid-utterance marker of sarcasm. The common occurrence of *hey* in non-humour is within reported speech, as in:

> you need to go back to your admixture supplier and say hey guys I've got a problem can you solve it (1013)

As part of irony/sarcasm, the marker normally accompanies a gesture of resignation (such as eye rolling or hand flailing). Like disparaging humour, the intention is to belittle. The distinction is through mode of operation.

References to thumb "twiddling" (1003, 1005) and the ineptitude of "professionals at work" (1026) are also part of the deprecating vocabulary that characterises this humour type. It is in descriptions of reactions, not people, that sarcasm and irony are most commonly employed.

Sometimes the incongruity in irony/sarcasm is obvious, as in:

> the formula for calculating bending stress M over Z don't worry it's only two weeks ago we were doing it I don't find it embarrassing at all you shouldn't (1009)
>
> there's delta and there's rectangular and obviously those words are quite similar so it's easy to er be confused when you look at the words (1029)

Only very occasionally is irony/sarcasm explicitly signalled, however, for example by the use of *not* in "it's really important I can tell you're very excited to do this group work *not*" (emphasis added. 3028) and the metastatement at the end of "the code gives us a very useful set of equations *I say useful cynically*" (emphasis added. 1020).

Generally irony is easy to misinterpret because there is no change in tone or facial indicator to signal that the meaning is not literal. Moreover in the UK lectures 90% of instances of

irony/sarcasm did not receive a laughter response (Table 6.5). From context, we can surmise that the function of ironic utterances such as "we have been working hard on those questions in the book haven't we" (1002), "whenever you're ready thank you" (1006), "no question oh very good yeah" (2008) is to correct behaviour. The lecturer is really telling his audience to *work harder, pay attention, ask questions*.

Aside from the student audience, irony is also used to mock other professionals, as in:

> every time there's an earthquake in the world lots of engineers fly out there on gallivanting holidays in all the distraught local population taking photos of how their structures have failed (1026)

The implication of questionable ethics through irony elevates the lecturer above his peers.

To a lesser extent, irony/sarcasm is also used to establish solidarity, especially through implied syllabus critique. Ironic descriptions of current lecture content as "the sexy stuff […] the good stuff" (1022), and statements such as "today you're going to get another portfolio question from me whoohoo" (1005) re-position the lecturer from authority figure to sympathiser.

6.5.4. Teasing/mock-threat

Nesi (2012a: 88) identifies lecturer-student teasing either as a face-threatening act that challenges competence or as flattery that rewards *typical* student behaviour: drinking and hangovers, over-sleeping, a lack of hard work. For Eisenberg (1986), teasing is a device used by adults to either control or have fun with children. Similar patterns exist in the ELC where teasing a specific audience member both confirms their adherence to the *typical* student script and also admonishes this as the lecturer assumes the role of impatient parent:

> I can sort of sense I can see that you're drifting away people are not listening very much so I'm going to ask you some questions an try and wake you up I'll ask you a question and then I'll wait and I'll pick on somebody who looks most sleepy the most sleepy one I can see that's the one that's going to get asked the question today […] that guy over there who's got the ah T-shirt on (3011)

Teasing tends to take the form of gentle mockery of an audience member or the audience, whereas mock-threat manifests in exaggerated threats made in jest to underline a requirement or concept.

An analysis of keywords indicates that pronoun usage is particularly significant in this humour type. Table 6.8 shows that *we* is most key in non-teasing/mock-threatening language and that *I* is more salient to this type.

positive			**negative**		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
stir	4	39.90	we	2	9.90
riding	3	35.51	the	21	9.75
sleepy	3	35.51			
bike	3	33.28			
i	30	28.91			
sadistic	2	26.67			
tendencies	2	26.67			
wakey	2	26.67			
my	9	25.98			
wake	2	21.13			
pain	2	19.05			
ask	4	17.94			
quiz	2	16.26			
question	6	15.91			
no	7	15.55			
m	9	14.98			
athletic	1	13.34			
baseball	1	13.34			
bat	1	13.34			

Table 6.8: 20 most highly ranked positive and negative keywords in teasing/mock-threatening

A comparison of relative occurrence (see Figure 6.6) additionally reveals that: *we* is three times less likely to occur in teasing/mock-threat compared to its average occurrence across all humour types (3145 vs. 10375 pmw), *you* is much more likely to occur in teasing/mock-threat (58176 vs. 34933 pmw), and so is *I* (47170 vs. 31716 pmw). Particularly in this type of humour, then, lecturers use *you* more than in non-humour and more than in other humour types, and *we* a lot less.

Pronoun choice can be explained by the targeted nature of teasing/mock-threat. Teasing can be closely associated with play through features such as mock challenge, insult through

the detraction of valued abilities, and exaggeration. In the ELC it is differentiated from gentler playful humour if an intended victim/group and some element of exaggerated threat or punishment exists. A common formula is: conditional action -> consequence. For example:

> if you're struggling getting any of those please get my help I'll come and bring a stick (1001).

Perhaps the most colourful exemplification of this type occurs in one of the UK lectures:

> kilonewton dot M if I see you write kilonewton slash M I'm going to come round your house with a baseball bat and break your fingers because that is not the unit of moment that is the unit of load per metre length do not confuse the two (1002)

This type of humour also occurs in response to non-hypothetical situations, such as:

> your first C-A-L tutorial is due next Monday except for the person who's phone's going off cause they're not gonna be able to sit down for about a month (1004)

The mock-threat of physical punishment in response to academic or behavioural failings is a recurring theme in the UK lectures.

In terms of function, behaviour modification is usually the aim if the recipient of the teasing is present (Eisenberg 1986: 185, Kotthoff 2007: 275). In the ELC the threat is almost always direct and not overheard, as humour is delivered through monologic speech and the lecturer rarely co-opts student conspirators.

Most common is the group-threat in which the message is applied to the entire audience, or to sections of it. The message in this type is always aimed at behaviour modification, as in:

> so any any thoughts on that before I inflict a bit of pain on you (1017)

> if you give me the correct value but it has a plus sign then all my sadistic tendencies come out and I will give you nil points (1029)

Teasing/mock-threatening offers an alternative to reprimand as a means of achieving participation.

The dominance of *you* and the targeted nature of the incidences of teasing and mock-threat (aimed both at individual students and at the group) also have the effect of rapport-building. The rudeness of the attacks can be understood as a compliment; singling out individual students or a particular group can be understood as a mark of intimacy.

6.5.5. Disparaging humour

Disparaging humour is linked to power. In the BASE corpus, the disparagement of out-group members is identified as a particular form of humour; a form which encourages bonding in the lecture theatre and reinforces in-group identity (Nesi 2012a: 86).

In the ELC, lecturers categorise an *other* outside the lecture theatre who is the butt of disparagement. The use of *he* and *him* is important, as the keyword analysis indicates (Table 6.9). These *others* are largely male, such as the textbook writer who is ridiculed because "he just can't type" (1007) and the truck mixer driver who is among those who "aren't noted for their care and attention" (1013). When independently calculating the volume or timing of the addition of superplasticiser to concrete mix, premeasured slow-release capsules should be given. The lecturer explains that this is because:

> it would a lot easier rather than have th- the driver measuring things out and everything else just to give him six of these little things saying five minutes before you discharge throw them in um there's only six he can count to six he can get em in right and er so you're ok um or more I say he can be bothered to count to six i- it's all a matter of motivation (1030)

The demarcation of *he* and *him* within this type of humour is placed in stark contrast with *we*, the most negative keyword. A specified target is required to achieve superiority in this type of humour.

positive			negative		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
who	17	46.42	we	21	20.07
t	48	33.92	the	129	18.90
i	87	31.76	is	39	18.58
he	14	30.71	so	26	12.96
wallet	3	30.66	point	3	12.08
driver	4	28.57	when	2	7.46
mouldings	3	26.17	beam	1	6.96
him	5	24.36	that	53	6.78
me	14	24.35			
typo	3	23.95			
admit	3	22.37			
don	21	21.04			
attitude	2	20.44			
consistently	2	20.44			
discount	2	20.44			
games	2	20.44			
jumble	2	20.44			
miracle	2	20.44			
pois (points)	2	20.44			

Table 6.9: 20 most highly ranked positive and negative keywords in disparaging humour

The ELC definition of disparaging humour, however, is most commonly applied in the form of attacks on in-group members. It is distinguished from teasing by the lack of threat and strongly linked to establishing (lecturer) position and superiority in the field of engineering rather than academia:

> task three was very badly done even [*name removed*] didn't get the right answer and I would like to explain to him why he didn't get the right answer the rest of you are allowed to listen a lot of you won't understand what I'm saying but what else can I do here (1030)

The general attack on the audience evokes the script of the unintelligent or lazy students, those beyond help. Even the best student must be addressed and his flaws explained. Humour lies in the outrageousness of the implication. Continuing the theme, the same lecturer explains that:

> um you gave me the four equations that you had worked out and then you gave me the four values and they were right but unfortunately two of the equations were what we in engineering would refer to as crap (1030)

The "we in engineering" refers to the wider field outside the lecturer theatre, where mistakes are not tolerated. There is little pedagogic value to the identification of a mistake in terms of correcting it, but the incident may be memorable. The lecturer departs so far from the script of nurturing educator that the shock of his rudeness causes amusement.

This seemingly rough form of disparagement is not uncommon, especially in the UK lectures. Disparagement in general is around three times as likely to occur in the UK subcorpus. Students particularly are admonished for crimes such as giving "stupid" answers (1022), mocked for "brown-nosing" (1029) and for being "a bunch of dippy nonces" (1002). Even pre-emptive disparagement can be found:

> I was waiting for someone to ask if I would kindly derive these two equations and um my answer to that was going to be er A we can't spare the time B you wouldn't understand it so there really isn't any point (1029)

Yet if there is an element of seriousness in the *tough love* approach, it seems that the lecturers are playing out the script of site humour – as the abrasive foreman – rather than trying to gain status over the students. The same kind of low register references were evident in the mock-threat humour of UK lectures. When disparagement is directed to others outside the lecture theatre, the level of aggression is maintained in references to other nations (1012, 1013) or other members of staff - one of whom is humorously described as a "wicked man" (1014).

6.5.6. Bawdy humour

Bawdy humour in the ELC also accords with notions of mixed or inappropriate lexis/register, which juxtaposes the high and the low (Fillmore 1994, Lee 2006, Nesi and Ahmad 2009, Nesi 2012a). It relates specifically to the vulgar or lewd (direct or implied), often referencing sex, as particularly indicated by the first five positive keyword entries (see Table 6.10). Bawdy humour constitutes only 0.08% in lectures from New Zealand and 0.06% in those from the UK, with no representation in the Malaysian subcorpus.

positive			negative		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
naked	3	40.51			
language	3	30.98			
bush	2	22.37			
shaft	3	22.01			
pert	2	21.64			
review	2	21.02			
yeah	8	20.65			
evaluation	2	20.01			
trust	2	18.86			
programme	2	17.97			
technique	2	17.24			
ambiguity	1	15.00			
arriving	1	15.00			
blurry	1	15.00			
contrast	1	15.00			
inserted	1	15.00			
obsessed	1	15.00			
orifice	1	15.00			
whack	1	15.00			

Table 6.10: 20 most highly ranked positive and negative keywords in bawdy humour

Much like disparaging and mock-threatening humour from the UK, the bawdy type seems to echo a type of site humour and relates to behaviour that is out of place in the university setting.

Ringing phones make the student particularly fair game for the reception of lewd rebukes, as in:

> are we done switched off or inserted on vibrate (1004)
>
> it's like it's trying to say please insert me in orifice (1005)

Instances of bawdy humour are also often unprompted, for example:

> <student>I need glasses though they're on order</student>
> <lecturer>it's a bit blurry because of th- there's not enough contrast with the light on but</lecturer>
> <student>but I'm just totally blind</student>
> <lecturer>ok we won't go into the possible causes for that we'll just let it lie</lecturer> (1024)

The script shifts seamlessly from housekeeping to masturbation. Jokes often have an element of shock – the taboo or personal attacks – which may cause offence (Ross 1998: 4). The exchange in 1024 above echoes the set-up and punchline of a joke format, made at the expense of the unknowing co-constructor. There is also an element of teasing as the lecturer directs the humorous attack at a single victim.

6.5.7. Black humour

Nesi (2012a: 88) identifies black humour as a means through which the taboo and embarrassing can be dealt with, particularly in the life sciences. Lecturers in the ELC use it to satirise dark topics, such as the lecturer who explains that:

> in Japan they call it karoshi that mean death attributed to uh stress at the work place so just like me come here and teach and collapse and pass away (2010)

Death, injury, and pain are recurring themes within black humour.

This type of humour largely occurs in discussions of health and safety, such as lecture 2010, which accounts for most of the 0.43% of black humour in the Malaysian subcorpus. The lecturer provides commentary on a series of images of events with grave consequences:

> ok this this slide show how children go to school in India yah and then again from here we can see that er this is quite a hazardous way to cross the river if say er if this is in Malaysia then Sarawak for instance then there'll be some crocs waiting underneath here so if the any one of the children fall then the croc will surely have a very heavy meal (2010)

Black humour is largely used when discussing life-threatening situations, such as:

> immediately after you release the radioactive you must run as fast as possible within one hundred yards around ten second so if you are fat fat ok big size yeah if the the construction area ok not so nice ok it's not easy to run hundred metre within ten second (2007)

Making light of potential tragedy through black humour emerges as a useful tool for coping with and illustrating the more serious aspects of engineering.

6.5.8. Wordplay

The purpose of wordplay is dual: to gratify the listener and to establish the superiority and skill of the teller. Included in this category are aspects of wit and unusual turns of phrase that cleverly twist the familiar, and incongruous/ambiguous comparisons and contrasts (Lee 2006: 62, Ross 1998: 4).

Wordplay in the ELC constitutes a display of wit for amusement where meaning centres on word choice. Some lexis in the keyword analysis (Table 6.11) is discipline-specific (*fasteners*, *hole*, *reversible*), and some appears to be unrelated (*apple*/*s*, *brewery*, *tie*). Most important in wordplay is what the lecturers *do* with words they choose.

Bisociation is followed by a punchline that both surprises and resolves the conflict. Koestler's (1989[1964]: 65) concept of "two strings [that] are tied together by an acoustic knot" is enacted, as in:

> normally this gets gets machined out you know bored out with a boring tool not a boring tool but is er bo- bores out a hole (3019)

The homonymic pun relies on ambiguity between a tool that creates a hole and a tool that causes/is attributed with tedium. Humour lies in the absurdity of the clarification between equipment and tedium planes. It could be argued that *funniness* is limited by the extent of incongruity, which is perhaps reflected by the absence of a laughter response to both this example and to the type in general (63%, see Figure 6.8).

positive			negative		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
apple	5	59.24	to	8	10.68
peace	3	38.79			
doctor	3	25.92			
brewery	2	25.86			
fasteners	2	25.86			
grandiose	2	25.86			
hoop	3	25.32			
hole	4	22.77			
apples	2	20.32			
beside	2	19.14			
boring	2	18.23			
r	6	17.98			
u	4	17.00			
t	16	16.96			
graduated	2	16.88			
phrase	2	16.88			
tie	2	16.34			
maybe	4	15.44			
reversible	4	15.28			

Table 6.11: 20 most highly ranked positive and negative keywords in wordplay

Similar examples of wordplay receive similar responses; as one lecturer laments regarding his pun on the concept of "mating" parts, there's an "opportunity to laugh", but "no laughers today" (3020). Students may be more likely to remember the distinction when the pun is notably bad, and this may offer a pedagogic motivation for its use.

The quantitative data suggest that the cognitive load in formulation and processing is not a barrier to the use of wordplay in the L2 context. The Malaysian lecturers seem to gain a certain amount of pleasure from accurately delivering complex sequences, such as:

> if a heat engine is a reversible head engine if you reverse that reversible heat engine this is a tongue twister here if you reverse a reversible heat engine then it will become a reversible refrigerator (2018)

Words become a game in which the engineering lecturer demonstrates their grasp of the language and its usage by taking on the role of knowledgeable English teacher:

> so circumferential or as call the hoop stress ok stress American call it hoop the British call it circumferential so it is the same thing it is in the circular direction remember hula hoop (2012)

Common sayings are carefully reconfigured:

> an apple a day keeps the doctor away right but what do I say what do I say in your first year not your standard one an apple a day keeps the doctor away but if the doctor is handsome keep the apple away that's for the girls (2017)

It was hypothesised that wordplay would be a less common humour type in lectures delivered by and to L2 speakers of English because it requires a high level of proficiency to decode the language in the first instance and then to resolve the inherent incongruities – which may also demand culture-specific knowledge. Wordplay occurs with most relative frequency in the Malaysian subcorpus, however (0.28%, UK 0.10%, NZ 0.13%, see Table 6.1 and Figure 6.1). In the ELC, the complexity of the humour type primarily offers lecturers an opportunity to demonstrate linguistic skill.

6.5.9. Self-deprecating humour

One of the strategic advantages of self-deprecation/denigration is that the revelation of weakness can paradoxically garner respect and trust. Self-denigrating Humor Schema (SHS) is identified as a culturally-specific strategy for developing relationships and dealing with non-serious threats to self-esteem (Niwa and Maruno 2010), or face. It allows tension reduction and creation of a friendly atmosphere, the construction of cohesion through group laughter, and the approachability of and trust in the teller is increased through risking ridicule (Niwa and Maruno 2010: 80).

Such functions are identified in self-effacement and in-jokes in MICASE (Lee 2006) and lecturer self-deprecation and lecturer self-aggrandisement in BASE (Nesi 2012). Although linked to a feeling of non-seriousness in the absence of humour by Chafe (2007), in the ELC self-deprecation functions as a non-serious defence mechanism which normally involves a negative reference to the self for comic effect.

The type occurs for the greatest overall percentage (tokens) in lectures from New Zealand (0.56%, UK 0.33%, MS 0.20%, see Table 6.1 and Figure 6.1). The instances from New Zealand

tend to be more fleshed out; their average token length is 33 compared to 22 UK and 25 MS (Table 6.2 and Figure 6.4). There are more occurrences of this humour type in lectures from the UK, the average of which is 1.33 per lecture compared to 0.56 MS and 0.96 NZ (Table 6.2 and Figure 6.3).

positive			**negative**		
kw	*freq*	*keyness*	*kw*	*freq*	*keyness*
i	108	123.95	the	74	21.94
hair	8	61.55	we	12	16.62
my	21	49.16	so	13	14.89
me	17	47.30	at	2	12.33
radii	4	39.19	and	30	8.26
spell	4	39.19	this	14	7.57
retire	3	33.14			
mistakes	4	30.78			
handwriting	3	22.60			
infallible	2	22.09			
rigidity	2	22.09			
god	3	20.98			
english	3	19.72			
damn	2	18.28			
m	18	17.58			
lecturer	3	17.41			
comb	2	16.57			
lunch	2	16.57			
proud	2	15.39			

Table 6.12: 20 most highly ranked positive and negative keywords in self-deprecating humour

The positive keyness of *I* and negative keyness of *we* within humour of this type is unsurprising, given the inward-focus of the self-deprecation (see Table 6.12). Lecturers cover mistakes in calculations with exaggerated jabs at their own inadequacy. One lecturer, for example, repeats in two lectures that "no man is infallible" (1029), "I keep telling you that no man is infallible" (1028) as students point out errors. Lack of preparation is also acknowledged then glossed over with good humour:

> I'm not prepared today just I haven't I have been working for the past two weekends […] make sure it's the correct one I may do the mistakes er purposely or accidentally (2020)

Self-deprecation is commonly linked to both cognitive abilities and physical attributes. As highlighted in the keyword analysis (Table 6.12), a particular thematic preoccupation is hair

(or lack of), particularly in the New Zealand lectures. Reference is made to "the few places I I do have hair left" (3001), "fond memories" of combing hair, (3001), the wonderful hair of others (3002), and pulling out the "few" hairs left (3019).

Explicit reference to the pedagogic role of the lecturer is also made. Lecturers express doubt in their linguistic ability, as in "I hope I didn't pronounce it wrongly" (2018). They also convey doubts about their own discipline-specific knowledge, as in:

> the only time you shouldn't do that is if you're a hundred per cent certain you're gonna get the right answer and I don't think anyone's at that position yet in this room I include me in that (1004)

The LR rate for this type of humour is less than half overall (UK 33%, MS 60%, NZ 44%, see Figure 6.8), but the themes of the self-deprecation are throughout light-hearted.

6.6. Conclusion

Text classified as humorous constitutes just under three per cent of the ELC and occurs most frequently in UK subcorpus (Table 6.1 and Table 6.2). Playful humour is most common across the corpus. Noticeable variation occurs in the distribution of attributed types across subcorpora; disparaging and sarcastic types, for example, are much more common in the UK lectures (Figure 6.3 and Figure 6.4).

There is some variation in pronoun usage across types (Figure 6.6). Although overall *you* is most commonly employed, the second most common pronoun across types (*I*) occurs relatively more frequently in bawdy, self-deprecating and mock-threatening types. *I* is also significantly more common to humour compared to non-humour. Pronouns feature heavily in the 3-gram analysis, cementing their place in language commonly used to deliver humour. Broadly, the dominance of the pronoun *I* in humour reflects greater emphasis on the lecturer explaining repercussions; *you* (as a single or plural pronoun) is linked to targeted attacks, as is *he* and *him*. The use of *we* is less key because a butt of humour is

necessary, which can be the self, in-group member/s or out-group member/s, but rarely the lecture participants (including lecturer) as a whole.

Humour episodes often depend on incongruous slips in expected register as the lecturer departs from the academic to the *everyday* script. The same abruptness that echoes throughout Hobbes' "sudden" pleasing act, Koestler's "sudden bisociation" and Schopenhauer's "suddenly perceived" incongruous planes, and Kant's "sudden transformation" to unexpected outcome is apparent in the ways in which the lecturers modify the behaviour of their students, establish roles and coherence (Hobbes 1996[1651]: 43, Kant 2007[1790]: 161, Koestler 1989[1964]: 51, Schopenhauer 1966[1959]: 59). Such switching causes surprise and amusement, and can result in laughter. Yet laughter was not found to be a reliable indicator of humour, particularly of self-deprecation and sarcasm/irony types.

Kotthoff suggested that teasing usually elicits laughter (2007: 274), but teasing/mock-threat in the ELC was met with a laughter response in less than half of all episodes across subcorpora (see Figure 6.8). The element of threat identified in teasing sets it apart from the more playful banter that elicited high laughter responses. The weight of the implicit messages carried within teasing perhaps inhibits student/lecturer laughter. The same may be said of black humour.

The playful type of humour emerged as a catchall for all episodes that are humorous but serve no particular function other than amusement. Unlike other types, when engaging with playful humour the lecturer does not appear to be positioning themselves, the audience or a third party, or playing to a script. The listener needs no particular skill to decode the humour, and it is not exclusive. Through jokes, as in playful humour, the lecturer intends only to amuse: the telling is not overly pedagogic, complex, or motivated by control or positioning. This contrasts sharply with black humour, which provides a means of talking about potentially dangerous or tragic aspects of engineering.

Irony/sarcasm type humour episodes are the lowest in average length across the corpus (29 tokens) (see Figure 6.4). As illustrated by Figure 6.9, these episodes regularly punctuate the lectures, especially in the UK subcorpus. Largely the lecturers weave in the short, sharp digs to make subtle corrections to behaviour and to reinforce a position of authority/superiority.

In the lectures from Malaysia and New Zealand, where disparaging humour does occur, it tends to be aimed outside the group, or towards gentler aspects of the *typical* student script. Focus is on mistakes or poor examples in textbooks (3019, 3021, 3023), bad money management (2007) or lack of experience (2009). As with the teasing and ironic types of humour, direct and biting attacks on students happen more in the UK lectures, where it seems that any attempt to create in-group cohesion adheres to the wider script of what is to be a young engineer (a form of resilience training), rather than an academic.

In bawdy humour, the conversational tone is lowered sharply as responses are fired back to impromptu situations. By shifting the register quickly between *site-banter* and the serious work of principles of civil engineering, the lecturer displays full control over language and audience. Connection can be made at the level of the low and inappropriate, and this can also be used to embarrass and establish hierarchy.

In lectures, Andeweg et al. (2011: 762) note that speakers who use such humour are "walking a fine line" in terms of presentation of expertise and effect on audience interest. In the ELC lectures fault-picking with the self through self-deprecation is delivered with a confidence that suggests power not weakness. The same feeling of expertise is generated through linguistic wit. There is little evidence of culturally-specific themes in wordplay, but some indication of culturally-specific usage exists, and a definite nod towards the employment of wordplay to establish the script of the linguistically skilful academic is clear.

Notwithstanding the variation in function and usage of types identified, at the macro-level the ELC lecturers use humour for the broad purposes of rapport-building, establishing expert position, and classroom management.

CHAPTER 7. STORY

7.1. Introduction

The structure and purpose of stories have long been topics of sociolinguistic discussion, often with reference to models of narrative structure. This chapter looks at the purposes of storytelling in the ELC lectures, and the ways in which various types of stories are realised linguistically. In light of the wider concept of *telling stories*, definitions of *narrative* and the surrounding terminology are explored, followed by a discussion of transmission in the context of the monologic academic lecture. The discussion draws primarily on Labov and Waletzky's (1967) structural model for oral narratives of personal experience, and Martin's (2008) four categories of story: *anecdote*, *exemplum*, *narrative*, and *recount*. The umbrella *story* element is first examined quantitatively, followed by closer analysis of the four attributed genres.

7.1.1. Theories of storytelling

The term narrative commonly describes a text-type category; one of four rhetorical modes in written and spoken language, along with *description*, *argumentation* and *exposition*. The distinction between the modes has been contested - for example, by the suggestion that description lies within the boundaries of narrative (Genette and Levonas 1976: 6), especially within the sciences (Herman 2009: 101). Narrative, however, is usually identified as realising a separate and specific function within written and spoken text, especially when used synonymously with the term *story*.

Definitions have historically split story and narrative along paradigmatic and syntagmatic lines: the raw materials (a linear synopsis of events and characters) versus the possible ways in which events, interpretations and perspectives form a chain. The divide is roughly expressed as *story* and *discourse* (Chatman 1978: 17-20, Toolan 1988: 9) and has been traced to Plato's *The Republic* (1973[380BC]) and Aristotle's *Poetics* (1996[335BC]) which

together introduced the notion of narrative (objective) and discourse (subjective) as two opposing camps (Genette and Levonas 1976: 8).

In other descriptions narrative functions as an umbrella term. Chatman (1978: 19) acknowledges the same Aristotelian debt, but suggests that "the story is the what in a narrative that is depicted, discourse the how". Cobley (2001: 5-7) treats story as the *what*, but substitutes narrative for the *how* that Chatman describes as discourse. He adds *plot* to the mix, extending Chatman's notion of story (as comprised of events and existents) by adding a separate consideration of causation (the *why*).

It appears that the terms story and narrative are used either synonymously with reference to a whole concept, or specifically to describe an element within that concept. Looking more closely at narrative as a particular element as the "showing or the telling" of events as Cobley suggests (2001: 5) requires a more fine-grained examination of structure.

Based on a study of tape-recordings of about 600 interviews, Labov and Waletzky (1967) identified the foundational units of narratives of personal experience and the pattern in which they occur. The oft-cited Labovian model divides narratives of personal experience into six stages: 1. *abstract*, 2. *orientation*, 3. *complication*, 4. *evaluation*, 5. *resolution*, and 6. *coda*. According to this model the abstract is a summary of the events and the orientation functions "to orient the listener in respect to person, place, time, and behavioural situation", the complication stage describes the series of events that comprise the complicating action, possibly over a number of cycles, and the resolution concludes the narrative (1967: 93). The floating evaluation stage can come before or after the resolution or coincide with it, and is regarded as "the significance or the point" because "a narrative that contains only an orientation, complicating action, and result is not a complete narrative" (1967: 94). These stages are obligatory. An optional coda acts as "a functional device for returning the verbal perspective to the present moment" (1967: 100).

Willis (2003: 196) identified this sequence as the "basic routine" of telling a story, the components of which cleanly map onto Labov and Waletzky's narrative model: 1. an utterance indicative of content, 2. a description of the situation, 3. a complicating factor, usually a problem, 4. an accompanying evaluation, 5. a resolution, and 6. a review and conclusion. Common to both understandings of narrative sequence is the relation between the basic units of narrative and the pattern in which they occur; an explicit equation of the order in which events are recapitulated and the temporal order in which they originally occurred.

The association of the central complication-evaluation-resolution structure with the narrative form is also established in Plum's sociolinguistic research (1988, cited in Eggins and Slade 1997: 236) and Martin's (2008: 43) association of a "narrative proper" with "the familiar equilibrium disturbed then equilibrium restored motif" identifiable in Labov and Waletzky's (1967) model. If narrative is considered to be a specific genre with an identifiable structure, it cannot be distinguished from story (the raw elements) and plot (the causation), and must be seen as one manifestation of the whole, drawing together historically divided notions of the syntagmatic and paradigmatic.

Martin (2008: 45) does not consider Labov and Waletzky's model as a satisfactory description of all the ways in which stories are told. He uses story as an umbrella term, and convincingly points out the problems of placing all types of story under a single narrative banner. Martin proposes an extension of the notion of a single narrative structure by identifying four genres of story: 1. *recount*, 2. *anecdote*, 3. *exemplum,* and 4. *narrative*.

In Martin's model, the four story genres are differentiated by the relationship between events and reactions. The first distinction is arrived at through opposing recounts (based on unproblematised events) with the other story genres:

> In narratives, the problem is resolved […]. In anecdotes and exemplums on the other hand, reacting to extraordinary experience is the point of the story, so resolution is not required. […] The point of these stories is to react - emotionally for anecdotes, ethically for exemplums. (2008: 45)

Martin's model thus develops Labov and Waletzky's notion of the narrative, identifying a network of possible pathways through which four possible story genres are differentiated, as shown in Figure 7.1.

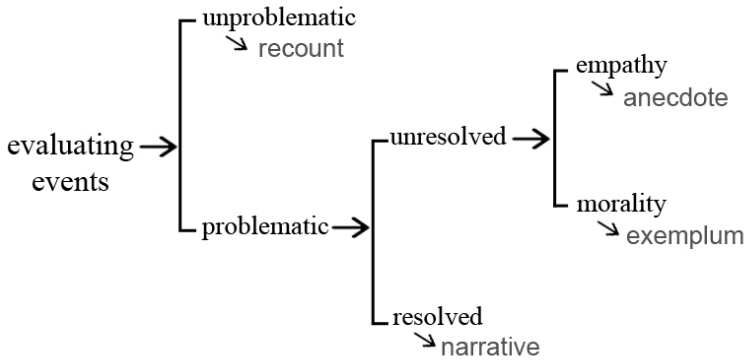


Figure 7.1: A choice network of story genres (Martin 2008: 44)

In Figure 7.1 only the narrative genre is associated with disturbed and restored equilibrium, as described in the Labovian model. Recounts narrate unproblematic events, and anecdotes and exempla narrate problematic events that are not resolved. Table 7.1 illustrates Martin's (2008: 43) claim that "the structure and function of the different stories derives from the relations between events and feelings".

genre	events	reaction
recount	unproblematic	running commentary
anecdote	unexpected disruption	emotional empathy
exemplum	noteworthy incident	moral judgment
narrative	complication resolved	build and release tension

Table 7.1: Events and feelings in four story genres (Martin 2008: 44)

This use of the term narrative in the description of a specific story genre still maps onto the structural features outlined in earlier uses; the whole narrative is still constituted of elements that combine sequentially in an organised, non-random way (Chatman 1978: 21, Toolan 1988: 7). The well-formed narrative structure is also familiar as one that "maintains and closes itself" (Chatman 1978: 21), including only what is relevant: "narrative selects some events and omits others" (Cobley 2001: 7). Whether used as an umbrella term or a specific term, sequence and a sense of contrivance are important to the notion of narrative – in terms of both the *what* and the *how*.

According to Labov and Waletzky strict temporal sequence is "the defining feature of narrative", because it can "recapitulate past experience in the same order as the original events" (1967: 81, 84). Temporal sequence is thus often used as a formal means of identifying story elements within larger units of discourse such as the lecture. Simpson-Vlach and Leicher define narrative in MICASE as a "story of two or more sequential clauses using the past tense or the historical present" (2006: 69), and Deroey and Taverniers (2011: 6) class as recounts those sections of the lecture where, often using past tense and time indications, "the lecturer presents information about past actions, events or situations".

Within the discussion of boundaries of terminology, one aspect that is not well elucidated is the role of the teller in the transmission of the tale. This is particularly pertinent to a consideration of *why* lecturers choose to convey information using particular pragmatic functions or types of functions. Basic definitions do not emphasise the narratorial role. Eggins and Slade (1997: 239), for example, discuss "stories which are concerned with protagonists who face and resolve problematic experiences". At one extreme, Genette and Levonas (1976: 9) explain that the ideal of narrative objectivity is the apparent absence of a narrator/narration. Although it seemingly has more in common with Martin's definition of a recount, the term story is used by Genette and Levonas synonymously with narrative to describe the ideal of diegesis: a simple narrative that contains only an account of events with no discernible narrator.

Alternatively, Scholes and Kellogg (1966: 4, cited in Toolan 1988: 6) emphasise the role of the narratorial voice by suggesting that all literary works "are distinguished by two characteristics: the presence of a story and a storyteller". Barring the exclusionary delineation of the field of literary works, this definition usefully highlights the issue of the *presence* of the storyteller. Avoiding such extremes, story can be characterised as

> […] a basic description of the fundamental events […], in their natural chronological order, with an accompanying and equally skeletal inventory of the roles of the characters in that story. (Toolan 1988: 9)

From this perspective, stories commonly contain a teller, the presence of whom is key to the discussion of narrative transmission.

7.1.2. Storytelling in academic discourse

In the educational context, emphasis is put on the function of narrative to convey information, construct new knowledge and make sense of experience (McDrury and Alterio 2002). The integration of storytelling into the academic setting has been discussed in relation to various disciplines, from law (Steslow and Gardner 2011) to dentistry (Keiser, Livingstone and Meldrumi 2008), largely in terms of fostering reflective thinking in students. Looking at the functions of stories in the lecture context requires consideration of the motivation behind narrative observation and evaluation, thus demanding that more attention is paid to the teller.

Martin's (2008) model of story genres suggests that storytelling might realise a variety of pedagogical purposes, and indeed a number of researchers have identified story as an important pedagogical feature in spoken academic discourse (Deroey and Taverniers 2011, Dyer and Keller-Cohen 2000, Maynard and Leicher 2007, Simpson-Vlach and Leicher 2006). Although there has been no systematic annotation of textual functions across large spoken academic corpora, attempts have been made to isolate and define story elements in small samples taken from corpora. Deroey and Taverniers (2011) consider recounts in their

functional analysis of 12 BASE lectures, for example, and Maynard and Leicher (2007) include narrative as a pedagogically interesting feature in the MICASE taxonomy.

Labov and Waletzky (1967: 95) highlight the way in which narrators position themselves in a favourable light, "a function which we may call self-aggrandizement". Looking at self-presentation in digital storytelling, Nelson, Hull and Roche-Smith (2008: 418) distinguish between "self-presentation" as the creation of an image (cf. Blumer 1969) as opposed to a more simple understanding of publicising the self (cf. Goffman 1959). The notion of image creation illuminates the relation between narratorial self-aggrandisement and the construction of personal identity.

Davies and Harré (1990: 50) discuss the concept of "positioning" in the exploration of the psychology of selfhood, and identify two types of positioning: the *interactive* and the *reflexive*. Dyer and Keller-Cohen (2000) apply both types of positioning to monologic narratives of personal experience within the academic lecture. Story elements in the lectures analysed by Dyer and Keller-Cohen are defined not only as reports of events in the past, but also as reports of events in which the lecturer (the first person narrator) partook.

Such narratives are described as a means by which lecturers position themselves as experts, and distance themselves from non-expert *other* characters. Dyer and Keller-Cohen (2000: 294) identify five devices through which the narrator's professional identity (as an academic) is established:

1. a lack of technical terminology
2. establishing the self as expert and the other as non-expert
3. the use of pronouns and other referring expressions
4. evaluation of the self and others on a "dual landscape of action and consciousness", in which the lecturer/narrator depicts him- or herself as the protagonist and controls the action, not merely describing their own role but also arguing for their expertise through self-justification and evaluation of themselves and others
5. "unequal egalitarianism" where self-mockery is used to democratise the discourse of expertise

They also note the importance of maintaining a culturally acceptable balance of self-aggrandisement and self-mockery (2000: 288).

Through the notion of narrative positioning, storytelling can be viewed as a means of revising and reflecting on the self, and constructing personal identity. The way in which the lecturer/narrator chooses to tell a story – which events they choose, the order in which they retell them and what is left out – can be understood not as an objective recount of historical events, but as a subjective choice that reveals the teller's process of meaning-making about the world.

7.2. Identifying stories

Each instance of story identified in the ELC annotation was initially manually broken down into its component parts at the level of sequence units, on the basis of Labovian rules (1967, 1972). The sequence units identified (abstract, orientation, complication, evaluation, resolution, and coda) map easily onto many of the examples of story found in the ELC, as in Figure 7.2. Sequence units are given in angled brackets.

<orientation>
once there was a really great story
it happened in my in this class in the first year
a student said to me
well I said to the students
I said
I was talking about DC motors
and I said you can't make a DC motor which doesn't have a commutator
it has to have segments to make it work
we'll see about that in the second semester
</orientation>
<complication>
and a student said
well he came to me the next week
and he said I don't think that's true what you said last week
and he um showed me a diagram
and I said oh that will never work
that's no good

the next week he turns up
and he's built one
and he says look
and um take it into the lab
</complication>
<evaluation>
and sure enough he was right
I was wrong
and it was a completely new idea that he'd thought of
</evaluation>
<resolution>
and it turned over
it worked
</resolution>
<coda>
and if he'd get a patent on it that's an amazing story
</coda>

Figure 7.2: A narrative annotated with Labovian sequence units (non-ELC annotation added. 3010)

The traditional Labovian model, however, does not map comfortably onto every instance of story identified. The inventory of component parts offered by Labov and Waletzky's (1967) model enables the identification of both that which is a narrative and, by proxy, that which is not a narrative, based only on the presence of certain sequence units. If, for example, a text contains a complication but lacks a resolution, using Martin's (2008) extended classification system it should be categorised as an anecdote if an unexpected disruption elicits emotional empathy, or as an exemplum if a noteworthy incident causes a reaction of moral judgment (as outlined in Table 7.1 and Figure 7.1).

Two stories about the same topic delivered in different ELC subcorpora exemplify this principle. The first retelling of the crane incident story (Figure 7.3) meets the criteria of a Labovian narrative.

<orientation>
it's not as embarrassing as the one I saw on YouTube
where some guy I presume it was a guy drove his little Ford Fiesta into the harbour off a quayside

that's not the funny bit
that's just sad
</orientation>
<complication>
some guy brings along a crane like this
tries to lift the car out
doesn't think about the fact
that if the car doors are shut the car will be heavier
because it's carrying water
so the crane topples into the harbour
</complication>
<resolution>
so they then have to bring another crane in to get the first crane and the car out
that they actually didn't make the same mistake twice
</resolution>
<evaluation>
have a look on YouTube
see if you can find the video
it's a hoot
so things should be in moment equilibrium
if they don't nasty things start to happen
</evaluation>
<coda>
and this is ok a little bit of a joke
and think yeah only a small crane
but its unfortunately very common
</coda>

Figure 7.3: A narrative annotated with Labovian sequence units (non-ELC annotation added. 1001)

The second version (Figure 7.4) feels like a story, but lacks a resolution stage.

<abstract>
this video sh- show the crane accidents
</abstract>
<orientation>
you notice this crane
er actually the workers were doing some lifting
I think there's a bit ok
</orientation>
<complication>
as what you can see here

start to tilt and splash into the water
</complication>
<evaluation>
ok so because of overloading that mean the the crane is not in equilibrium
that is why you have to know your free body diagram before you do anything
</evaluation>

Figure 7.4: A non-Labovian story (non-ELC annotation added. 2010)

Although the event in Figure 7.4 is problematised (as the crane falls into the water), it is not resolved. This is in contrast to the example in Figure 7.3, where the crane is retrieved. The string in Figure 7.4 cannot therefore be classified as a Labovian narrative. It does, however, accord with Martin's (2008) exemplum pathway, which is highlighted in Figure 7.5. The intended reaction to the event in the exemplum is judgment, rather than empathy, as emphasis is put on the need to "know your free body diagram before you do anything".

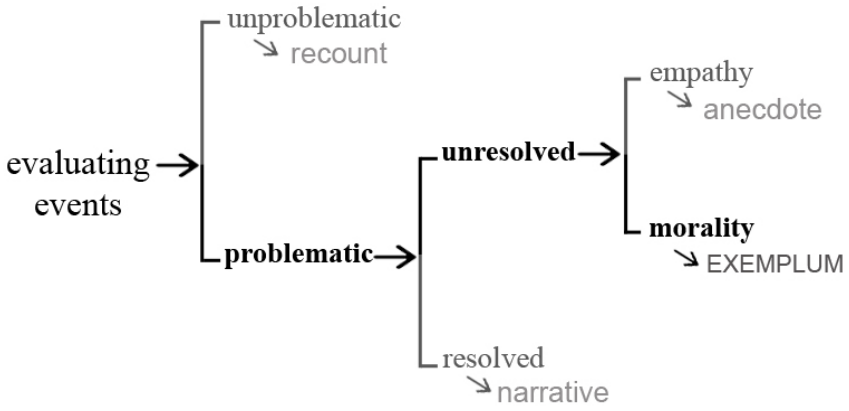


Figure 7.5: A choice network showing the path of an exemplum in bold type (cf. Martin 2008)

As the stories in the ELC are often used to illustrate an engineering principle rather than a moral, in the ELC taxonomy Martin's definition of exempla has been adjusted to refer to a reaction of scientific judgment.

7.3. Macro-level patterns in storytelling

All instances of story identified were analysed using Labovian rules in the same way, and then assigned to a genre classification based on their structural components (whether they contained a complication and/or resolution, and the type of reaction expected). Although oral recounts of personal experience underlie key discussions in the literature, narratives in the ELC also contain oral accounts of the experience of others in the field; something that may be considered as a particular feature of the discipline. Both types have been included.

One hundred and fifty three instances of story were identified and classified (Table 4.2). Raw and normalised token count data are given in Table 7.2 and the number of occurrences of all types across subcorpora in Table 7.3. The normalised token duration data are shown in Figure 7.6 and the normalised occurrence per lecture data are shown in Figure 7.7.

	UK		MS		NZ		all	
	tokens	%	tokens	%	tokens	%	tokens	%
anecdote	2190	0.87	179	0.15	521	0.33	2890	0.55
exemplum	2506	1.00	1952	1.62	193	0.12	4651	0.88
narrative	2348	0.94	1446	1.20	1830	1.17	5624	1.06
recount	1374	0.55	2360	1.96	1319	0.84	5053	0.96
all story	8418	3.35	5937	4.94	3863	2.46	18218	3.45

Table 7.2: Token duration (raw and % subcorpus/corpus) of story types

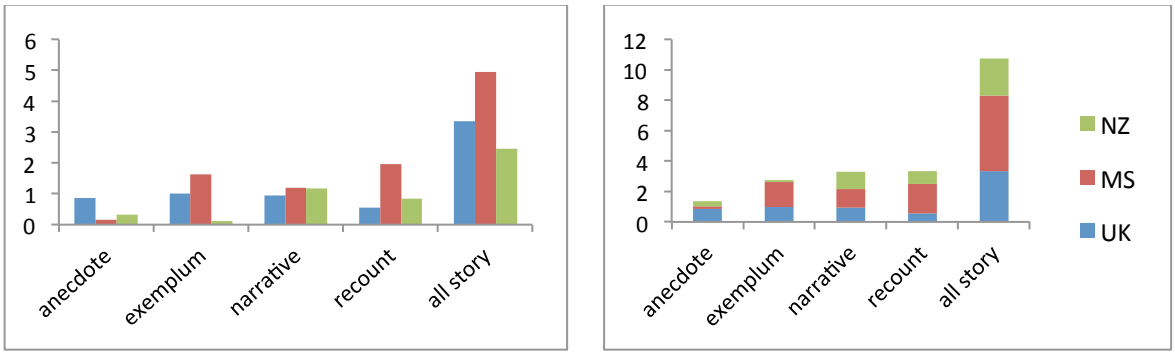


Figure 7.6: Token duration (%) of story types – cluster view (left) and stacked view (right)

The data in Table 7.2 and Table 7.3 show minimal variation in token count and occurrence patterns. In other words, whichever way it is viewed the picture remains constant: as the stacked views (right hand side, Figure 7.6 and Figure 7.7) show, the largest type (in normalised token duration and normalised occurrence per lecture) is recounts, followed by narratives, exempla, and then anecdotes. When viewed by subcorpus, story is most common by both measures in the Malaysian component, particularly the recount and exempla types. Narrative is approximately equal across all subcorpora, and anecdotes are more common in the UK lectures.

	UK		**MS**		**NZ**		**all**	
	total occurrence	per lecture	total occurrence	per lecture	total occurrence	per lecture	total occurrence	per lecture
anecdote	16	0.53	2	0.11	9	0.32	27	0.36
exemplum	15	0.50	15	0.83	2	0.07	32	0.42
narrative	16	0.53	10	0.56	15	0.54	41	0.54
recount	15	0.50	21	1.17	14	0.50	50	0.66
all story	62	2.07	48	2.67	40	1.43	150	1.97

Table 7.3: Occurrence (raw and per lecture) of story types

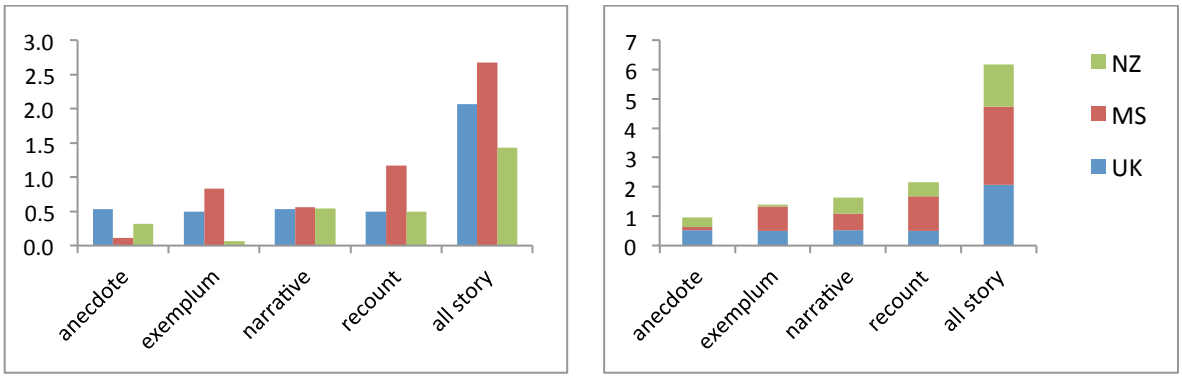
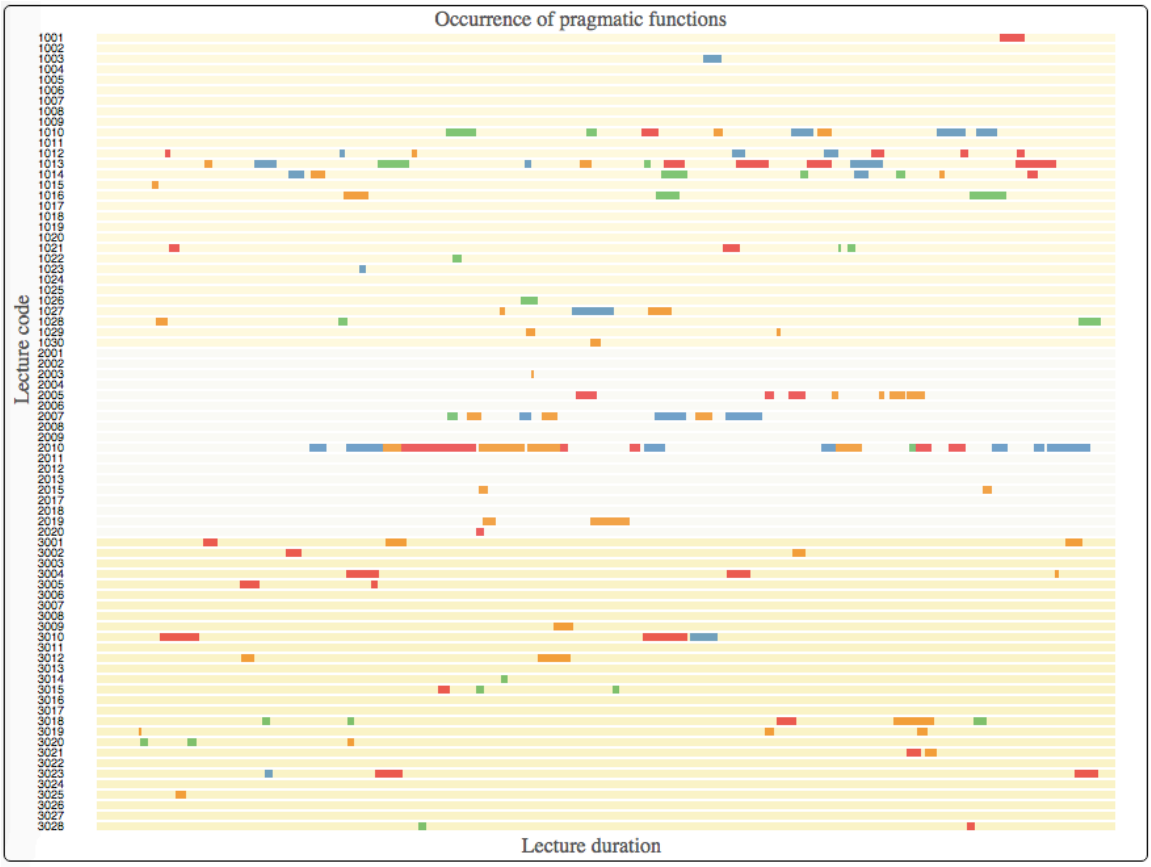


Figure 7.7: Occurrence (per lecture) of story types – cluster view (left) and stacked view (right)

This pattern of distribution is made clear by the visualisation of the occurrence and duration of story types across the corpus and in subcorpora in Figure 7.8.



KEY: **anecdote** | **exemplum** | **narrative** | **recount**

Figure 7.8: Occurrence and duration of story types

Table 7.4 shows the average token count of each instance of the four genres of storytelling.

	UK	MS	NZ	all
anecdote	137	90	58	107
exemplum	167	130	97	145
narrative	147	145	122	137
recount	92	112	94	101
all story	136	124	97	121

Table 7.4: Average token count (per instance) of story types

Recounts tend to be the shortest of the story genres and exempla and narratives are the longest. Recounts are unproblematised and therefore the story events are not resolved or evaluated, which may explain their shorter length. Narratives must include a complication and a resolution stage and can optionally include evaluation. Their longer length perhaps reflects the greater number of stages they typically contain. Anecdotes prompt a reaction of empathy and are almost as short on average as the unproblematised recounts. Exempla are the other genre in which the problem is unresolved (see Figure 7.5). They typically result in a judgment reaction. The high average token count of exempla may be linked to the seriousness of the events reported, as graver or more complicated stories require more detailed retelling. More detailed quantitative analysis of the patterns of token length and occurrence is undertaken in the discussions of story genres (7.4.1-7.4.4).

The keyword analysis in Table 4.7 showed that hesitation markers (*um* and *er*) are also a feature of ELC storytelling. An analysis of common hesitation markers (pmw) confirms that their presence is to an extent a feature of storytelling, particularly in narratives (Figure 7.9).

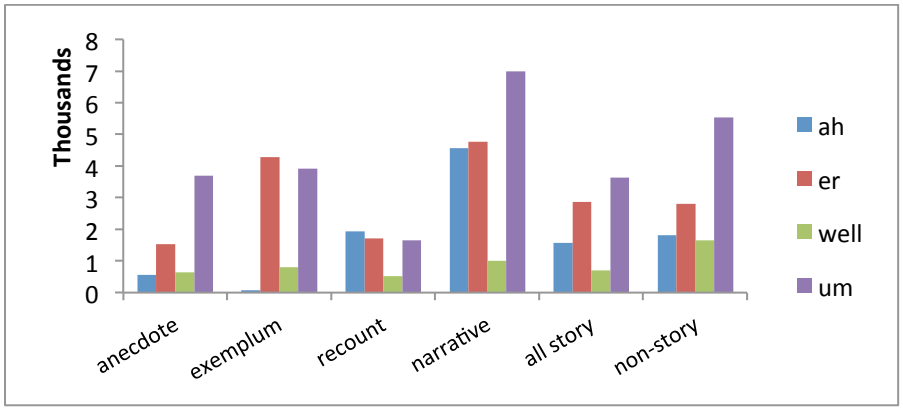


Figure 7.9: Hesitation markers *ah*, *er*, *well*, and *um* (pmw) in story, story types and non-story

Norrick (2001) claims that *well* is especially common in the vicinity of boundary markers of storytelling, but this is not the case in the ELC storytelling component. The primary function

of *well* in the ELC as a whole is as a hesitation marker in the delivery of technical information, as in:

> why haven't I mentioned the direction yeah g- g- good good question again<gap reason="pause"/> well<gap reason="pause"/> you see it comes down to the lines of force […] (3002)

As Figure 7.9 shows, the use of *well* (which is predominantly as a hesitation marker) is more common to non-story than story.

The stories in lectures analysed by Dyer and Keller-Cohen (2000) noticeably lacked technical terminology, which indicates that they are doing something other than delivering technical content. The same applies to the ELC stories. It is in the negative keywords that technical language about calculations (*point*, *value*, *times*, *minus* along with eight numerical references), equations (*x*, *f*), and structures (*beam*, *stress*, *moment*, *force*, *section*) dominates (see Table 4.7). The keyword analysis also shows that *they* and *he* are more salient in *story* and *we* and *you* are more salient in non-story text.

	anecdote		**exemplum**		**narrative**		**recount**		**all story**		**non-story**	
	freq	*pmw*	*freq*	*pmw*	*freq*	*pmw*	*freq*	*pmw*	*freq*	*pmw*	*freq*	*pmw*
i	36	12457	27	5805	81	14403	75	14843	219	12021	5242	10280
we	25	8651	34	7310	58	10313	36	7124	153	8398	5785	11344
you	53	18339	71	15266	82	14580	82	16228	288	15809	12856	25211
they	31	10727	50	10750	80	14225	58	11478	219	12021	1162	2279
he	4	1384	12	2580	40	7112	21	4156	77	4227	188	369

Table 7.5: Occurrence (raw frequency and pmw) of *i*, *we*, *you*, *they* and *he* in story, story types, and non-story

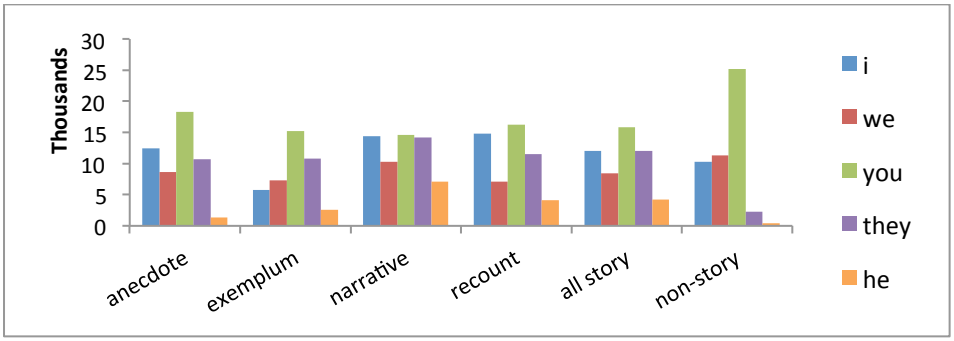


Figure 7.10: Occurrence (pmw) of *i*, *we*, *you*, *they* and *he* in story, story types, and non-story

The focus of stories is on external participants in the events described (*they* and *he*). In non-story, on the other hand, the lecturer makes more reference to *we* and *you*, as participants in the lecture theatre or as members of the professional community of engineers.

The most frequently occurring 3-grams were calculated based on all story, story genres and non-story (Table 7.6 and Figure 7.11). A 3-gram view of the lexis was generated because the frequency of 4-grams was too low to inform any useful conclusions. For example, the most frequent 4-grams in anecdotes only occurred twice.

anecdote			**exemplum**			**narrative**		
3-gram	*raw*	*pmw*	*3-gram*	*raw*	*pmw*	*3-gram*	*raw*	*pmw*
a lot of	6	2076	a lot of	8	1720	it s a	9	1600
it s a	4	1384	ok this is	8	1720	and that s	6	1067
we don t	4	1384	some of the	8	1720	and so on	5	889
we ve got	4	1384	you ve got	8	1720	one of the	5	889
and if you	3	1038	you can see	7	1505	have a look	4	711
don t have	3	1038	call it er	6	1290	it wasn t	4	711
in a microwave	3	1038	of the bridge	6	1290	that s a	4	711
on the other	3	1038	this is the	6	1290	the hard shoulder	4	711
poisson s ratio	3	1038	this is what	6	1290	to get the	4	711
problem with it	3	1038	we call it	6	1290	you can see	4	711
re going to	3	1038	what we call	6	1290	all the manholes	3	533
the channel tunnel	3	1038	as what you	5	1075	and he said	3	533
you ve got	3	1038	can see here	5	1075	and he was	3	533
a a lump	2	692	is what happen	5	1075	and i said	3	533
a look at	2	161	it s a	5	1075	and they ve	3	533
recount			**all story**			**non-story**		
3-gram	*raw*	*pmw*	*3-gram*	*raw*	*pmw*	*3-gram*	*raw*	*pmw*
a lot of	11	1222	a lot of	26	493	you ve got	896	1757
you can see	9	1000	it s a	23	436	re going to	657	1288
this is the	8	889	you can see	21	398	it s a	468	918
can see here	7	778	this is the	17	323	we ve got	461	904
all over the	5	555	can see here	13	247	going to be	446	875
in hong kong	5	555	you ve got	13	247	we re going	421	826
it s a	5	555	and so on	10	190	that s the	403	790
it s not	5	555	it s not	10	190	it s not	380	745
so that s	5	555	one of the	10	190	you have to	367	720
there s a	5	555	so that s	10	190	so it s	337	661
and it s	4	444	some of the	10	190	i m going	323	633
and that is	4	444	ok this is	9	171	you don t	318	624
collapse like this	4	444	so it s	9	171	m going to	317	622
i m going	4	444	there s a	9	171	so that s	313	614
it s very	4	444	we call it	9	171	s going to	304	596

Table 7.6: 3-grams (raw frequency and pmw) in story, story genres and non-story

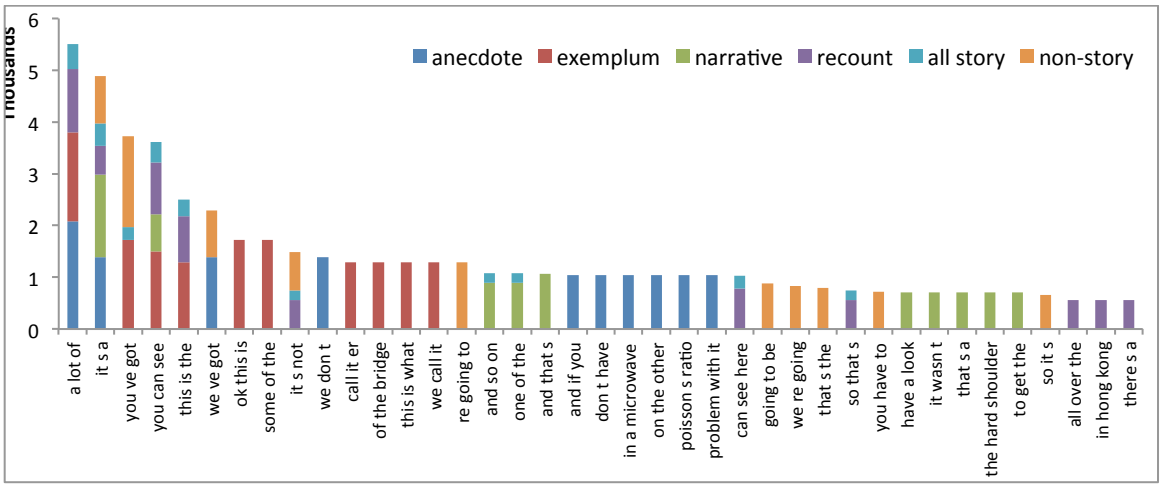


Figure 7.11: 3-grams (pmw) in story, story genres and non-story

Concordance searches show that frequently occurring 3-grams tend to cluster either in the orientation stage when lecturers are setting the scene for the complication, or in the complication/coda when impact is being described. For example, in the case of the most common 3-grams across all story genres, lecturers talk generally about *a lot of* "granite" (1010), "students" (1028, 2010), "contractors" (2007), "people" (2010), "money" (2010), "other scientist" (2019), and "errors" (3018). Similarly, *it s a* precedes: "highly fertile country", "well known phenomenon" (1013), "very dangerous act" (2010), "very simple machine" (3009). The 3-gram *you ve got* precedes: "the guy at the quarry" (1010), "loads of properties" (1013), and "all those little pores in concrete full of water" (1013). Formulaic sequences tend to orient the listener to the main themes/actors in the story, putting emphasis on their centrality.

The 3-grams *that s a*, *and that s*, *this is what*, and *this is the* seem to function as indicators of consequence in certain parts of certain story genres, for example in the resolutions of narratives, and in the codas of genres that contain a complication. However, *so that s* also occurs in recounts, which do not contain a resolution or explicit lesson, and *that s the* is also common in non-story lecture text.

The use of formulaic (n-gram) language is not a dominant feature of stories. The STTR analysis showed that the text of story had a higher lexical diversity than non-story (Table 4.4). There is a tendency for higher usage in recounts and lower usage in narratives, but overall even the 3-grams identified as most frequent in story still occur more commonly (pmw) in non-story, as illustrated in Figure 7.12.

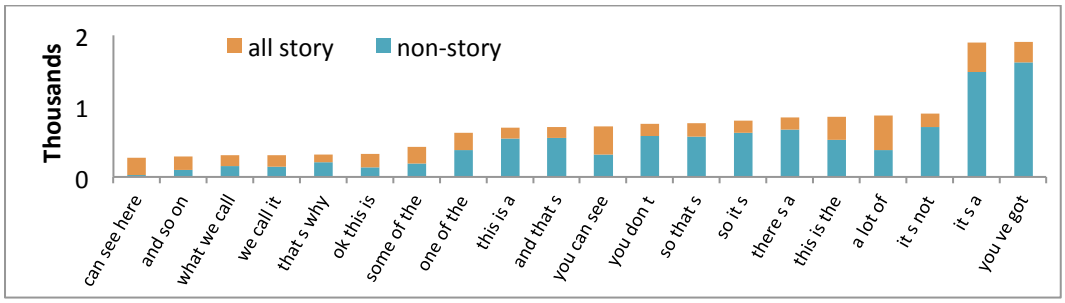


Figure 7.12: 20 most frequent (pmw) 3-grams in story and their frequency in non-story

Formulaic language is language that comes to mind quickly, and is therefore used most commonly when speaking off-the-cuff, as in spontaneous conversation. The limited use of formulaic language indicated in Figure 7.12 suggests that some of the stories in ELC lectures may have been told on previous (perhaps numerous) occasions, with the result that speakers are well-practiced in their delivery and have less need to rely on formulae.

The following sections (7.4.1-7.4.4) analyse the four story types in more detail, and a new type is introduced, *story-like* (7.4.5).

7.4. Story genres and story-likes

7.4.1. Narratives

Narratives are the most complete of all storytelling genres, as an orientation, complication and resolution are mandatory. Labov and Waletzky (1967) differentiate between two functions that explain the sequence of the identified narrative units: the *referential*, a

means of relaying the events of the story as they happened, and the *evaluative*, the point of the story. These functions are realised through a spectrum of clause types, "the smallest unit of linguistic expression which defines the functions of narratives" (1967: 75). An alphabetical series of annotations is used to differentiate each new clause, with a numbered scale either side of the letter to indicate how far the clause could potentially move either upwards or downwards (or, be *displaced*) in the narrative. Figure 7.13 is a version of Figure 7.3 that has been annotated according to this system. Clause rules are marked by a three digit identifier in curly brackets, sequence units are given in angled brackets and clause type is given in square brackets.

<abstract>
(0a0) it's not as embarrassing as the one I saw on You Tube [FixC]
(0b0) where some guy I presume it was a guy drove his little Ford Fiesta into the harbour off a quayside [FixC]
</abstract>
<orientation>
(0c1) that's not the funny bit [CoC]
(1d0) that's just sad [CoC]
</orientation>
<complication>
(0e0) some guy brings along a crane like this [FixC]
(0f0) tries to lift the car out [FixC]
(1g0) doesn't think about the fact [ResC]
(1h0) that if the car doors are shut the car will be heavier [ResC]
(1i0) because it's carrying water [ResC]
(0j0) so the crane topples into the harbour [FixC]
</complication>
<resolution>
(0k0) so they then have to bring another crane in to get the first crane and the car out [FixC]
(0l3) that they actually didn't make the same mistake twice [ResC]
(12m7) have a look on YouTube [FreC]
(12n6) see if you can find the video [ResC]
(13o5) it's a hoot [ResC]
</resolution>
<coda>
(15p4) so things should be in moment equilibrium [FreC]
(16q3) if they don't nasty things start to happen [FreC]

(0r0) and this is ok a little bit of a joke [FreC]
(8s0) and think yeah only a small crane [ResC]
(9t0) but its unfortunately very common [ResC]
</coda>

Figure 7.13: An example of a narrative story marked up to identify Labovian sequences and clauses (non-ELC annotation added. 1001)

At the referential end of the spectrum, "only independent clauses are relevant to temporal sequence" (Labov and Waletzky 1967: 82). These *fixed clauses* [FixC] cannot be moved; they preserve the order of events as they occurred. In Figure 7.13, line 0k0 "so they then have to bring another crane in to get the first crane and the car out" is a fixed clause because it has zero displacement potential either way – it cannot move in the narrative sequence. Clauses 0a0 and 0b0 are also fixed, and make up the abstract unit.

The primary function of the abstract is summative, and can be realised by a variety of clause types. 0a0 and 0b0 perform three functions: 1. they summarise the content (a video clip illustrating the type of embarrassing scenario previously referenced in the lecture), 2. they validate the relevance of the previous lecture content (the lecturer explained that he saw a clip on a popular website that illustrated in the real world the sort of embarrassing situation he had just outlined theoretically, making the point that these things can and do happen), and 3. they engage the listener, triggering a desire to hear the expanded narrative, to know what happened next.

Clauses that have no fixed place in the temporal sequence can shift without affecting the order of events (Labov and Waletzky 1967: 84). They are termed subordinate, or *free clauses* [FreC], and perform the evaluative function. In Figure 7.13, clause 0r0 "and this is ok a little bit of a joke", is purely evaluative and could be placed anywhere in the narrative – a displacement potential of the maximum eighteen clauses upwards or two clauses downwards – without disturbing the temporal sequence.

Between fixed and free are two clause types that can be reordered in some places, but not without limit: *restricted clauses* [ResC] and *co-ordinate clauses* [CoC]. In Figure 7.13, clause 0l3 "that they actually didn't make the same mistake twice" is restricted. It could move down a maximum of three clauses, but not up. It is evaluative, but not completely free, as it would not make sense if moved prior to the mention of the mistake in clause j. Co-ordinate clauses relate to the referential, such as 0c1 ("that's not the funny bit") and 1d0 ("that's just sad"). Clause c could be moved one line downwards, but not upwards, and the inverse is true of clause d; they have identical displacement sets, and so could be exchanged without altering the temporal sequence. Clauses c and d make up the orientation stage of this narrative. In conjunction with the scene-setting function, clauses within the orientation represent a significant majority of the fixed narrative clauses and coalesced coordinates that Labov and Waletzky (1967: 92) identified as constitutive of the primary sequence.

Co-ordinate clauses play a greater role in the complication than in any other unit in ELC narratives, and this indicates an emphasis on determining the referential aspects of the narrative. For example, in Figure 7.14, the lecturer relays a narrative about an unsuccessful wiring incident.

<abstract>
(0a0) in fact in the building I was in [FixC]
(0b1) at university years ago [ResC]
(1c2) they used aluminium wiring [ResC]
(1d0) because of a thing [ResC]
(0e0) called then the Rhodesia copper crisis [FixC]
</abstract>
<orientation>
(0f1) there was a point in history [CoC]
(1g0) a while back [CoC]
(0h0) when we ran out of copper [FixC]
(0i0) because um then Rhodesia now Zimbabwe declared itself independent [FixC]
(0j0) and stopped shipping copper [FixC]
(0k0) so we ha- we wired a whole load of buildings in aluminium [FixC]
</orientation>
<complication>
(0l0) with um disastrous results [FixC]

(0m1) basically um it wasn't very successful at all [CoC]
(1n0) it corroded at the terminals [CoC]
(0o0) and they were forever having to rewire everything [FixC]
</complication>
<resolution>
(0p0) and um after a while the copper came back [FixC]
(0q0) and the- they rewired them all in copper [FixC]
</resolution>
<coda>
(0r0) but i- it was indicative [FixC]
(0s0) as to what the hell are we going to do [FixC]
(0t0) when the copper does run out [FixC]
</coda>

Figure 7.14: A narrative annotated with Labovian sequence units (non-ELC annotation added. 1021)

After the listener has been oriented to the setting by a string of concise referential clauses, the complication delivers the, often humorous, punchline via interchangeable clauses with an almost playful tone. For example, "disastrous" rewiring is described as "[not] very successful at all" (Figure 7.14). This tone is unsurprising given that these narrative sequences provide a time-out from the larger lecture narrative, a more light-hearted interlude to illustrate an (often dry) academic lesson.

Labov and Waletzky (1967: 87) use the term "temporal juncture" to describe two clauses whose positions cannot be reversed. There are many narrative relations that cause a temporal juncture, but every narrative must use a semantic equivalent to the temporal conjunction "then" at least once. By definition, a narrative must contain this feature; "the x-then-y relation is the fundamental one in narrative" (1967: 91). In Figure 7.13, the complication is densely populated by referential clauses, almost half of which are fixed clauses. The defining temporal juncture in this example is explicit and occurs wholly within the complication (clauses e-j) as a crane is brought to recover the car and then topples into the water. This pattern continues across narratives in the ELC, as the key x-then-y temporal junctures occur largely around the orientation/complication units; those which perform the most significant referential function.

Following the extended work of the orientation to set the scene prior to the point of conflict (the complication), the resolution is swift, concise, and even sometimes unforgiving. Its function is to tie up the narrative interlude in no uncertain terms.

The oral bridge back to the real world of the lecture theatre is established by the coda of each string; Labov and Waletzky's "functional device for returning the verbal perspective to the present moment" (1967: 100). It can be argued that the coda provides more than a bridge – that it offers the *moral* of the narrative as well as making explicit its relation to the larger lecture content. The clustering of free and restricted clauses appears to aid this added pedagogic function, as in Figure 7.13, clauses m-t, where the free clauses p and q ("so things should be in moment equilibrium / if they don't nasty things start to happen") could have appeared anywhere in the string, particularly perhaps in the abstract.

In the coda in Figure 7.13 the free clauses function to signal the winding up of the narrative illustration through reference to the third party setting that binds the classroom lecture content to the time-out narrative: YouTube. The softer story-world notions that "it's a hoot" and "a little bit of a joke" are brought into sharp focus by an intervening reference to the "nasty things" (q) that happen if the correct outcome is not achieved – a result that the concluding observation reinforces is "unfortunately very common". Likewise, in Figure 7.14, the jovially retold rewiring failure contrasts with the context of the bigger question of "what the hell are we going to do when the copper does run out". The codas mark a return to the serious business of learning. The evaluative clauses that can be understood as wholly or partially free to displace across the segment can be viewed as more restricted when the coda unit is considered in isolation.

Labov and Waletzky suggest that evaluation can occur at any point, but it is usually found around the climax of the narrative:

> Multicoordinate clauses or groups of free or restricted clauses are located at the break between the complicating action and the resolution of these complications. (1967: 95)

Evaluation is "the point of the narrative, its raison d'être" (1967: 94). It explains why the resolution to the complication occurred, evaluating how appropriate this is given the expectations set up in the orientation. In my analysis I have not identified evaluation as a distinct unit of narrative sequence. Referring back to Labov and Waletzky's dictum that the temporal sequence of a narrative is in accordance with the sequence of the narrative units, it follows that evaluation cannot be considered as a sequentially bound unit because its place is not fixed, nor fixable.

In some narratives, the retold events are not explicitly situated in the past, as in: "the crane topples into a harbour" and they "have to bring another crane" (Figure 7.3). The historic present is used, as in Figure 7.3, when the sequence of events is clear and there is no need to indicate the temporal context. McCarthy (2010: 61) suggests that such use emphasises the relevance of the *now* to the narrated past events. Pastness – which was flagged up as a feature of keywords across stories (Table 4.7) – is, however, clearly marked grammatically in other examples, as in "there was a point in history" when there was a copper shortage which "caused" problems, and then "the copper came back" (Figure 7.14).

The keyword analysis of narrative (Table 7.7) is very similar to the keyword analysis of all story in the corpus (Table 4.7). In narratives, past tense forms (*was*, *were*, *said*), references to locations (*university*, *warwick*, *sea*, *carriageway*) and people (*he*, *they*, *graduates*) are important. Numbers (indicating equations) and technical language (*point*, *stress*, *beam*) have more salience in the inverse reference corpus, as indicated by the negative keywords.

The keyword analysis of narratives (Table 7.7) also shows that hesitation markers (*ah*, *er*, *um*) are salient in narrative storytelling. The comparison of hesitation markers (pmw) across all story types in Figure 7.9 confirms that *ah*, *er*, and *um* are a particular feature of narratives. Hesitation may be indicative of moments of consideration, given that storytelling functions as a time-out from conveying and receiving correct and factually dense content. The more extended narrative structure may offer a reprieve for both lecturer and students.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
was	74	206.613	point	1	41.003
he	46	155.515	you	93	39.297
they	97	149.446	is	67	37.084
lift	14	68.609	if	14	28.014
were	24	68.047	going	9	15.560
said	23	62.407	ok	2	15.248
ah	41	57.309	three	4	13.345
had	24	56.785	five	4	12.948
guy	10	40.755	two	13	12.535
university	7	36.809	ll	2	12.404
warwick	4	36.134	here	11	12.068
er	44	36.098	zero	1	11.980
bleed	5	34.649	s	70	10.955
shoulder	5	34.649	there	19	10.834
accident	6	34.209	value	1	10.655
tank	10	32.863	need	2	10.457
years	10	28.070	stress	1	10.365
sea	5	27.117	take	1	9.960
carriageway	3	27.101	twenty	1	9.652
flukes	3	27.101	six	1	9.518
graduates	3	27.101	moment	2	9.379
manholes	3	27.101	four	3	9.076
retarded	3	27.101	beam	3	8.881
and	191	24.674	hundred	2	8.787
um	63	24.279	put	2	8.787

Table 7.7: 25 most highly ranked positive and negative keywords in narratives

Although people and places are more salient in the lexis of narrative than in all text that does not perform this function, the lecturers do not extend this story type beyond a skeletal outline of events, existents, and evaluations. The didactic function of the key messages underpinning the complicating action – for example, that aluminium causes wiring problems, or that understanding equilibrium matters when dealing with heavy machinery – is clear. The transmission, or *how*, of the narrative, however, offers the potential for other types of education to take place.

In Figure 7.14, for example, space within the narrative structure is not given to an explanation that corrosion (the complicating action) happened because dissimilar metals were touching (the aluminium wiring and the steel terminals), or that if the terminals had also been made out of steel, unless moisture was present, corrosion would not have occurred. Technical terms, to describe for example the process of galvanic corrosion, are

not used. The lecturer explains why the wiring was put in, but not why it failed to work, which suggests that the emphasis of the narrative is on something other than explaining a technical problem.

More technical contextual details are not given in the rest of lecture content either. At the start of lecture 1021, the lecturer states that corrosion theory will come "a lot later in the course", and there is brief mention of the material properties of aluminium directly following the narrative. The resistance of copper wiring is considered in depth, and a brief reference is made to the fact that "aluminium doesn't make very good wiring" (1021). But it is clear that the narrative does not function to either convey or reinforce factual information that is crucial to the lecture.

In the narrative in Figure 7.15 also, the lecturer presumes either that his audience has a formative background similar to his own and comprehends key concepts that are not explicitly outlined in the lecture, or that full understanding of the technical aspect is not the main priority of the narrative. As in the previous narratives, little extra information is given beyond the chain of facts of person, place, and event. There is no elaboration. The idea that standards of any kind exist to prevent tragedies such as the death described is not referenced, and there is no mention of relevant quality management fail-safes, such as the mechanism which interlinks modern sets of lift doors so that one door cannot open without the other.

<orientation>
er this accident occurred in Bayan Lepas one of the lift shaft so this the victim here he was not aware that the lift door was open but there is no lift
</orientation>
<complication>
so taking for granted that there is a lift so once he step in step in the lift pit without any lift car so he just fall off
</complication>
<resolution>
and this is the cause of death
</resolution>

<coda>
that is why for this type of er equipment or er what is called the lifting machines it has to have a certificate of fitness from J-K-K-P or th- D-O-S-H department of occupational safety and health and all lifts must be registered yah you should go inside our lift here at C twenty three you can see the P-M-A number [pendaftaran mesin angkat] registration hoisting machines two one seven zero five something like that er on top of the lift so that is it means that our lift has been registered with J-K-K-P but I have yet to see the certificate of fitness so whether the lift is satisfied er certified or not that is another issue here but at least the lis- the lift is registered
</coda> (2010)

Figure 7.15: A narrative (non-ELC annotation added. 2010)

However the narrative in Figure 7.15 does relate content to the present context, not only in terms of the lecture theme and the field, but also to the audience through the explicit observation that "our lifts here" have the required registration number. This contrasts markedly with the more historical narratives, such as Figure 7.14.

The very small space that narratives occupy in the ELC suggests that they are not an important means of conveying or constructing new information; it could be that they function more as a showcase for subject expertise, in terms of both theoretical knowledge and practical experience. References to the past display the narrator's world knowledge; by narrating his personal experience of significant events in the history of engineering, the lecturer in Figure 7.14 concretises his own long-term involvement in the field.

Expert positioning is achieved through distancing (cf. Dyer and Keller-Cohen 2000). The distance, however, is not from the inexpert *other*, but from an uncontrollable external factor. The lecturer in Figure 7.14, for example, jovially admits his inclusion in the "we [who] wired a whole load of buildings in aluminium". It is clear that the emphasised outcome that the wiring "wasn't very successful at all" was the result not of poor workmanship, but of the fact that aluminium corrodes terminals, a key *lesson* of the narrative. The underlying constraint – the unavailability of a more suitable material – also precludes the "disastrous results" from being associated with the engineers; the Rhodesian copper crisis is the cause.

Despite the apparent blunder, wiring "a whole load of buildings in aluminium" can be understood as quite an achievement, especially as by the time of the onset of problems the lecturer is distanced from the messy scenario.

Further distance is achieved by switching from *we* to *they*; the lecturer moves from active participant (in the wiring) to detached observer (of the rewiring), and finally in the coda to an omniscient commentator voicing a universal concern: "what the hell are we going to do when the copper does run out". The coda provides both another key lesson and a reinforcement of the lecturer's status. In this part of the narrative he shifts position from expert to visionary, predicting a crisis some time in the future.

Narratives tend to be the most personal and involved/involving of the story genres. Out of the 16 UK narratives, for example, 12 refer to first-hand experiences – typically events that took place on a site visit or during testing or more mundane events that took place at the university (see Table 7.8).

narrative type	UK		MS		NZ	
	raw	*%*	*raw*	*%*	*raw*	*%*
personal experience	12	75	2	20	11	73
experience of others	4	25	8	80	4	27

Table 7.8: Types of experience (raw frequency and %) within narratives

Table 7.8 shows a clear distinction between the subcorpora. Although the UK narratives draw heavily on personal experience (as in Figure 7.16) those from Malaysia describe the experiences of others (as in Figure 7.17).

<complication>
I hate to admit to this one but one site I was on we had cube failures and the reason was that when I'd been sending the cubes off I'd been having to break the ice on the top of the tank before I could get them out and um the tank had a heater in we just hadn't bothered to get the spark to wire it in
</complication>
<resolution>

and ah fairly obviously by the time the area manager appeared to ah come and have a look and see what had gone wrong it was all wired in and working fine and we said oh no no problem with that would we do a thing like that
</resolution>
<evaluation>
and ah but ok sort of nevertheless it caused endless hassle the fact that we'd had these cube failures
</evaluation>
<coda>
if you keep them too cold they'll go down a low strength
</coda>

Figure 7.16: A narrative of personal experience (non-ELC annotation added. 1012)

<orientation>
this accident occur in Port Dickson in Negeri Sembilan so the house is located very close to the T N B transmission line and during this time some of the workers were installing the high tension cable
</orientation>
<complication>
and perhaps the cable that is holding this pulley it was broken and hit one of the houses
</complication>
<resolution>
er luckily nobody got injured in this incident
</resolution>

Figure 7.17: A narrative about the experience of others (non-ELC annotation added. 2010)

Such narratives about experiences in the workplace, or *site narratives*, are common. Further examples are given in Figure 7.18 and Figure 7.19.

<orientation>
years ago I had that I was doing some ah engine assessments in dynamometers and there are huge fluctuations in dynamometers um it was the first job that I had I did some measurements and I took them to my boss and I said look these two engines this one produces seven per cent more power than that engine
</orientation>
<complication>
and he said to me oh what's the uncertainty and he sent me back to do this I mean I had learnt this but it wasn't something I really wanted to do
</complication>

<resolution>
and I found that the uncertainty was far larger than the gap between the two engines and there you need to get into statistics to start deciding you know whether it's significant or not et cetera et cetera
</resolution>
<coda>
but without knowing the uncertainty or at least that the uncertainty is small enough you are actually the er data is quite dangerous on its own
<coda>

Figure 7.18: A site narrative (non-ELC annotation added. 3023)

<orientation>
again one site I was on we were um pouring some very fiddly columns th- they had steel columns and we were effectively cladding the concr- cladding the steel columns in concrete so it's very difficult to get the it was to increase the load bearing of an existing steel frame so um it was very difficult to get the concrete in it was taking hours and hours they were having to virtually push it in by hand er and so what we did was we got the concrete four hour retarded so we could get it delivered um in the middle of the day and have the whole afternoon to try and get it in and we were doing a few of these columns very day and we had this four hour retarded concrete
</orientation>
<complication>
and that was fine until one day we had a very cold night and they came along the next morning took the shutters off the previous day's concrete and were um just about to er tidy up and put them up for the following set of columns and they noticed that the columns that we'd poured the previous day when the shutters had come off they'd slowly slumped down basically because it had been a cold night what had been intended to be a four hour dose of retarder had retarded it right through the night and it hadn't set
</complication>
<resolution>
so obviously we had to wash them out and start again on the previous day's work
</resolution>
<coda>
so do watch out with retarders get a cold day and the stuff won't set it would have set eventually but er basically they are very sensitive to temperature
<coda>

Figure 7.19: A site narrative (non-ELC annotation added. 1013)

Acknowledgment of fault is again evident in the blunder theme of uncalculated engine measurements and slumping columns. It is the resolution in these narratives of personal experience that seems to enable such frank ownership of blame. The lecturers document their failures, in narrative form, for pedagogical purposes – to illustrate to students potential problems, and perhaps more importantly to model the correct attitude for dealing with mistakes. As the lecturer in Figure 7.19 makes clear in the use of "obviously", if an error is made on site, it is fixed and lessons are learnt. Engineers do the required calculations, even if they do not want to (Figure 7.18), or wash out the failed concrete and start again (Figure 7.19), or rewire everything correctly (Figure 7.14). In each case, a coda containing the lesson follows.

7.4.2. Recounts

In their sample of lectures from the BASE corpus Deroey and Taverniers (2011) broadly define recounting as a subfunction of informing. The ELC recounts adhere to this description as the lecturers convey information about events in the past in chronological order. The recounts contain no complication and so no resolution, which means that this genre has the lowest average token length (101, Table 7.4). Recounts are simple and brief.

Deroey and Taverniers also note a "stark contrast" in the use of story genres between the disciplines (2011: 6). They report that there were few recounts in the physical sciences, but numerous instances in the arts and humanities. In the ELC, recounts are the most commonly occurring story genre (0.66 per lecture, Table 7.3 and Figure 7.7) and constitute the second largest token count (0.96%, Table 7.2 and Figure 7.6). By both measures they are most common to the Malaysian subcorpus.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
lego	19	143.788	you	89	30.473
was	38	70.935	point	4	22.509
he	26	66.485	we	42	20.477
years	16	59.460	two	8	19.472
celsius	10	56.718	one	21	14.753
they	58	54.547	right	3	14.368
scientist	7	52.848	force	1	14.260
ah	34	42.290	at	12	12.377
hong	5	41.047	moment	1	11.364
kong	5	41.047	ve	12	11.050
malaysia	9	41.044	got	10	9.505
london	6	39.201	that	90	8.709
collapse	9	38.820	twenty	1	8.636
used	16	36.697	out	6	8.335
bridge	7	33.923	re	14	7.539
scaffolding	7	33.923	minus	1	7.214
conveyor	4	32.162	same	2	6.964
made	13	30.333	five	6	6.763
worker	5	30.269	look	3	6.689
asian	3	27.860			
buckingham	3	27.860			
holland	3	27.860			
investigation	3	27.860			
palace	3	27.860			
wembley	3	27.860			

Table 7.9: 25 most highly ranked positive and negative keywords in recounts

As shown in the keyword analysis (Table 7.9), in recounts emphasis is put on the centrality of places (*Malaysia*, *London*, *Hong*, *Kong*, *Buckingham, Palace*, *Holland*, *Wembley*), people (*he*, *scientist, they*, *workers*, *worker*, *conveyor*) and structures (*bridge*, *scaffolding*). The general formula is to include any combination of the three components within some form of retold incident or point of interest. For example:

> it may have been before the start of term but when they er demolished the buildings where they're doing that big construction site just opposite John Laing they did what they always do now and they put all the er demolition arisings through a crusher and um crushed it down but they didn't actually ship it away to make concrete out of it they just used it to make hallroads and things like that (1010)

> this is another one similar to what happen in Malaysia also this is in Shanghai which is far away from Malaysia but it can happen also in the in in in here this building totally collapse you know immediately after completion you see this almost similar to Highland Tower case in Malaysia where part of the buildings ok or the blocks yeah one of them suddenly collapse (2007)

Important historical names in the field also make special appearances as the subject of recounts, as in:

> Faraday the scientist Faraday was ah a guy who bound books and he went to a lecture by a famous scientist he went to a lecture series by a famous scientist and after the lecture series was over he said to the scientist um he wrote down the notes and bound them because he was a book binder and presented the lecturer with this bound book and said here's this bound book I've done for you ah will you let me use your lab I'd like to be a scientist and work in your lab and luckily the scientist said yes you can do that and Faraday went into the lab and over the next twenty years he invented the subject of electrical engineering as well as doing a heap of work on chemistry and other subjects a remarkable guy (3012)

Carnot, Kelvin (2019) and Coulomb (3001) also feature. Lecturers flesh out the characters of the type of "clever brilliant scientists" (3001) through recounts of small details of their lives, such as place of birth and occupation, seemingly to make the concepts to which their names were given more memorable.

Some differences were noted across subcorpora. Recounts in the lectures from New Zealand are mainly used to explain *how* something was carried out or achieved. In only four out of 14 instances is the recount based on personal experience; in most instances it describes or explains a process typically used in a specific industry, for example the steel industry:

> yeah rim on a steel wheel you know th- the good old horse carts yeah that's how they put the rims on there they heated up the rims and hammer them on and then le- just let them cool down and you could never get them off the wooden yeah wheels yeah that's how that was done just basically shrunk on there (3019)

Or the shipping industry:

> if you're loading a ship up you do not want the centre of gravity to get above the meta centre in fact you don't even want to get it close if you get it above as that ships sails off it's going to turn over and these things have happened in the past and they do happen due to bad engineering and sometimes bad captaincy (3021)

Recounts in the Malaysian lectures, on the other hand, often accompany a visual aid and provide further contextual information relating to the situation depicted in the image. For example, one lecturer describes in detail a trip to Legoland in spring, showing slides of the structures made from Lego, including Big Ben, Wembley Stadium, and various castles (2005). Another lecturer recounts construction practices in Asia:

> er this type of scaffolding is widely used in Hong Kong in fact this this picture was taken in Hong Kong as what we can see here this the material for the scaffolding is made from bamboo and this is a standard practice in Hong Kong er in in in Asian countries except for Malaysia and Singapore most other countries use this type of scaffolding in China in Hong Kong in Indonesia in India in Vietnam in Thailand but in term of safety we are as what we can see slightly better than them and this is standard practice it is not against the law to use this type of scaffolding in Hong Kong but these are very skilled people they know what they are doing (2010)

As with the New Zealand lectures, these recounts do not express personal experience. Even where the lecturer is referring to pictures he has personally taken (such as at Legoland), the purpose of the recount is not to talk about the visit itself or what happened there, but to describe the layout of the place and its various structures.

More of the UK recounts are based on relating personal experience (eight out of 15 instances), as in:

> when I was down in London the other day I noticed that the the Gherkin th- the funny shaped building um in the middle of the city had got a couple of panels missing um that would be almost certainly a thermal effect that's taken them out (1014)

Lecturers also describe their first-hand experience of the behaviour of students and other staff, as in:

> one of my colleagues was um was checking your exam paper for me um last week and he said what's this infinitesimal shouldn't you say infinitesimal element and I said no no it's very common just to call it an infinitesimal um they they're used to it being called infinitesimal if I call it an infinitesimal element in the exam they'll all go I wonder what that is so there you go we've got the infinitesimal element (1028)

> I've been teaching this subject for many years and it's about ten years ago I think that a student in a class twigged to the fact that these formulas exist and hence he er used that formula in an exam and I thought well why should I not er teach that this formula does exist on the basis that some students would want to use it but at the end of the day it will only give you the principle strains (1029)

Overall, recounts tend to be explanatory and descriptive in nature, typically referring to a situation which does not personally affect the speaker, or from which he or she is personally removed.

7.4.3. Exempla

According to the broader ELC definition which extends judgment to matters which are scientific, exempla are the third most common form of storytelling in terms of token count across the corpus (0.88%, Table 7.2) and also occur with the third highest frequency by occurrence (0.42 per lecture Figure 7.6 / Table 7.3). They are relatively most common in the Malaysian subcorpus.

Notably, exempla have the highest average token length (145 tokens, Table 7.4), which suggests that lecturers give more detail when retelling stories which require a judgment reaction. The events and complications are perhaps more complex, or serious, in this story genre.

The keywords in Table 7.10 reflect that exempla are located outside the lecture theatre (*bridge*, *aircraft*, *station*, and *dam*). The majority of recounts in the Malaysian subcorpus occur in lectures 2007 and 2015, which are both delivered by the same lecturer, both on the theme of occupational health and safety. The lecturer provides a commentary on a series of video clips, which is why *video* appears in the keyword list.

positive keywords			negative keywords		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
bridge	26	180.619	point	1	32.713
er	62	98.966	you	85	23.475
was	42	91.423	i	33	22.165
they	63	74.983	we	38	19.321
video	11	74.712	five	2	14.857
happen	17	52.733	be	11	14.137
aircraft	5	47.089	your	5	13.201
ok	41	46.537	beam	1	12.391
station	7	45.342	know	3	12.084
were	17	44.073	going	8	11.614
students	11	42.523	two	10	11.580
footing	5	38.749	s	56	10.312
popouts	4	37.671	that	78	10.256
chloride	8	37.234	ah	1	9.148
she	5	36.558	say	2	8.953
had	17	36.135	don	2	8.895
built	9	35.584	m	4	8.339
calcium	6	35.518	re	12	8.153
he	17	34.610	now	6	7.626
years	10	31.580	need	2	7.523
dam	4	25.396	six	1	7.160
clip	3	23.773	three	5	6.985
faculty	3	23.773			
overloading	3	23.773			
um	54	19.500			

Table 7.10: 25 most highly ranked positive and negative keywords in exempla

Unlike anecdotes, exempla often have markedly negative consequences. The stories in lecture 2010, for example, draw on scenarios such as severe burns from a pot of boiling dalca, and an accident with a forklift truck:

> so from the video you can see that the the girl was hit by the forklift because because of very very simple reason she did not hear anything because of her i-Tune normally when you use i-Tune you listen the music very very loud so it will cut you off anything from outside so even though the forklift driver he use the horn or whatever so the the the girl in this video yeah even though it's acting she did not hear anything and hence she was hit by the forklift this type of accident actually occur sometimes (2010)

The themes of death and mutilation are strongly linked to situations in which students may find themselves during their future careers. For example:

> so if we talk about safety let me give you some scenario about safety for example this is an engineering perspective yeah so this is what happen last time ok er you still remember not long ago ok er this bridge foot bridge yeah that connect the two area

> here recreational area yeah ok this one definitely design by an engineer ok this is future engineers probably you will design the company the company that design this one actually donated the bridge to this the body that run this place yeah and unfortunately the bridge collapse yeah some student fall into the river and died how many of them died a few yeah a few students ok a few students and ok this is what er er they had ok the data about the case ok twenty two students fell into the river so actually a lot of students yeah on that bridge when the tragedy happen ok it was a suspension bridge ok the weight of the bridge is not able to support the the foundation is not able to support the weight of the bridge ok the weight of the bridge ok this is engineering elements yeah this is what happen ok some part of the er er problem here ok failure that happen in the middle of the bridge this is the support which is not enough to to support the weight of the students on top of the bridge when the tragedy happen yeah so we as an engineer normally we design this thing can last how many years one hundred years and this thing last how many years not even one year not even one year yeah not even one year right this is what happen some of the footing ok that hold the bridge yeah give way and we look at the construction here is very very not professional right the footing here doesn't look like a footing ok doesn't look like a footing this is not an engineer's design this is a contractor's design probably they want to save money they just design a simple footing which is not sufficient to carry the load hopefully none of our our students will involve in matters like this (2007)

The examples are made pertinent to the audience, for example in 2007 by expressing the hope that "none of our students will involve in matters like this".

In the Malaysian lectures particularly, lecturers use exempla as a way of asking students to imagine themselves in industry, normally in a position of responsibility for others, as in:

> what we can see here this Indonesian worker he's a contract worker working er doing some work at this faculty so he's climbing on the ledge of this building trying to pick up a bunch of rambutan and he's extremely dangerous thing to do it's a very dangerous act as what you can see here the width of the ledge is just er perhaps one and a half feet and he can just fall off if he fall it is the faculty's responsibility the faculty is actually answerable to to the Department of Safety and Health (2010)

In a later description of refining working practices, the same lecturer comments that "we have contribution to make t- to the nation" (2010). The theme of responsibility runs throughout this story genre.

Another recurring theme in exempla is the financial impact of failures in health and safety, as in:

> this bridge as I as I mention to you this bridge cost around two hundred eighty million the cost to do investigation there are many parties who involve in this investigation ok and we were part of that ok and my my duty at this time was to evaluate the conditions of contract yeah whether the the contractor was following what we call it the right what we call it er er procedures the right requirement ok and so on and they have to pay the university using the tax payer money ok and eventually they have decided that the bridge was not er good enough to withstand the loading and then repair need to be carry out and if you read the newspaper the cost to repair this bridge was around seventy million ringgit which is very expensive (2010)

In these Malaysian exempla students are primed to be responsible, as engineers, people and citizens. The same implication that the engineering students in the room will join those who have responsibility for overseeing the safety of others, either directly or indirectly, can also be found in the UK exempla, as in:

> that's why we're interested in making the lo- compression members shorter length because it makes them stronger and when you get it wrong you kill people you will notice that there is not one diagonal bracing member in that scaffolding system they're all horizontal and vertical tubes so the wind blew it's a mechanism the thing collapses if you walk down the road here and you think you think you're seeing some sort of G-MEX centre with all this like green cladding screening to the scaffolding where they were actually repairing the façade of the old BT building in town most scaffolding systems like this now are actually clad to protect members of the public from passing by so they're basically acting like their own separate building with all the wind forces blowing on it it's not just a case of in this one where the only wind load was coming was what actually hit the tubes it wasn't screened off at all so it's incredibly important to get the bracing system correct because the wind forces that are going on these systems now are huge I'm sure the person who designed a- and built that spent a lovely time in prison (1003)

No examples of stressing professional responsibility were found in the exempla from New Zealand. In general, the potential severity of the consequence of poor practice is stressed.

7.4.4. Anecdotes

According to Martin, both anecdotes and exempla are stories that contain an event(s) that is problematised, but not resolved. The distinction is made at the level of reaction:

anecdotes elicit emotional empathy, whereas exempla elicit a "moral judgment" (2008: 44) (see Table 7.1 / Figure 7.1).

Anecdotes are the least common form of storytelling across the corpus. They comprise 0.55% of all tokens (Table 7.2) and occur on average 0.36 times per lecture (Figure 7.6, Table 7.3). They are most common to the UK subcorpus, and the average token length of each occurrence is second shortest, after recounts, of all story types (Table 7.4).

The keywords (Table 7.11) indicate that anecdotes may have quite serious topics; *fire*, *dramatic*, *explodes* rank highly.

positive keywords			**negative keywords**		
keyword	*frequency*	*keyness*	*keyword*	*frequency*	*keyness*
they	39	45.678	point	1	18.608
was	23	45.085	is	36	17.750
excess	6	40.953	five	1	10.422
um	47	37.913	your	3	8.732
claims	4	36.346	do	6	8.043
microwave	4	36.346	you	62	7.947
dunk	3	31.004	going	5	7.420
fire	8	29.108	three	2	7.168
learned	4	27.947	go	2	6.699
dramatic	3	26.517			
countries	4	26.154			
tunnel	3	22.721			
clause	6	21.789			
carriageway	2	20.669			
digress	2	20.669			
dries	2	20.669			
explodes	2	20.669			
flower	2	20.669			
topsoil	2	20.669			
unnecessary	2	20.669			
went	5	20.515			
had	10	20.274			
years	6	18.631			
year	7	17.920			
channel	3	17.073			

Table 7.11: 25 most highly ranked positive and negative keywords in anecdotes

On closer analysis, however, the costs of the described complications are not grim. For example, the explosion (1014), the "real disaster" (1016), and the "dramatic scenes" (1028) referenced in the following examples do not have serious negative consequences. In

anecdotes, theatrical language adds to the entertainment value, rather than the solemnity, of the story. For example:

> you can demonstrate this in fact if you put a a lump of concrete in a microwave oven just take a uh a little well it works best wi- with actually grout y- you don't do it with the aggregate um good strong mix um fully saturated put it in a microwave put it on full heat and you'll probably break the plate because it explodes um certainly I did once much to the disgust of the owner of the microwave at Leeds University (1014)

> so if you add a bit of extra water to your concrete you are not only making it a lot weaker but you're making it vastly more permeable and if you want to see the effect of that take a step outside and take a look at the ring road um basically where in the nineteen sixties and seventies they tended to work with rather higher water cement ratios than they should and so in went all the fluorides and started rusting all the reinforcement although it should be said it was probably only designed with a thirty year design life so it's probably done what it was designed to do but whatever it's costing a fortune to sort out and if you want to see a real disaster go up and have a look at the Midland links up er where the M-6 goes through Birmingham where they're spending um over a thirty year period the- they've spent several billion pounds (1016)

> it has led to some can I say quite dramatic scenes in examination rooms in the past where students have stuck their hand up and I've gone to speak to them and they've said this question needs Poisson's Ratio and you say yes and just before you dr- walk away they say you haven't given us Poisson's Ratio and you say yes and you're just about to walk away and th- they sort of grab you by the sleeve and it can be quite you know dramatic c- could we try not to have that in this year's exam because um what I notice is a lot of students on their initiative assume what the Poisson's Ratio might be but naturally get it wrong no it is better to work it out from that excellent formula (1028)

Like narratives, anecdotes often do not contain detail about engineering concepts, and this suggests that they can perform a role other than purely to convey technical information. However, although they do not contain much technical language, anecdotes often demonstrate significant verbal skill, for example through the use of interesting analogies. It could be that a role of anecdotes is to introduce concepts in an entertaining and memorable way, rather than in elaborate or technical detail.

7.4.5. Story-likes

Some strings of text in the ELC retell events, but do not quite meet the criteria for classification as story because these events are not situated in the past. For example:

> <abstract>I don't know if any of the part timers ha- ha- have observed this</abstract><orientation> but if you go out on a er a very large concrete pour the morning after it's done you put your hand down on it you can feel it's quite warm um even if it's sort of snowing or something all around you um there's a reasonable amount of heat given off by concrete</orientation><complication> it can cause problems</complication> (non-ELC annotation added. 1013)

> <orientation>in Malaysia the one that we normally order in the lab is the granite because it is I think for the normal construction like a building like a bridges it is ok for use for you to use granite</orientation><complication> but if you are thinking of making long span light weight concrete that is different you need to deal first with a different set up of aggregate if you are using if you are thinking of building the special structures strong structures to contain nuclear for example you also need to use different type of aggregate</complication><coda> so the type of aggregate based on the specific gravity</coda> (non-ELC annotation added. 2003)

> <orientation>what if you are involve in a project you graduate and you work in a power plant you are involve in a project where the power plant wants to expand and they want to build a new power plant and the contractors come to you and say look I have a power plant the efficiency is sixty per cent you know sounds good very good sixty per cent is very good for a power plant</orientation><complication> ok so you take the data you calculate the maximum you can do that you calculate Carnot efficiency the maximum efficiency and you found out oh my god the maximum is only fifty five per cent and this guy is claiming the efficiency is sixty per cent so is that possible no no it's impossible</complication><coda> why because it's violating the second law of thermodynamics</coda> (non-ELC annotation added. 2015)

These strings, which will be referred to as *story-likes,* occur quite frequently in the ELC and seem to be pedagogically important; they generally use hypotheses or predictions of future events to make general claims of relevance to engineering, as in "*if* you go out on a er a very large concrete", "*what if* you are involve in a project", "*if* you are thinking of making long span lightweight concrete" (emphasis added. 1013, 2015, 2003). In general *story-likes* present common problems that are not resolved (as in the examples above from 2015, 1013

and 2003). They mirror the structure of exempla, and the main focus is on scientific judgment.

Story-likes often make analogies in order to explain engineering concepts, as in:

> resistance occurs when the electrons are moving through a conductor and they bash into an atom when the electron hits an atom it gives off energy it causes there to be power loss and the more collisions that occur the greater the resistance so you can compare a conductor to a crowded room if you have a crowded room and you want to walk across a crowded room it's very likely you're going to bump into somebody and the more crowded the room is the more collisions will occur and that's like high resistance you see high resistance is when the room is really crowded and you get heaps of collisions a low resistance is when there's not many people around and you can walk through with only very few collisions so tha- that's a sort of um er a very non physical physics people would hate what I've just said but I think it gives you an idea of what resistance actually is it's um it's the power loss that occurs due to collisions between electrons (3005)

The party setting is familiar to students, but the generalised scene of bodies colliding in a crowded room is hypothetical – the lecturer uses *if* rather than locating the experience as personal (for example, *when I was at a crowded party* […]). Later in the lecture the resistance analogy is extended:

> remember my party my people in the room and having the combina- um and the person moving through a room if you hit somebody then that means you have a higher resistance no- mo- mo- the more collisions you have as you walk through a crowded room the higher the resistance now if you have higher temperature that's like everybody in the room beginning to move around I don't know dance or something let's say you had some dance music people were dancing around moving about a lot then you're much more likely to have collisions it's quite easy to avoid people who are standing still but if people are moving you're going to have more collisions and that's how it is in a resistance wire (3005)

In this example another familiar script, dancing, is used to explain the effect of temperature on resistance. Engineering principles are made accessible and memorable through equation to familiar experiences (parties and dance music), and the lecturer who is able to draw such analogy positions himself simultaneously as an expert educator and a contemporary.

Another analogous story-like occurs in lecture 2012 where students are learning about project plans and the importance of assigning tasks, monitoring progress, and taking corrective action. The lecturer explains:

> very simple analogy of this is a map on a journey I'm taking my daughter this weekend to Keele University well I've got a vague idea that Keele University is up the M6 Junction Fifteen or Sixteen but what I really need is to know well when have I gone too far up the M6 I know that Keele is now sixty-seven miles away from home so I've got an idea of where I should be and I've glanced at the map so I've got an idea that if I've been travelling up the M6 for two hours I've probably gone wrong and that's the purpose of a project plan to know whether you're ahead of schedule behind schedule can I stop for a coffee break is she going to make me stop for a coffee break perhaps or do we need to speed up to get there a lot of the tools I'm going to show you today are designed to be used by project teams (3004)

Here the lecturer does not draw on a hypothetical scenario as in the previous examples, but applies a concrete (albeit future) experience to illustrate a general point by analogy. It is not a story because the illustrative journey and possible adjustments have not (yet) happened. A car journey that does not run to plan, however, is a commonly understood experience; its comparability to the process of project planning is easily grasped.

Story-likes have not been annotated systematically across the corpus. However, in the examples provided here, a noticeable feature is the use of *you*. Figure 7.20 extends the comparison of pronouns (pmw) in stories given in Figure 7.10 to include story-likes.

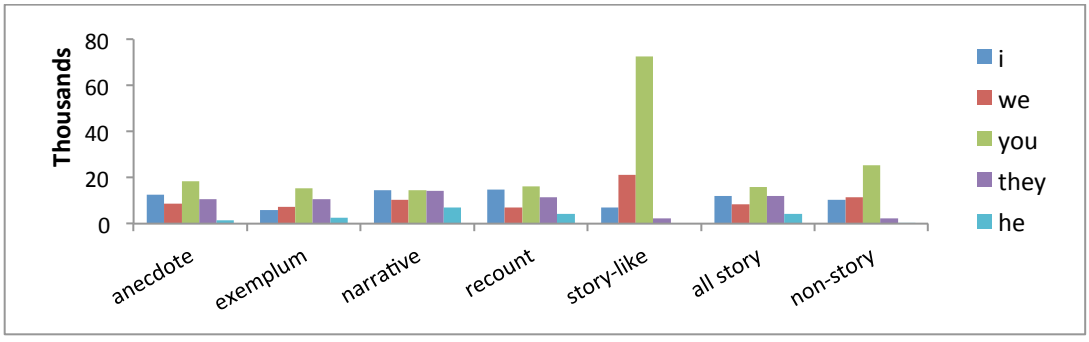


Figure 7.20: Occurrence (pmw) of *i*, *we*, *you, they* and *he* in story, story types, story-likes and non-story

The contrast is marked: *you* is by far the more common to story-likes than story or non-story. Sometimes the *you* in ELC story-likes is second person plural and directly addresses the student body, as in "the tools I'm going to show you today" (3004). Ädel (2010: 81) points out that lecturers use *you* in specific reference to audience members and in generic reference. In story-likes, however, its usage is predominantly generic and conditional, as in:

> *if you* are arranging building a new building it's worth doing it properly *if you* um or even just installing a new heating system in a building um *if you* just get a plumber in and tell him to put in a heating system he'll normally put one in that's about twice as powerful as is actually needed um *if you* actually work it out properly and all the heating suppliers give you little programs on their websites to enable you to do this you can get a system that's properly balanced for the building and ah it shows you that in fact *you* don't need a huge amount of heat to heat a building (emphasis added. 1014)

In general in ELC stories and summaries, *you* is a negative keyword (see Table 4.7 and Table 4.5). In humour, *you* is positively key (see Table 4.6) and largely refers specifically to the audience. In the story-likes identified, *you* is by far the most common pronoun (see Figure 7.20), but it functions as an informal, generic alternative to *one.*

7.5. Conclusion

Like humour, the retelling of events in the ELC stories marks a short punctuation in the overall lecture speech event. Stories take up no more than a few minutes of the roughly hour-long lectures (3.5%, Table 4.1). Each occurrence constitutes on average only 120 tokens (Table 7.4). Yet stories undoubtedly perform a particular function in the lecture discourse discussed.

Using Labov and Waletzky's (1967) structural model to classify story according to Martin's (2008) genres, it emerged that narrative type storytelling occupies the most token space (Table 7.2) and recounts occur most commonly across the corpus (Table 7.3). Variation in average token length and distribution across subcorpora exists across all types.

The completeness of the narrative structure allows lecturers to relay details of real life engineering situations: they set the scene, give the problem and talk about how it was resolved, often with a postscript comment in coda form. Narratives are fairly evenly spread across all three cultural components. Personal narratives allow the lecturer the opportunity to model the role of an expert engineer, in the manner described by Dyer and Keller-Cohen (2000). It was noted that the UK narratives rely heavily on personal experience, whereas the Malaysian narratives rely heavily on the experiences of others.

One possible explanation for the difference in retelling personal experience in stories, suggested by a Malaysian colleague, is the different career trajectories of lecturers in the two countries. Engineering lecturers in the UK have often spent several years in industry before entering academia, whilst their Malaysian counterparts tend to enter academia at an earlier stage, pre-experience.

It is also possible that the Malaysian lecturers rely more heavily on pre-prepared course materials, perhaps because they are less confident about their own and their students' knowledge of English, and are therefore less willing to extemporise, or because in the Malaysian context there is a greater expectation that different lecturers delivering the same programme will cover the same ground.

Anecdotes and exempla are on average the least common storytelling genres in engineering lectures (based on normalised token and occurrence counts), but are also subject to the most culture-specific variation. On both counts, exempla are used more heavily in the Malaysian lectures, and are notably lacking in the New Zealand component. Anecdotes are far more common in the UK component.

Differences may possibly be due to differing beliefs about the role of lectures. Exempla illustrate points of information, so are more likely to be used when the lecture has a primarily informing role. Anecdotes perform a more entertaining function and appeal to the emotions; they may serve as a means of modelling attitudes towards incidents that are

likely to occur in the professional life of an engineer. In the UK there may be a greater emphasis on student autonomy, and if students are expected to discover key information for themselves, the purpose of the lecture changes; there is more space for the expression of thoughts and opinions more loosely related to the program of study.

Story-likes are informed by the experience of the lecturer in the same way as stories and often carry scientific judgment, but their presentation emphasises more directly, for pedagogic purposes, the potential pitfalls students may encounter.

ELC stories in general lack technical terminology. Although lecturers do make use of personal experience, there is a strong focus on external participants in the events described. The relatively high STTR across stories in comparison to non-story (see Table 4.4) suggests that the retelling of events is at least partially rehearsed or scripted.

CHAPTER 8. CONCLUSIONS AND FUTURE DEVELOPMENTS

8.1. Conclusions

8.1.1. Pragmatic features of lectures

In response to the aims laid out in 2.5, the data-driven process of pragmatic annotation developed to inform this thesis has led to the identification of linguistic features that typically realise various purposes of lecture discourse, namely humour, story and summary. In support of these aims, a tool for analysing the occurrence and duration of pragmatic features within lecture discourse – *ELVis* – has been created. Variation in the realisation of these functions across cultural/educational contexts has also been identified.

Each function sheds light on the specific nature of lecture discourse, is not easily recoverable from context, and occurs more than once in the corpus. Encoding and then visualising these features has enabled comparison of their location, duration and relative frequency in EMI lectures delivered in different cultural contexts. Variations in data patterns across the subcorpora point to the role of cultural difference in the delivery of the ELC lectures, regardless of consistency of language medium (English), discipline (engineering), and education level (undergraduate). Certain broad functions emerged as common to these engineering lectures, but the extent and duration to which they are used varies, as do the attributed types through which they are more finely expressed.

From a pedagogical perspective, students need to be able to interpret the functions of different parts of the lecture in order to understand the *message* that is being delivered. Analysis has shown that humour, storytelling and summarising do not occur in predictable positions. Therefore, students cannot assume that at certain points in the lecture the lecturer is going to do certain things. They can, however, learn to interpret cues relating to the linguistic and paralinguistic features of the lecture discourse. Likewise, teachers of EAP can use the identification of such features to support learning.

The annotation framework has identified and enabled quantitative and qualitative analyses of the language functions that are fundamental to engineering lectures. The breakdown of quantitative data extracted from the ELC annotation shows that the token count of the pragmatic features identified accounts for 16.41% of the corpus: humour (2.88%), story (3.45%), and summary (10.08%) (Table 4.1). Individual occurrences of each feature occur on average 27 times per lecture: humour (2), storytelling (8) and summarising (17) (Table 4.2). The average token count of each occurrence is: humour (36), story (121), and summary (42) (Table 4.3).

The overall distributions are split between the four types of summary identified (review content of previous lecture, review content of current lecture, preview content of current lecture, and preview content of future lecture), the nine types of humour (bawdy, black, disparaging, irony/sarcasm, joke, playful, self-deprecating, teasing/mock-threat, and wordplay), and the four types of story (anecdote, exemplum, narrative, recount). Variation in duration, occurrence and length is also evident at the level of attribute type and cultural subcorpora, as laid out in Table 4.1, Table 4.2, and Table 4.3. Appendix VI summarises the findings of this thesis in terms of the purpose, content, distribution and lexicogrammatical features of the pragmatic elements and attributes discussed. The table does not cover every point raised in the analysis chapters, but provides an overview of most of the main features.

Summary emerged as an important discourse function, as it accounts for on average a tenth of the total tokens in the ELC and some form of summarising occurs on average 17 times per lecture (Table 4.2). The general trend is for UK lecturers to summarise in short bursts moderately frequently, and for the Malaysian lecturers to summarise in slightly shorter bursts slightly more frequently. The lecturers from New Zealand give summaries least often, but each summary is on average much longer than in the other subcorpora.

Specific features of summaries have been identified, both in terms of the umbrella category and the four attributed types. Discourse structuring aids the processing of content because

new information can be more easily comprehended when it is predicted (Young 1994: 168). Various micro-structuring and signalling language devices are used in the ELC. The reviews repackage information to reinforce learning, and the previews predict new information to scaffold learning – both in the lecture theatre, and for later revision through note-taking.

ELC reviews and previews function to help the audience keep pace with the lecture programme. Introductory reviews of previous content, and in-lecture reviews and previews bring everyone up to speed. In the largely monologic lectures, student input is limited, but a sense of togetherness is encouraged through the language choices of the lecturer, such as strategies for minimising imposition and the use of the inclusive *we*. The frequent use of the inclusive *we* noticed within reviews and previews seems to mark the point in the lecturer where the audience is brought together. Students may progress at different speeds when working through calculations, for example, and it is during these periods, where non-summative (largely technical) language is used, that the more imperative and didactic *you* is more salient.

The lecturers who deliver the summary types make clear the pedagogical intent of the particular discourse function. Previews or reviews that refer to content outside the current lecture are characterised as supporting the incremental development of understanding. As one lecturer explains:

> the theory builds on very er easily very naturally from one lecture to the next each thing leads each lecture leads onto the next one (3004)

Previews of upcoming content in the current lecture are also presented as shoring up the foundations needed to contextualise and grasp the next step, as in:

> we're now going to hop into two side issues that we require to move forward (3024)
>
> now we're going to get into the equations this is probably going to be where everything starts fitting together (3026)

The point of delivering reviews of current content fits into this philosophy of equipping students with the knowledge (and reference material) required to progress, and ensuring that there are no gaps in understanding. One lecturer explicitly asks his students to make sure that their notes are accurate and complete, explaining why he has engaged in repeated types of review:

> folks can you please make sure that in the notes that you took off the board that you are when I spoke to you and re- re-emphasised it and summarised at the end those <event desc="points to board"/> are the three different situations depending on the material and depending on the cross section those are the formulas we use to calculate the shear stress in different types of beam (1007)

The thrust of all types of summary is to aid comprehension. Specifically, lecturers link knowledge acquisition to degree success, as in:

> for the academic purposes there are so many things that we need to discuss as far as aggregate is concerned uh what is aggregate what is the er the what is natural aggregate what is cast aggregate eh what is the effect of the shape of aggregate the texture the grade et cetera so these are the things that I would like to cover this morning (2003)

The significance of repackaging information to ensure that it can be or has been absorbed is clear; it is neatly expressed by the lecturer who states "I tried my best to help you understand what E was" (3010). The value placed upon this discourse function is not only made evident by its quantitative occurrence, but also by the way in which it is presented by lecturers.

The inclusion of various ways of summarising the same information increases the probability that different learner types, or learners who are paying more or less attention at certain points, can access at least one version or iteration of the given content. The general pattern of regular occurrence distributed across lectures supports the premise that gradual and repeated exposure to information enables familiarity and absorption. The discourse choices made by the ELC lecturers illustrate that information must be revised – students must be exposed to it several times, and in several ways (hence summaries). The macro-

level visualisations make clear this drip effect, which may be especially pertinent to disciplines that are heavy in complex and incrementally acquired concepts, such as engineering.

The primary function of using humour, on the other hand, does not relate as strongly to content acquisition or scaffolding key concepts. The nine types of humour discussed in this thesis often flout Grice's (1975) maxim of manner; use of language to communicate clearly is not the aim as speakers deliberately subvert conversational expectations. The humour of the ELC lecturers relies heavily on switching and incongruity.

The breakdown of humour types identified within these lectures is weighted towards more hostile forms and the establishment of superiority. It is students, not outsiders, who are most commonly subjected to ironic or sarcastic jibes, disparaging remarks, or attacks in the form of teasing or mock-threat. A definite orientation towards industry is detectable in utterances that seem more native to exchanges on a building site, which I have referred to as *site humour*. By transferring this type of *site-talk* to class, there is a sense that the lecturers are putting students through a rite of passage.

Entrance into the professional world is also, however, bestowed through the use of humour types that attack an out-group member, or entail self-deprecation. The testing/education of students through humour is tempered by the show of togetherness established as the lecturer elevates the student recipients of humour above the butt of attacks on the self and others. If students are cast in the role of apprentices on the receiving end of workplace banter, they also have a protected place within the community.

Both humour type distribution and laughter response rate differ across subcorpora. The analysis of the co-occurrence of humour and laughter shows that the two are not co-extensive. Overall, laughter followed immediately in only half of all instances of humour annotated.

The findings explain why certain problems with humour are flagged up in the preparatory literature given to international students when they come to UK universities, but they also suggest that humour culture-shock may be multi-directional. Humour is identified as more common in lectures from the UK subcorpus of the ELC. Irony and sarcasm in UK lectures, for example, may be difficult for non-UK students to distinguish. The use of wit and wordplay, however, is most common in the Malaysian lectures, whilst self-deprecation occurs for the longest duration in the New Zealand lectures.

As in the use of humour, when lecturers tell stories, conveying information is not the primary intention. Storytelling also performs some kind of social, or socialising, function. Each lecturer has made decisions about which raw materials of characters and events to outline, and how these are best transmitted. Significant expertise in pedagogy, if not in storytelling, underlies these decisions.

The story interludes in ELC lectures offer students something they are unlikely to find in their written course materials: a vicarious experience of real-world engineering problems. The stories discussed offer an insight into the type of world in which many of the audience will spend their future careers. Some examples can be read as typical on-site learning-by-doing (such as rewiring, pouring concrete, and checking legislation). The ELC stories may be read as an inclusive gesture; an invitation into, rather than just description of, the club.

In this sense, the lecturer's position is one of both experienced professional and gatekeeper. Echoes of the emphasis on problem-solving and the influence of professional practice discussed in Chapter 1 are evident throughout the instances of storytelling, humour and summarising identified.

8.1.2. Pedagogical implications

Around the world, engineering lecturers and students are being encouraged to deliver or receive lectures in the medium of English when they have little experience, and access to

little material that can help them. An important impact of the findings of this thesis is in the area of ESP/EAP and staff development. There are many participant groups who could benefit from this research. For example, new university staff who are proficient in English may need help in constructing lectures in their field. Novice students may also struggle to understand the lecture genre. Teachers of EAP might benefit from better understanding the nature of lectures. Students and lecturers who are not proficient in English may know how to deliver or listen to lectures, but might find translation of the instruction language difficult. Table 8.1 indicates some possible variables in relation to this matrix of genre and language proficiency.

	engineering student	engineering lecturer	EAP teacher
not proficient in EMI	x	x	
not proficient in delivering/ receiving lectures	x	x	x

Table 8.1: Potential impact of research

Students and lecturers from contexts where informing is the prime purpose of lectures may have difficulty adapting to freer styles of lecture delivery. The inclusion of storytelling, summarising and humour may be problematic for those who are accustomed to treating all parts of the lecture in the same way. For example, students may be used to making notes when the lecturer provides key facts, and also when he/she uses the type of pragmatic language discussed. Lecturers may not be used to incorporating pragmatic language into their delivery.

These groups may benefit from exposure to samples of, for example, authentic narratives of personal engineering experience or instances of sarcasm, so that they can become acquainted with the features of the lecture genre and learn to interpret their purpose, relating experiences to their own prior knowledge and their future circumstances. Examples of authentic material (such as stories, humour and instances of summarising) are difficult to

source from published EAP/ESP materials, however, as lecture extracts in published materials are often scripted, and lack many of the pragmatic features noted in authentic lectures (see, for example, Nesi 2012a).

In the increasingly globalised field of English-medium lecture delivery, understanding differences between pragmatic features is valuable to both students and staff. Knowing more about what to expect – as deliverer or receiver – can only enhance the positive communicative effects intended, and reduce unwanted effects such as alienation. These features seem to play an important role in lectures across a range of cultural contexts, and it is therefore important not to neglect them, for example, when teaching academic listening skills in the EAP/ESP classroom or when training new lecturers.

It is impossible to come to any final conclusions about the reasons for the types of variation noted across the subcorpora because it is likely that the lecturers themselves cannot fully account for their selection of pragmatic language. Moreover, this is a small sample – of which pragmatic features account for fractions of the overall token count – and there are doubtless individual differences between lecturers regardless of the context in which they are operating. Nevertheless, it does seem clear from this small study that there are differences in lecturing style – specifically differences relating to the type and amount of pragmatic language delivered – that are useful to identify and discuss.

8.2. Future developments

8.2.1. Scale of raw data

This research could be extended by expanding the corpus to incorporate further subcorpora of EMI engineering lectures from different parts of the world, and data from a greater range of institutions in the geographical contexts discussed. Patterns identified in the current corpus represent only the language choices of a particular (and narrow) set of lecturers in three universities and cannot with any certainty be generalised beyond this scope. An

increase in both overall token count and the range of contexts from which data was collected would enable stronger quantitative claims and more nuanced qualitative comparisons to be made.

A bigger dataset would also facilitate greater statistical certainty, which in turn would enable a more refined and representative description of categories. Currently, potential distinctions at the attribute level cannot be confirmed due to a lack of data. The non-taxonomical potential *negative* type attribute, for example, was identified as a possible refinement to the review/preview attribute of the summary element (see 5.2.3), as was a potential *story-like* attribute (or perhaps element) in relation to story (see 7.4.5). There may also be a case for distinguishing between summaries that are purely metatextual, as in:

> we've done axial stress both in materials and we did a little bit of it last term (1008)

Compared to summaries that contain information which is repeated elsewhere in the text, as in:

> so remember I told you yesterday that if you have a current flow you always get magnetic fields there's no exception (3001)

Finer adjustments to the hierarchy of the current ELC taxonomy and/or the identification of new elements or attributes might be possible with more, and more diverse, data.

The breadth of subcorpora could also be expanded to include, for example, all lectures on a course attended by a single degree cohort, which would enable better description of pragmatic features related to course structure (such as reviews of previous lectures and previews of future lectures). Access to the full range of preceding, proceeding and parallel modules would result in the construction of a more detailed matrix of such relational language structures.

Enhancement of the scale of data would also allow confirmation of cross-lecture references within the pragmatic language identified. For example, a finding of this thesis was that a

common formula amongst lecturers is to accompany a review of previous information with a chronological reference, as in:

> so remember I told you yesterday that if you have a current flow you always get magnetic fields (3011)

Without access to the referenced lecture, the summarised content cannot be validated, and the relations of use of pragmatic language at a higher structural level (within the module or course) cannot be examined.

8.2.2. Supporting auxiliary material

Discussion in this thesis was limited to patterns identified in transcribed lecture data. Confidence in the interpretation of this data would be increased through the collection of auxiliary data, especially if the researcher who gathered the supporting information was also the person who analysed the corpus (Flowerdew 2005: 329, Handford 2010: 37-38). For example, Othman (2010: 678) identified a gap between researcher interpretation and lecturer perception of the intended usage of micro markers in lecture discourse. Although three stages of IAR were performed on the ELC, the results have not been triangulated with auxiliary data, such as interviews with lecturers or students. Both the initial coding and IAR test results are therefore, to an extent, based on plausibility rather than certainty (cf. Mann and Thompson 1988: 246). A gap may exist, for example, in whether a lecturer intended to use sarcasm and whether their students interpreted it as such. Adding ethnographic information such as interviews of the type undertaken by Straker Cook (1975) may clarify intended pragmatic usage in some cases and so increase the reliability of the annotations.

8.2.3. Enhancement of header metadata

In addition to scaling up the amount of raw text, an increase in the amount of sociolinguistic metadata collected would also enable more informed conclusions to be reached. The ELC headers currently include limited description. The inclusion of further metadata – such as specific figures for the L1 of students and lecturers, the gender of students, the level of

experience and educational background of lecturers – would also enable more insightful analysis of the reported findings.

8.2.4. Enhancement of annotation

Some elements in the 2014 ELC taxonomy would be more accurately characterised by more detailed description in their annotation. Currently, strings of text are annotated as one element, and in some cases a single attributed type, such as <summary type="review content of previous lecture">. Multiple attributes per element are not allowed within TEI conventions. However, in the case of the summary element, for example, several instances were found where the annotated string primarily constitutes a preview of the current lecture content, but also inseparably other types of summarising (see Table 3.4).

Further predefined legal key/value pairs would need to be added to the element hash to refine the description and more precisely guide resultant qualitative analysis. One method would be to add another key named *subtype* (in addition to *type*) to the DTD, which could be paired with existing or new attributes of elements (such as summary), enabling more complex descriptions, as in:

> <summary type="preview content of current lecture" subtype="review content of previous lecture">
>
> <summary type="review content of current lecture" subtype="metatextual">
>
> <summary type="preview content of current lecture" subtype="negative">

Analysis at the deeper level of attribute combinations through such improvements would better inform understanding of teaching techniques, such as information structuring and managing expectations.

The current method of storing the subjective linguistic annotation inline alongside the structural markup makes the ELC unnecessarily difficult to parse, and also somewhat conflicts with AHDS guidelines for corpus construction (Sinclair 2005). Although this did not

pose problems for the completion of this thesis due to programmatic workarounds, it does create a barrier to the wider project aims of continuing to add to the corpus, and to making it publicly accessible.

A planned future improvement is to separate the markup and annotation: to convert the inline pragmatic annotations indices into stand-off form and store them in separate XML files. Although the event descriptions have been categorised as a type of structural markup, it could be argued that they better fall into the category of annotation as the information recorded is to an extent subjective interpretation. For this reason, ideally a layer of event descriptions would also be stored separately.

To implement this development, all raw transcribed text must be static in order that the indices of the annotations in the stand-off files are correct; the original transcripts must be completely accurate before stand-off files can be created, and the transcripts cannot be edited post-annotation. Adjustment to the workflow model (Figure 3.2) would be necessary in order that the requirement for static text in Phase 1, with no further amendments in Phases 2-4, was met. Conversion of the current corpus to this format could be achieved using a Python script. The XML would be parsed and an internal representation could then be manipulated to output the desired stand-off XML files.

The use of stand-off annotation would enable various forms of linguistic annotation to be applied in addition to the identification of pragmatic features. The next stage in the annotation process for the ELC is to part-of-speech (POS) tag all word forms so that some linguistic structures within pragmatic features can be automatically extracted. The addition of prosodic and semantic annotation layers would also offer an additional means of interrogating and describing the character of the pragmatic text identified in this thesis. As well as examining their interrelation, the implementation of various levels of annotation would make the ELC a more useful resource for a wider range of research interests. As the

corpus grows, it would be most efficient to use dedicated software for creating, editing and exporting all levels of metadata, as described in the next section (8.2.5).

8.2.5. Enhancement of structural markup

A significant limitation of the current project is the absence of time codes; analyses in this thesis are based on token measures. Adding some form of time codes to the transcribed speech would increase the levels of analysis that can be undertaken. Chronological measurement, for example, would enable particular spans of the lecture to be investigated, such as the first ten minutes, rather than first 500 tokens as in example of lecture introductions (discussed in 4.7). The most powerful application of time codes is that they enable text and corresponding audio/video data to be aligned. This function is particularly useful for pragmatic interpretation, for example through being able to see gestures and facial expressions. Alignment can be simply achieved if the text is time-stamped with self-closing tags at regular, arbitrary, intervals. Another option would be to add time codes as an attribute of tags that enclose some kind of meaningful unit, such as an utterance or clause.

The addition of some form of meaningful segmentation of transcribed speech would also broaden the range of possible analyses. Currently the ELC contains no punctuation, as the project team initially considered the process to be too subjective to apply to spontaneous speech. Distinct utterances are structurally marked up, but in the largely monologic ELC lectures utterance units can span the entire lecture uninterrupted.

Ideally, the ELC would follow the example of corpora that have segmented transcribed speech. The Nordic Dialect Corpus and Database (Johannessen et al. 2009), for example, uses the Glossa web interface, which returns both search concordances and corresponding media files (transcripts and audio/video fragments). The videos and transcripts are aligned based on time encoded *segments*, which are analogous (as far as possible in speech data) to sentence-level structures; they are the largest string of words that can be syntactically parsed. In extended speech, the division can be based on small pauses, intonation, or the

completion of a grammatical unit/sentence; an *ideal segment* is approximately ten seconds in length (Hagen and Khachaturyan 2015). The retrieved search term/s can be returned in the context of the video segment in which it occurred. Acoustic analysis is also enabled based on these time-encoded segments (Kosek et al. 2015).

The level of detail of the descriptions within the structural markup of the ELC text is currently very basic and would be enhanced through alignment with the video files. The ELC lecturers draw on kinaesthetic or visual data to reinforce key ideas. One lecturer from the UK, for example, demonstrates the effect on viscosity that superplasticisers have on a stiff concrete mix (1013). The demonstration involves passing around a container of the combined mixture for the audience to stir. Within the transcribed speech, the event is simply encoded in the text through tags such as <event desc="stirs mixture"/> and <event desc="passes container amongst audience"/>. In another case, a Malaysian lecturer illustrates the effect of forces on bridges by describing the features of a projected image (1002). He examines in detail the operation of structural components such as pin joints and roller bearings through reference to the image, which is recorded at various points in the transcript as <event desc="lecturer points to board"/>. Such events are currently encoded with limited description through self-closing tags and would benefit from a more detailed textual account and/or visual access to the recorded event. Another beneficial adjustment would be for opening and closing tags to enclose the entire duration of the event described.

If the ELC transcripts and their corresponding lectures were time-aligned and accessed via an interface, ideally functionality for displaying referenced material would also be incorporated. For example, the original JPEG image of the bridge could be available for inspection in a separate window. Other information external to the lecture could also be made accessible, such as the relevant parts of course syllabi, or extracts from interviews with lecturers and/or students that augment understanding of particular passages of lecture discourse.

In the process of constructing the ELC, an area that did not receive sufficient attention was potential options for aligning the multimedia files. Such alignment seems to be common in corpora of naturally occurring speech that focus on dialectical variation and privilege metadata such as regional origin, ethnic and social background. This is not the case in corpora of academic speech; neither BASE nor MICASE transcripts are aligned to the original recorded lectures, for example. The software to align files is, however, available to corpus linguists.

Dedicated annotation tools for creating, editing, visualising and searching annotations for audio/video data also offer extensive functionality. Specialist corpora can be built with aligned audio files and transcripts using corpus query software, such as WordSmith Tools (Thompson 2010). The Multi-purpose Annotation Environment (MAE) (Stubbs 2011) offers a straightforward option for corpus annotation using a simple DTD to describe attributes and elements, colour-encode selected strings of text, and export the annotation in stand-off format. More tailored, freely available, options include: Glozz (Widlöcher and Mathet 2012), WebAnno (Yimam et al. 2014), CAT: Content Annotation Tool (Bartalesi Lenzi, Moretti and Sprugnoli 2012), FoLiA Linguistic Annotation Tool – Flat (van Gompel and Reynaert 2013), and eMargin: A Collaborative Textual Annotation Tool (Kehoe and Gee 2013). Perhaps one of the strongest options is ELAN (EUDICO Linguistic Annotator) (The Language Archive 2015). ELAN displays speech and/or video signals (together with their annotations), links different layers of annotations together and to media streams, and allows an unlimited number of annotation layers to be constructed, along with various export options. Open-source corpus interfaces that display concording text and video are also increasingly available and well-supported, as in the case of the Glossa interface (Johannessen et al. 2009, Kosek et al. 2015).

An expanded ELC corpus would benefit significantly from: 1. the use of dedicated transcription software in which to automatically generate stand-off annotation layers and time codes, and 2. the use of an interface through which to display the aligned files, as well

as auxiliary material. The interface option could be achieved through improvements to the *ELVis* range of interactivity (such as linking the source view to the corresponding video fragment) or through using a pre-existing system such as Glossa.

8.2.6. Storage of the ELC

Corpus storage is another area that warranted greater consideration in the planning stages of the ELC's construction. Particularly in light of its planned expansion, the corpus should be stored in a web-based repository, such as GitHub (GitHub Inc 2015). Problems in collating raw data and metadata from collaborators were experienced during the initial corpus construction phase of this thesis, and revisions during the four annotation phases have not been systematically documented. One of the biggest advantages of a repository to an expanding project is the version control system, which documents all changes. In terms of collaboration, GitHub also allows write access for multiple users (such as a project team). The deposited material can also be transferred to new accounts via *forking*, which would enable distribution of the corpus. GitHub is free (if the stored project is made publicly available) and open source, which adheres to the principles of shared data that underpin the development of the ELC.

As unmaintained corpora are susceptible to being lost, for preservation and dissemination purposes, the finalised ELC could also be deposited in field-specific stable and relevant archives. For example, depositing the ELC with the Oxford Text Archive (OTA) (University of Oxford 2015) and/or the Common Language Resources and Technology Infrastructure (CLARIN) (Hinrichs and Krauwer 2014) would enable the holdings and metadata to be easily accessed by the scientific community.

To facilitate replicability and comparability, all documentation regarding precise guidelines – what text to annotate, what tags to apply in what circumstances, and how to deal with special cases (cf. Petrillo and Baycroft 2010: 3) – should be collated into a publicly available handbook along with transcription and metadata protocols. Such a manual will be published

when the corpus is made publicly available. Looking to the future, freely accessible and well-documented instances of pragmatic annotation would also help to build a model that contributes to the TEI system and that can be integrated into other similar resources.

8.2.7. Reliability testing

A valuable extension of the IAR testing process would be more detailed comparison of annotations. In the first iteration of testing (a master annotator versus three partial annotators) (3.5.2), the calculation of reliability was only done at the level of elements, not attributes. The practical reason for this was that at the attribute level, changes occurred as the master annotator cycled through the entire corpus and made adjustments in terms of both the presence and the hierarchy of certain attributes (see 3.4.6). Comparison of annotations at the attribute level was therefore not possible; the only comprehensive result was at the element level, which did not change by the third pass at annotation.

I do not expect this process of adjustment to cease, especially as further subcorpora are added. It is likely that continued adjustment will occur as various annotators feed back into the content and hierarchy of the working list. Despite the impact on the feasibility of comparison at the finer attribute level, this is a desired process of cyclical adjustment (cf. Wallis 2014). It is also an issue that is addressed both in the procedures adopted in the version of the corpus discussed (2014) (as laid out in the workflow model in Figure 3.2) and recommendations for the future expansion of the corpus in this section. A requirement of any ongoing adjustment is the implementation of a system of version control for referencing purposes.

Section 3.5.2 describes a script written to compare intersect values for IAR testing. If an alternative method of IAR testing was sought (for example for future collaborators who did not want to run the script or amend it to output a different type of calculation), it would be efficient to incorporate automated IAR testing into the interface proposed in section 8.2.5. An alternative/additional way of comparing the indices would be to modify the functionality

of *ELVis* and visually investigate annotation pairs which return low or no intersection. Re-purposing *ELVis* in this way would enable manual post-correction via the source text view, but would require significant modification to functionality.

8.2.8. Developing ELVis

Visualisation of the annotation metadata gives both a starting point from which to explore, and means of communicating/disseminating, evidence that supports the research questions (2.5). *ELVis* currently meets these criteria for internal project use only. It would need significant modification for either public release or to be fit for purpose if the ELC was expanded.

If the dataset was scaled up, showing every lecture on the timeline would be unfeasible. Some form of data compression or filtering would need to take place to avoid overload on cognitive processing. The implementation of filtering options by variables of lecture, subcorpus or pragmatic feature would offer a first step in data view reduction. The design of *ELVis* is currently limited to the principles of small multiples and various types of bar-chart representation. The addition of other visualisation types would augment the macro-level entry point into lecture data and enable data compression through clustering techniques. For example, the ELC data could also be visualised using techniques from correspondence analysis (CA) applied to spoken corpora (see, for example, Nakamura 2002). CA is a multivariate statistical technique that can be applied to categorical data for both communicative and exploratory purposes. CA would better take into account the effect of confounding variables on the variable under investigation (such as pragmatic feature), which would be particularly pertinent if the proposed extension of the header metadata (8.2.3) was implemented.

One of the strengths of using *ELVis* is that outlying results which may impact on the reliability of statistical analyses can be easily seen. This enables the interpreter to either discard deviant profiles or tailor the choice of statistical measure towards those that are

robust to outliers. The identification of deviant profiles also flags up the need for further data compression, and/or flexibility in approach to data visualisation. For example, the addition of an option to render a box and whiskers plot would show distance from the median value of all results, giving further perspective to the range of outlying results by describing central tendencies.

All *ELVis* data and libraries (D3 and jQuery) are currently stored locally for offline use. For easy access, the ability to upload new data should be added. This would involve adapting the pipeline, which is at present based on a static data structure. Currently, the annotated and marked-up ELC XML files are read into a string token by token rather than parsed, which was a workaround necessitated by problems with invalid XML source data at the time of development (as described in 8.2.4). Although the source data is now valid, the workaround has not yet been updated. Parsing the XML and counting tokens in this way would be more robust. Then, a simple web server could be set up to parse any data of the same format (the commonplace TEI-compliant XML) and serve a JSON structure from which the visualisation could be generated. The addition of a front-end interface (suggested in 8.2.5) would allow documents to be uploaded and libraries to be linked to online.

In terms of data validation, basing the visualisation on dynamic data would also be useful for the development of the corpus in terms of data modification. For example, due to the subjective nature of the annotations, errors are an inevitable phenomenon. Rather than modifying and re-uploading the currently static source data, it would be efficient to have dynamic content that could be modified via the source view text in *ELVis*. Modification could occur either following IAR tests, or in earlier stages of general annotation checking, depending on whether stand-off annotation files were generated (see 8.2.4).

A final consideration is that of interoperability between corpus tools of the type investigated by Nesi et al. (2012). ELVis is currently intended to perform a very specific, quantitative function, which is augmented in this thesis by other types of corpus linguistic

analysis. The growth of the ELC will require more robust processes for structuring, storing and analysing the data. One solution would be to incorporate *ELVis* into a tailored interface alongside linguistic analysis tools for various types of automatic parsing, annotation alignment, annotation agreement measurement, text mining and text querying. Alternatively – to avoid reinventing the wheel – integration with the architecture of existing open-source software could be explored.

8.2.9. Other questions to explore

Expanding the corpus and refining the systems of annotation, markup and storage will also better place me to investigate questions that were outside the remit of this thesis, but remain central to understanding the particular character of engineering lecture discourse.

With more data from contributors whose L1 is not English, the role of code-switching in lectures can be explored. In the current corpus, there was too little code-switching in the Malaysian component to draw any conclusions regarding patterns of usage. With more examples, the question of whether lecturers are more likely to switch to their L1 when using pragmatic language can be interrogated – and if so, I can begin to look at why.

The ELC lecturers' use of humour, story, and summary accounts for about 16% (tokens) of the overall corpus; the other 84% is visualised as the blank spaces between colour-encoded blocks in *ELVis*. The macro and micro features of the text referred to in shorthand as *non-annotated* have not been explored in detail. I hypothesise that these blank spaces denote the main informational lecture content, where lecturers explain technical concepts and work through problems. Systematic analysis of this main content would indicate what sort of language typifies the rest of the lecture; it would fill in the blanks, as it were.

References

Ädel, A. (2010) 'Just to Give You Kind of a Map of Where we are Going: A Taxonomy of Metadiscourse in Spoken and Written Academic English'. *Nordic Journal of English Studies* 9 (2), 69-97

Aijmer, K. and Rühlemann, C. (eds.) (2015) *Corpus Pragmatics: A Handbook.* Cambridge: Cambridge University Press

Allen, J. and Core, M. (1997) *Draft of DAMSL: Dialogue Act Markup in several Layers*. University of Pennsylvania: Discourse Research Initiative

Allison, D. and Tauroza, S. (1995) 'The Effect of Discourse Organisation on Lecture Comprehension'. *English for Specific Purposes* 14 (2), 157-173

Alsop, S. and Nesi, H (2009) 'Issues in the development of the British Academic Written English (BAWE) corpus'. *Corpora* 4 (1), 71-83

Ammon, U. and McConnell, G. (2002) *English as an Academic Language in Europe: A Survey of its use in Teaching*. Frankfurt: Peter Lang

Andeweg, B., Gagestein, S., de Jong, J., and Wackers, M. (2011) ''Poke Fun at Yourself': The Problem of Self-Deprecating Humor'. *SEFI Conference Global Engineering Recognition, Sustainability and Mobility*, 759-764

Andrade, M. S. (2006) 'International Students in English-Speaking Universities: Adjustment Factors'. *Journal of Research in International Education* 5, 131-154

Anthony, L. (2011) *AntConc (Version 3.2.2)* [online] available from <http://www.antlab.sci.waseda.ac.jp/> [15/03 2012]

Aristotle (1996[335BC]) *Poetics*. trans. by Heath, M. London: Penguin

Artstein, R. and Poesio, M. (2008) 'Inter-Coder Agreement for Computational Linguistics'. *Computational Linguistics* 34 (4), 555-596

Attardo, S. (2003) 'Introduction: The Pragmatics of Humor'. *Journal of Pragmatics* 35, 1287-1294

Attardo, S. (1994) *Linguistic Theories of Humour*. New York: Mouton de Gruyter

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C., and Wilcock, S. (2000) 'A Comparative Evaluation of Modern English Corpus Grammatical Annotation Schemes'. *ICAME Journal* 24, 7-23

Austin, J. (1962) *How to do Things with Words*. Cambridge, MA.: Harvard University Press

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum

Baker, P., Hardie, A., and McEnery, A. (2006) *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press

Bamford, J. (2004) 'Gestural and Symbolic Uses of the Deictic *here* in Academic Lectures'. in *Discourse Patterns in Spoken and Written Corpora*. ed. by Aijmer, K. and Stenström, A. Amsterdam and Philadelphia: John Benjamins, 113-138

Bartalesi Lenzi, V., Moretti, G., and Sprugnoli, R. (2012) 'CAT: The CELCT Annotation Tool'. *Proceedings of LREC 2012*. held 23-25/05/12 at Istanbul, Turkey. European Language Resources Association

Berry, R. (2005) 'Making the most of Metalanguage'. *Language Awareness* 14 (1), 3-20

Bhatia, V. K. (1997) 'Genre-Mixing in Academic Introductions'. *English for Specific Purposes* 16 (3), 181-195

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman

Biglan, A. (1973) 'The Characteristics of Subject Matter in Different Academic Areas'. *Journal of Applied Psychology* 57 (3): 195-203

Blakemore, D. (1992) *Understanding Utterances*. Oxford: Blackwell

Blumer, H. (1969) *Symbolic Interactionism: Perspective and Method*. Englewood Cliffs, NJ: Prentice Hall

BNC Consortium (2007) *The British National Corpus. Version 3 (BNC XML Edition) (Distributed by Oxford University Computing Services)* [online] available from <http://www.natcorp.ox.ac.uk/> [01/07 2011]

Brenn-White, M. and Faethe, E. (2013) *English-Taught Master's Programs in Europe: A 2013 Update*. New York: Institute of International Education, Center for Academic Mobility Research

Brewer, C. A. (1999) 'Color use Guidelines for Data Representation'. *Proceedings of the Section on Statistical Graphics*. Alexandria, VA: American Statistical Association

Brinton, L. J. (1996) *Pragmatic Markers in English: Grammaticalization and Discourse Functions*. Berlin: Mouton de Gruyter

Broeder, D. and Wittenburg, P. (eds.) (2001) *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources. Vol 15*. 'Interaction of Tools and Metadata-Descriptions for Multimedia Language Resources'. Stroudsburg, PA, USA: Association for Computational Linguistics

Brown University (2011) *Brown University: The Plan for Academic Enrichment*. Rhode Island, US: Brown University

Brown, P. and Levinson, S. (1987) *Politeness*. Cambridge: Cambridge University Press

Bublitz, W. and Norrick, N. R. (eds.) (2011) *Foundations of Pragmatics (Handbooks of Pragmatics).* Berlin and Boston: De Gruyter Mouton

Butcher, A. and McGrath, T. (2004) 'International Students in New Zealand: Needs and Responses'. *International Education Journal* 5 (4), 540-551

Camiciottoli, B. C. (2005) 'Adjusting a Business Lecture for an International Audience: A Case Study'. *English for Specific Purposes* 24 (2), 183-199

Capozzoli, M., McSweeney, L., and Sinha, D. (1999) 'Beyond Kappa: A Review of Interrater Agreement Measures'. *The Canadian Journal of Statistics* 27 (1), 3-23

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (eds.) (1999) *Readings in Information Visualization: Using Vision to Think.* San Francisco, CA: Morgan Kaufmann

Cardiff University Careers Service (2009) *Career Planning and Identifying Employers for International Students* [online] available from <www.cardiff.ac.uk/carsv> [01/01 2014]

Carnap, R. (1942) *Introduction to Semantics*. Cambridge, MA.: MIT Press

Carpendale, S. (2003) *Considering Visual Variables as a Basis for Information Visualisation. Research Report 2001-693-16*. Calgary, Canada: Department of Computer Science, University of Calgary

Carrell, P. L. and Eisterhold, J. C. (1983) 'Schema Theory and ESL Reading Pedagogy'. *TESOL Quarterly* 17 (4), 553-573

Chafe, W. (2007) *The Importance of Not being Earnest: The Feeling Behind Laughter and Humour*. Amsterdam; Philadelphia: John Benjamins

Chafe, W. (1985) 'Some Reasons for Hesitating'. in *Perspectives on Silence*. ed. by Tannen, D. and Saville-Troike, M. Norwood, NJ.: Ablex, 21-30

Chatman, S. (1978) *Story and Discourse: Narrative Structure in Fiction and Film*. Ithaca: Cornell University Press

Chaudron, C. and Richards, J.R. (1986) 'The Effect of Discourse Markers on the Comprehension of Lectures'*. Applied Linguistics* 7 (2), 113-127

Cheng, S. W. (2010) 'A Corpus-Based Approach to the Study of Speech Act of Thanking'. *Concentric: Studies in Linguistics* 36 (2), 257-274

Chi, E. H. (2000) 'A Taxonomy of Visualization Techniques using the Data State Reference Model'. *Proceedings of IEEE Symposium on Information Visualization (InfoVis'00)*. IEEE Computer Society Press

Chipman, G. (1996) *Review of High Speed Visual Estimation using Preattentive Processing* [online] available from <http://www.cs.umd.edu/class/spring2002/cmsc838f/preattentive.ppt#267> [15/01 2014]

CLARIN (2009) *Common Language Resources and Technology Infrastructure (CLARIN) Short Guide: Standards for Text Encoding* [online] available from <http://www.clarin.eu/files/standards-text-CLARIN-ShortGuide.pdf> [01/07 2011]

Cobley, P. (2001) *Narrative*. London: Routledge

Codrops (2014) *Codrops* [online] available from <http://tympanus.net/codrops/2012/05/23/understanding-the-rule-of-thirds-in-web-design/> [10/01 2014]

Cohen, J. (1960) 'A Coefficient of Agreement for Nominal Scales'*. Educational and Psychological Measurement* 20 (1), 37-46

Cohen, L., Manion, L., and Morrison, K. (2000) *Research Methods in Education*. 5th edn. London: Routledge Falmer

Collins, P. C. (2004[1987]) 'Cleft and Pseudo-Cleft Constructions in English Spoken and Written Discourse'. in *Corpus Linguistics: Readings in a Widening Discipline*. ed. by Sampson, G. and McCarthy, D. London and New York: Continuum, 85-94

Collins, P. C. (1991) *Cleft and Pseudo-Cleft Constructions in English*. London: Routledge

Connor, U. and Upton, T. A. (2004) 'The Genre of Grant Proposals: A Corpus Linguistic Analysis'. in *Discourse in the Professions: Perspectives from Corpus Linguistics*. ed. by Connor, U. and Upton, T. A. Philadelphia: John Benjamins, 235-256

Coventry University (2016) *British Academic Spoken English (BASE) Corpus: Holdings* [online] available from <www.coventry.ac.uk/base> [01/03 2016]

Crawford Camiciottoli, B. (2004) 'Interactive Discourse Structuring in L2 Guest Lectures: Some Insights from a Comparative Corpus-Based Study'. *Journal of English for Academic Purposes* 3 (1), 39-54

Creer, S. and Thompson, P. (2005) 'TEI Mark-Up of Spoken Language Data: The BASE Experience'. *Reading Working Paper in Linguistics* 8, 149-174

Culpeper, J. (2012) 'The Metalanguage of IMPOLITENESS: Using Sketch Engine to Explore the *Oxford English Corpus*'. in *Contemporary Corpus Linguistics*. ed. by Baker, P. London and New York: Continuum, 64-86

Culpeper, J., Archer, D., and Davies, M. (2008) 'Pragmatic Annotation'. in *Corpus Linguistics: An International Handbook*. ed. by Kytö, M. and Lüdeling, A. New York: Mouton de Gruyter, 613-642

Culy, C. (ed.) (2013) *Arbeitstreffen: Netzwerk Lexicographie*. 'Visualization of Language and Linguistic Information'. held 20/11/13 at IDS Mannheim

Dafouz, E., Núñez, B., and Sancho, C. (2007) 'Analysing Stance in a CLIL University Context: Non-Native Speaker use of Personal Pronouns and Modal Verbs'. *The International Journal of Bilingual Education and Bilingualism* 10 (5), 647-662

Davies, B. and Harré, R. (1990) 'Positioning: The Discursive Production of Selves'. *Journal for the Theory of Social Behaviour* 20 (1), 44-63

DCMI (2010) *Dublin Core Metadata Element Set, Version 1.1* [online] available from <http://dublincore.org/documents/dces/> [01/07 2011]

De Felice, R. and Deane, P. (2012) *Identifying Speech Acts in E-Mails: Toward Automated Scoring of the TOEIC® E-Mail Task (ETS RR–12-16)*. Princeton, New Jersey: ETS

de Wit, H. and Jones, E. (2012) *Five Years of Changing Internationalisation Agendas* [online] available from <http://www.universityworldnews.com/article.php?story=20121009165209797> [10/04 2014]

Dearden, J. (2014) *English as a Medium of Instruction – A Growing Global Phenomenon*. London: British Council

Deroey, K. L. B. (2013) *Relevance Markers in Lectures: A Corpus-Based Study*. PhD thesis. Gent: Universiteit Gent

Deroey, K. L. B. (2012) '"What They Highlight Is...": The Discourse Functions of Basic "Wh"-Clefts in Lectures'. *Journal of English for Academic Purposes* 11 (2), 112-124

Deroey, K. L. B. and Taverniers, M. (2012) 'Just Remember this: Lexicogrammatical Relevance Markers in Lectures'. *English for Specific Purposes* 31, 221-233

Deroey, K. and Taverniers, M. (2011) 'A Corpus-Based Study of Lecture Functions'. *Moderna Språk* 105 (2), 1-22

Dickinson, M. and Lee, C. M. (2009) 'Modifying Corpus Annotation to Support the Analysis of Learner Language'. *CALICO Journal* 26 (3), 545-561

Dipity (2011) *Dipity* [online] available from <http://www.dipity.com/> [10/01 2014]

Doiz, A., Lasagabaster, D., and Sierra, J. M. (eds.) (2013) *English-Medium Instruction at Universities: Global Challenges.* Great Britain: Short Run Press

Dubois, S. and Sankoff, D. (2001) 'The Variations Approach Toward Discourse Structuring Effects and Socio-Interactional Dynamics'. in *The Handbook of Discourse Analysis*. ed. by Schiffrin, D., Tannen, D., and Hamilton, H. E. Oxford and Massachusetts: Blackwell, 282-303

Dunkel, P., A. and Davis, J., N. (1994) 'The Effects of Rhetorical Signalling Cues on the Recall of English Lecture Information by Speakers of English as a Native Or Second Language'. in *Academic Listening*. ed. by Flowerdew, J. Cambridge: Cambridge University Press, 55-74

Dunning, T. (1993) 'Accurate Methods for the Statistics of Surprise and Coincidence'. *Computational Linguistics* 19 (1), 61-74

Dyer, J. and Keller-Cohen, D. (2000) 'The Discursive Construction of Professional Self through Narratives of Personal Experience'. *Discourse Studies* 2 (3), 283-304

EAC (2016) *About Us* [online] available from <http://www.eac.org.my/web/about_EAC.html> [01/03 2016]

Eggins, S. and Slade, D. (1997) *Analysing Casual Conversation*. London: Cassell

Eisenberg, A. (1986) 'Teasing: Verbal Play in Two Mexicano Homes'. in *Language Socialization Across Cultures*. ed. by Schieffelin, B. and Ochs, E. Cambridge: Cambridge University Press, 182-198

Elliott, J. and Elliott, D. (2003) 'The Human Language Chorus Corpus (HULCC)'. *Proceedings of the Corpus Linguistics 2003 Conference*. held 28-31/03/03 at UCREL. Lancaster: Lancaster University

Engineering Council (2016) *Accreditation of Higher Education Programmes* [online] available from <https://www.engc.org.uk/education-skills/accreditation-of-higher-education-programmes/> [01/03 2016]

European Council (2000) *Presidency Conclusions, Lisbon European Council, 23-24/03/2000*. Lisbon: European Council

Fillmore, C. (1994) 'Humour in Academic Discourse'. in *What's Going on here? Complementary Studies of Professional Talk*. ed. by Grimshaw, A. and Burke, P. J. Norwood, NJ: Ablex

Fillmore, C. J. (1976) 'Frame Semantics and the Nature of Language'*. Annals of the New York Academy of Sciences* 280 (1), 20-32

Flowerdew, J. and Forest, R. W. (2015) *Signalling Nouns in English*. Cambridge: Cambridge University Press

Flowerdew, J. (2013) *Discourse in English Language Education*. New York: Routledge

Flowerdew, J. (ed.) (1994) *Academic Listening: Research Perspectives.* Cambridge: Cambridge University Press

Flowerdew, L. (2005) 'An Integration of Corpus-Based and Genre-Based Approaches to Text Analysis in EAP/ESP: Countering Criticisms Against Corpus-Based Methodologies'*. English for Specific Purposes* 24, 321-332

Fortanet, I. (2004) 'The use of 'we' in University Lectures: Reference and Function'. *English for Specific Purposes* 23 (1), 45-66

Fowler, R. (1991) *Discourse in the News: Language and Ideology in the Press*. London and New York: Routledge

Francis, G. (1986) *Anaphoric Nouns*. Birmingham: ELR

Fraser, B. (2009) 'Topic Orientation Markers'. *Journal of Pragmatics* 41 (5), 892-898

Fraser, B. (1999) 'What are Discourse Markers?'. *Journal of Pragmatics* 31, 931-952

Fraser, B. (1996) 'Pragmatic Markers'. *Pragmatics* 6 (2), 167-190

Fraser, B. (1988) 'Types of English Discourse Markers'. *Acta Linguistica Hungarica* 38 (1-4), 19-33

Freud, S. (1976[1905]) *Jokes and their Relation to the Unconscious*. trans. by Strachey, J. New York: Penguin Books

Fried, M. and Östman, J. O. (2005) 'Construction Grammar and Spoken Language: The Case of Pragmatic Particles'. *Journal of Pragmatics* 37 (11), 1752-1778

Gabrielatos, C. and Marchi, A. (2011) 'Keyness: Matching Metrics to Definitions'. *Corpus Linguistics in the South: Theoretical-Methodological Challenges in Corpus Approaches to Discourse Studies - and Some Ways of Addressing Them.* held 05/11/2011 at University of Portsmouth

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

Genette, G. and Levonas, A. (1976) 'Boundaries of Narrative'. *New Literary History. Readers and Spectators: Some Views and Reviews* 8 (1), 1-13

GitHub Inc. (2015) *GitHub* [online] available from <https://github.com/> [01/01 2015]

Glenn, P. (2003) *Laughter in Interaction*. Cambridge: Cambridge University Press

Goffman, E. (1967) *Interaction Ritual. Essays on Face-to-Face Behavior*. New York: Anchor

Goffman, E. (1959) *The Presentation of Self in Everyday Life*. Garden City, NY: Doubleday Anchor

Google Developers (2013) *Annotated Chart* [online] available from <https://developers.google.com/chart/interactive/docs/gallery/annotationchart> [21/01 2014]

Gotti, M. (2015) 'Code-Switching and Plurilingualism in English-Medium Education for Academic and Professional Purposes'. *Language Learning in Higher Education* 5 (1), 83-103

Grant, L. E. (2011) 'The Frequency and Functions of just in British Academic Spoken English'. *Journal of English for Academic Purposes* 10 (3), 183-197

Grice, P. (1975) 'Logic and Conversation'. in *Syntax and Semantics 3: Speech Acts*. ed. by Cole, P. and Morgan, J. New York: Academic Press, 41-48

Grönqvist, L. (2003) *TEI Or XCES? Porting the Göteborg Spoken Language Corpus to XML.* PhD thesis. Sweden: Graduate School of Language Technology (GSLT)

Hagen, K. and Khachaturyan, E. (2015) *Draft Guidelines for the use of ELAN in Speech Corpora.* Unpublished guidelines. Oslo: University of Oslo

Halliday, M. A. K. (1993) 'Language as Social Semiotic'. in *Language and Literacy in Social Practice: A Reader*. ed. by Maybin, J. Cleveland, UK: Multilingual Matters Ltd., 23-43

Halliday, M. A. K. (1985) *Spoken and Written Language*. Geelong, Vic.: Deakin University Press [republished London: Oxford University Press 1989]

Handford, M. (2010) *The Language of Business Meetings*. Cambridge: Cambridge University Press

Hay, J. (2000) 'Functions of Humor in the Conversations of Men and Women'. *Journal of Pragmatics* 32, 709-742

HEA (2014) *Teaching International Students Project* [online] available from <http://www.heacademy.ac.uk/assets/documents/internationalisation/Teaching_International_Students_Project_2014.pdf> [10/07 2014]

Healey, C. G. (1996) 'Choosing Effective Colours for Data Visualisation'. *Proceedings of the IEEE Visualization '96*. San Francisco, CA: IEEE

HEBRG (2011) *Professional, Statutory and Regulatory Bodies: An Exploration of their Engagement with Higher Education* [online] available from <http://www.universitiesuk.ac.uk/highereducation/Documents/2011/HEBRG_ProfessionalBodies.pdf> [01/03 2016]

Herman, D. (2009) *Basic Elements of Narrative*. Malden, MA.: Wiley-Blackwell

Hinrichs, E. and Krauwer, S. (2014) 'The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars'. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. held 26-31/05/2014 at Reykjavik. European Language Resources Association

Hobbes, T. (1996[1651]) *Leviathan*. Cambridge: Cambridge University Press

Hoey, M. (1983) *On the Surface of Discourse*. London: George Allen and Unwin

Holmes, P. (2004) 'Negotiating Differences in Learning and Intercultural Communication: Ethnic Chinese Students in a New Zealand University'. *Business Communication Quarterly* 67, 294-307

House of Commons (2009) *Engineering: Turning Ideas into Reality. Fourth Report of Session 2008–09.* London: The Stationary Office (HC-50 I) [online] available from <http://www.publications.parliament.uk/pa/cm200809/cmselect/cmdius/50/50i.pdf> [01/03 2016]

Hunston, S. and Francis, G. (2000) *Pattern Grammar. A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam and Philadelphia: John Benjamins

Hunston, S. and Thompson, G. (eds.) (2000) *Evaluation in Text: Authorial Stance and the Construction of Discourse.* Oxford: Oxford University Press

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press

Hunston, S. (1994) 'Evaluation and Organization in a Sample of Written Academic Discourse'. in *Advances in Written Text Analysis*. ed. by Coulthard, M. London: Routledge, 191-218

ICEF Monitor (2014) *Summing Up International Student Mobility in 2014* [online] available from <http://monitor.icef.com/2014/02/summing-up-international-student-mobility-in-2014/> [01/06 2015]

IEA (2016) *The Washington Accord* [online] available from <http://www.ieagreements.org/washington-accord/> [01/03 2016]

IPENZ (2016) *About Us* [online] available from <https://www.ipenz.nz/> [01/03 2016]

ISO (2011) *International Organization for Standardization* [online] available from <http://www.iso.org> [01/07 2011]

Jenkins, J. (2014) *English as a Lingua Franca in the International University: The Politics of Academic English Language Policy*. London and New York: Routledge

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., and Vangsnes, Ø. A. (2009) 'The Nordic Dialect Corpus - an Advanced Research Tool'. *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*. held 14-16/05/2009 at Odense, Denmark. NEALT Proceedings Series

Jordan, R. R. (1997) *English for Academic Purposes: A Guide and Resource Book for Teachers*. Cambridge: Cambridge University Press

Joyce, M. (2013) *Picking the Best Intercoder Reliability Statistic for Your Digital Activism Content Analysis* [online] available from <http://digital-activism.org/2013/05/picking-the-best-intercoder-reliability-statistic-for-your-digital-activism-content-analysis/> [27/06/14 2014]

Kachru, B. B. (1994) 'Englishization and Contact Linguistics'. *World Englishes* 13 (2), 135-154

Kallmeyer, L., Meyer, R., and Wagner, A. (2009) [online] available from <http://www.sfb441.uni-tuebingen.de/c1/tusnelda-guidelines.html> [01/07 2011]

Kanoksilapatham, B. (2005) 'Rhetorical Structure of Biochemistry Research Articles'. *English for Specific Purposes* 24, 269-292

Kant, I. (2007[1790]) *Critique of Judgment*. Oxford: Oxford University Press

Keats, P. A. (2009) 'Multiple Text Analysis in Narrative Research: Visual, Written, and Spoken Stories of Experience'. *Qualitative Research* 9 (2), 181-195

Kehoe, A. and Gee, M. (2013) 'eMargin: A Collaborative Textual Annotation Tool'. *Ariadne* (71)

Keiser, J., Livingstone, V., and Meldrumi, A. (2008) 'Professional Storytelling in Clinical Dental Anatomy Teaching'. *Anat Sci Ed* 1, 84-89

Kilgarriff, A. (n.d) *Sketch Engine: Glossary of Sketch Engine Terminology (Jargon Buster)* [online] available from <https://trac.sketchengine.co.uk/wiki/SkE/Help/JargonBuster> [01/07 2011]

Kirkpatrick, A. and Sophiaan, S. (2014) 'Non-Standard Or New Standards Or Errors? the use of Inflectional Marking for Present and Past Tenses in English as an Asian Lingua Franca'. in *The Evolution of Englishes*. ed. by Buschfeld, S., Hoffman, T., Huber, M., and Kautsch, A. Amsterdam: John Benjamins, 386-400

Klein, M. (1999) 'Standardisation Efforts on the Level of Dialogue Act in the MATE Project'. *Proceedings of the ACL Workshop 'Towards Standards and Tools for Discourse Tagging'*. held 21/06/1999 at Maryland. New Brunswick, NJ.: ACL

Knott, A. and Dale, R. (1994) 'Using Linguistic Phenomena to Motivate a Set of Rhetorical Relations'. *Discourse Processes* 18 (1), 35-62

Koestler, A. (1989[1964]) *The Act of Creation*. London: Arkana/Penguin

Kosek, M., Nøklestad, A., Priestley, J., Hagen, K., and Johannessen, J. B. (2015) 'Visualisation in Speech Corpora: Maps and Waves in the Glossa System'. *Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*. held 11-12/05/2015 at Vilnius. Linköping University Electronic Press

Kotthoff, H. (2007) 'Oral Genres of Humor. On the Dialectic of Genre Knowledge and Creative Authoring'. *Pragmatics* 12, 263-296

Krippendorff, K. (2004) 'Reliability in Content Analysis: Some Common Misconceptions and Recommendations'. *Human Communication Research* 30 (3), 411-433

Labov, W. (1972) *Language in the Inner City*. Philadelphia: University of Pennsylvania Press

Labov, W. and Waletzky, J. (1967) 'Narrative Analysis: Oral Versions of Personal Experience'. in *Sociolinguistics: The Essential Readings (2003)*. ed. by Bratt Paulston, C. and Tucker, G. R. Oxford: Wiley-Blackwell, 74-104

Lee, D. (2006) 'Humour in Spoken Academic Discourse'. *Journal of Language, Culture and Communication* 8 (3), 49-68

Lee, J. J. (2009) 'Size Matters: An Exploratory Comparison of Small- and Large-Class University Lecture Introductions'. *English for Specific Purposes* 28 (1), 42-57

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

Leech, G. and Weisser, M. (2003) 'Generic Speech Act Annotation for Task-Oriented Dialogues'. *Proceedings of the Corpus Linguistics 2003 Conference*. held 28-31/03/2003 at Lancaster. Lancaster: UCREL Technical Papers

Leech, G. N. (1983) *Principles of Pragmatics*. London: Longman

Levinson, S. C. (1983) *Pragmatics*. Cambridge: Cambridge University Press

Lewis, M. (1986) *English Verb: An Exploration of Structure and Meaning*. London: Language Teaching Publications

Li, D. C. S. (2013) 'Linguistic Hegemony Or Linguistic Capital? Internationalization and English-Medium Instruction at the Chinese University of Hong Kong'. in *English-Medium Instruction at Universities: Global Challenges*. ed. by Doiz, A., Lasagabaster, D., and Sierra, J. M. Great Britain: Short Run Press, 65-83

Lightstone Software LLC (2014) *Preceden* [online] available from <http://www.preceden.com/> [10/01 2014]

Lindemann, S. and Mauranen, A. (2001) '"It's just Real Messy": The Occurrence and Function of just in a Corpus of Academic Speech'. *English for Specific Purposes* 20, Supplement 1, 459-475

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002) 'Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability'. *Human Communication Research* 28 (4), 587-604

Lombard, M., Snyder-Duch, J., and Campanella Bracken, C. (2010) *Intercoder Reliability: Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects* [online] available from <http://matthewlombard.com/reliability/> [27/06/14 2014]

Lyons, J. (1977) *Semantics. Vol. 2*. Cambridge: Cambridge University Press

Mahlberg, M. and Smith, C. (2012) 'Dickens, the Suspended Quotation and the Corpus'. *Language and Literature* 21 (1), 51-56

Mann, W. C. and Thompson, S. (1988) 'Rhetorical Structure Theory: Toward a Functional Theory of Text Organization'. *Text* 8 (3), 243-281

Maringe, F. (2010) 'The Meaning of Globalization and Internationalization in HE: Findings from a World Survey'. in *Globalization and Internationalization in Higher Education: Theoretical, Strategic and Management Perspectives*. ed. by Maringe, F. and Foskett, N. London: Continuum, 17-34

Martin, J. R. (2008) 'Negotiating Values: Narrative and Exposition'. *Bioethical Inquiry* 5, 41-55

Martin, J. R. (2009) 'Genre and Language Learning: A Social Semiotic Perspective'. *Linguistics and Education* 20 (1), 10-21

Martineau, W. H. (1972) 'A Model of Social Functions of Humor'. in *The Psychology of Humor*. ed. by Goldstein, J. H. and McGhee, P. E. New York: Academic Press, 101-125

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

McAllister, P. G. (2015) 'Speech Acts: A Synchronic Perspective'. in *Corpus Pragmatics: A Handbook*. ed. by Aijmer, K. and Rühlemann, C. Cambridge: Cambridge University Press, 29-51

McCarthy, M. (2010) *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press

McDrury, J. and Alterio, M. G. (2002) *Learning through Storytelling: Using Reflection and Experience in Higher Education Contexts*. Palmerston North: The Dunmore Press

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

Meyer, J. C. (2000) 'Humor as a Double-Edged Sword: Four Functions of Humor in Communication'. *Communication Theory* 10 (3), 310-331

Moreland, K. (2013) 'A Survey of Visualization Pipelines'. *IEEE Transactions on Visualizations and Computer Graphics* 19 (3), 367-378

Morreall, J. (1983) *Taking Laughter Seriously*. Albany, NY.: SUNY Press

Morris, C. (1938) 'Foundations of the Theory of Signs'. in *International Encyclopedia of Unified Science*. ed. by Neurath, O. and Carnap, R. Chicago: University of Chicago Press, 77-138

Morrison, J. W. (1974) *An Investigation of Problems in Listening Comprehension Encountered by Overseas Students in the First Year of Postgraduate Studies in Sciences in the University of Newcastle upon Tyne, and the Implications for Teaching*. Unpublished MEd thesis. Newcastle: University of Newcastle upon Tyne

Mouly, G. J. (1978) *Educational Research: The Art and Science of Investigation*. Boston: Allen and Bacon

Nakamura, J. (2002) 'A Galaxy of Words: Structures Based upon Distributions of Verbs, Nouns and Adjectives in the LOB Corpus'. in *English Corpus Linguistics in Japan*. ed. by Saito, O., Nakamura, J., and Yamazaki, S. Amsterdam and New York: Rodopi, 19-42

National University of Singapore (2012) *NUS: International Prospectus for International Students*. Singapore: NUS

Navarro, B., Civit, M., Martí, M. A., Marcos, R., and Fernández, B. (2003) *Syntactic, Semantic and Pragmatic Annotation in Cast3LB* [online] available from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.4101&rep=rep1&type=pdf > [02/09 2013]

Nelson, M. N., Hull, G. A., and Roche-Smith, J. (2008) 'Challenges of Multimedia Self-Presentation: Taking, and Mistaking, the show on the Road'. *Written Communication*, 415-440

Nesi, H. (2012a) 'Laughter in University Lectures'. *Journal of English for Academic Purposes* 11 (2), 79-89

Nesi, H. (2012b) 'ESP and Corpus Studies'. in *The Handbook of English for Specific Purposes*. ed. by Partridge, B. and Starley, S. Oxford: Wiley-Blackwell, 407-426

Nesi, H. and Ahmad, U. (2009) 'Pragmatic Annotation in an International Corpus of Engineering Lectures'. *American Association for Corpus Linguistics Conference*. held 8-11/10/09 at University of Alberta, Canada

Nesi, H. and Moreton, E. (2012) 'EFL/ESL Writers and the use of Shell Nouns'. in *Academic Writing in a Second Or Foreign Language: Issues and Challenges Facing ESL/EFL Academic Writers in Higher Education Contexts*. ed. by Tang, R. London: Continuum, 126-145

Nesi, H., Moreton, E., Rayson, P., Sharoff, S., and Stewart, C. (2012) *JISC Final Report: Increasing Interoperability between Corpus Tools*: JISC

Nesi, H. and Thompson, P. (2006) *The British Academic Spoken English Corpus Manual* [online] available from <www.coventry.ac.uk/base> [01/07 2011]

Neuendorf, K. A. (2002) *The Content Analysis Guidebook*. Thousand Oaks, CA: Sage

Niwa, S. and Maruno, S. (2010) 'Strategic Aspects of Cultural Schema: A Key for Examining how Cultural Values are Practiced in Real-Life Settings'. *Journal of Social, Evolutionary, and Cultural Psychology* 4 (2), 79-91

Norman, D. A. (1993) *Things that make Us Smart: Defending Human Attributes in the Age of the Machine*. Cambridge, MA.: Perseus Books

Norrick, N. R. and Spitz, A. (2008) 'Humor as a Resource for Mitigating Conflict'. *Journal of Pragmatics* 40, 1661-1686

Norrick, N. R. (2001) 'Discourse Markers in Oral Narrative'. *Journal of Pragmatics* 33 (6), 849-878

Norrick, N. R. (1986) 'A Frame-Theoretical Analysis of Verbal Humor: Bisociation as Schema Conflict'. *Semiotica* 3 (4), 225-246

Nunan, D. (2013) *Learner-Centered English Language Education: The Selected Works of David Nunan*. New York: Routledge

OECD (2013) *Education Indicators in Focus: How is International Student Mobility Shaping Up?* [online] available from <http://www.oecd.org/education/skills-beyond-school/EDIF%202013--N%C2%B014%20(eng)-Final.pdf> [02/03 2014]

Olsen, L. A. and Huckin, T. H. (1990) 'Point-Driven Understanding in Engineering Lecture Comprehension'. *English for Specific Purposes* 9 (1), 33-47

Othman, Z. (2010) 'The use of Okay, Right and Yeah in Academic Lectures by Native Speaker Lecturers: Their 'Anticipated' and 'Real' Meanings'. *Discourse Studies* 12, 665-681

Partington, A. (2006) *The Linguistics of Laughter: A Corpus-Assisted Study of Laughter-Talk*. London: Routledge

Peirce, C. (1934) *Collected Papers: Vol. V. Pragmatism and Pragmaticism*. Cambridge, MA.: Harvard University Press

Peters, P. (2004) *The Cambridge Guide to English Usage*. Cambridge: Cambridge University Press

Petrillo, M. and Baycroft, J. (2010) *Fairview Research: Introduction to Manual Annotation* [online] available from <https://gate.ac.uk/teamware/man-ann-intro.pdf> [17/08 2012]

Plato (1973[380BC]) *The Republic*. trans. by Jowett, B. New York: First Anchor Books

Plaza, S. M. and Álvarez, I. A. (2013) 'University Large Lectures in MICASE: A Systemic Functional Analysis'. *Revista Española De Lingüística Aplicada* 1, 183-207

Plo Alastrué, R. and Pérez-Llantada, C. (eds.) (2015) *English as a Scientific and Research Language: Debates and Discourses English in Europe.* Berlin and Boston: Walter de Gruyter

Plum, G. (1988) *Textual and Contextual Conditioning in Spoken English: A Genre-Based Approach*. Unpublished PhD thesis. University of Sydney: Department of Linguistics

QAA (2015) *Subject Benchmark Statement* [online] available from <http://www.qaa.ac.uk/en/Publications/Documents/SBS-engineering-15.pdf> [01/03 2016]

Ramsden, P. (1992) *Learning to Teach in Higher Education*. London: Routledge

Raskin, V. (1985[1944]) *Semantic Mechanisms of Humor*. Dordrecht: D. Reidel Publishing Company

Rayson, P. (n.d.) *Log-Likelihood Calculator* [online] available from <http://ucrel.lancs.ac.uk/llwizard.html> [01/10 2014]

Rayson, P. and Mariani, J. (2009) 'Visualising Corpus Linguistics'. *Proceedings of the Corpus Linguistics Conference (CL2009)*. held at University of Liverpool, UK

Reershemius, G. (2012) 'Research Cultures and the Pragmatic Functions of Humor in Academic Research Presentations: A Corpus Assisted Analysis'. *Journal of Pragmatics* 44, 863-875

Richards, J. C., Platt, J. T., and Platt, H. K. (1992) *The Longman Dictionary of Applied Linguistics and Language Teaching*. Essex: Longman

Römer, U. and O'Donnell, M., B. (2011) 'From Student Hard Drive to Web Corpus (part 1): The Design, Compilation and Genre Classification of the Michigan Corpus of Upper-level Student Papers (MICUSP)'. *Corpora* 6 (2): 159–177

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

Rounds, P. (1985) *Talking the Mathematics through: Disciplinary Transaction and Socio-Educational Interaction*. Unpublished PhD thesis: The University of Michigan

Rowley-Jolivet, E. and Carter-Thomas, S. (2005) 'The Rhetoric of Conference Presentation Introductions: Context, Argument and Interaction'*. International Journal of Applied Linguistics* 15 (1), 45-71

Rühlemann, C. (2010) 'What can a Corpus Tell Us about Pragmatics?'. in *The Routledge Handbook of Corpus Linguistics*. ed. by O'Keeffe, A. and McCarthy, M. London: Routledge, 288-301

Savy, R. (2010) 'Pr.a.T.i.D: A Coding Scheme for Pragmatic Annotation of Dialogues'. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*. held 05/2010 at Malta. European Language Resources Association

Schiffrin, D. (1987) *Discourse Markers*. Cambridge: Cambridge University Press

Schmid, H. (2000) *English Abstract Nouns as Conceptual Shells*. Berlin: Mouton de Gruyter

Scholes, R. and Kellogg, R. (1966) *The Nature of Narrative*. New York: Oxford University Press

Schopenhauer, A. (1966[1959]) *The World as Will and Representation*. *Vol 1*. trans. by Payne, E. F. J. USA: Dover Publications

Schorup, L. C. (1985) *Common Discourse Particles in English Conversation: Like, Well, Y'Know*. New York: Garland

Scott, M. (2014) *WordSmith Tools Help. Lexical Analysis Software* [online] available from <http://www.lexically.net/downloads/version6/HTML/proc_tag_handling.htm> [09/07 2014]

Scott, M. (2012) *WordSmith Tools (Version 6)* [online] available from <http://www.lexically.net/wordsmith/> [01/02 2013]

Scott, M. (1997) 'PC Analysis of Key Words – and Key Key Words'. *System* 25 (1), 1-13

Scott, W. (1955) 'Reliability of Content Analysis: The Case of Nominal Scale Coding'. *Public Opinion Quarterly* 19 (3), 321-325

Searle, J. (1978) 'Literal Meaning'. *Erkenntnis* 13 (1), 207-224

Searle, J. (1976) 'A Classification of Illocutionary Acts'. *Language in Society* 5, 1-23

Searle, J. (1969) *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press

Shamsudin, S. and Ebrahimi, S. J. (2013) 'Analysis of the Moves of Engineering Lecture Introductions'. *Procedia - Social and Behavioral Sciences* 70, 1303-1311

Shneiderman, B. (1996) 'The Eyes have it: A Task by Data Type Taxonomy for Information Visualizations'. *Proceedings of IEEE Visual Languages (also Maryland HCIL TR: 96-13)*, 336-343

Simpson, R., Briggs, S., Ovens, J., and Swales, J. (2002) *The Michigan Corpus of Academic Spoken English*. Ann Arbor: The Regents of the University of Michigan

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

Sinclair, J. (2005) 'Corpus and Text - Basic Principles'. in *Developing Linguistic Corpora: A Guide to Good Practice*. ed. by Wynne, M. Oxford: Oxbow Books, 1-16

Sinclair, J. (2004) *Trust the Text: Language, Corpus and Discourse*. London: Routledge

Sinclair, J. and Coulthard, R. (1975) *Towards an Analysis of Discourse: The English used by Teachers and Pupils*. Oxford: Oxford University Press

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

Sport England (n.d) *Sport England Infographic* [online] available from <http://infographics.sportengland.org/> [16/01 2014]

Stebbins, R. (2012[1980]) 'The Role of Humour in Teaching: Strategy and Self-Expression'. in *Teacher Strategies: Explorations in the Sociology of the School*. ed. by Woods, P. London: Croom Helm, 84-97

Steslow, D. M. and Gardner, C. (2011) 'More than One Way to Tell a Story: Integrating Storytelling into Your Law Course'. *Journal of Legal Studies Education* 28 (2), 249-271

Straker Cook, R. H. (1975) *A Communicative Approach to the Analysis of Extended Monologue Discourse and its Relevance to the Development of Teaching Materials for English for Special Purposes*. Unpublished M. Litt thesis. University of Edinburgh

Stubbs, A. (2011) 'MAE and MAI: Lightweight Annotation and Adjudication Tools'. *Proceedings of the Linguistic Annotation Workshop V*. held 23-24/07/2011 at Portland, Oregon. Association of Computational Linguistics

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

Swales, J. M. (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press

SyncRO Soft SRL (2014) *<oXygen/> XML Editor 15.2* [online] available from <http://www.oxygenxml.com/>

Tableau Software (2014) *Tableau* [online] available from <http://www.tableausoftware.com/> [10/01 2014]

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

Taylor, C. (2003) *An Introduction to Metadata* [online] available from <http://www.library.uq.edu.au/iad/ctmeta4.html> [22/07 2011]

TEI Consortium (2011) *TEI P5: Guidelines for Electronic Text Encoding and Interchange* [online] available from <http://www.tei-c.org> [01/07 2011]

TEIWiki (2010) *Text Encoding Initiative (TEI) Wiki* [online] available from <http://wiki.tei-c.org/index.php/Main_Page> [01/07 2011]

The Language Archive (2015) *ELAN - Linguistic Annotator Version 4.9.1* [online] available from < http://www.mpi.nl/corpus/html/elan/> [02/08 2015]

Thompson, P. (2010) 'Building a Specialised Audio-Visual Corpus'. in *The Routledge Handbook of Corpus Linguistics*. ed. by O'Keefe, A. and McCarthy, M. London: Routledge, 93-103

Thompson, S. E. (1994) 'Frameworks and Contexts: A Genre-Based Approach to Analysing Lecture Introductions'. *English for Specific Purposes* 13, 171-186

Tibco (2014) *Spotfire* [online] available from <http://spotfire.tibco.com/> [10/01 2014]

Timeglider/Mnemograph LLC (2008) *Timeglider* [online] available from <http://timeglider.com/?tab=free> [10/01 2014]

Toolan, M. J. (1988) *Narrative: A Critical Linguistic Introduction*. Cornwall: TJ Press

Treisman, A. (1985) 'Preattentive Processing in Vision'. *Computer Vision, Graphics, and Image Processing* 31, 155-177

Tufte, E. (1997) *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press

Tufte, E. (1990) *Envisioning Information*. Cheshire, CT: Graphics Press

UCAS (2016) *Find an Undergraduate Course* [online] available from <http://search.ucas.com/> [01/03 2016]

UNESCO (2015) *UNESCO Institute for Statistics. Outbound Students* [online] available from <http://data.uis.unesco.org/index.aspx?queryid=163&lang=en#> [01/08 2015]

University College London (2012) *Global Vision: UCL International Strategy 2012-2017*. London: UCL

University of Northampton Student Services (2012) *International Student Handbook* [online] available from <http://www.northampton.ac.uk/Downloads/2551-international_students_handbook.pdf> [01/01 2014]

University of Oxford (2015) *The University of Oxford Text Archive* [online] available from <http://ota.ox.ac.uk/> [01/01 2015]

University of Sheffield International Student Support (2013) *British Humour* [online] available from <http://www.sheffield.ac.uk/ssid/international/news/british-humour-1.292335> [01/01 2014]

University of Western Australia (2014) *Achieving Success: Strategic Goals. Strategic Plan 2014-2020*. Perth: University of Western Australia

van Dijk, T. A. (2012) 'Discourse and Knowledge'. in *The Routledge Handbook of Discourse Analysis*. ed. by Gee, J. P. and Handford, M. Oxon and New York: Routledge, 587-603

van Dijk, T. A. (1977) 'Semantic Macro-Structures and Knowledge Frames in Discourse Comprehension'. in *Cognitive Processes in Comprehension: Proceedings of the Twelfth*

Carnegie-Mellon Symposium on Cognition. ed. by Carpenter, P. A. and Just, M. A. Hillsdale, N.J.: Lawrence Erlbaum, 3-32

van Gompel, M. and Reynaert, M. (2013) 'FoLiA: A Practical XML Format for Linguistic Annotation – a Descriptive and Comparative Study'. *Computational Linguistics in the Netherlands* 3, 63-81

visual.ly (2013) *NBA Visualization* [online] available from <http://visual.ly/nba-draft-top-picks-v-top-performers> [16/01 2014]

W3C (2011) *World Wide Web Consortium (W3C)* [online] available from <http://www.w3.org/> [01/07 2011]

Wächter, B. (2008) 'Teaching in English on the Rise in European Higher Education'. *International Higher Education* 52 (3), 3-4

Wächter, B. and Maiworm, F. (2008) *English-Taught Programmes in European Higher Education: The Picture in 2007*. Bonn: Lemmens

Wallis, S. (2014) 'What might a Corpus of Parsed Spoken Data Tell Us about Language?'. presented at The Olomouc Linguistics Colloquium (Olinco) held 05-07/07/14 at Palacký University, Czech Republic. London: Survey of English Usage

Walsh, P. and Crawford Camiciottoli, B. (2001) 'Lecturing to an Unfamiliar Audience: Some Functions of Interaction in Business Lectures Given by Visiting Academics'. *Quaderni Del Dipartimento Di Linguistica Del'Universitá Degli Studi Di Firenza* 11, 171-183

Wang, T. (2014) 'Humor in British Academic Lectures and Chinese Students' Perceptions of It'. *Journal of Pragmatics* 68, 80-93

Ware, C. (2008) *Visual Thinking for Design*. Burlington, MA.: Morgan Kaufmann

Ware, C. (2004) *Information Visualization: Perception for Design*. San Francisco, CA: Morgan Kaufmann

Weisser, M. (2015) 'Speech Act Annotation'. in *Corpus Pragmatics: A Handbook*. ed. by Aijmer, K. and Rühlemann, C. Cambridge: Cambridge University Press, 84-116

Widlöcher, A. and Mathet, Y. (2012) 'The Glozz Platform: A Corpus Annotation and Mining Tool'. *DocEng '12: Proceedings of the 2012 ACM Symposium on Document Engineering*. held 04-07/09/12 at Paris, France. New York: ACM

Wijasuriya, B. S. (1971) *Occurrence of Discourse Markers and Inter-Sentence Connectives in University Lectures and their Place in the Testing and Teaching of Listening Comprehension in English as a Foreign Language*. Unpublished PhD thesis. Manchester: University of Manchester

Wilkinson, B. (2005) 'Where is English Taking Universities?'*. The Guardian* 18/03/2005

Willis, D. (2003) *Rules, Patterns and Words*. Cambridge: Cambridge University Press

Wittenburg, P., Broeder, D., and Sloman, B. (2000) *Meta-Descriptions for Language Resources - EAGLES/ISLE - A Proposal for a Meta-Description Standard for Language Resources* [online] available from <http://www.mpi.nl/ISLE/documents/papers/white_paper_11.pdf> [01/07 2011]

Wittgenstein, L. (1922) *Tractatus Logico-Philosophicus, with an Introduction by Bertrand Russell.* trans. by Ogden, C. K. Edinburgh: Edinburgh Press

Wynne, M. (2004) 'OLAC the Open Language Archives Community'. *CERN Workshop on Innovations in Scholarly Communication: Implementing the Benefits of OAI (OAI3)*. held 12-14/02/2004 at Geneva, Switzerland

XCES (2008) *XML Corpus Encoding Standard Document XCES 1.0.4.* [online] available from <http://www.xces.org/> [01/07 2011]

Yaakob, S. (2013) *A Genre Analysis and Corpus Based Study of University Lecture Introductions*. Unpublished PhD thesis. University of Birmingham Research Archive: University of Birmingham

Yaakob, S. (2011) 'Disciplinary Differences in Social Science and Physical Science BASE Lecture Introductions'. *Corpus Linguistics Conference*. held 20-22/07/2011 at Birmingham University

Yeo, J. and Ting, S. (2014) 'Personal Pronouns for Student Engagement in Arts and Science Lecture Introductions'*. English for Specific Purposes* 34, 26-37

Yi, J. S., Kang, Y. A., Stask, J. T., and Jacko, J. A. (2007) 'Toward a Deeper Understanding of the Role of Interaction in Information Visualization'*. IEEE Transactions on Visualization and Computer Graphics (InfoVis '07)* 13 (6), 1224-1231

Yimam, S. M., Eckart de Castilho, R., Gurevych, I., and Biemann, C. (2014) 'Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno in: Proceedings of ACL-2014, Demo Session, Baltimore, MD, USA (Pdf) (Bib)'. *Proceedings of the 52nd Annual Meeting of*

the Association for Computational Linguistics (ACL). held 22-27/07/14 at Baltimore, MD., USA. Association of Computational Linguistics

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

Yule, G. (1996) *Pragmatics (Oxford Introduction to Language Study ELT)*. Oxford: Oxford University Press

Yusoff, Y. M. (2010) 'Demographic Differences among International Undergraduate Students at a Malaysian Public University'. *Global Journal of Management and Business Research* 10 (2), 36-41

Zhang, Q. (2005) 'Immediacy, Humor, Power Distance and Classroom Communication Apprehension in Chinese College Classrooms'. *Communication Quarterly* 53 (1), 109-124

Živkovi, B. (2014) 'A Corpus-Based Study of the Structure of University Lecture Introductions'. *Mediterranean Journal of Social Sciences* 5 (13), 107-112

Appendices

Appendix I: Transcription protocols (condensed version)

1. General:

1.1 Uncertain spellings to be looked up in Online Oxford Reference or the OED Online

1.2 Where the spelling does not appear in either of these reference books, use Google for suggestions and use the most consistent occurrence noted.

1.3 Words that are always *–ise* rather than *–ize* are:

advertise	dis(en)franchise	merchandise
advise	disguise	prise (open)
apprise	enfranchise	revise
chastise	enterprise	supervise
circumcise	excise	surmise
comprise	exercise	surprise
demise	franchise	televise
despise	improvise	
devise	incise	

1.4 Other variable spellings and/or words not found in the dictionary:

aagh	in so far - not insofar	straight away
adviser	judgement	swap
analyse	linchpin	technobabble
biggie	medieval	thingy
chock-a	naive	ticklist
combating	nineteen-o-five	ton - *rather than tonne*
e-commerce	oi	trade-off
e-mail	okey-dokey	weblog
encyclopedia	oof	web page
et cetera	per cent	web site
filestore	piccies	whoops
focused	shear stress	woo-hoo
huh	so-called	yeah
hurray - not hooray	shear stress	yep – *not yup*
infrared	so-called	

1.5 Finite list of fillers:

Um [but no em]
Er [short hesitation, i.e. as it sounds]
Erm [longer hesitation, i.e. as it sounds]

1.1 Finite list of South-Asian particles:

ah	lor	mah
yah	wah	meh
lah	leh	dah
orh	hor	

2. Capitalisation

2.1 Book/film titles to be capitalised: e.g. The Joy of Engineering, The Adventures of an Engineer

2.2 Personal names to be capitalised (but see 2.10 with measurements): e.g. Isaac Newton

2.3 Names of methods to be capitalised: e.g. Mohr's Stress Circles

2.4 Names of laws to be capitalised: e.g. Boyle's Law

2.5 Names of Engineering phenomena to be capitalised: e.g. Poisson's Ratio, Young's Modulus

2.6 Names of departments are capitalised: e.g. Department of Engineering (but not Engineering department)
2.7 Directions are only capitalised if they form part of a proper name: e.g. South Yorkshire
2.8 Otherwise NO capitalisations except for: individual letters in initialisms, individual letters in acronyms, spelling a word aloud, web addresses
2.9 No chemical substances capitalised (except chemical symbols, e.g. K, Fe)
2.10 All units of measurements to have lower case letters even if they are the names of scientists, e.g. newtons, daltons, pascals

3. **Hyphenation**

3.1 In general, hyphens are to be used sparingly. Conventions from Oxford Online Reference and OED Online are to be followed and/or follow these conventions:
3.2 Hyphens to be used:
3.2.1 For formulae: e.g. alpha-squared-plus-beta-squared-over-six
3.2.2 For web addresses: e.g. W-W-W-dot-NATO-dot-org
3.2.3 For connecting non-word spoken noise components: e.g. dah-di-dah-di-dah
3.2.4 For initialisms: e.g. C-U online [but acronyms such as ELC, BASE do not have hyphens]
3.2.5 For false starts: e.g. Coven-
3.2.6 For spelling words out: e.g. C-O-V-E-N-T-R-Y
3.2.7 For 'non-something' words: e.g. non-specific. See also:

post-something words	something-shaped words	socio-something words
something-like words	counter-something words	semi-something words
something-related words	anti-something words	pro-something words
something-specific words	quasi-something words	pseudo-something words
mid-something words		

3.2.8 For 'pre-something' words, hyphenated if the word following 'pre' begins with 'i' or 'e', or if the resultant 'word' could be ambiguous: e.g. pre-experimental, pre-position (cf. preposition)
3.2.9 Specific others:

middle-sized	first-hand	one-legged
one-sided	un-British	touchy-feely
south-east	oft-cited	two-hundred-and-something
P-value	okey-dokey	fifty-odd
twelve-pounds-ninety-nine	wire-free	arch-example

3.3 Not hyphenated:

cosomething e.g. coworker	personalitywise
resomething e.g. reread	protosomething e.g. prototype
somethingish e.g. yellowish	missomething e.g. misemphasizing
subsomething e.g. subgroup	overdo
somethingwise e.g.	underuse

4. South-East Asian Particles:

4.1 Common South-East Asian particles are not hyphenated, but are transcribed as a separate word:

ah	lor	mah
yah	wah	meh
lah	leh	dah
orh	hor	

5. Apostrophes:

5.1 Use apostrophes when words are contracted: e.g. we're, they're, I'm
5.2 Use apostrophes with possessives: e.g. Newton's idea was to, Mohr's Stress Circles, Boyle's Law

Appendix II: Example ELC lecture consent form

LECTURE FILMING CONSENT FORM

The collaborative project 'A study of lecturing styles in Malaysia and the UK' will process your personal data, in the form of film recordings of your lectures, for the purposes of linguistic analysis and for teaching and staff development. In addition, the recordings may be stored in a data archive and may subsequently be re-used by researchers at Coventry University or at other institutions for the same purposes.

Clips or complete recordings may be published on the project website, and will be shared with project research partners in New Zealand and Malaysia.

Transcripts of the lectures will also be produced by the project, but these will not contain personal identifiers unless individual lecturers choose to be identified.

Ideally we would like to make as wide a use as possible of the recordings; however, participants will be offered the opportunity to opt out of certain uses of the personal data. You are therefore asked to indicate your preferences as requested below, by deleting as applicable.

I consent to the use of my personal data for all the purposes outlined above, including publication on the project website and possible re-use by researchers at other institutions

I consent to the use of my personal data as outlined above with the exception of publication on the project website

I consent to the use of my personal data as outlined above but I only wish it to be used within Coventry University

SIGNED PRINT NAME DATE

Appendix III: Ethical approval

Certificate of Ethical Approval

Applicant:

Siân Alsop

Project Title:

An anatomy of the academic lecture, with special reference to engineering

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Medium Risk

Date of approval:

27 September 2011

Project Reference Number:

P1179

Appendix IV: Example of ELC header metadata (1001)

```
<?xml version="1.0" encoding="UTF-8"?> <!DOCTYPE TEI.2 SYSTEM "ELC.dtd">
<TEI.2>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Structures</title>
      </titleStmt>
      <publicationStmt>
        <distributor>ELC: Coventry University</distributor>
        <idno>1001</idno>
        <availability>
          <p>The Engineering Lecture Corpus (ELC) was developed at Coventry University under the directorship of Hilary Nesi with contributions from ELC partner institutions. The original recordings are held at Coventry University and at the relevant contributing universities where they may form part of other corpora and be distributed independently. The ELC is freely available to researchers who agree to the following conditions:</p>
          <p>1. The recordings and transcriptions should not be modified in any way</p>
          <p>2. The recordings and transcriptions should be used for research purposes only; they should not be reproduced in teaching materials except by ELC partner institutions</p>
          <p>3. The recordings and transcriptions should not be reproduced in full for a wider audience/readership, although researchers are free to quote short passages of text (up to 200 running words from any given speech event)</p>
          <p>4. The corpus developers should be informed of all presentations or publications arising from analysis of the corpus</p><p> Researchers should acknowledge their use of the corpus using the following form of words: The recordings and transcriptions used in this study come from the Engineering Lecture Corpus (ELC), which was developed at Coventry University under the directorship of Hilary Nesi with contributions from ELC partner institutions. The UK component of the ELC was developed under the directorship of Hilary Nesi. Corpus development was assisted by funding from the British Council.</p>
        </availability>
      </publicationStmt>
    </fileDesc>
    <sourceDesc>
      <recordingStmt>
        <recording dur="01:28:00">01:28:00</recording>
        <date value="20/10/2008">20-Oct-08</date>
        <equipment>
          <p>Sony HVR-A1E Video Camera</p>
          <p>Standard DV Tape</p>
          <p>Sennheiser ew100 g2 Microphone</p>
        </equipment>
        <resp>
          <persName>Dean Butlin</persName>
        </resp>
      </recordingStmt>
      <respStmt>
        <resp>original transcription</resp>
        <persName>Jan Rhodes</persName>
      </respStmt>
      <resp>transcription correction, markup and annotation</resp>
      <persName>Siân Alsop</persName>
      </respStmt>
```

```
    </sourceDesc>
    <encodingDesc>
      <p>Unpunctuated transcription of speech</p>
    </encodingDesc>
    <profileDesc>
      <langUsage>
        <language id="en">English</language>
      </langUsage>
      <particDesc>
        <person id="cm1001" role="camera person" n="n" sex="m" TEIform="person">
          <p>"cm1001, camera person, male"</p>
        </person>
        <person id="nm1001" role="main speaker" n="n" sex="m" TEIform="person">
          <p>"nm2001, main speaker, non-student, male"</p>
        </person>
        <person id="sm" role="participant" n="s" sex="m" TEIform="person">
          <p TEIform="p">sm, participant, student, any male student</p>
        </person>
        <person id="sf" role="participant" n="s" sex="f" TEIform="person">
          <p TEIform="p">sf, participant, student, any female student</p>
        </person>
        <personGrp id="ss" role="audience" sex="x" size="50-100" TEIform="personGrp">
          <p TEIform="p">ss, audience, group</p>
        </personGrp>
      </particDesc>
      <textClass>
        <keywords>
          <list>
            <item n="speechevent">Lecture</item>
            <item n="acaddept">Undergraduate</item>
            <item n="partlevel">Civil Engineering</item>
            <item n="module">102BE and H54BE Structures</item>
          </list>
        </keywords>
      </textClass>
    </profileDesc>
  </teiHeader>
  <text>
    <body><!-- marked up and annotated transcript here--></body>
  </text>
</TEI.2>
```

Appendix V: Elements and attributes within ELC and MICASE pragmatic taxonomies

ELC pragmatic tagset + *attributes*	**MICASE: final pragmatic tagset + *further definition*** (Maynard and Leicher 2007)	**MICASE: full pragmatic inventory + *definitions*** (Simpson-Vlach and Leicher 2006)
	Advice (giving, soliciting): *includes suggestions and recommendations only*	Advice/Direction, Giving, or Soliciting: *includes advice about learning, studying, test taking, interpreting course material, general academic advising, as well as instructions or directions*
Housekeeping	Assigning Homework	Assigning Homework: *any mention of assignments to be turned in at a later time; only applies to classroom events*
		Logistics / Announcements: *other than upcoming topics, homework, or test*
		Returning or Going Over Homework or an Exam: *talk about homework or an exam after it has been graded and returned*
		Reviewing for an Exam: *preparation for a test or exam before it is taken*
Explaining: *defining*, *equating*, *translating*	Defining / glossing terms	Definitions: *short phrases or more lengthy explanations used to define or gloss terms or concepts*
	Directives: *tell someone to do something (sometimes politely) and cannot be declined if the addressee wants to maintain face*	
		Disagreement: e*xchanges in which one speaker contradicts or refutes another's position or statement. These are primarily interactional to differentiate them from one-sided or unanimous criticism of other work or people (which would be coded as negative evaluation)*
		Discussion: *open exchange of ideas or opinions when all present are allowed to speak, and several speakers do so, but not strictly in a question/answer format. The focus is on one or more content-related topics; conversational chatting is not included*
		Dramatization: *any utterance or stretch of discourse in which a speaker animates another voice, assumes another role, or uses a theatrical or flamboyant style*
	Evaluation (positive, negative): *restricted to unexpected or unusual adjectives, and phrases which are metaphorical, uncommon, or otherwise of interest pedagogically*	Evaluation (positive and negative): *includes praise or criticism of self, others, or any subject matter; generally an expression of opinion, rather than a factual report*
		Examples: *includes general examples, personal examples, and examples from readings or other outside sources*
		Group/Pair Work: a*pplies only to classroom events when students are asked to carry out an activity or discussion in pairs or small groups. In such instances, the transcript usually captures the conversation of one of these groups; although in at least one case, two separate groups were recorded and transcribed*
Humour: *bawdy; black; disparagement; irony/sarcasm; joke; playful/mock-threat; self-denigration; teasing; wordplay*		Humor: *although humor is often noted by XML laugh tags in the transcripts (and can be used as a search term), laugh tags are not always indicative of humor, and attempts at humor do not always elicit laugh tags*
Summary: *preview content of current lecture; preview content of future lecture;*	Introductory Road Map	Introductory Road Map: *at least two or more statements or phrases outlining or announcing the topics or course of the class or events*

review content of current lecture; review content of future lecture		
		Large Group Activity: *any exercise or activity involving the entire class; only applies to classroom events*
Story: *narrative; anecdote; exemplum; recount*	Narrative	Narratives: *a story of two or more sequential clauses using the past tense or the historical present*
	Questions	Problem Solving: o*n board or in groups, usually in science/math courses, study groups, office hours, etc.*
		Questions: t*he default for this rating refers to literal questions asked by the instructor; rhetorical questions or questions by students are both noted separately*
		Referring to Handouts: *mention of a handout or other written material or paper (e.g., a student's work)*
	Requests: *generally require some kind of action to be performed*	Requests: *any request for help, a particular action or favour, an appointment, etc.; usually phrased as a question and distinct from directives in that they can be refused*
	Speaker introductions	Speaker Introductions: *includes those in classroom presentations. Does not include introduction of MICASE researchers or fieldworkers unless those segments are included in the transcription*
	Tangents	Tangents, Personal Topics: *a digression from subject matter, unrelated to the lecture*

Appendix VI: Summary of findings

A summary of the purpose, content, distribution and lexicogrammatical features of ELC elements and attributes.

	purpose, content and distribution	lexicogrammatical features
summary	• most common element (token count and instances) • recurs throughout lectures • some chaining of attributes: reviews of previous content and previews of current content early in lectures, also reviews of current content and previews of future content towards the end of lectures	• lowest STTR of all elements • high formulaicity: widest range of 4-gram types of all elements, clustering of a small number of types in the high frequency range • *we* is the most common pronoun (pmw) and highly salient compared to non-summary • temporal deixis (e.g. *week*, *today*, *next*, *later*, *tomorrow*, and *now*) • evaluative language (e.g. *key* and *important*) • shell nouns summarise complex concepts or processes • enumeration (exact and inexact)
review previous lecture content	• longest average token count (per instance) of all attributes • some clustering towards the start of lectures • scope can be brief or expanded	• *you* is the most common pronoun (pmw) • *we* is salient and predominantly inclusive • strongly characterised by the pattern: temporal deixis + pronoun + simple past tense verb + topic reference (e.g. *last week we looked at shear design*) • discourse functions are frequently named (e.g. *recap* and *summary*) • *remember* is salient • some negative evaluation (e.g. *a bit more difficult*)
review current lecture content	• repackages recently given information to aid absorption • scope tends to be brief	• *you* is the most common pronoun (pmw) • non-specific chronological references to information delivery, which are usually fronted (e.g. *I said earlier that …*) • distance from original delivery to review tends to be short • *mentioned* is salient (e.g. *I just mentioned*) • minimising language (e.g. *a little bit* and *just*) • logical connectors, often following the pattern: conjunction (acting as causal logical connector) + anaphoric demonstrative reference to topic (or vice-versa) (e.g. *so that is how we measure things on site*)
preview current lecture content	• shortest average token count (per instance) • most common attribute (instances and token count) • scaffolds learning in the current lecture • can be characterised as *negative*, where the scope of upcoming information is explicitly limited (including outlines of material that has not been/will not be covered)	• 4-grams are common (pmw), especially *we're going to* and *I'm going to* • *we* is the most common pronoun (pmw) • simple pseudo-clefts are used to front important information (e.g. *what we are going to move on is look at a more powerful technique called the method of sections*) • down-toning strategies are employed to minimise imposition (e.g. *just* and *a little bit*) • pronoun + modal auxiliary (e.g. *I will do a demonstration*) is more common than pronoun + semi-modal (e.g. *I'm going to do a demonstration*) • distance to the delivery of previewed content is commonly identified (e.g. *move onto now* and *last five or ten minutes*)
preview future lecture content	• least common attribute (token count and instances) • higher-level information structuring function • some clustering towards the end of lectures • relatively minimal detail is given about upcoming content	• *we* is the most common pronoun (pmw) • the pattern pronoun + semi-modal is more common than pronoun + modal auxiliary • common pattern: temporal deixis + pronoun + auxiliary verb + lexical verb (e.g. *in your third year you're going to start dealing with fatigue*)

	purpose, content and distribution	lexicogrammatical features
humour	• least common element (token count) • most common in UK lectures • recurs throughout lectures • 50% laughter response rate: humour and laughter are not co-extensive	• highest lexical variety (STTR) of all elements • overall *you* is the most common pronoun (pmw) • an average range of 4-gram types that occur with average frequency • *you* is positively key and used with specific reference to the audience • non-discipline-specific lexis is key (e.g. references to people, places, and food)
bawdy	• juxtaposes the high and the low • not present in MS lectures	• non-discipline-specific lexis is key (e.g. *naked* and *orifice*)
black	• satirises topics related to death, injury and pain • common in lectures relating to health and safety	
disparaging	• three times as likely to occur in the UK subcorpus • highly targeted: aimed at in-group members and external *others* • generally used to establish the lecturer's position in the field and encourage in-group cohesion	• *he* and *him* are positively key • *we* is negatively key
irony/ sarcasm	• significantly most common to the UK subcorpus • mainly functions as a means of behaviour modification or self-aggrandisement • rarely explicitly signalled and so easy to misinterpret or miss • lowest laughter response rate of all types	• comparative absence of pronouns – largely operates through evaluating events not people • positive tropes are key (e.g. *interesting*, *hurray*, and *fun*) but function as negative evaluations
joke	• occurs only in one lecture • longest average token count per instance • highly structured: contains a set-up and punchline • functions as a time-out from the lecture	
playful	• most common attribute (token count and instances) • largely integrated and uncritical with no punchline or resolution • builds the positive face of the group • tends to draw on lecturer's own experience • highest laughter response rate	• pronouns *i* and *he* are particularly salient
self-deprecating	• targeted at the self • common in NZ lectures • hair is a recurring theme	• *I* is positively key • *we* is negatively key
teasing/ mock-threat	• targeted and often critical • mockery of an audience member or the audience • the mock-threat of physical punishment in response to academic or behavioural failings is most common in the UK lectures • group threats are common • an alternative means of reprimanding students for the purpose of behaviour modification	• *I* is the most salient pronoun • *you* is more common (pmw) than in other attributes • *we* is three times less likely to occur than in general humour • common pattern: conditional action -> consequence (e.g. *if you write something down but get the wrong answer* [you buy me a] *family pack Mars bar*)
wordplay	• most frequent in MS lectures • establishes the verbal skill of the lecturer • heavy reliance on incongruity	

	purpose, content and distribution	lexicogrammatical features
story	• least common element (instances) • recurs throughout lectures • focus on people and places • based on both personal experience and the experience of others • little expansion of technical concepts • little delivery of new information • all attributes may be partially scripted	• second highest STTR of all elements • an average range of 4-gram types that occur with average frequency • past tense verb forms are key • absence of technical terminology • *you* is the most common pronoun (pmw) • least use of formulaic sequences of all elements • hesitation markers are salient
anecdote	• least common attribute (tokens and instances) • an unresolved complication and a moral judgment reaction • showcases verbal skill, e.g. use of analogy	• *we* is the most common pronoun (pmw)
exemplum	• longest average token count per instance • an unresolved complication and a scientific judgment reaction • serious themes (e.g. death and mutilation) • contains markedly negative consequences • references to field-specific responsibility	• *we* is the most common pronoun (pmw)
narrative	• most common attribute (token count) • includes a complication that is resolved • often includes a coda • relatively high reference to personal experience, especially in relation to industry	• only attribute (not including *story-like*) where *I* is more common than *you* (pmw)
recount	• most common attribute (instances) • most common in MS lectures (tokens and instances) • shortest average token count per instance • does not include a complication • emphasis on explaining concepts	• *we* is the most common pronoun (pmw) • keywords about people, places and structures
story-like	• based on hypotheses or predictions of future events • tend to be unresolved and focus on scientific judgment • often involves an analogy	• *you* is by far the most common pronoun (pmw) of all attributes • *you* is almost always generic (not specific) and conditional

non-annoated	• constitutes the majority of lectures (84%) • interspersed with humour, story and summary	• lower STTR than all elements • roughly equal number of total 4-gram types to humour and story, but nine times less than summary • wide range of 4-gram types that occur infrequently • more boundary markers than in elements (e.g. *so*, *yeah*, *ok*, *right*) • absence of highly key pronouns • numerical references and technical terminology are more salient than in pragmatic text (indicating formulae and workings out)