Interaction and Engagement in Problem-Based Learning Sessions: A Corpus-Based Analysis

Carole MacDiarmid

Submitted in fulfilment of the requirements for the

Doctor of Philosophy in English Language

School of Critical Studies

College of Arts

University of Glasgow

July, 2017

Abstract

This research is motivated by the need for a better understanding of the nature of student-centred interactions in university settings. Although there is now a considerable amount of research into written academic English, studies of spoken academic English, particularly of student-centred, disciplinary-specific events, are still relatively few in comparison. This work aims to go some way towards redressing the balance.

The study provides a description of a variety of linguistic features of one type of speech event, problem-based learning sessions (PBLs), within the context of a postgraduate programme in Medical Genetics. PBLs are underpinned by a very clear pedagogy driving their incorporation into academic programmes: through a cycle of tutorials, individual research and presentations, students develop content knowledge and the skills thought essential for the professional practitioner. Although common within the field of medicine, there has been relatively little research into how the discipline and pedagogy are realised linguistically.

This study analyses a specially compiled corpus of five complete PBL cycles, each with two stages. It comprises over 12 hours of speech, approximately 115,000 words and is searchable as a whole and for each stage. By applying a variety of approaches, including Conversation Analysis (CA), Corpus Linguistics, and aspects of Discourse Analysis, this allows for a more detailed and fine-grained analysis of student discourse than one approach alone.

Applying CA, the study identifies features of the overall organisational structure and the different patterns of talk found in each stage. Academic functions common to the stage two presentations are also identified. The corpus-based analysis investigates three specific linguistic areas: keyword analysis is used to explore vocabulary as a marker of the discipline and approach, personal pronouns as markers of engagement, and the structural and discourse functions of lexical bundles. The investigation into how the interactions unfold and the consideration of keywords reflect the discipline and underlying epistemology of PBL sessions. Clear differences in the frequency and use of personal pronouns

and lexical bundles are evident in each stage, indicating that both the mode (spoken) and the nature of each speech event (highly interactive exchanges or presentations) affect linguistic choices.

This study of a bespoke corpus provides an in-depth analysis of a disciplinary-specific, student-centred speaking event. This may be useful for EAP teachers and task and materials designers working with students on pre-sessional programmes who need academic language support. Methodologically it adds to the growing number of studies taking a multidimensional approach (i.e. in methodology and focus) to understanding spoken academic discourse.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AC	Academic Core
AT	Academic Technical
AVL	Academic Vocabulary List
AWL	Academic Word List
CA	Conversation Analysis
CL	Corpus Linguistics
CLIL	Content and Language Integrated Learning
DO	Discourse Organisers
DOFH	Discourse Organisers Floor-Holding
EAP	English for Academic Purposes
FPP	First Pair Part
GSL	General Service List
HE	Higher Education
HF	High Frequency
IRE	Initiation, Response, Evaluation
IRF	Initiation, Response, Feedback
IR(I)	Initiation, Response, (Initiation)
LF	Low Frequency
LO	Learning Objective
NP	Noun Phrase
NS	Native Speakers
NNS	Non-Native Speakers
PBL	Problem-Based Learning
PBLs	Problem-Based Learning Sessions
PoS	Part of Speech
RE	Referential Expressions
SAEp/i	Stance A Epistemic (personal/impersonal)
SBA	Stance B Attitudinal
SFL	Systemic Functional Linguistics
SL	Semi-Lexical
SPP	Second Pair Part
Corpora:	
BASE	The British Academic Spoken English Corpus

BNC	British National Corpus
CANBEC	Cambridge and Nottingham Business English Corpus
CANCODE	Cambridge and Nottingham Corpus of Discourse in English
CEC	Cambridge English Corpus
COCA	Corpus of Contemporary American English
COCA-A	Corpus of Contemporary American English-Academic
MICASE	Michigan Corpus of Academic Spoken English
T2K-SWAL	The TOEFL 2000 Spoken and Written Academic Language Corpus

Acknowledgements

First and foremost, my thanks go to my supervisors, Dr. Wendy Anderson and Dr. Catherine Emmott. I have benefitted enormously from their guiding questions, forensic attention to detail and generosity in time. They have been outstanding. Thanks also to Professor John Corbett for his encouragement and supervision in the early stages.

There are many other people to thank in the University of Glasgow. This work would not have been possible without the help of the staff and students on the MSc in Medical Genetics. Particular thanks go to Dr. Maria Jackson and Dr. Leah Marks and, of course, to the students and tutors who allowed me to sit in and record their classes.

The School of Modern Languages and Cultures kindly funded this research. Thanks also to colleagues and friends in English for Academic Study. Tim Rowland gave invaluable assistance with recordings, Kerry and Andy Sharkey helped with a lengthy transcription and Susan Finlay with early editing. Others provided encouragement and kept me going including Anneli Williams and Dr. Esther Daborn, and Anía Rolinska and Brían Doonan by joining me in intensive writing sessions. Thanks also to fellow PhD students for their moral support and companionship along the way, particularly Dr. Richard Hill Davis, Sally Zacharias for conversations on classroom talk, and Steve Kirk and Grace Poulter.

Thanks to my family and friends for being there throughout and last but not least, thank you to my husband Sertac for technical support, lunches and dinners and taking me for walks. Thank you for getting me to the end.

Chapter 1 Introduction to Spoken Academic English and Problem-Based Learning

1.1 Introduction

Oral interactions are recognised as playing an important role in academic study, providing crucial opportunities to develop academic skills, knowledge and identity. Yet perhaps because speaking does not routinely form a major part in the assessment of many taught degree programmes in many parts of the world and undoubtedly because of the complexities of compiling spoken corpora, research in this area is still relatively limited, certainly in comparison to written academic discourse. An informed description of the types of interactions students will engage in and how these are linguistically realised is therefore required in order to support academic language and literacies development.

The number of non-native speaker students (NNS) studying degrees in English in both majority English-speaking and in non-English speaking contexts is increasing. English for academic purposes (EAP) pre-sessional courses, on which I teach, prepare and develop students with academic skills, language and literacies to enable them to successfully participate in higher education (HE) settings. This has led to research into academic discourse which informs teaching, course design and assessment. There is now a considerable amount of research into written academic English and a wide range of teaching materials and resources to support EAP teaching in this area. However, studies of spoken academic English as a distinct set of registers, of disciplinarily-specific spoken genres and of particular student-centred learning sessions are still relatively few in comparison (reflected also in the limited range of EAP teaching materials and guidance available). One such type of speech event is problem-based learning (PBL), a student-centred learning approach common in medicine and spreading to other disciplines. The research aims to provide a description of a variety of linguistic features of PBL sessions (PBLs) within the context of a postgraduate programme in medical genetics at a UK university. PBL is well-established and widely used in the School of Medicine with native speaker (NS) and NNS students in the context of this study. As an approach to learning, it also forms one of the components of our disciplinary-specific summer pre-sessional course for students

going on to medicine and life-science programmes. The specific discipline of medical genetics was selected because not only is the PBL approach applied in the programme but also because each year we teach students going on this specific postgraduate programme who need to develop their academic English. In doing so, it can help with a practical issue: how can we best prepare international students for oral interaction in English-medium HE institutions?

The rest of this chapter is organised as follows. An overview of research into spoken academic English summarises key findings in relation to this area and particularly to small group teaching where relevant. I then introduce problem-based learning as a pedagogy before finally considering the approaches and methodologies that will be employed in this study.

1.2 The Importance of Speaking in University Settings

Much of the research into spoken academic English has focussed on lecturer discourse (e.g., Fortanet, 2004; Rounds, 1987a, 1987b; Thompson, 1994); there are some studies of other more expert genres such as conference presentations and talks (Du Bois, 2007; Ventola, 2002; Wulff et al., 2009) and of features of tutor-student interactions (e.g., Limberg, 2010). Lectures and tutor instruction are an important means by which academics impart knowledge and model academic expectations and practices, for example in the ways of critiquing and thinking. The process of learning occurs, however, not only through lectures and associated reading, but also through interacting in seminars, tutorials and other student-centred academic speech events. These spoken interactions are an integral part of learning in the academy. Dialogic interactions are opportunities for the exploration and co-construction of knowledge, for the development of academic and disciplinary identity (Benwell & Stokoe, 2002; Hyland, 2000) and where students are socialised into the academy and discourse communities of their discipline, particularly at higher levels of study (Duff, 2010; Mauranen, 2006).

Interactions can take a variety of forms with varying degrees of interactivity and levels of contributions, from the more interactive lecture to a variety of types of classroom interactions (see Aguilar, 2016, and Basturkmen, 2016, for a

summary). The latter range from tutorials, discussions or seminars, study-groups, presentations (which may be part of seminars) to specific types of small group teaching, such as lab groups and problem-based learning sessions (PBLs). In a recent review of "dialogic interactions", Basturkmen (2016, p. 154) comments on the "disparate nature of events" with varying formats, of the perceived importance by students (more so at graduate level) and how "participation in dialogic speaking in class settings is understood to be an important means by which students can develop their understanding of and ability to articulate disciplinary matter." (p. 158). It is these occluded student-centred learning events that are the less well-researched area of spoken academic English. This is even more noticeable when we consider disciplinary-specific speaking.

1.2.1 Studies of Spoken Academic Discourse

The majority of research into academic discourse to date has concentrated on a range of written academic genres and registers, from expert to novice and in different disciplines (e.g., Hyland, 2000, 2005b, 2009; Nesi & Gardner, 2012). These studies have identified the distinctiveness *of* academic registers and the variations *between* various academic registers, genres and disciplines. The research includes the identification of a range of specific linguistic items, features related to rhetorical organisation and to disciplinary variation. This provides useful reference points and frameworks for research and more recently studies have also compared written and spoken academic outputs, most notably the work by Biber (2006a, 2006b) and Biber et al. (2004).

The works mentioned above, along with the development of large spoken corpora and a relatively small number of studies into patterns of interaction in distinct types of speech events, have helped to identify general characteristics of spoken academic English. This section will provide an overview of the current understanding of this area, focussing when possible and relevant on student-centred speech events.

The studies that will be considered here can broadly be grouped into two areas. The first group are those exploring patterns of interaction and organisation

within specific types of academic speech event. These included studies applying the exchange structure model developed by Sinclair and Coulthard (1975) (e.g., Basturkmen, 1999, 2002; Tapper, 1996), Conversation Analysis (CA), (e.g., Walsh et al., 2011) and Genre Analysis and Systemic Functional Linguistics (SFL) (Legg, 2007; Tanguay, 2015; Woodward-Kron & Remedios, 2007). The second group aim to identify and analyse specific linguistic features of spoken academic registers or genres. Many (but not all) take Corpus Linguistics (CL) as the starting point and identify specific items either from a large corpus of a variety of spoken academic registers, or of a specific register or genre (e.g., lectures or seminar discussions).

Although grouping in this manner is a convenient way to present this review, there are obviously overlaps. Studies of interaction patterns may also identify specific linguistic items and vice versa. For example, studies based on SFL may identify patterns of organisation and the specific linguistic features associated with these. Walsh et al. (2011) is an example where CA and CL are combined, with CA used primarily for describing sequential organisation and CL providing data to inform the analysis of the specific types or sections of talk which were identified in their study.

1.2.2 Academic Speech Events and Patterns of Interaction

One strand of research has investigated how learning unfolds by considering patterns of organisation and interaction in small group teaching events (e.g., seminars and tutorials).

Weissberg (1993) investigated the graduate seminar (students presenting their research). This qualitative study involved observing, describing and analysing a number of seminars of NNS and native speaker (NS) graduate students in animal sciences and agronomy and interviewing instructors and a number of the NNS participants. This research included identifying the stages within the seminars along with styles of talk. These ranged from very planned, scripted talks to those consisting more of “fresh-talk”, where “the text is formulated by the animator from moment to moment” (Goffman, 1981, p. 172). In this work, Weissberg identified a number of aspects of the talks including “topic shifters”

and “advance organisers” (p. 28). His discussion includes an interesting account of how students make use of slides to structure their talks. Although this is an earlier study and is based on extracts rather than complete recordings, this work highlights the importance of “first hand investigation” (p. 34) rather than relying on idealised forms of spoken academic English when preparing students for academic study.

Tapper (1996) considered exchange patterns in four university teaching contexts (lectures, labs, writing classes and writing conferences), basing her research on Sinclair and Coulthard’s exchange structure framework (1975), with a specific focus on international students. To explain briefly, Sinclair and Coulthard use Halliday’s notion of rank scale as a basis for designing a model for categorizing elements of turns and exchanges that occur within a lesson. The smallest unit is the act, followed by a move, an exchange and then a transaction. Transactions combine to make the overall lesson. Sinclair and Coulthard identified two-part exchanges of initiation-response (IR) and the three-part initiation–response-feedback (IRF) pattern, common in teacher-fronted classrooms. Tapper contrasts this with two-part adjacency pairs identified through CA, an approach which explores the sequential organisation and development of turns-in-talk as it unfolds. She found overall that two-part exchanges were most common in her data, but also that these were initiated by both students and teachers in all settings. Three-part exchanges were found to be more common in teacher-dominated classes but also that on occasion students initiated these. She also found that the third part was not always evaluating responses but might be acknowledging information. This study gives some indication of the format of exchanges. However, while this work covers a number of speech events, it concentrated on NNS and only analysed data from eight students.

Basturkmen’s (1999, 2002) research on postgraduate MBA seminars included the study of sequential organisation, also employing Sinclair and Coulthard’s (1975) exchange system. Basturkmen (2002) noted two main patterns of organisation, simple exchanges and more complex exchanges. She illustrated how the complex exchange patterns allow ideas to emerge, how interlocutors scaffold the talk and how knowledge is co-constructed through interaction. This research focussed only on seminars within Business Studies.

Combining Conversation Analysis and Corpus Linguistics in a "CA-CL" approach, O'Keeffe and Walsh (2012), Walsh et al. (2011), and Walsh and O'Keeffe (2010) report on studies of small group teaching in HE. Investigating how students and tutors manage and negotiate the learning events, they identified patterns of organisation relating to the specific pedagogic aims of the relevant section of talk. This included "instructional" and "argumentative" talk, with specific keywords and multi-word units used, for example, to direct attention.

In relation to seminars, Tanguay (2015) studied the rhetorical and functional organisation of phases and moves in student seminars from the British Academic Spoken English (BASE) corpus. Using a SFL framework, she identified overall phases, or discourse macro genres such as responding, and related moves e.g., "description", "evaluation" and "interpretation". This study compared stretches of talk in different seminars from different disciplines. However, each seminar is necessarily viewed as an 'isolated' event.

Although obviously dominated by a focus on lecturer talk, a number of studies have also attempted to identify overall patterns of organisation of lectures. Thompson (1994, p. 176) identified specific steps, for example "setting up lecture framework", while Young (1994, p. 166) identified macro-phrases in lectures, including "discourse structuring" and "conclusion". Deroey and Taverniers (2011) provide more detail by identifying a range of academic functions in lectures from the BASE corpus, including "reporting" and "describing". Their study provides more information about what students do, rather than general stages. These studies provide an indication of how longer stretches of talk may be organised.

The studies above, particularly those of student dialogic events, suggest that sequential organisation is far from simple. Whatever the framework employed, interaction patterns are found to be complex endeavours, with learners and tutors co-constructing knowledge as the learning events unfold. The studies of lectures and seminars also suggest there may be more or less defined phases or stages that a 'typical' speech event may progress through, or which can at least be distinguished by certain academic functions. These functions can be aligned to the pedagogic aims of the speech event (e.g., Deroey & Taverniers, 2011;

Walsh et al., 2011). However, much of this work has been concerned with how learning is facilitated by teaching staff, aiming to discover, or describe, how best to enhance learning opportunities.

1.2.3 Corpus-Based Studies of Spoken Academic English

Another strand of research has focussed on specific language items. The development of spoken academic corpora has facilitated an increasing number of studies of spoken academic English, employing the tools of CL to identify a range of features that distinguish academic registers. Different software tools can be used to identify the frequency of words, keywords (significant in relation to another reference corpus), lexical bundles (sequences of fixed strings of words) and other grammatical items, the latter depending on how or whether the corpus is tagged. In addition to this quantitative data, qualitative analysis of the items provides complementary information on functions and use. These corpus-based studies have provided insights into a number of characteristics of spoken academic English.

While there are some studies of smaller corpora of spoken academic English, (e.g., Farr, 2003), three major spoken academic English corpora have provided the foundation for studies of specific features of spoken academic discourse. Both the Michigan Corpus of Spoken Academic English (MICASE) and the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) include a range of registers and speech events, from class sessions (seminars and lectures) to office hours and study groups, while the BASE corpus comprises seminars and lectures across four broad disciplinary groupings. All three include speech events from across a range of levels, from undergraduate to postgraduate. Research based on these corpora include comparisons of spoken and written registers (e.g., textbooks, classroom teaching and study-groups, Biber & Barbieri, 2007), or explorations of individual types of speech events (e.g. seminars, Tanguay, 2015; lectures, Fortanet, 2004). These have provided us with information on the linguistic characteristics of spoken academic English. Prominent among these are markers of involvement, stance and discourse organisation.

Biber et al. (2002) explored the T2K-SWAL corpus employing multidimensional analysis, a quantitative corpus-based technique which identifies the frequency and distribution of linguistic features in texts and across texts (and by statistical analysis identifies patterns common to specific registers). This research has highlighted factors which differentiate speech from writing and different types of academic registers from each other. Research has indicated the involved nature of classroom teaching compared to the more informational focus in writing, characterised by its high use of nouns (Biber et al., 2002). They note that "in contrast, the spoken registers - again regardless of purpose - are characterized by features of involvement and interaction, situated reference, more overt persuasion and fewer features of impersonal style" (p. 41).

One such marker of involvement, orientation to the audience and engagement is the use of first and second person pronouns (Ädel, 2010; Biber, 2006a, 2006b; Biber et al., 1999; Carter & McCarthy, 2006; Fortanet, 2004; Hyland, 2005b, Rounds 1987a, 1987b). Although found in many forms of spoken academic discourse, differences between lecturer and student use of personal pronoun use has also been identified (O'Boyle, 2014).

Stance markers are another marker of engagement, indicating attitudes and assessments towards propositional content. Biber (2006a) found stance markers to be more overtly marked in spoken registers, although this study did not look specifically at study group interactions.

Although speaking has been found to be less 'informational' than writing, it is seen as somewhere on a cline between conversation and writing (e.g., Csomay, 2006). This is reflected, for example, in the structural categories and discourse functions of lexical bundles. From corpus studies, these fixed strings of words have been shown to have important discourse organising and stance functions, varying not only across spoken and written varieties (Biber et al., 2004), with, for instance, classroom teaching using more bundles than textbooks, but also between spoken registers. Biber and Barbieri (2007, p. 265) found that bundles "serve important discourse functions in both spoken and written texts". Nesi and Basturkmen (2006) also found that lectures contain a mix of 'oral' and 'literate' bundles in their study of the cohesive role of lexical bundles in organisation of

lectures. These uses are important as it indicates how speakers may organise their talk in order to enable listeners to follow.

Other phraseology related studies have included Simpson's (2004) investigation of formulaic expressions in MICASE, finding them to have two broad functions. The first was organisation and structuring of discourse, the second function related to interactivity, with these expressions more common in interactive speaking than in monologues and lectures.

Csomay (2007) compared teacher and student talk in relation to the level of students and disciplines. She identified differences in length and number of turns (students take more but shorter turns) and a range of linguistic items that show differences in orientation depending on level and discipline. Again, she found classroom language to include "features associated with both informational focus (as in academic prose) and involved discourse (as in face-to-face conversation)" (p. 117).

The use of vocabulary has also been identified. Nesi (2002, p. 354) suggests that academic words in BASE reflect the "interactive and interpersonal nature of spoken academic discourse". In addition, grammatical structures in large corpora have been identified (e.g., Biber, 2006b; Zareva, 2009), although as the PBL corpus is not tagged, I will only consider a limited number of grammatical features in the present study.

These studies suggest that spoken academic English is quite distinct from written registers, that lecturers and students differ in the language they use and that the type and mode of the communicative event can affect language choice. However, there is relatively little research into student-centred speaking nor of disciplinary-specific, student-centred speech events.

1.2.4 Disciplinary-Specific Speaking and Medical English

In addition to the insights of spoken academic English provided by studies mentioned above, it is worthwhile considering medical English studies in particular to see what insights they may be able to offer. Much of the research into medical English appears to focus on professional interactions (e.g., doctor–

patient exchanges) or institutional discourse in terms of, for example, ward-round talk, (e.g., Ferguson, 2001; Hoekje, 2007; Sarangi & Roberts, 1999). In terms of medical English specifically, research has focussed primarily on lexis and on written English (no doubt for the very obvious practical issues relating to collection, collation and analysis). In pedagogic terms, Wang et al. (2008) make the case for developing a specific medical academic word list. Hyland (2008b) makes the case for a more discipline-specific academic vocabulary and also highlights the importance of lexical bundles for creating (albeit written) academic discourse. While there a small number of studies related to PBL which will be introduced below, speaking in the disciplines is under-researched.

1.2.5 Research Gap: Student-Centred, Disciplinary-Specific Speaking

Studies to date have identified a range of features that mark out spoken academic registers from written academic English; some have compared spoken registers or the language use of different participants and, to some extent, have identified features of speech events in specific disciplines. These features have included patterns of interaction and linguistic items that signal overall organisation, academic functions and vocabulary, and pronouns and lexical bundles as markers of engagement, interaction and discourse organisation. However, the majority of work has concentrated on lectures or the academic seminar. Few have focussed on one disciplinary area with a view to exploring an aspect of the discipline as their starting point, and few have specifically considered student-centred learning events. However, as with lectures, the seminar, far from being homogenous, can vary in format, purpose and organisation (Aguilar, 2016; Basturkmen, 1999, 2002, 2016) and may reflect different disciplinary practices and identities. Basturkmen shows how seminars vary in purpose, while research indicates that writing in the disciplines varies. Corpus studies have shown how written discourse can differ in organisation and language use across a number of fields (e.g., Nesi & Gardner, 2012). Because of this, I would suggest that it is fair to assume that different types of speaking may vary in format and language, and that this may also reflect the academic discipline. Indeed Hyland (2009) makes a case for more disciplinary-specific research, as does Basturkmen (2016), who points out the need for more of both a disciplinary focus and of student-student talk.

One example of a student-centred speech event, common in the field of medicine is problem–based learning (PBL). This is a distinct type of learning event, underpinned by a very specific pedagogy, as discussed in the next section.

1.3 Problem-Based Learning: An Educational Approach

This section provides an introduction to the key features of problem-based learning and the specific pedagogy underpinning the approach. It also explains how it is applied in the context of this research.

The emergence and implementation of PBL in medical education is commonly associated with the opening of McMaster University in Canada in the 1960s (Barrows & Tamblyn, 1980). Dissatisfaction with traditional modes of medical education involving the transmission of knowledge via lectures and rote learning converged with developments in knowledge and understanding of the cognitive processes involved in learning.

The aim of the original model of PBL was to introduce a more student–centred form of learning that reflected the practical skills a doctor needs to acquire and develop for future practitioner life (Barrows, 1986). These were clinical reasoning skills, problem-solving, an ability to apply knowledge, and self-directed learning. PBL was seen as an educational approach that would help facilitate the development of these skills and at the same time ensure an ability to deal with the ever-growing knowledge in the field. This focus on the development of skills through practice also highlighted the importance of the process of learning (Barrows, 1994). Given the focus on participation, it obviously privileges speaking as part of the learning process, or “speaking to learn” (Basturkmen, 2016, p. 161).

In support of this active approach to learning, Barrows (1994, p. 25) discusses the problems associated with rote-learning, including lack of recall and the benefits of learning in context, and later issues related to ensuring information is available both as procedural (skills learned during in problem-solving activities) and declarative knowledge (available for recall) (pp. 71-73).

Intellectually, this approach has been linked to philosophical, psychological and educational developments at the time of its emergence, including the influence of cognitive theories focussing on the mental processes involved in learning, and educationally to the links to experiential learning and constructivism (Savin-Baden & Major, 2004).

1.3.1 Applications of Problem-Based Learning

In its purest form, PBL is fully integrated into the curriculum. The 'problem' or scenario is the basis of a cycle of study. Barrows' (1986) model for undergraduate medical students mirrors the cycle of work and the clinical reasoning skills doctors engage in in the following way: after the problem (based on actual patients) is introduced (without prior input or study), students brainstorm, sharing what they know of the issue, what they need to find out, and then generate initial hypotheses. During the first tutorial one student acts as a scribe, recording ideas (hypotheses), facts learned about the patient related to the hypotheses, and learning objectives (areas for study). At the end of the first session, the objectives for individual study are agreed on and resources to be used are identified. This stage is followed by self-directed learning, which may involve individual or collaborative work. In the second tutorial, students apply what they have learnt, reassessing the problem in light of new information. The final decision-making is based on the evidence they have accumulated.

Barrows includes two final phases where the participants first summarise and then reflect on what they have learnt. This, he states, is necessary in order for the learning to move from procedural to declarative knowledge for it to be available again at a later stage. To enable this, students need to verbalise what they have learned. The cycle finishes with a self- and peer-evaluation to encourage learners to self-monitor their own performance, another important skill for life-long self-learning (Barrows, 1994, pp. 71-73). As an approach, it is based on the assumption that skilled physicians use deductive reasoning processes and takes the view that repetition and practice are necessary in order for doctors to develop these skills (Barrows, 1986).

Camp (1996) comments on the widespread adoption of PBL, including its variants. These are primarily concerned with the level of integration of the sessions into the curriculum as a whole, and what and when additional information is provided (for example parallel lectures may be given, fixed resource sessions used; see, for example, Alavi, 1995). Barrows (1999) himself suggests six variations and Savin-Baden and Major (2004, pp. 35-45) describe eight modes.

What most frameworks do have in common is the PBL tutorial, the ethos of collaborative learning, the requirement for student discussion and hypothesising, followed by self-directed work and subsequent reassessment of the problem. In describing the educational basis for PBL, Bligh (1999, p. 6) identifies seven steps in PBL, many of which, if they are in fact evident, may have associated language patterns. These are clarifying terms and concepts, defining problem(s), analysing the problems, listing possible explanations, formulating learning objectives and setting priorities, reporting back, synthesizing and testing information.

Apart from the stages involved, it is clear that collaboration and interaction are key components of the approach. The role of the student and tutor are possibly very different from that which students may have experienced before. Not only do both tutors and students collaborate in knowledge construction, but the student as the main contributor, certainly in terms of quantity of talk, is frequently the norm.

1.3.2 Evaluating the Effectiveness of Problem-Based Learning

As an educational approach, PBL is not without criticism, not least because variations in implementation make assessment of the effectiveness of PBL problematic. The level of integration into the curriculum, the expertise of tutors, and modes of assessment have all been questioned. Colliver's review of the literature concludes there is no convincing evidence that it is a superior mode of education overall in improving students' knowledge base or clinical performance (2000, p. 259). However, more promising investigations have been conducted in a number of areas, including the acquisition of cognitive skills

(Hmelo, 1998), the development of meta-cognition (Downing et al., 2009), the relationship between case quality, group functioning and test results (Nieminen et al., 2006) and on the level of 'deep' learning it promotes (Dolmans et al., 2001; Mok et al., 2009).

In relation to implementation, the facilitator and the questions and strategies used to facilitate and scaffold knowledge construction are a key part of PBL. As such, the effectiveness of the facilitator has also been the focus of attention. Hmelo-Silver and Barrows (2006) discussed the role of the PBL tutor in scaffolding learning. They identify, for example, "re-voicing", "summarising" and "elaborating", as some of the strategies used. They also considered the opportunities afforded for knowledge construction and the different ways tutors used questions (Hmelo-Silver & Barrows, 2008). For example, short-answer questions were used to focus student attention, long-answer questions to push for clarifications and elaborations and, by far the most common, meta-questions to evaluate hypotheses and check understandings. These studies highlight a number of the roles the facilitator takes and the complexity of the PBL tutorial, and indicate the potential impact facilitators may have on student engagement and active participation. Chapter 3 considers interaction and engagement in different 'types of talk' in more detail.

1.3.3 Problem-Based Learning in the Context of this Study

The postgraduate MSc in Medical Genetics that this research draws its data from employs problem-based learning as part of its curriculum, as with many courses within the School of Medicine for both undergraduate and postgraduate students at the University of Glasgow and is a well-established mode of study.

Students on this particular course follow a programme incorporating lectures, labs, 'problem' sessions and problem-based learning sessions (PBLs). Although not a fully integrated PBL model, it does have a number of similarities to the McMaster model, as we shall see below. In terms of Savin-Baden and Major's (2004) suggestion of modes of PBL in curricula, it is most similar to Mode 5 which is a two-strand approach where PBL modules draw from other modules in a mixed approach. PBL "is seen by tutors as a vital component … designed to

maximise the use of both problem-based learning and other learning methods" (pp. 41-42).

The PBLs play a significant and regular part in the approach to teaching and learning on the programme, with final PBL assessments contributing 30 credits of the 120 credits awarded for the taught component of the degree. Although the School of Medicine does not follow the 'pure' PBL McMaster model (Barrows, 1994, 1986), partly due to institutional constraints in module and curriculum design, the model employed is similar in that learners are presented with scenarios, work through a series of stages to share knowledge, identify key terms and issues, and then set learning objectives. The first session (one hour at the beginning of the week) is then followed by independent learning in preparation for the subsequent class (from one and a half to two hours at the end of the week) where students report back and make final suggestions for courses of action. It differs from the pure Barrows PBL model in that the self-directed learning requires students to look at only one objective each, which can mean that students may have less to contribute and comment on at the reporting back stage. Also, a number of the PBLs are supplemented with what Alavi (1995) refers to as "fixed resource" sessions, in this case some of the lectures and the 'problem' sessions.

In the first two semesters, while the PBL cycles have a key role in the programme delivery, they are formative in nature and unassessed. The focus is very much on developing skills and learning how to do PBL. Term one has seven PBLs, each delivered in a weekly cycle of work. In semester two, the groups change around. New groups will include some participants from semester one along with students they have not yet worked with. The four PBLs in semester two are followed by a 'mini-big' PBL. This PBL cycle is carried out over two weeks, involves more detail and requires each group to compile a report and give a group presentation to the whole cohort. In essence, this is practice for the assessed PBLs in semester three.

In summary, within the context of this study, we see an adaption of the PBL 'classic' model. The specific profile of the PBLs to be included in the corpus in this study is provided in Chapter 2.

1.3.4 Linguistic Research into Problem-Based Learning

While the educational value of PBLs and the role of the facilitator have been the focus of research, to date there has been much less investigation into the linguistic aspects of PBL sessions, nor of the (potentially) differing roles and interaction patterns displayed by participants. Although this is surprising given that at differing points in the cycle participants may engage in longer and shorter turns and with different purposes, a small number of studies have been carried out.

Legg (2007) and Woodward-Kron and Remedios (2007) are the most informative in terms of sequencing and the linguistic realisation of stages. Woodward-Kron and Remedios report on an initial small-scale investigation into sequencing and scaffolding of a first year undergraduate physiotherapy session, offering insights into the corresponding linguistic elements of a PBL tutorial. Using Bernstein's model for pedagogic discourse and SFL, they first identified generic stages. They found the process more dynamic than guideline procedures for that context suggest, with stages merging or at times emerging earlier than expected and fewer contributions of students from non-English speaking backgrounds. Remedios et al. (2008) also investigated students' cultural background and its relationship to collaborative behaviour in undergraduate physiotherapy students in Australia (both Asian background and Australian). The Asian students were identified as "silent participants", participating by listening and also doing required work, but verbally contributing less. Legg (2007) takes a genre analysis approach to studying seven first year undergraduate PBL tutorials (the second tutorials in a three-tutorial PBL cycle), identifying constituent structure, and the identification of obligatory and optional stages.

Three studies apply CA to study segments of talk in PBLs. Glenn et al. (1999) identify organisational frameworks, while Koschmann et al. (1997, 2000) show how problems emerge, and how tutors guide and help focus topics. These studies, although at undergraduate level and of a limited number of tutorials or only of a part of a cycle, suggest the dynamic unfolding of PBL tutorials. Da Silva and Dennick (2010) are the only study that considers vocabulary use in PBLs, although they only cover one PBL cycle.

In summary, problem-based learning is characterised by interaction and engagement between participants and with the subject matter. This involves learning by doing and actively engaging with content. The research to date has indicated the complexities of the PBL cycle, however there has been little research into whole cycles or into postgraduate groups, and only one (Da Silva & Dennick, 2010) employing tools from CL. This is also reflected in the lack of teaching resources and guides available to EAP teacher working with students preparing for this type of learning.

1.4 Emerging Themes and Research Aims

To date it appears that research into PBL from a linguistic perspective, at least within the medical and science fields, has primarily focussed on undergraduate learners, frequently in one stage of a PBL cycle, or has concentrated specifically on learners from specific cultural and linguistic backgrounds. My research aims to look at PBL in the context of a postgraduate taught programme of international students, both NN and NNS, and to develop a collection of whole PBL cycles. As the approach and the specific model of PBL reported on in this research is quite established within the academic School and the Medical Genetics programme, the primary aim of this research is to describe how PBL is enacted as a way to inform EAP teaching and materials design, rather than to critique the approach to teaching and learning.

The research aims to explore three broad themes. The first is to explore how patterns of interaction unfold and contribute to the pedagogic aims of PBL. This will build on previous studies of academic interactions of small group teaching, but with a specific focus on a distinct pedagogical approach selected precisely because is it situated within one discipline. The second is a related point in that I am interested in investigating how students engage with the subject matter and their peers, and how this reflects the discipline. The third is to consider how, along with the pedagogy and discipline, the variations in the format within this learning cycle are reflected in language choices. To study these aspects of PBL, I have built a corpus of PBL tutorial sessions.

In doing so the study aims to provide an account of how the specific pedagogic aims of PBL and of the discipline are realised by employing different methods to provide a wide-angle view of the event. In order to clarify the overall rhetorical patterns and interactions, Conversation Analysis is drawn upon. In order to investigate specific linguistic features that occur in both stages of the cycle, methods from Corpus Linguistics are used along with detailed analysis of examples, thus combining quantitative and qualitative approaches. This combined approach aims to provide an in-depth study of this type of student-centred speaking.

1.5 Chapter Summary and Outline of the Thesis

This chapter has highlighted the importance of speaking in academic settings as an integral part of the learning process. It has provided an overview of studies of linguistic characteristics of spoken academic discourse, including specific linguistic items and patterns of organisation. It has also pointed out the lack of disciplinary–specific, student-centred research into spoken academic English. It then introduced PBL, a specific approach used extensively in medical education and its application in the context of this research. This pedagogic approach is the focus of the thesis. The research aims to demonstrate how the learning unfolds across the stages of a PBL cycle and how interaction and engagement are realised through a number of linguistic items.

The rest of the thesis is arranged in the following way. Chapter 2 introduces the PBL corpus compiled specifically for this research. It also provides an overview of the methodologies that will be used in the subsequent analysis. Chapter 3 explores the overall rhetorical organisation and interaction patterns in PBLs, along with specific academic functions identified in the student presentations. Chapter 4 identifies keywords in each stage of the PBL cycle. This is followed by two chapters on specific linguistic features that characterise interaction and engagement. Chapter 5 shows how personal pronouns are used as a marker of engagement and disciplinary identity. Chapter 6 then goes on to investigate how lexical bundles contribute to the organisation of the discourse, in conveying stance and engagement, and the extent to which they contribute in exemplifying

the specific discipline. The work concludes with the implications of the findings for further research and for EAP.

As each subsequent chapter covers a distinct aspect, the relevant literature and methodology will be presented in that chapter.

Chapter 2 Methodology: Building and Analysing the Corpus

This study involved the compilation and analysis of a bespoke corpus of PBL tutorials. The qualitative and quantitative analysis involves applying a CA-informed approach for the study of interaction patterns and tools from Corpus Linguistics (CL) to identify a number of linguistic items. This chapter first discusses issues in corpus design and transcription. It then presents the PBL corpus and introduces the specific methodological approaches employed in the research.

2.1 Building a Corpus: Representation and Size

Over the last half century, the growth of Corpus Linguistics and its contribution to our understanding of texts has been considerable. This section introduces key issues in design, including the selection of texts and achieving representativeness.

A corpus is essentially a collection of naturally occurring language data in machine readable form, allowing for automatic processing, and which is representative of a register, genre or mode (Anderson, 2006; Biber et al., 1994; Flowerdew, 2004; Sinclair, 1991; Stubbs, 2004). The main advantage of a corpus is that by analysing and commenting on attested examples, it provides empirical evidence of language use rather than researchers relying upon limited or invented examples, or intuition and introspection alone. The use of various computer-based tools for both corpus compilation and analysis also allows for investigations on a much larger scale than was previously feasible, reflected in the number and size of corpora now available for research.

However, while a corpus is a collection of texts, it should also be compiled in a principled and systematic manner to ensure it is representative. In fact, representativeness is often cited as much more important than size. In a discussion of the development of larger scale corpora and generalisability of findings, Biber (1993) states:

> a corpus must be representative in order to be appropriately used as the basis for generalizations concerning a language as a whole ….

> Representativeness refers to the extent to which a sample includes the full range of variability in a population from both linguistic and situational perspectives. (p. 243)

Biber (1993) and McEnery et al. (2006) both stress the importance of clearly defining the target population at the outset. This then allows for the development of a sampling frame, i.e. the target population from which samples can be drawn, the identification of sampling units and the subsequent collection of representative samples of the given area (Biber, 1993). The resulting corpus should include a proportional, suitably balanced variety of texts from a range of 'authors' to ensure that no one type of text is disproportionately over or under represented or excluded. Because of this, stratified sampling is recommended. This is particularly important when developing large-scale corpora or when more than one type of communicative event or genre is being collected. Douglas (2003) also notes the requirement to be flexible in order to cater for information that may come to light during the compilation process, and stresses the importance of insider knowledge. An example of this is the use of both English as a foreign language professionals and named events (e.g., seminars), in order to identify appropriate sample units in MICASE (Lindemann & Mauranen, 2001).

It should be noted that the question of representativeness is in itself open to some degree of interpretation and subjectivity (Douglas, 2003; Stubbs, 2004). As McEnery et al. (2006, p. 16) note, "any claim of corpus balance is largely an act of faith rather than a statement of fact".

Although representativeness is important, size is also a factor to be considered. There are now many "mega-corpora" (Koester, 2010, p. 66) of collections of spoken, written, and both spoken and written language, comprising different registers, genres, and quantities of texts (depending on the researchers, they may describe the holdings as examples of registers and/or genres). These are substantial in size and can run into millions of words, even for some of the spoken corpora. MICASE for example, is around 1.8 million words, the T2K-SWAL just under 1.7 million, and BASE over 1.6 million. Many large-scale investigations and findings based on such corpora have greatly added to our understanding of features of language used in these registers and of register variation (e.g., Biber

et al., 1999; Biber et al., 2004; Nesi, 2002; Thompson, 2006). Large-scale corpora are also needed for studying, for example, features such as vocabulary which might not typically re-occur frequently in smaller corpora and in order to be able to generalise findings. However, size can bring its own challenges and the extent to which a large-scale corpus is necessarily the most advantageous is also worthy of comment.

O'Keeffe and Farr (2003) acknowledge the benefits of large-scale corpora but also note that the decision to use a large general corpus or a small specialised one depends on needs and that "it is often the design of the corpus as opposed to its size that determines its suitability" (p. 410). Smith et al. (2008) also note the challenges of frequency analysis of large corpora. This includes the fundamental need for further qualitative work in order to adequately interpret and categorise instances of language use, and the potentially large number of irrelevant instances that need to be discarded. Biber et al. (1994) comment that analysing large sets of data also needs additional support as the sheer scale of findings can be overwhelming (and, indeed, potentially unmanageable).

J. Flowerdew (2001) states that smaller corpora of language from specific purpose areas are more appropriate for studying features of a specific genre and register. He does, however, state that for a specialised corpus to be useful for frequency, it is necessary to show how it varies, i.e. it "must be demonstrated that the specialist corpus has a different make up to a general corpus" (p. 76). Koester (2010) also comments that while smaller corpora can bring problems related to limitations in size for certain types of investigation such as lexis and phraseology, the advantage as noted by others is that with a small corpus it is easier to consider all examples of a linguistic feature. She also states that as "specialised corpora are carefully targeted, they are more likely to reliably represent a particular register or genre than general corpora" (p. 69). Moreover, with a large varied corpus and large quantities of data, less can be said about specific contexts.

While large-scale general corpora include a range of registers, L. Flowerdew (2004) points out that they may not include large numbers of smaller, less well-represented areas within the corpus. For example, while the three large corpora

of spoken academic English do include seminars and study groups, these are from a range of disciplines. MICASE includes only seven seminars, eight study groups and 11 student presentations (411,720 words in total) from across four academic divisions. The BASE corpus has 40 seminars but only ten within each of the four disciplines, and in the T2K-SWAL corpus study groups from across six disciplines account for only 141,100 words. A corpus may be large, but findings from specific areas may draw on a relatively small number of samples.

While specialised texts and occluded genres may offer less generalisable findings and may be less readily available, with a smaller-scale corpus researchers are more able to consider specific aspects from across the whole corpus in the qualitative analysis. The compiler may also know more about the situational context, aiding a more fine-grained analysis and interpretation of the data.

Thus, while large corpora bring certain advantages, for example in relation to representativeness of a register or genre and in generalisability, a small, specialised corpus is not without benefit. This is most noticeable in relation to providing relatively large samples of specialised genres, the knowledge of the context the researcher may bring to the analysis, and the likelihood of being able to qualitatively analyse much, if not all, of the data.

2.2 Developing a Transcription System

Included in the process of designing and building a corpus of speech is the development of an appropriate transcription system. This section provides an overview of key points that should be taken into consideration.

Far from being a simple case of recording speaking in written form, transcription of speech involves many challenges and is not a standardised process (Cameron, 2001; Johansson, 1995). However, while a homogenous system may not be available (or even desirable), the issues surrounding transcription can be identified and then used to inform the establishment of a workable system for individual researchers. These factors relate to the challenges of capturing and representing spoken discourse in such a way that it maintains as much of its integrity as possible in the written form; the tension between the need to

ensure readable and usable data (i.e., for each researcher, according to their needs); and of ensuring computer readability (Cook, 1995; Edwards, 1995; O'Connell & Kowal, 1999; Thompson, 2005). Each of these points is discussed below.

Transcribing, by its very nature involves changing the form of spoken language, in itself identified as an artificial process (e.g., Adolphs & Knight, 2010; Cameron, 2001; Meyer, 2002). While such a transformation is clearly necessary in order to ensure it is accessible and analysable (at its most obvious, the limitations of the working memory limit human capabilities), decisions have to be made as to what and how best to capture and represent spoken language.

In discussing the development of a transcription system for dealing with spoken data involving children, Ochs (1979) raises a number of points relevant to the transcription of any data. These include physical layout, the capturing and relevance of non-verbal data and whether turns are contingent on previous ones. Layout itself can bring cultural expectations, for example the order of information going from left to right. With vertical layouts, inferences are made about the temporal nature of messages: a top-down, vertical layout is typically interpreted as occurring chronologically in time, with utterances contingent and sequential (unless otherwise signalled). In relation to the representation of non-verbal information and the extent to which it is ignored or included, Ochs states that it is important to note relevant links between verbal and nonverbal information but also points out that too much detail and information can make transcripts difficult to read. Ultimately, the transcript should reflect the goals of the researcher.

This final point, reflecting the needs and goals of the researcher, is perhaps the main reason why a standardised system has not proved practical (Chafe, 1995). For example, a researcher studying aspects of phonology and prosodic features will require a detailed annotation system, but one that may make the transcript more difficult to read if the transcripts are used to research other linguistic features or for a corpus search. Similarly, the minute detail included in a strict Conversation Analysis system, with, for example, latching and parts of words transcribed (e.g., 'conver' instead of conversation), may also make the

transcript less reader or computer friendly. The level of detail of prosodic features included is also closely linked to the very practical limitations of expertise and time available for transcribers when compiling a spoken corpus, particularly large scale corpora (see, for example Crowdy, 1994, on the British National Corpus (BNC)). Sinclair (1995, p. 102) indeed makes a case for plain text, maintaining "people should be allowed and encouraged to get on with what they want to do and not try to second-guess the future by putting in lots of annotations they do not want for themselves". O'Connell and Kowal (1999, p. 112) summarise the issues neatly: "in short, the usefulness of a transcription system must be judged in light of the purposes for which it is used".

While levels of detail and conventions may vary, what can be agreed upon is the need to be systematic and transparent in the approach used. Edwards (1995, pp. 20-22) suggests that categories used in transcription systems should be "systematically discriminable", "exhaustive" and "systematically contrastive". In addition to layout conventions mentioned above, a range of other features of the system need to be set out, including decisions about spelling (of words and of vocalised pauses), capitalisation, whether linked expressions count as distinct lexical items or as merged words, e.g., going to for "gonna" (Meyer 2002, pp. 72-73), along with any indication of acronyms, numbers and abbreviations, timing, and whether and how to mark pauses.

Most challenging of all is how to indicate utterance boundaries. This can be done by indicating intonation contours, according to pauses, or by imposing a system more in line with written punctuation. The latter brings with it the possible problem of the transcriber imposing their interpretation on the text and representing through punctuation something that "isn't there" (Cameron, 2001, p. 34). Cameron also notes that "units may well coincide with grammatical constructs such as a clause but they do not have to; generally what the boundaries signal is the structure the speaker wishes to impose on the *information* s/he is giving" (p. 35).

Probably because of concerns about imposing an artificial structure or system, a number of transcription systems have based their use of full stops and commas on pauses and intonation. For Sacks et al. (1974), in their research into the

nature of the turn-taking system in conversation through CA, punctuation is used as an intonation marker (they also note that turns, or turn-constructional units, can be lexical, phrasal, clausal and sentential in nature and that the points at which turns can transfer to another speaker appear to be at syntactically relevant points). However, identifying intonational units is not a simple matter nor always practical. The BASE compilers made use of the wavesurfer software to mark pauses of 0.2 seconds within or between utterances (Nesi & Thompson, 2006). For this, not only is specific software required, but marking of pauses to this level can also make transcripts less easy to read. Crowdy (1994, p. 25) in contrast, describes how a "broad orthographic system" was used for the BNC, which included punctuation (commas and full stops) at pauses or syntactically appropriate points. In the MICASE corpus, there is little capitalisation and while commas and punctuation marks are used, these are aligned to length of pauses and intonation contours: "a full stop indicates the end of an utterance with a falling intonation contour" (Ädel et al., 2007, p. 14). The T2K-SWAL corpus also uses commas for short pauses, and full stops for sentence boundaries or final intonation (Biber et al., 2004). The guidelines for the Scottish Corpus of Texts and Speech (SCOTS) corpus allow for punctuation and capitalisation, with commas to indicate pauses when syntactically appropriate, advising transcribers to use intuition. Similar to the MICASE and BNC, this seems to be a practical, easily applicable approach.

In summary, decisions need to be made covering aspects from physical layout to how best to capture utterance boundaries and relevant non-verbal information. Most importantly, the transcription guidelines need to be systematic, transparent, practically applicable, and reflect the needs of the researcher.

2.3 Approaches and Tools for Corpus Analysis

Along with the development of large and small, general and more specialised corpora, the tools for analysis have also developed, many of which allow for easy analysis of a range of features. Approaches to identification of items may vary (e.g., by frequency, keywords, or searching for specific items) but studies generally combine both quantitative analysis and a qualitative analysis of the language in use. This section provides an overview of a number of tools used in

corpora research. Although some large-scale corpora and approaches are included, the focus is more on small-scale and, where possible, academic corpora, as these are of most relevance to my research and corpus.

Some of the approaches, particularly those employed with large scale corpora, make use of sophisticated software, not always available to the individual researcher. Multidimensional analysis, outlined in Chapter 1, has been used to illustrate how linguistic differences between registers can be identified (Biber et al., 1994), for example comparing features of academic registers in the T2K-SWAL (Biber et al., 2004). However, as Xiao and McEnery (2005, p. 68) note, multidimensional analysis is "computationally/statistically demanding". They compare results using this approach and those gained by using the keyword function of WordSmith Tools (Scott, 1996) and suggest that results are approximately similar. Other complex tools are available, including Sketch Engine (Kilgarriff et al., 2014) used with the BASE corpus, and Wmatrix (Rayson, 2009), used by Da Silva and Dennick (2010) to analyse vocabulary changes in one PBL cycle. Complex tools also allow a corpus to be tagged for grammatical features, or, for example, moves in collections of texts (e.g., Henry & Roseberry, 2001; Upton & Connor, 2001).

In addition to the more complex tools, there are less expensive or free and more accessible tools and software available, one of which is WordSmith Tools (Scott, 1996). This software has been used to explore a range of corpora and linguistic items. Features include a word list tool to identify words or fixed sequences of words by frequency, a keyword tool to illustrate statistically significant differences between corpora, and a concordancing tool which allows the researcher to view items in their surrounding co-text and thus facilitates close analysis of use. Specific examples include the exploration of frequency of personal pronouns in lectures (Fortanet, 2004), keywords in MICASE (Hyland, 2009), and individual words, for example the use of *just*, again in MICASE (Lindemann & Mauranen, 2001).

Sections 2.1, 2.2, and 2.3 have introduced factors related to the design, transcription and analysis of a corpus. The next section discusses how these have been applied in the development of the PBL corpus.

2.4 The Problem-Based Learning Corpus

This section presents the PBL corpus, describing its representativeness, data collection and transcription system, and its composition.

2.4.1 Representativeness

The PBL corpus will clearly not be representative of all university seminar discourse, nor even of all spoken interactions in the Medical Genetics programme that the participants are engaged in. However, it is an example of a small, specialist corpus from one discipline and of one very particular usually occluded speech event.

The PBLs in this corpus, as explained in Chapter 1, are a core but unassessed part of the course. There were two reasons for collecting data from the unassessed PBL component of the programme. One was the wish to concentrate on a truly occluded genre that is primarily focussed on the process of learning. Also, in practical terms, although it would have been possible to also record the longer assessed PBLs later in the academic year, this would have required a considerably longer period of time for both data collection and transcription.

In terms of representativeness, this corpus includes five of the twelve PBLs in this unassessed part of the course. Again as explained in Chapter 1, each PBL cycle covers one topic and involves two tutorials, with a period for independent study between the two. The PBL tutorials in this corpus took place at the end of the first semester and in the first half of the second. Although they were more heavily concentrated in the second part of this particular component of the programme for practical reasons, in terms of the sampling unit the collection includes whole cycles. Four of the five PBL cycles occurred within the same week, with the fifth PBL involving a longer period of intervening study, followed by a group presentation to the whole cohort (the final PBL12.2).

As each PBL session includes seven or eight students and one facilitator, and as different groups were recorded each week, the corpus includes a representative sample of the nationalities (14), genders, and educational backgrounds (science

or medicine).[1] Thirty-four of the forty-two students in the cohort are represented in the corpus, along with all of the facilitators on the programme. Although each group comprises students with a range of language levels, no attempt was made to identify or determine these. The group was taken as a representative sample of postgraduate students who had met the entrance requirements for the programme.[2]

2.4.2 Data Collection and Ethics

This section explains the data collection process, including ethical issues. The research was approved by the College of Arts Ethics Committee. Before the data collection commenced, in order to ensure that all students were aware of the purpose and subsequent use of the recordings, the aims of the research were first introduced to the whole programme cohort. Students were then given time to read a plain language statement and ask any questions. This was to provide assurances of anonymity and that the rights of students were made clear, including the right not to take part or to withdraw at any point. Before recording each PBL group cycle, all participants were asked to sign individual permissions for audio and DVD recording, and to allow for the use of transcriptions in publications. It was made clear that the video recordings were to facilitate transcription and analysis and would not be used in presentations of the research.

The audio and DVD recording of the tutorials were made with the researcher as a silent observer.[3] The DVD recordings were invaluable since transcribing solely from audio was extremely challenging. This was due in part to the number of speakers, overlaps, the range of accents and speed of speech, and because of frequent references that the participants made to white-boards, screens and handouts during the PBL sessions. In order to facilitate transcription, Praat

[1] In fact, the mix and balance of students in each group is determined by the teaching staff on the programme.
[2] All students also contributed at least linguistically effectively throughout the sessions, although participation levels varied. Levels of participation was not, however, obviously attributable to language level (some of the UK students were noticeably reticent for example, and some of the NNS speakers were notably dominant at points).
[3] Although an observer may appear to affect participation levels, this did not appear to be the case in any recordings. I have observed many hundreds of classes as part of teacher training and education courses and would say that students appeared unaffected by my presence.

software (Boersma & Weenink, 2013) was used, with audio recordings converted to wave format for compatibility. When the sound file is loaded into Praat, the transcriber allocates a separate tier or row for each speaker in a viewing pane; a tier for transcriber notes can also be added. The transcriber can play and re-play very small sections of recordings with ease. A specifically designed additional piece of software provided speaker and time stamps on transcripts. Although some software packages were available for DVDs, none could handle the size of PBL DVD recordings. The specific transcription system developed is described in the next section.

2.4.3 Developing the Transcription Guidelines

The initial stage of transcription was exploratory and iterative in nature to allow for the development of a system that would suit the focus of the research, be compatible with the software and which would also reflect the nature of the PBLs themselves. As I am interested in the organisation and unfolding of the talk and of specific linguistic items used to realise these discussions, the transcription needed to represent the highly interactive nature of the event as closely as possible, including utterance overlaps, false starts, pauses and laughter. In order to develop a sense of the relevant issues, the initial stage involved transcribing the first three hours and noting problems that arose and decisions that were made.

The areas that required most attention were those concerning how to 'punctuate' and mark utterance boundaries, layout, and deciding on the level of detail to include for overlaps and latching (where one turn is immediately followed without any perceptible gap by another speaker).

To identify boundaries of units of talk and in order to facilitate the reading of longer turns (particularly in the PBL2 presentations), I have used commas for pauses and full stops where it appears intuitively clear to me that the speaker has completed a section of talk (most notably with a falling intonation). I also decided to time and note pauses, as these occur relatively frequently and are significant in PBLs as they often indicate reference to board work or handouts; hesitations and pauses also characterise much unplanned speech. These are

marked when they are of two seconds or more. Although the decision on the length of pause to mark is to an extent arbitrary, marking a pause of this length was relatively easy to apply.

In terms of layout, the sequential and contingent nature of most turns allows for a top-down layout. Overlaps and latching are easily seen on the Praat tiers and can be noted in transcriber notes. They are also indicated by the Praat time-stamp and can be referred to in detailed transcription printouts when required. The aim of PBLs, i.e., to share and build knowledge through the multi-party discussion of issues largely amongst peers, is clearly evident in these overlaps and in false starts, notably in PBL1s and at the post-presentation question and answer sections in PBL2s. For false starts and overlaps which occur mid-word, the demands of a software programme for searching mean any mid-word truncation is followed by a hyphen to clearly mark this out, and overlaps are indicated only at the beginning of words, not mid-word. As the research does not focus on prosodic or phonetic aspects other than for signalling pauses and questions, these have not been incorporated into the transcriptions.

The nature of the PBL sessions in this corpus, where all but the facilitator are equal participants, along with the number of participants actively contributing, and the high level of turn-taking, brought its own challenges. This made transcribing a much lengthier process than initially anticipated.[4] The variety of accents and at times the speed of the speakers, coupled with the unfamiliar terminology[5] and subject matter also proved to be challenging. Because of these factors, not all parts were clearly audible. Wholly unintelligible words are indicated by a question mark in square brackets, [?]. However, on a number of occasions a 'best guess' was possible, for example when part of a word was clear, or where the surrounding talk, a handout or slides assisted. These have been included to help make sense of the utterance in the subsequent analysis. Square brackets are again used, but include a question mark before the word, [? guessed word]. Explanations of non-verbal information, for example *asides, gaze,* or *concurrent talk*, when viewed by the transcriber as significant to

[4] Cameron (2001) suggests an average of 10 hours to one hour of recording but this proved to be completely unrealistic. On average, each PBL cycle took between 20 to 25 hours to transcribe.
[5] Handouts from participants and specialist online dictionaries recommended by the Medical Genetics team were used for unfamiliar terms.

comprehension of the message are also included within square brackets if they were mid-utterance e.g., [*aside; inaudible*], or noted in the transcriber notes tier. Where overlaps and concurrent conversations occur, I decided to transcribe what I could in an attempt to represent as much of the talk as possible, but it again means that utterances may be incomplete. This all inevitably means that the word count can only be approximate.

The presentation of examples in the thesis include speaker codes with tutors identified with an additional *T* and number, the transcribed excerpt and the relevant PBL number and stage. Timestamps have been removed for ease of reading. An example is provided below:

> S24T3: ok chair do you want to
> S20: yeah sure (PBL10.1)

In Chapter 3, as a CA informed approach is taken to explore the unfolding of the interactions, the transcription system has been slightly modified to reflect the CA approach to transcription. This includes using parentheses for marking pauses and line numbering. This is explained in more detail in that chapter. The guidelines for the transcription system used for all other examples can be found in Appendix 2.

By its nature, transcription requires listening multiple times to recordings and sections of recordings.[6] This, along with cross-referencing to the DVDs, enabled a quite detailed transcript to be made. While every effort has been made to capture as much of the speech as possible, the multiple overlaps meant some parts are inaudible, which means it is not an absolutely complete representation.

2.4.4 Composition of the PBL Corpus

The data consists of recordings of five complete PBL cycles from a postgraduate taught degree in Medical Genetics.

[6] One PBL, PBL12.2, was transcribed by colleagues but carefully reviewed and checked by myself.

Table 2.1 provides the following information for each PBL session: the number of participants including the facilitators, the topic of each cycle, the length of each tutorial rounded up to the nearest minute, and exact word counts (tokens). The total number of hours and minutes for each stage is provided in brackets. For PBL12.2, the number of participants includes the 15 presenters and six facilitators. Additional voices were captured in the post-presentation question and answer section but are not counted here. Each PBL group typically included at least one native speaker (NS) student and one student with a background in medicine.[7] As students were re-grouped after PBL7.2,[8] the total number of main participants was 40,[9] including the six facilitators on the programme. The cohort is a mix of NS and NNS. This means that in examples from the transcripts, there may be grammatical inaccuracies. No attempt has been made to alter student utterances.

Table 2.1: Composition of the PBL Corpus

PBL and Stage	**No. of participants**	**Topic**	**Length in Minutes** (hour: min.)	**Tokens**
PBL7.1	8	Cytogenetics	68	8,065
PBL7.2	9		63	9,168
PBL9.1	9	Pre-natal screening	64	8,687
PBL9.2	9		109	17,686
PBL10.1	9	Cystic Fibrosis	58	7,632
PBL10.2	8		85	13,580
PBL11.1	8	Cancer	67	11,259
PBL11.2	9		89	14,805
PBL12.1	7	Dysmorphology	50	8,528
PBL12.2	15 + Q & A		72	16,083
PBL1s:	37		307 (5:11)	44,171
PBL2s:	40		418 (6:58)	71,322
Total:	40 speakers (34 students & 6 tutors)		725 (12:08)	115,493

Note. Minutes are rounded up.

[7] Students with a background in medicine have frequently studied medicine at this university and so have experience of PBLs.
[8] Students did seven PBL cycles in semester one, one of which was recorded. In semester two, some of the PBL groups included students who had been recorded in a group in semester one.
[9] Additional students were captured in the follow-up question and answer PBL12.2.

The first stage of the cycle, the PBL1s, accounts for 42% of the time and 38% of the tokens, the PBL2 sessions account for 58% of the time and 62% of the tokens.

The corpus is not large, certainly not by the standards of many of the current mega-corpora and conclusions cannot be said to be generalisable to academic speaking as a whole. However, at approximately 115,000 words and around 12 hours of one specific type of speech event in a very specialised area, it is not insignificant.

2.5 Analysing the Corpus

The overarching aim of this research is to provide a description of how the PBL as a disciplinary specific, student-centred learning event, underpinned by a specific pedagogy, is realised linguistically in this context. In order to provide as rich a description as possible, different approaches to analysis are employed. Researchers making a case for drawing on a range of approaches argue that this provides a more comprehensive picture of learning events and that they may usefully complement each other (e.g., Dalton-Puffer, 2007; Walsh et al., 2011; Morton, 2012). Green and Dixon (2002) state:

> No one approach … provides a full picture of what constitutes language-in-use, yet taken together they create a broader understanding of the challenges facing students and teachers. (p. 400)

In selecting areas for consideration, these were informed by the aims of PBL and relevant literature on the nature of academic and interactive talk: the PBLs are, as we have seen, about interaction and engagement both with the disciplinary subject matter, but also with peers and tutors.

In order to investigate how the PBLs unfold as a learning event, a Conversation Analysis-informed approach (different to a pure CA and explained in Chapter 3) was used to identify patterns of sequential organisation of the interactions, and of the overall organisational structure and development of the stages. As the PBL2 presentation reports are essentially a long turn, in order to identify more clearly what the students talk about, academic functions were identified. This

involved analysing the longer stretches of talk, i.e. the presentations, and identifying functional categories (e.g., cause and effect).

As noted, both the corpus itself and the software offer opportunities to identify a range of linguistic features depending on the interests of the researcher. A number of linguistic features have been identified as markers of a discipline, of interaction, engagement and in the organisation of discourse. In order to explore these features, corpus tools were used. WordSmith Tools (Scott, 2004) was selected for a number of reasons. As mentioned in Section 2.3, the software enables the compilation of word frequency lists, keywords and clusters (lexical bundles). The interface quickly and easily allows the researcher to view items from a concordance line and in the fuller context, necessary for the more fine-grained analysis when exploring aspects of use. It has been used quite extensively in studies of both written and spoken discourse and is thus well-tested and reported on (e.g. Adolphs, 2006; Ghadessy et al., 2001; O'Keeffe & McCarthy, 2010). This includes the volume of studies edited by Bondi and Scott (2010), almost all of which use WordSmith to explore keyness.

Anthony (2013) discusses four generation of corpus tools, of which WordSmith is third generation. The fourth generation tools, he notes, deal with very large corpora (not necessary for the small PBL corpus), and also enable tagging of, for example, part of speech (PoS) and in the case of Wmatrix (Rayson, 2009), semantic tagging. At an early stage, it was clear that the PBL corpus included many instances of comprehensible but not grammatically accurate utterances, mainly, but not exclusively, from NNS participants, and many incomplete utterances. As the main focus of the study was not going to be on parts of speech (PoS), a software tool such as WMatrix that tags in this way was not required. Da Silva and Dennick (2010) use the Wmatrix semantic tagger to consider the development of reasoning skills and subject knowledge over one cycle, with a view to future studies investigating lexical development of over time. As the PBL topics changed each week and indeed the composition of the groups changed during the data collection, this approach will not be used. The keyword analysis should also serve to show if there are differences in focus between the two stages.

As the analysis in Chapters 4, 5 and 6 each employs a different feature of WordSmith Tools, more detail will be provided in the relevant chapter, although a brief introduction is provided here.

Vocabulary is a recognised marker of disciplinary variation. In order to identify how the subject matter distinguishes PBLs from other examples of academic speech, keywords were identified and explored in Chapter 4. The word list and concordance functions were used to identify and facilitate the qualitative study of personal pronouns, a recognised marker of engagement and involvement in Chapter 5. Three- and four word lexical bundles have also been identified as markers of stance and discourse organisation, and were also identified using frequency tools. This is explained in Chapter 6. By considering a number of features, this provides a rich, multifaceted description of the PBL cycle.

This thesis, then, aims to make its original contribution to research through the development and exploration of a bespoke corpus of disciplinary-specific student-centred speaking, a speech event which privileges speaking as part of the *process* of learning, i.e. speaking to learn. This type of largely occluded genre is by definition less easy to come by. By adding to the descriptions of how PBLs are realised linguistically and how these features both contribute to and realise learning, this offers insights to a pedagogically rich tool for education, both in HE and, I would suggest, language learning.

Methodologically, by employing what I view as complementary approaches to analysis, this study contributes to the growing number of studies of spoken discourse that take a wide-angle, multifaceted perspective on a specific type of speech event.

2.6 Chapter Summary

This chapter has discussed factors involved in building and transcribing a representative corpus. I have discussed the advantages of developing small specialised corpora. I then presented the PBL corpus developed specifically for this research. The methodologies to be employed in exploring the corpus were also introduced. The next chapter presents the findings of the patterns of organisation and academic functions before moving on to subsequent chapters covering keywords, pronouns and lexical bundles.

Chapter 3 "Doing Problem-Based Learning": Organisational Structure and Types of Talk

"Why that now." (Schegloff, 1980, p. 147)

3.1 Introduction

Problem-based learning as we have seen is essentially concerned with the dual purposes of developing subject-related knowledge and collaborative problem-solving skills. It is ideally student-centred and led, with an overall guiding framework of steps to complete and it is through interaction that this learning is realised. How PBLs are enacted is the subject of this chapter.

The specific aims here are to identify and describe the different patterns of interaction and talk in the corpus in relation to the overall pedagogic aims of problem-based learning. The first point that one immediately notices from looking at the transcripts of the two stages of the cycle is that the PBL1 discussions are highly interactive, dynamic events with multiple participants contributing in quite fast–paced turns–at-talk; this is compared to large stretches of the PBL2s which are devoted to individual student presentations, each being effectively a long extended turn. In investigating the development of the multi-party interactive aspects, the transcripts were analysed drawing on tools and concepts from Conversation Analysis (CA), an approach that studies the unfolding of interactive talk. However, apart from the openings, closings and follow-ups to the student presentations, CA, with its focus on sequence and turn-taking, does not lend itself well to the analysis of the main body of the presentations, where turn-taking is largely suspended. In order to provide a fuller picture of talk in both stages of the PBL cycle, an exploration of academic functions represented in the presentations has also been carried out. Together this then provides a description of common interaction patterns and functions of talk over the PBL cycle.

The chapter is divided into four main areas. The first part provides an introduction to the principles and methodology underpinning CA, along with relevant applications to classrooms. This is followed by an explication of the

overall organisational structure of each stage in the PBL cycle. A discussion of specific 'types of talk' identified in the transcripts is then presented. The final part looks at academic functions identified in the stage two presentations.

3.2 A Brief Introduction to Conversation Analysis

The investigation of the unfolding of the PBL interactions is informed primarily by Conversation Analysis, an approach that studies the co-construction of talk-in-interaction. This section will provide an introduction to CA and aspects relevant to the current study.

CA studies the organisation of social action in interaction (Schegloff, 2007; Seedhouse, 2004; Sidnell, 2010). In explaining how interactions unfold, it has a very specific focus of "linking meaning and context to the idea of sequence" (Heritage, 1998, p. 3). Although originally concerned with understanding the mechanics of conversation (Sacks et al., 1974) and how "participants understand and respond to one another in their turns at talk" (Hutchby & Wooffitt, 2008, p. 14), it has been widely applied to a range of speech-exchange systems. This includes examples of institutional talk, characterised by specific goals and greater restrictions on the participants (e.g., Drew & Heritage, 1992; Heritage, 1998). Although the approach is still commonly known as Conversation Analysis, it is now also referred to as talk-in-interaction (Schegloff, 2007, p. xiii), more accurately reflecting the range of speech-exchange systems and contexts studied, i.e., conversation is not the only focus of research using this approach.

The purpose of this section and the brief introduction to CA is not to provide a full and comprehensive account of the approach; summaries can be found elsewhere (see, for example, Heritage, 1998; Hutchby & Wooffitt, 2008; Sidnell, 2010; ten Have, 2007). A brief overview of the initial influences on the development of CA is, however, useful in order to highlight key aspects of CA that will be drawn upon to inform the analysis of the interactions in the PBL corpus.

CA has its roots in sociology and ethnography. Pioneered by Harvey Sacks, and developed further with Schegloff and Jefferson, Sacks was influenced by

Goffman's work on conversation as an institution, i.e. an entity worthy of study in its own right (e.g., Goffman, 1983), regarded as the basis for social action and Garfinkel's (1967, p. 76) interest in the "common-sense knowledge" on which people base their social actions, i.e. on why this institutional order is understood and recognised. This is based on the assumption that we orient to norms to make sense of actions and so develop intersubjectivity, i.e. mutual understanding (Heritage & Clayman, 2010; Seedhouse, 2004; Sidnell, 2010). From Goffman emerges the significance of order in social interaction, and from Garfinkel's (1967) ethnomethodology, a focus on trying to explain why we know what to do and say, as a means to understanding social actions.

Sacks et al.'s (1974) work on CA, however, goes further. Taking an emic perspective, i.e. aiming to establish an insider's perspective, they provided a systematic approach to identifying and describing the organisation of turn-taking in conversation, at the heart of which is the notion of sequence. To establish this insider perspective, this requires evidence to be in the data, i.e. within the system, rather than interpretations being imposed. Their analysis, based on naturally occurring data, suggests a description of conversational order which is "context-free" (i.e. not bound by one situational context), where the order or norms set out should be applicable to a wide range of contexts and which is applied in a "context-sensitive" way, i.e. specific to the linguistic and situational context in which it occurs (Sacks et al., 1974). The concepts of sequence and turns, or turns-at-talk, are obviously linked. A turn, or first part pair (FPP), is normatively followed by a conditionally, contextually relevant second part pair (SPP), forming an adjacency pair and the basis of a sequence of actions. In this way, turns are seen as "context-shaped" and "context-renewing" (Heritage, 2013b, p. 3), in that turns are shaped by the immediate context they are produced in (i.e. with reference to preceding contributions, within a specific situational context) and in themselves provide a context for the next turn, thus forming part of the sequential environment and so being context-renewing. Heritage (2013b, p. 2) notes "Conversation Analysis (CA) involves an approach to language and interaction premised on the notion that the persons use language in order that their recipients will recognize their intentions in speaking", maintaining that participants focus on the actions of messages, i.e. what they understand the message to mean or what it requires them to do. Because of this,

sequence organisation (how series of turns are coherently developed) are key concerns of CA. Turns-at-talk are viewed as “documents” of evidence, showing an orientation to following norms, affiliation and a structural bias toward cooperation, i.e. the reciprocity of perspectives (Seedhouse, 2004, p. 9), again highlighting the notion of shared understanding. Talk-in-interaction is seen as a jointly accomplished activity, with the listener and the speaker the co-constructors of the emerging talk. At least in conversation, talk is “locally managed” and “party administered”, i.e. managed by the participants in relation to what is being said within that specific context, at that time (Sacks et al., 1974, pp. 725-726).

Since the interaction itself and the idea of sequences (of turns) are central to the construction of meaning and context (Heritage, 1998), CA studies may focus on a variety of aspects related to turn-taking, for example, the construction of turns (turn-constructional units), transition-relevance places (when another speaker may have the opportunity to take the floor), repair (dealing with trouble spots), when and how turns are allocated (e.g., a norm in conversation is not to overlap, at least not for any lengthy period of time), and the sequence of turn-taking through analysis of FPPs and SPPs. As the relationship and interaction between turns is central, Heritage (2013b, p. 10) suggests that “sequence organization is the ‘engine room’ of interaction”.

In summary, by studying how interactions unfold turn-by-turn in naturally occurring data, CA aims to show how each turn and sequences of turns are connected, how stretches of talk develop and thus how interactions are co-constructed and understood. It is essentially trying to address the question of “why that now” (Schegloff, 1980. p. 147).

In relation to problem-based learning, as it is said that through interaction that skills and knowledge are developed, an approach such as CA which has interaction at its heart is well-suited to investigating how PBLs are enacted.

3.2.1 Applied Conversation Analysis

While describing the 'institutional order' of interaction was the main concern of 'pure CA', with conversation taken as the benchmark for comparisons to other types of speech-exchange systems, "applied CA" examines "the operation of social institutions in talk" (Heritage, 2004, p. 104). Institutional talk is characterised by specific goals of participants, where there may be restrictions on types of interaction, for example in who can normally contribute, when, and in what manner (Heritage, 1998). Drew and Heritage (1992) and Heritage (1998), for example, provide a number of illustrations of how institutional identities are enacted and maintained through talk, and how institutional goals can shape and constrain interactions. From this, Heritage (1997, p. 225) suggests ways in which institutional talk may differ from conversation and which may form a locus of investigation:

- overall structural organisation of the interaction
- turn-taking organisation
- sequence organisation
- turn design
- lexical choice
- epistemological and other forms of asymmetry.

By studying interactions in detail, institutional CA can offer insights into how the discourse is created, i.e. how it unfolds in response to previous contributions and how the institution is "talked into being" (Heritage, 1984, p. 237).

3.3 Conversation Analysis Methodology

The methodology for applying CA very much reflects the dynamic nature of talk-in-interaction itself: in relying directly on the data, a CA methodology starts from a bottom up, unmotivated (i.e. where the analyst comes to the transcript with an open mind and is not constrained by a priori categories), detailed and systematic analysis of turns as they unfold. A number of handbooks provide practical guidance on carrying out CA-informed analysis (e.g., Heritage, 1997, 2004; Seedhouse, 2004; Sidnell, 2010; ten Have, 2007). The approaches are

summarised and synthesised here and can generally be found to include the following stages:

- collection of naturally occurring data (audio and/or video)
- preparation of detailed transcripts
- identification of overall structural organisation
- identification of "episodes" of talk and compilation of "collections"
- detailed analysis of phases/episodes (e.g., sequence organisation, turn construction, repair)
- comparisons (e.g., to other speech-exchange systems; to "deviant" examples).

3.3.1 Naturally Occurring Data and Transcripts

In order to apply CA as an approach, analysts make use of transcripts of naturally occurring data. Very detailed transcripts are produced based on repeated listenings and, if practical, may even include analyst conference discussions, where a number of researchers come together to discuss their descriptions (e.g., Glenn et al., 1999). Since the transcript is the primary resource for analysis, recording as much detail as possible is desirable. Nowadays audio can often be aided by visual recordings, allowing for the study of gaze and the use of other non-verbal artefacts (e.g., gesture). However, although this can provide for a fuller description, the use of one or a number of cameras in itself can affect the naturalness of the data and should be considered with caution (ten Have, 2007).

Nonetheless, as detailed a transcription as possible is encouraged. Jefferson (2004) produced a very detailed transcription system, which is often used as the basis for CA studies. Modified, 'slimmed-down' versions are also made use of, adapted to reflect the aims of specific research and practical limitations. For example, as in this research, the size of the corpus and the use of computer software for other aspects of the analysis may restrict orthographic modifications that might reflect how something is said and annotations of word or vowel lengthening. An example of a modified system used in order to apply CA to a corpus of already transcribed data can be found in Walsh et al. (2011)

(for example, they add pause information and line numbers to specific stretches to reflect CA presentation conventions). As ten Have (2007, p. 96) says, transcriptions are “always and necessarily selective”.

3.3.2 Identification of Patterns

CA starts with ‘unmotivated’ reading, from which can emerge specific aspects for more in-depth analysis. While an analyst will always have some purpose, this merely means not trying to impose patterns onto the data. The analysis could include overall organisational structure and “episodes” of talk, stretches of talk with a distinct pattern, that stand out. A number of examples of episode can be built into “collections”, allowing for similar patterns to be explored. During the analysis of specific aspects, comparisons to other speech-exchange systems and deviant examples (i.e. ones that do not appear to ‘fit’) might be made. Since the analysis and reporting should be evidenced in the transcripts, no a priori categories or additional contextual information is assumed: the core data should speak for itself, with all comments seen as locally relevant. Ten Have (2007, p. 31) explains that from a strict CA point of view, the perspective of others even by participants, e.g., post-event interpretations, may be just that: an interpretation rather than evidence of actual action.

3.3.3 Limitations of Conversation Analysis

While CA’s strength lies in focussing specifically on natural data and in taking a dynamic view to understanding interactions, limitations should also be acknowledged.

Although CA does not discount the larger context, it maintains that all relevant aspects (e.g., identity, shared knowledge) should be evident in the talk at some point; the evidence in the interaction is the only source of data for the strict CA researcher. However, there may be other layers and contextual information that might be relevant and indeed some applied or CA-informed studies do now provide additional contextual background information (see, for example, Dalton-Puffer, 2007; Mori, 2002; Morton, 2012; Peräkylä, 1997). Morton (2012), for example, notes that additional layers of information about participants may aid

understanding. Morton also notes that the pre-occupation with small, specific data sets and what can immediately be observed means it is not suited to documenting learning over time or internal cognitive processes. Beyond CA, Mercer (2008) also emphasises the importance of temporal aspects to be taken into consideration when investigating classroom learning, as learning typically progresses over time.

Requiring that the evidence be found directly in the talk, CA is a "militantly behavioural discipline" (Markee, 2007, p. 1023) in that it focusses on the manifestation of understandings through talk and actions. Although it cannot claim to show internal cognitive states or provide a comprehensive account of how learning takes place (e.g., in the case of education and language teaching), it does provide a perspective on how talk and actions may embody aspects of thought and learning and may be, in part at least, an observable indicator of learning or the application of skills (e.g., Firth & Wagner, 2007; Pekarek-Doehler, 2010).

The focus in CA is on social actions, rather than attempting any form-to-function mapping. As such, researchers are interested in what people do and how they co-construct talk but much less so on detailing language forms. That is not to say specific items are not studied. Beach (1993), for example, considers uses of 'okay' and Bolden (2009) looks at 'so' in turn openings. However, the focus is very much on situated interactional use. For an investigation of language use across collections or corpora, alternative additional approaches may need to be employed. For example, Walsh and O'Keeffe (2010) and Walsh et al. (2011) make use of Corpus Linguistics (CL) in conjunction with CA to shed light on the use of a number of linguistic items and how they are interactionally relevant.

Applying a 'full-blown' CA transcription system is only realistically possible with very small data sets due to the time-intensive nature of the process. Additionally, the very detailed analysis of small data sets means findings are not generalisable. That said, by using a modified transcription system, as is the case here, and by building collections of data sets, an indication of patterns can be provided.

Notwithstanding the limitations, the benefits of using a CA-informed approach lie in the insights it can provide into how interactions develop, and how participants respond and shape interactions to co-construct the discourse and achieve institutional aims. Examples of how this has been applied in classroom settings are provided in the next section.

3.4 Applications of Conversation Analysis in Classroom Settings

The analysis of talk-in-interaction has been applied to many institutional settings, for example courtroom language, doctor-patient interactions, and includes classroom settings both in schools (e.g., Dalton-Puffer, 2007; Jakonen & Morton, 2013; Koole, 2010, 2012; McHoul, 1978; Mehan, 1979; Sert, 2013; Sert & Walsh, 2013; Walsh, 2006) and HE contexts (e.g., Benwell & Stokoe, 2002; Dippold, 2014; Stokoe, 2000; Walsh, et al., 2011).

Of particular interest, of course, are the studies related to educational contexts. Three not unrelated areas will be highlighted in the following sections. The first introduces studies of 'types of talk' (sub-varieties of the speech-exchange system classroom talk), and how these are linked to the overall institutional goals of the learning event. This includes three studies focussing on stretches of talk in a PBL tutorial (Glenn et al., 1999; Koschmann et al., 1997, 2000). The second, the initiation, response, feedback/evaluation (IRF/E) sequence is discussed because of its prevalence in discussions of classroom discourse. The third area concerns epistemics. As knowledge development and sharing is one of the underlying aims of PBL, how knowledge imbalances are addressed again links to one of the overall aims of PBL.

3.4.1 Types of Talk

Within educational linguistics, Mercer (1995, 2000) has identified idealised "types of talk" in classroom settings which may contribute to learning, for example, "exploratory talk". The exact approach for identification is not altogether clear, however the concept of patterns of talk, often referred to as speech-exchange systems, can also be found in CA-informed approaches to HE settings.

Three studies apply CA to study segments of talk in PBLs. Glenn et al. (1999) look at the description of a "theory presentation and assessment" section of a PBL tutorial, occurring after a period of independent study. They identify two organising frameworks, a student-centred problem solving or decision making orientation and another of instructional, teacher-student interaction (p. 131). Overall, they find the PBL discussions "thoroughly interactional" (p. 131): students concur, discount or do not support contributions (e.g., for the latter responding with critical questioning, disaffiliative laughter, or even with silence).

Koschmann, et al. (1997, 2000) consider how learning issues (objectives) emerge along with the tutor and students' role in a specific segment in a PBL (a "knowledge display segment" which is part of their "identification of issue" phase). They observe that as participants collaborate, "students may expand, clarify restrict or alter topics", which are "dynamic and emergent" (Koschmann, et al., 1997, p. 7). They also identify the tutor role as guiding and focussing by using questions, at times giving some information and asking for confirmation (2000).

A number of studies are now exploring how CA can be used to inform or be combined with other research methodologies, providing complementary information on talk. This includes those which combine Conversation Analysis and Corpus Linguistics (CA-CL). Employing a CA-CL approach, O'Keeffe & Walsh (2012), Walsh & O'Keeffe (2010) and Walsh, et al. (2011) study small group teaching in HE, investigating how students and tutors "manage the complex relationship between pedagogic goals and the talk used to realise them" (Walsh & O'Keeffe, 2010, p. 144). Applying CL to identify specific, recurring linguistic features and CA to study the development of interactions in the Limerick-Belfast Corpus of Academic Spoken English corpus, they identify different types of talk or speech-exchange systems, categorised in O'Keeffe and Walsh (2012) as *organisational talk* (for organising), *instructional talk* (for teaching/informing), *argumentative talk* (developing criticality and discussion) and *discursive talk* (for sharing and reflecting on learning), each with different purposes and patterns of interaction but all serving the wider pedagogic, institutional goals. For example, they found the IRF sequence (discussed in the next section)

prominent in instructional talk where the tutor is controlling the discourse, but also in more symmetrical exchanges in argumentative talk as tutors and students "give and take" (p. 176).

Seedhouse (2004, p. 124) explores language classrooms and illustrates how different trajectories relate to pedagogic goals. He highlights the reflexive nature of turn-taking constraining tasks and task-constraining turn-taking.

Within PBLs it is likely that different phases may include different 'types of talk' and patterns of interaction, depending on the pedagogic focus.

3.4.2 The IRF/E Sequence

As noted, CA sets out a basic two-part turn sequence, where a FPP or initiation move requires a conditionally relevant response. However, one sequence of interaction that has been found to be prevalent in traditional classroom settings is the initiation, response, feedback/evaluation sequence (IRF/E), identified from both a linguistic discourse analysis perspective (originally by Sinclair & Coulthard, 1975) and applied in CA-based studies, the focus in the latter more on institutional identities and the purpose of turns (e.g., McHoul, 1978; Mehan, 1979). In the Sinclair and Coulthard model, the typical triadic pattern is teacher initiation (I), student response (R) and teacher feedback (F), resulting in an IRF sequence. The final F is often to evaluate student contributions. Mehan (1979) uses the term evaluation (E) in place of feedback; he also suggests that an IR adjacency pair then becomes the first part of the IR+E sequence, thus maintaining an overall paired sequence patterning. Both variations (IRF/IRE) are now commonly referred to in language teacher education and discourse.

In typical classroom settings, this pattern has been seen to allow the teacher to control (and dominate) the discourse. McHoul (1978, p. 211) notes that at least in a traditional, teacher-fronted classroom "differential participation rights and obligations" give the teacher more rights, for example in nominating speakers, highlighting the institutional identity of teachers and students. Although criticized for privileging teacher control and constraining student discourse, the final evaluation/feedback move has also been found to have a useful scaffolding

function. For example, the evaluation move may be a prompt which will lead to further responses (e.g., Mehan, 1979; Seedhouse, 2004). Alternatively the response may be unsatisfactory and may be followed by a second initiation, leading to an IR^n (i.e., recurrent) sequence ending in a final E (obligatory as a terminal move in institutional sequences for Mehan; optional in Sinclair and Coulthard).

In relation to HE settings, although not from a CA perspective, as noted in Chapter 1, Basturkmen (1999, 2002) applied the Sinclair and Coulthard exchange structure model to the analysis of a corpus of postgraduate MBA seminars, finding both IRF and IR^n patterns, although with the most common pattern to be IR^n. Although the two-part IR pattern was the more prominent in her corpus, complex exchanges involving IRF/I appeared to allow "ideas to emerge and be negotiated in interaction" (Basturkmen, 2002, p. 233).[10] The F/I move was the ignition for further exchanges, often as a result of dissatisfaction with the first response. Her work on seminars highlights not only the two patterns but also the relevance of complex exchanges as the locus of learning and discussion.

Whether these patterns, particularly the IRF/I, are prevalent in a very student-centred learning event will be of interest.

3.4.3 Epistemic Imbalances

Although again not exclusive to institutional talk, Heritage (2013a, p.16) suggests that "institutional interactions are overwhelmingly characterized by asymmetries". These may be due to access or rights to access of knowledge and roles ascribed and, as we have seen in classroom settings, can be played out through an IRF sequence.

In attempting to redress knowledge imbalances, Heritage (2013a) highlights the role of epistemics, which he describes as concerned with the knowledge claims that "interactants assert, contest and defend in and through turn-at-talk and sequences of interaction" (p. 370). He maintains that interactants orient towards the knowledge of others and work to remove knowledge imbalances

[10] Where F/I indicates that the feedback move also becomes the initiating (I) move for a following turn.

(asymmetries) by working towards epistemic congruence where status and stance are compatible (p. 379). Heritage's differentiation between epistemic status and epistemic stance is as follows: one's relative access to knowledge provides higher or lower epistemic status, positioning the person as more or less knowledgeable, i.e. more "knowing" (K+) or "unknowing" (K−) (pp. 376-378). This involves not only knowing but also the "rights to possess and articulate" (p. 377) this knowledge. For example, even though they may ask questions (initially suggesting K−) to elicit knowledge from students, a tutor would normally be expected to be more knowledgeable (K+) than students. Epistemic stance indicates commitment to knowledge, with parallels to epistemics and evidentiality in linguistics; see for example Gray and Biber, 2012.

Heritage (2012b) summarises the work on epistemics and its role in driving sequences forward as follows:

> Underlying the conception … is a kind of "hydraulic" metaphor according to which any turn that formulates a K+/K− imbalance between participants will warrant the production of talk that redresses the imbalance. Here the proposal is that asserting something from a K+ position can be the basis for initiating or expanding a sequence, and that positioning oneself in a K− position can likewise motivate sequences or their expansion. (p. 49)

In relation to classroom discourse a number of studies have identified the role epistemic imbalances play, examples of which are summarised in the next section.

3.4.3.1 Epistemic Imbalances in Classroom Settings

In their investigation of epistemic search sequences Jakonen and Morton (2013) look at student-student interactions in bilingual Finnish Content and Language Integrated Learning (CLIL) classrooms, specifically student information requests, types of responses and the interaction patterns that unfold. Drawing on Heritage's distinction between epistemic stance and epistemic status, they explore how epistemic rights and responsibilities are managed and how the

epistemic status indexed by a turn does not necessarily match the epistemic status of the speaker (e.g., teacher questions). They find that the information request confers potential epistemic status to the recipient, who may or may not have access to the knowledge and can respond in a number of ways: ones that resolve the request with a "knowing" answer, ones that provide an "unknowing response" and those that then result in "contested knowing" (K+) responses. They found that unknowing (K−) responses and responses where knowledge is then contested led to further sequence expansions, noting students "rarely abandon an epistemic search" (p. 11). They also found that responses that resolve a knowledge gap may lead to chained sequences at a later point (as the K+ responder to the initial information request is then seen to be knowledgeable and worthy of further requests).

Sert (2013) explores epistemic status checks and how participants work to co-construct knowledge and reduce information imbalances. He finds, for example, that a K− status can motivate a sequence of elaboration. Sert and Walsh (2013) also study claims of insufficient knowledge and how these are managed in two English as an additional language classrooms in Luxembourg. They adopt a multi-modal conversation-analytic approach, specifically identifying how the teacher manages these interactions, e.g., by utilising deictic gestures, embodied vocabulary explanations, code-switching and designedly incomplete utterances. Their study found few instances where such claims were accounted for (i.e. explained) by students and that the teacher moved on to another student. They also show how teachers guide students in building knowledge, for example by eliciting, questioning and prompting.

These studies suggest different sequencing patterns and actions depending on the epistemic stance and status of the speaker and recipient(s) and their access to knowledge (e.g., extended sequences may be required in the search for a K+ response). However, they have primarily focussed on the teacher's interactional competence and considerations of student interaction patterns, and have been carried out in school contexts.

Epistemics as an area is relevant in the study of the PBL tutorials as dealing with knowledge gaps and imbalances are a core element of PBL tutorials. That is, by

working together the aim is to develop knowledge and shared understanding through problem-solving in the pursuit of learning. How these are dealt with in turns-at-talk will be considered in the analysis.

3.4.4 Summary of Classroom Applications

The three areas above all show how institutional contexts can be realised in action and are interconnected as follows. The types of talk research shows how sequences of actions can be aligned to pedagogic and institutional goals (e.g., to inform, to collaborate, to organise). The specific work on IR (F/E) indicates how control can be exercised and how exchanges may be sequenced, and how certain patterns may become more prominent depending on the goal and focus of parts of a lesson (e.g., teacher checks of knowledge). Finally, studies of epistemics provide an indication of how knowledge imbalances may be dealt with in the pursuit of the learning goals and how sequences may develop. These areas will be drawn upon where relevant in the study of the PBL interactions.

As we have seen, much of the research has been on school learning contexts, and the majority of the work has concentrated on teacher interactional competence. The work on the PBL corpus will expand previous studies by looking for collections of talk across the whole PBL cycle. It also aims to bring the student interactions to the fore, particularly with relevance to types of talk in relation to the aims of the PBLs, which as we have seen is a particular type of student-centred and managed learning event in a specific discipline.

3.5 Applying Conversation Analysis to the Corpus

A CA-informed approach is employed with the aims of describing the overall structural organisation and the unfolding and development of PBL turns-at-talk. These aspects are then discussed in relation to the overall pedagogical aims of the PBL of developing both problem-solving skills and the development of subject knowledge.

The initial transcription process in the preparation of the corpus necessarily required careful and repeated listening to the audio recordings and reference to the DVD recordings for additional clarification. Once a full transcript was

available, an initial reading was carried out and the overall organisational structure of each part of the PBL cycle was identified. In the close reading and study of transcripts, “episodes” indicating distinct patterns of talk emerged. Further examples of similar sequences were compiled to make “collections”, which were then analysed (Heritage, 2011; Markee, 2007).

To follow is a description of these two aspects. First, the overarching organisational structure (the main phases) is presented along with a brief summary of the aims of each phase and prominent *types of talk*. This section is followed by a detailed analysis of extracts representing the six main types of talk that I have identified. For the detailed discussion and presentation of extracts of types of talk, the transcription system used in other parts of the thesis has been slightly modified to reflect more closely the transcription practices found in CA approaches. This involves adding in additional pause information and explicitly showing both latching (the next turn commencing immediately with no perceptible gap) and words with emphatic stress underlined if found; additionally, the layout and line numbering was amended. Where additional transcriber notes are relevant, these are added in double brackets ((*notes*)), and referred to in the commentary. While this is a modified and ‘slimmed-down’ version, as ten Have (2007, p. 31) says “transcripts are unavoidably incomplete”. It is however, I believe, sufficient to provide a basis for analysis of the types of organisational patterns found.

3.6 Overall Organisational Structure: Phases in PBLs

Although CA studies typically focus on specific episodes or phases of talk (speech-exchange systems) within a communicative event or type of institutional talk, it is recognised particularly in institutional discourse that an overall organisational structure or internal shape to the whole interaction may be evident (Heritage, 2013b). Zimmerman (1984, 1992), for example, identifies the phase structure in 911 emergency calls, and Limberg (2010) does so in the organisation of academic office hour consultations.

As we have noted, the nature of the institutional talk and the purpose of specific phases may constrain the types of actions and turn allocations of participants.

Studying the ways in which they unfold in the context of specific phases and the aims of the "overall interactional project" (Seedhouse, 2004) can help in developing an overall picture of how students 'do' PBL.

Although CA eschews the notion of a priori categorisation, that PBLs have an overall organisational structure has been acknowledged by a number of studies, including specific CA-orientated studies, for example by Glenn et al. (1999) and Koschmann et al. (1997, 2000). Notwithstanding the concerns of assigning a priori categories, some awareness of context and aims should be recognised: students are presented with information and expectations on the running of PBL sessions at the beginning of the academic year and so come to the PBL with certain expectations; these guidelines are also known to the researcher (see 'The Steps of the PBL Process' in Appendix 1). While they may or may not adhere to the exact stages, the aims of PBLs are clearly set out and the cycle repeated each week.

The expectations that participants have all been made aware of are as follows. To recap from Chapter 2, in the first meeting of the PBL cycle the students are expected to concentrate on understanding a specific scenario presented via a handout and then to identify learning objectives for self-directed independent study before the second part of the cycle. Within the first stage students are expected to identify any unknown terms, both scientific and language related, and where possible explain these to each other. This should be followed by an identification of the main issues and then a discussion or brainstorming phase where group members share what they know and from this develop a set of questions or learning objectives (LOs) to be addressed in independent study (planning their learning). In order to ensure full group participation, learning objectives are not assigned until the end of the discussion. Each group has a Chair, who should lead the group through the stages, and a Scribe, who keeps a record of issues, questions/LOs and who they have been allocated to on the board. The role of the tutor is to facilitate but not lead the PBL session. The PBL guidelines indicate that the tutor's role is primarily to ensure the students are on the right track, rather than to provide information. In PBL2s students take turns in reporting back from their independent study. This information should be discussed in relation to the scenario.

The following sections present the overall organisational structure of first PBL1s and then PBL2 tutorials. The phases are described in relation to how they are identifiable, their pedagogic aims, and how they are characterised in terms of turns-at-talk. Examples of sequences that are not included in the analysis of specific types of talk are provided here, although the bulk of the analysis of patterns comes in Section 3.7. Although there was some preliminary and final social chat in some but not all PBLs, this typically included many overlaps and concurrent conversations and could not all be captured. Because of this, these sections are not referred to in any detail in the discussion of overall organisational structure.

3.6.1 Phases in PBL1s

While the students fulfilled the aims of the first part of the PBL cycle covering all points set out in the PBL guidelines (see Appendix 1), at some points in the tutorial the stages were not always quite as clear cut as the guidelines suggest. However, the following specifically delineated phases were identified:

- (opening chat)
- introduction to the Scenario
- (explanation of unknown terms)
- definition of problem and discussion of main issues
- identification of learning objectives
- allocation of learning objectives
- (closing social chat).

The parentheses ( ) indicate phases that were not always evident as a distinct phase and (apart from the opening social chat) were combined into the preceding bracketed phase. A brief description of the phases is provided below.

3.6.1.1 Introduction to the Scenario

Although in one scenario there was a short phase of some preliminary chat, this was kept quite short.[11] Longer exchanges about social matters occurred after the

[11] About an eyelash in one participant's eye.

PBL work had come to a close. The students have a full programme, time is limited and they start concentrating on the scenario straightaway.

All PBL1s included an introduction to the scenario, typically dominated by organisational issues. This phase and specific tasks are specifically signposted. Although this may include an informal comment on the work to be done (e.g., S1: *it's a long one*, PBL7.1), more typically the initial exchanges involved issues concerning the management of the PBL process, for example relating to distribution of handouts, assigning the Chair and Scribe if not already assigned and signalling time to read the scenario. (This reading is usually done silently but on one occasion the scenario was read aloud.) In three of the five PBL1s the signal to start work and the allocation and checking of roles was initiated by the tutor as in example (1) to follow, indicating their higher status and their right to direct the class.

> (1) S24T3: can we have volunteers scribe and chair or will i just choose you? (PBL10.1)

Further examples of *organisational talk* provided by both teachers and students occurring in the PBL cycle are discussed in section 3.7.1 below. Once the scenario has been read and the roles established, the students move on to discuss unknown terms and the main issues.

3.6.1.2 Unknown Terms

In four of the five tutorials, the opportunity to identify and explain unknown terminology was explicitly signalled by the Chair and in this next example, a specific participant, S25, nominated:

> (2) S28: thank you very much, er S25 is there any terminology problem? (PBL11.1)

When this is a distinct, delineated phase, students deal with explicit knowledge gaps in a number of ways: this may include a listing of unknown terms with no follow-up, brief clarifications through *simple knowledge exchange* sequences, or occasionally a longer discussion of a term (e.g., *lip pits* in PBL12.1). These have

been identified as a type of talk within the group I have called *dealing with knowledge gaps* and are explained in more detail in 3.7.2 below.

In two of the PBLs the clarification of terminology stage merged into the identification and discussion of main issues as students explored and shared what they know.

3.6.1.3 Definition of Problem and Discussion of Main Issues

Although in four of the PBL1s the students do start trying to separate out the explanation of terms and the definition and discussion of issues phases (in one they go straight into the issue), in two of the PBL1s the discussion of unknown terms and of the definition of the problem and key issues merge together: there is no distinct signalling of a new phase and explanations of terminology and the definition of the problem and discussion merge together.

The phase concerning the discussion of the problem is by far the most complex section of talk. This stage is also at the core of the PBL approach both in terms of developing skills and content knowledge collaboratively. The exchanges in this phase are typically exploratory in nature, with the discussion of one issue extending over many turns and frequently involving multiple responses to an initial statement or request for information. The examples below show sequences of exchanges signalled by the introduction of specific terms (3) and sometimes by a direct signal to move on (4):

> (3) S36: we we should, er consider about the *abnormal, chromosomal abnormalities*. (PBL 12.1)

> (4) S1: *right so shall we look at something else* cos we're getting bogged down. (PBL 7.1)

A fuller example of this type of *problem-solving exploratory talk* is included in 3.7.3.

Although in the PBLs in this corpus the facilitators take a backseat for much of the time allowing the students to lead and manage the discussion, they do

intervene on occasion. It is within the discussion of main issues phase that the facilitator is likely to become actively involved by prompting and probing for more information. This type of speech-exchange system, *instructional talk*, is also elaborated upon in Section 3.7.4.

3.6.1.4 Identification of Learning Objectives

When the main issues have been discussed and as the session nears an end, the students check they have set out the specific questions or learning objectives (LOs) to be researched in subsequent independent study. This section is usually signalled (e.g., line 1 in Example 5 below) and primarily consists of students suggesting an area, followed by confirmations (see line 6 to 10 below). The focus is usually on identifying a sufficient number of questions (one per participant) and clarifying if and when topics might be split up (e.g., lines 6 to 16). At this point, students are getting close to being allocated a specific question and so it is important that they are clear on the focus and scope. Moving on to the next point (line 19) signals consensus that the previous point has been dealt with.

(5)

01	S18:	i think you need to go on to questions now
02	S10:	yeah
03	S14:	ok
04	S15:	yeah
05		(9)
06		maybe (,) what's neural tube defects and (,) its types?
07		(5)
08	S8:	should we focus on Spina?
09	S13:	mhum
10	S14:	mhum yeah
11	S8:	so what's Spina
12		(12) ((scribe writing))
13	S12:	shall we say neural tube defects?
14		Spina Bifida
15	S14:	say eh
16	S8:	no make it Spina
17	S10:	yeah
18		(8)
19	S14:	and types (PBL9.1)

This type of exchange continues until there are a sufficient number of learning objectives, i.e. one for each participant the group.

3.6.1.5 Allocating Learning Objectives

The PBL guidelines used in this programme require the independent learning objectives to be assigned at the end of the PBL tutorial, with each participant taking one LO. In this stage students for the most part self-select. This is again quite straightforward and is characterised by a series of individual requests, all responding to the preceding turns (i.e. students are cognisant of which LOs have been taken as the allocation of LOs is noted on the board by the scribe). This self-selection is illustrated in Example 6. Prior to this extract, the students have agreed on the number and focus of questions. No overt signal is provided to start selecting but it is clear from the list on the board that no more LOs are required and that that phase of the discussion is over.

(6)

01	S4:	i'll have a look at developmental delay
02	S7:	so does that mean i'll get dysmorphia
03	S5:	dysmorphia
04	S6:	i'll do the FISH (PBL7.1)

In two tutorials the end of the PBL session is specifically signalled (7):

> (7) S9T1: well that's not bad at all for a complex scenario that's good. i'm going to look at the snow. (PBL7.1)

On other occasions there is an implicit understanding that the PBL work has been completed.

3.6.1.6 Closing Chat

Although the students frequently leave to go to other classes or because the room needs to be vacated, there is some *closing social chat* as students leave the room (e.g., about parents visiting). As much of this was concurrent talk, it was not possible to capture and transcribe clearly.

3.6.2 Phases in PBL2s

Identifying the overall organisational structure in the PBL2s was more straightforward, with the stages much more clearly delineated. In the second part of the cycle, the Chair is the one who generally directs the proceedings, with each student taking it in turns to report their findings from the independent study of their specific learning objective. The report takes the form of a short presentation (around eight minutes on average) and is accompanied by a written summary handout which may include references and visual support (e.g., diagrams, tables, pictures); depending on the room available students may make use of PowerPoint. After each presentation, members of the group should ask the speaker for clarification of any points that remain unclear and discuss any problematic issues. According to the PBL guidelines for this context (see Appendix 1), and indeed the Barrows model (1994), there should also be discussion of the scenario in light of the information presented, although this was not found as a distinct phase in any of the PBLs in this corpus and not often within the reports.

Of the five PBL2s in the corpus, four are in small group format (with usually eight students and a facilitator). The final PBL2 tutorial in the corpus (PBL12.2), the 'mini-big' PBL which is practice for the assessed PBLs in the following semester, takes the format of plenary presentations with all students from the programme and all facilitators present. Two or three representatives from each PBL group select a section of their research to present and give talks lasting around ten minutes, followed by a short question and answer section; all groups have researched the same scenario. For the 'mini-big' PBL, it is expected that all presentations will be supported by PowerPoint slides and for this longer PBL each group also submits an extensive, detailed group report. The main phases identified in PBL2s are as follows:

- preliminary talk and orientation to scenario
- <student report>
- <follow-up discussion>
- closing talk

< > indicates a recursive phase

As with stage one of the PBL cycle, the phases can be mapped on to the PBL guidelines (Appendix 1) with varying degrees of success. It is of course easy to identify the "present new knowledge" phase, and the follow-up phase clearly provides an opportunity to debate, or at least raise and clarify issues. However, it is also immediately obvious that in comparison to the PBL guidelines in Appendix 1, certain aspects of stage 8 *("Combine old and new information and gain understanding of its application to the problem…the group should revisit the scenario"* and *"see how they would handle the case with the benefit of their new knowledge")* and all of stage 9 (*summarize what was learned, etc.*) of the guide are minimal (stage 8) or absent (stage 9). A fuller description of each of the phases in PBL2s is provided below.

3.6.2.1 Preliminary Talk and Orientation to Scenario

As the students enter the room and settle in, there may be a period of preliminary talk. This may be very general, for example about *crackers* (PBL7.2), *paper cuts* (PBL11.2) or course related work e.g., *feedback sheets* (PBL10.2), but as with PBL1s this phase is short.

Following any initial chitchat, the first phase is concerned with organisational issues for the PBL session, such as distributing handouts and organising the order of speakers. This focus on organisational issues is often signalled. In three of the five sessions, the role of Chair in leading the proceedings also becomes evident as they organise the order of speakers and topics and give a brief signal to start. For example:

> (8) S5: *ok so shall we start with em* the the what is dysmorphology.
> (PBL 7.2)

Students then launch in to their report.

3.6.2.2 Student Reports

The report on learning objectives is essentially a planned presentation based on independent study with limited exchanges with the audience. It is clearly orientated to the explanation of the topic under review, i.e. content knowledge,

and has been named *presentational talk (*Barnes, 2008). Barnes (2008, pp. 4-5) states that in presentational talk "the speaker's attention is primarily focussed on adjusting the language, content and manner to the needs of an audience" and is a "'final draft' for display and evaluation … providing expected answers". Similarly, in the PBL2s the aim is to present the findings from the learning objectives, thus contributing to the building of shared subject knowledge. The reports are long, usually uninterrupted turns-at-talk. The presentations confer on the presenter the right to take a long turn: even though hesitations and pauses occur, these are not usually regarded as turn-relevance places until the speaker signals the end. Interruptions are for most talks relatively infrequent (only in 14 out of 43 talks; six within one PBL). The beginnings and endings of the presentations are signalled, usually by the presenter themselves. A fuller discussion of the student reports (presentational talk) is provided in Sections 3.7.6 and 3.8.

3.6.2.3 Follow-up Discussion

The post-presentation follow-up discussion is similar to the stage one exchanges in terms of the potential for multiple speaker contributions and for extended series of turns. The questions relate directly to the presentation content and once a point is clarified, a new question can be proffered or the next speaker takes the floor. This is in contrast to stage 1, where topics merged. It includes primarily a mix of *clarification checks* and following some presentations *problem-solving exploratory* and *instructional talk*.

3.6.2.4 Closing Talk

Once all the presentations are completed and questions finished, there is a final closing phase. The final closing talk may be related to practicalities for the next PBL (e.g., at the end of PBL9.2 the tutor informs them of the room and opportunity to be videoed), or more general chat on future PBL topics, or general conversation (e.g., an exchange about sweets in PBL7.2). Concurrent conversations again made it difficult to capture all parts of the final closing talk.

3.6.3 Summary of Phases in PBL1s and PBL2s

This section has introduced the overall structural organisation of the PBL1 and PBL2 stages of the cycle. Overall the students were orientated to specific aspects of the PBL requirements and were very much 'on-task' throughout in terms of the overall aims, evidenced by the limited digressions into social talk.

Perhaps not surprisingly given the guidelines and the weekly PBL cycle, there is a general overall organisational structure which aids students in working towards the overall aims of problem-based learning. However, the number of distinct phases is not entirely fixed or distinct, particularly in PBL1s where the actual boundaries are often more fluid. This fluidity and fuzziness in boundaries was also found by Legg (2007) and Woodward–Kron and Remedios (2007), albeit employing a different, genre and Systemic Functional Linguistics (SFL) based methodology. In saying that, we have seen that a number of specific phases can be identified by overt signalling.

The second stage of the cycle is notably different from the first stage in respect to both its format and the length of the turns: phases are clearly marked and the reports are extended individual turns. However, the post-presentation follow-up discussion phase is similar to the phase one exchanges in terms of the potential for multiple speaker contribution and for extended turns-at-talk on one topic. The facilitator's role can also be more prominent in the second stage of the cycle. This is particularly noted in the 'mini-big' PBL12.2, where one facilitator organises the groups and leads and directs the post-presentation questions, taking on the role of Chair.

Students orient to specific tasks in different phases, each of which (i.e. both phases and tasks) contributes towards the overall aims of problem-based learning. These provide opportunities for both sharing and building knowledge and for effective organisation of the event. The different orientations in the phases also evidenced different types of talk. Although these were not necessarily confined to one stage only, different phases were characterised by a predominance of certain types. These specific types of talk show more clearly

how the phases are characterised and how the aims of PBL are realised. The next section describes the most prominent patterns.

3.7 Six Types of Talk in PBLs

Although the phases in the PBLs are not always clearly delineated, the close reading of the PBL transcripts also indicates different types of speech-exchange systems, which are characterised by distinct patterns of interaction and sequencing (Markee, 2007).[12] These are collectively described as 'types of talk' in this work. To investigate how these unfold and how they contribute to the dual aims of collaborative student-centred learning and of knowledge building, I will look at sequencing patterns, with a focus on how knowledge gaps are dealt with and how the roles and rights to speak in PBLs are managed and distributed.

Collections of six types of talk seen as most relevant to the achievement of PBL goals were identified. Within these, there were two main overarching orientations: those relating to the organisation and management of the PBL tutorial, identified as *organisational talk*, and those relating *to dealing with knowledge gaps*. This second orientation, *dealing with knowledge gaps*, has been separated into the following sub-categories:

- *simple information exchanges*: an information request followed by a response
- *clarification checks*: requests (from students or tutors) for clarification or elaboration of points introduced in discussions or presentations
- *instructional talk*: tutor-led talk, guiding students to learning points
- *problem-solving exploratory talk*: longer episodes where group members exchange information, question and explore an issue (drawing on Mercer, 1995, 2000)
- *presentational talk*: pre-prepared talks where the speaker reports back on information related to their individual learning objective (Barnes, 2008).

[12] The term "speech-exchange system" is, I have found, poorly delineated. Examples can range from openings in an emergency call (e.g. Zimmerman, 1992), to classroom talk generally or to specific segments of a lesson (e.g. Walsh et al., 2011).

In PBL2s, the aim of the student report, named here *presentational talk*, is to share content knowledge. In this way, it could be seen as a type of instructional talk; however, I have kept this as a separate category. In instructional talk the tutor has both higher epistemic stance and status and is checking and guiding student understanding. In the reports while the student has researched the area, they are still essentially presenting their findings for consideration by the group, and indeed by the tutor. These types of talk all contribute to the building of shared knowledge. Each type of talk is presented below.

The sub-categories are not necessarily mutually exclusive. For example, problem-solving exploratory talk may include a simple information exchange or clarification check (e.g., asking the previous speaker to provide more information on a point made), but each of the above types can also be identified as separate, independent sequences at various points in the transcripts.

It should be noted that this section primarily focusses on those parts of talk which are evidenced by a high level of interactive exchanges. Although presentational talk will be discussed in relation to openings, closings and interruptions, the content of the long turns in PBLs will also be considered from the point of view of academic functions in 3.8. A description of each of the types of talk is now presented.

3.7.1 Organisational Talk

Organisational talk here refers to sequences where participants are orientated to the procedures for carrying out the PBL tasks at hand. It is similar in purpose to the regulative register identified in Christie (2000, p. 184) (authoritative talk used by teachers to direct students in activities) and to an extent Walsh and O'Keeffe's (2010) procedural talk (information on procedural issues). What is different in these PBL tutorials from a traditional classroom environment is that the management of rights and responsibilities to turns is held not only or primarily by the tutor, but also by the students themselves. In this way, classroom management of the PBL session is a shared responsibility, highlighting the student-centred nature of the learning event. However, in the PBL tutorials students also take on specific roles of Scribe and Chair, status roles which confer

rights to initiate and direct the proceedings (Drew & Heritage, 1992). The examples below aim to illustrate how the PBL cycle is managed by both tutors and students.

3.7.1.1 Tutor-Led Organisational Talk

This section introduces extracts of tutor-led organisational talk. Prior to Example 9 below, students had settled in and started to read the scenario silently. The tutor looks around the group before starting the role allocation.

(9)

01	S31T4:	so who's chair this week?
02		*((laughter))*
03		(9)
04	S26:	silence
05		(3)
06	S28:	i can be ((puts hand up))
07		((tutor nods))
08		(5)
09	S31T4:	scribe?
10		(4)
11	S27:	i can do it but =
12	S25:	=you've done it yeah (2)
13	S26:	i can do it if people want me to
14		((looks around the room; other students nod; one claps))
15	S25:	Ok
		((Students then go on to talk about use of the white board)) (PBL11.1)

In line 1 the tutor starts the PBL tutorial off with a first pair part (FPP) asking who will take on the role of Chair. This is followed by laughter from the group and an extended silence of nine seconds at line 3 where participants look at each other or their papers and the tutor looks around the room. This laughter may be to mitigate the lack of a preferred response (Glenn & Holt, 2013), i.e. a student offering to be Chair. The lack of a verbal response, which would normatively be expected, is eventually noted explicitly at line 4, in another response, or another second pair part (SPP) when S26 says *silence*. While this complies with the normative expectation to provide a verbal SPP response, the

speaker does not themselves offer to take on a role and, in this sense, it is a dispreferred response (Schegloff, 2007, p. 69). A further pause of three seconds occurs before S28 eventually offers to take on a role, indicating this with a verbal response (line 6), a gesture (putting their hand up) and looking to the tutor. The action following this is in the form of a non-verbal nod by the tutor. The overall sequence here is essentially an initiating FPP question (line 1), followed by four 'responses' in lines 2 to 6 (including the actions of actual silence and laughter), only one of which is preferred, and followed finally with a non-verbal acknowledgement at line 7, i.e. an IRF sequence, found in traditional classrooms (Sinclair & Coulthard, 1975; McHoul, 1978; Mehan, 1979). In this extract, we see multiple ways to respond between the I and F/E. Turns-at-talk or actions in lines 2 to 6 are all related to not only the immediately preceding turn but all need to be seen in the context of line 1, when the sequence is initiated.

This IRF pattern is repeated when the tutor then starts a new sequence with a one word prompt for the role of Scribe, said with rising (i.e. questioning) intonation (line 9). Again, there is a notable silence before another student appears to self-nominate (line 11), while at the same time looking at the tutor. However, this comes with a qualification initiated by the disjunct *but*, signalling a reluctance of some form. This initiates another response when S25 quickly, with no discernible gap, suggests S27 should not need to take on this responsibility noting *you've done it yeah* (line 12), indicating the *but* of the previous speaker suggests a reluctance to take on the role and the shared understanding (by S25 at least) of why. This prompts S26 to offer and solicit support by use of gaze, looking around the room to other participants. The third place F/E move in the form of non-verbal agreement comes with others looking to her (i.e. to S26), one clapping, others nodding and one acknowledging with a verbal *ok* at line 15.

After line 9, the tutor relinquishes control as the students resolve the situation themselves. Thus, we see the students' understanding of the need to take on roles, albeit reluctantly at the outset, share them out over the course of the term and to proactively contribute. Students are expected to offer rather than be nominated. Through this we see students responding to previous turns and

also their recognition of the larger interactional project of the PBL element of the curriculum occurring over the academic year. In the second part of the Example, while we see an overall IRF pattern, instead of the teacher leading this it is a student who provides the final response at line 15 with the acknowledgement token *ok* (Beach, 1993).

Another aspect of organisational talk involves ensuring the group is 'on track', moving through the stages in a timely manner as they have only one hour to complete stage one of the cycle. This prompting can be carried out either by the tutor or by a student (typically the Chair), and often involves a longer initiating prompt and a shorter response, illustrated in Example 10.

(10)

01	S18T2:	i think you need to go on to questions now
02	S10:	yeah
03	S14:	Ok
04	S15:	yeah (PBL9.1)

After a discussion of issues and the content of each potential learning objective the tutor moves the group along in a simple statement-acknowledgement sequence. This is done merely with a declarative statement directed to the whole group, acknowledged with three agreement responses from students in lines 2 to 4. Students then move on.

Along with an awareness of timekeeping, this signalling also marks a transition in phases in the PBL. The tutor's control also occurs post-presentation, as seen in Example 11, where they initiate the proceedings with a question for follow-up questions (line 1). After one question is asked (lines 4 to 10) and just as S27 starts to answer (line 11), the tutor again intervenes. Here the tutor overtly signals the interruption (line 12) and goes on to give an instruction for the question to be repeated to aid hearing in the lecture theatre. We see here a short initiation-response pair, with the tutor leading the proceedings.

(11)

01	S38T6	right any questions for the group three?
02		(3)
03		((laughter))
04	S39:	i just want to know eh (,)
05		you know you were saying about the ethical issues
06		do you think it's (,)
07		do you think it's right
08		that somebody is going to terminate a pregnancy
09		just due to the cleft lip palate Van de Woude syndrome (.)
10		yes (.) yes (.)
11	S27:	for me
12	S38T6:	so- sorry to interrupt um (,) (PBL12.2)

3.7.1.2 Student-Led Organisational Talk

While it is more common for the tutor to take an organisational role at the beginning of the tutorial, students soon take over this role. In Example 12 below, although at the outset of this PBL the tutor gave the signal to start, in subsequent phases it is the Chair (S1) who directs the proceedings. In line 2, S1 signals the start with *ok* and uses the topic *dysmorphology* along with the speaker's name (S5) to indicate the start of the presentation. S5 confirms this with *ok* in the SPP at line 3 and then launches into their report. After the presentation and follow-up discussion (removed), the Chair checks that there are no further follow up questions, to which there is silence of four seconds. This is understood as a 'no' because S1 then indicates the next speaker by topic at line 6 with *the first techniques i have down is FISH*. S6 confirms this with a short *yep* in response and then S1 gives a further instruction to S6 to pass out handouts, made less direct by the question form. In the next example (12), we see a series of question and answer adjacency pairs.

(12)

01	S9T1:	right well who's first?
02	S1:	ok so shall we start with em the what is dysmorphology S5?
03	S5:	ok
		((presentation and questions removed))
04	S1:	anymore questions on developmental delay?

05 (4)
06 i think the first techniques i have down is FISH,
07 S6: yep
08 S1: do you want to pass out
09 S6: ok ((gives H/Os)) (PBL7.2)

In the collections of this type of talk, the prompts to start and to move proceedings along were found to be in the form of questions, or hedged suggestions (*i think you*) and with short responses. Both of these downplay the directness of the instruction. Instructions followed by a short or minimal response were also found in Walsh and O'Keeffe's (2010) procedural talk.

3.7.2 Simple Knowledge Exchanges

In *simple knowledge exchanges*, an information request is made indicating the requester's lower (K−) epistemic stance. Examples (13) and (14) below show how other participants respond with a knowing (K+) response.

In Example 13 to follow, before this exchange all the students have had time to read the scenario and the Scribe has drawn a visual representation of the pedigree on the board. The Chair (S1) opens the phase at line 1 with a FPP question about any unknown terms, asserting their right to direct the discussion in the role of Chair. The question is directed to the whole group, opening up the floor to all to make requests for information (i.e. to self-identify knowledge gaps) and to any participant who can, to define. Participants self-nominate to respond.

(13)

01 S1: are there any (,) words and phrases anyone (,)
02 doesn't understand wants to define? (2) ((looking round))
03 S3: what does it mean hemizygosity? ((looks at S1))
04 (4)
05 S4: i think it's when there's only like one copy (2)
06 ((S7 also nods in agreement at the same time))
07 S3: isn't it homozygosity? (1)
08 S4: i think its cos it's only one there's only one copy or something
09 (11)
10 S1: are there any other?

11	S5:	C (,) GH
12		(3)
13		yeah
14	S3:	is it not a techniques that's (,) [=com
15	S7:	[yeah
16	S4:	=comparative geno=
17	S7:	= genomic hybridisation (PBL7.1)

After the opening question in line 1, there is a brief silence followed by a SPP question in line 3, when S3 makes an information request for the meaning of *hemizygosity*. In fact, this SPP becomes the FPP for the following response. S4 looks first at the Chair who instead of answering directly looks around the room allowing another participant to speak. S4 provides a response, downplayed with an *i think* (line 5). S7 agrees with this with a nonverbal nod. However, S3 queries the verbal response with a similar term (*homozygosity*). S4 repeats the information (*one copy*), again downplayed with *i think* and also adds *or something* at the end and, despite this tentativeness, this appears to be taken as the term. The discussion then moves on to a new question at line 10 with no further acknowledgement or action by S3. (Sometime later in the PBL, however, S3 makes use of this definition and at that stage there is evidence that it has been accepted.)

Here we have an initial question (line 1) which serves as an instruction, shaping the following sequence of question and answer in response, i.e. the response in the form of a question opens a new FPP and question and answer sequence (lines 3 to 8). This series is amongst students and is not always followed by a confirmation nor is an acceptance or change of state token (such as *ok* or *oh*) provided. The lack of evaluation may be because the questioner has not positioned themselves as K+, and so is also not in a position to offer an evaluation on the veracity of the response.

This pattern is repeated in lines 10 to 17 in a second sequence of an initiatory question followed by a response by S5 at lines 11 to 13 with *C, GH*, the speaker themselves confirming the abbreviation at line 13 with *yeah*. This time S3 makes a tentative suggestion, downplayed in the form of a question (*is it not a techniques*). This is confirmed by S7 at line 15, immediately followed by S4, who

starts to explain the abbreviation, finished off by S7 at line 17. Here three participants contribute to fulfilling the information request. This time we have a question and answer with the SPP answer started by S3 (line 14), confirmed and developed by S7, with S4 also confirming and overlapping with S7, but again with no acknowledgement or confirmation by the requester (S5).

Essentially, we see a pattern of an initiatory question which opens the floor to subsequent question and answer sequences, with a question or term given (e.g., *C, GH*) as an information request followed by a short response. In both these extracts the responder downplays their K+ epistemic status with a hedge, *i think*, or indirect response, *is it not a*. The sequences are quite short and no further discussion of the terms occurs here. This occurs primarily when the phase of identifying unknown terms is signalled and kept as a distinct stage. The students appear clearly orientated to the aim of explaining when they know the meaning. In a number of extracts such as these, there is no further discussion of these terms and the tutorial then moves on. As such they are 'standalone' examples of information requests. However, this type of information exchange may also start off longer stretches of problem-solving exploratory talk. This is discussed in 3.7.3.

In one PBL, clarification of a linguistic item was requested[13]:

(14)

01	S33:	what about assured? ((looking down at papers))
02	S34:	assured? like if i keep telling you it's going to be ok
03		i've given you assurance
04	S33:	ah assurance ok (PBL9.1)

Prior to this extract, students have either been listing unknown terms or explaining ones they know to each other. In Example 14, S33 asks specifically about the meaning of *assured*. S34 responds looking directly to S33 with a confirmation check question by repeating *assured* with a rising intonation but immediately provides an explanation in lines 2-3. S33 acknowledges understanding, or rather makes a claim to understand (Koole, 2010) with a

[13] An allowance for this type of specific language related request is built into the guidelines and directed to students whose first language is not English (see Appendix 1, point 1).

confirmation acknowledgement *ah,* which I would suggest here is similar to 'oh', operating as a change of state token (Heritage, 1984, p. 299), a repeat of *assurance* and the confirmation token *ok*. Here we see an IRF pattern, in the form of a 'genuine' question but rather than a teacher asking, we see a student information request, a student response, and a student acknowledgement of change of state (i.e., the third move acknowledges).

In Example 15 below from PBL11.1, we see how a known and an unknown term might be dealt with.

(15)

01	S28:	thank you very much (,) er S25,
02		is there any terminology problem?
03	S7:	FAP [?]
04	S29:	yeah MAP
05	S25:	MAP i've never heard of.
06		FAP (,) i think, Familial ((looks to S5))
07	S5:	Familial Adenomatous Polyposis
08	S25:	mm ((laughs) yeah
09	S7:	ok (PBL11.1)

S28, the Chair, thanks S25, who has just read out the scenario and then asks the group for unknown terms (line 2). FAP and MAP are both called out (lines 3 to 4). The unknown term (MAP) is merely listed, with one student making it clear they do not know the term (line 5) and no one else offers a definition. Most likely due to the specific purpose of this stage, but unlike the findings of Jakonen and Morton (2013), students do not continue exploring the term. In line 6, S25 in a tentative tone starts to offer the full term for FAP, which is completed by S5. S25 then acknowledges this with *mm*, laughter and *yeah* acknowledging the help, as does S7 with *ok*. S5 later goes on to provide more terminology in the PBL, who now seems to be regarded as K+ in this area (see Jakonen and Morton, 2013 for a similar pattern).

This phase of checking terms and language, which is intended to be a point where a 'knower' can share information, is typically quite short when kept as a distinct section. There is either an exchange of information, or a listing of unknown terms. When a participant can answer, they do; typically, when the

sequence of questions and answers is complete, the students proceed on to a new term. Not all information requests and responses are finished off by an evaluation or acknowledgement.

However, the fuller discussion of terms typically occurs in the discussion of main issues or when the unknown terms phase merges into the "discussion of issues" and becomes part of *problem-solving exploratory talk*.

3.7.3 Problem-Solving Exploratory Talk

Problem-solving exploratory talk exemplifies how the sharing and displays of knowledge unfold in usually complex sequences of turns. The term draws on Mercer's (1995, 2000) exploratory talk, which involves students identifying and speculating on specific content issues, and sharing knowledge and reasons in the joint pursuit of the learning goals. In this way, the participants obviously also co-construct the interaction. As Mercer also states, the three categories he refers to, which also include cumulative and disputational talk (talk where claims are either agreed or disputed uncritically), are idealised and unlikely to be found "in their pure form" (p. 102). Problem-solving exploratory talk might also include listing (similar to Mercer's cumulative talk) and agreeing on practical matters (e.g., number and specification of questions). For this reason, the term *problem-solving exploratory talk* is preferred here as this covers a slightly wider range of points than Mercer.

While students may request and provide knowledge in simple information exchanges in the orientation stage, the exploration of terms frequently occurs within a longer series of turns where students work towards identifying issues for independent study; it also occurs at some points in the follow-up discussion to student reports in PBL2s. This type of talk, i.e. the collaborative, learner-centred development of skills and content knowledge, is central to PBL pedagogy. This is done through a process of questioning, offering ideas, and requesting and providing further clarification or elaboration, thus working towards an intersubjective, shared understanding. The following example (16) shows how students collaborate to build knowledge and illustrates how the talk unfolds.

Prior to this extract, students have identified the need to find out about inherited causes of cancer and related genes. The excerpt and more detailed discussion of testing discussed below shows how the students offer suggestions, politely challenge and provide reasons for this and share subject knowledge. Multiple contributors add to the discussion.

(16)

01	S27:	and i think the problem is that (,)
02		no one has been tested before so (,)
03		they should test someone who is alive and have the (,) cancer
04	S25:	mhum ((agreement))
05	S27:	not [? both]
06	S25:	like P yeah
07	S5:	P
08	S29:	i think F's mother must be tested first (.)
09	S26:	the mother?
10	S27:	the mother will
11	S26:	the mother's got cancer
12	S27:	the [mother's alive
13	S29:	[she's alive
14	S25:	yeah
15	S27:	yeah but she's from the other side (.)
16		the [mother has
17	S26:	[she might not have any [?] (.)
18		it's not i don't think it's a recessive thing (,)
19		well we need to look in
20	S27:	[?]
21	S29:	yeah [we
22	S28:	[no, maybe no
23	S5:	i think usually the cancer (,)
24		it is recessive as a gene but [he's] here it's dominant (.)
		(PBL11.1)

At line 1, S27 identifies a problem that no one has been tested and although there are short pauses at the end of lines 1 and 2, no one takes the opportunity to interject, thus allowing S27 to continue and suggest a solution in line 3. S25 responds with agreement. S27 then starts a not fully audible response but is anyway superseded by S25 who agrees with and builds on the point by suggesting a possible candidate for testing, *like P yeah* (line 6), as does S5, thus both showing agreement with S27. However, S29 then suggests another person, *F*, at

line 8 (downplaying this with *i think),* indirectly contesting S25's earlier suggestion of *P*. This then provokes a series of turns contesting and explaining reasons. S26 questions the choice of *F* with a questioning *the mother*?, as does S27 both starting turns (lines 9 and 10) but with S26 taking over the floor to add a reason (*cancer*) at line 11. This follows with overlapping contributions from S27 and S29 (*she's alive*). From the reason provided later by S27 at line 15, we see that the S27 contribution at line 12 is actually agreeing with S26's questioning of the choice of mother, also demonstrated in *yeah* spoken in an uncertain tone by S25 at line 14. S27 provides a fuller explanation at line 15 (*from the other side*), adding to the case for not taking up S29's suggestion of *F,* a family member mentioned in line 8 (here the students are also referring to the pedigree, a map of the family tree, on the board and referring to the other side of the family). S26 goes on to add a further reason (*don't think it's a recessive thing,* line 18*).* S26 does show some concession to S29 suggesting further research *(well we need to look in,* line 19*),* in an attempt to mitigate the disaffiliative impact (Lindström & Sorjonen, 2013, p. 351). Although the turns at lines 21 to 22 are not fully audible, we can see that there appears to be both agreement *(yeah,* line 21*)* and disagreement *(no,* line 22*),* followed by S5 modifying the subject knowledge in lines 23 to 24. By referring to both recessive and dominant genes, S5 acknowledges S26's earlier point (from lines 17 to 18) but adds information to support the case. This continues for 25 more turns, with students continuing to add ideas, confirm and question issues related to who to test, before they eventually move on to other causes of cancer.

This problem-solving exploratory talk is characterised by a long series of interrelated turns-at-talk. It is not necessarily a chained question and answer series of sequences (Schegloff, 2007, p. 207) but as we can see can include a question, participants agreeing or disagreeing and adding ideas in response in the form of confirmation or contradictory suggestions, and further points to elaborate earlier responses. Each turn-at-talk shapes the following contribution and helps to build up what is essentially a response to the first suggestion. This is the centre of this sequence of turns, provoking the most discussion.

We see multiple contributors and contributions as students share knowledge and provide reasons for their decisions and information. Students appear to need to

build (at least some) group consensus to move forward. The opening suggestion in lines 1 to 3 can lead to multiple responses (lines 4 to 6) and which in turn lead to further responses, for example in line 7, S5 confirms the response of the preceding speaker with P, another possible candidate for testing. Thus, the FPP and SPP are not clear cut or distinct as they become intertwined with new FPP and SPP sequences. The role of knower (K+) is distributed amongst participants and shifts as the dialogue develops.

The long exchanges of inter-related initiations and responses indicate the complexity of this type of talk and demonstrate indeed how the knowledge imbalances (e.g., in the form of knowledge about genes here) drive the sequences forward (Heritage, 2012b). In this study, this also seems to indicate the powerfulness of this type of talk for engaging students and allowing for the development of reasoning and the sharing and building of subject knowledge.

3.7.4 Instructional Talk

While not the most prominent or common type of interaction, we do see instances in five PBLs where the tutor specifically intervenes and asks questions to elicit and scaffold student understanding. (This seems dependent on the individual tutor.) This type of talk is named here *instructional talk*. This speech-exchange system is most likely to occur in the discussion phase, although it can also be found in the post-presentation follow-up discussion phase in PBL2s.

The series of turns are tutor-initiated, either adding to or interrupting student discussions. Although the tutor typically uses questions, they have both a higher epistemic status and stance. This type of interaction is as Walsh and O'Keeffe (2010, p. 149) found "tightly controlled" by the tutor and in this way is more like a traditional classroom exchange.

In the following episode in Example 17, we see a tutor-led question and answer series. Before this intervention, the students were discussing deletions in the polymorphic and non-polymorphic regions and had started to identify areas they were unsure of (lines 1 to 11).

(17)

01	S1:	there’s a lot in it.
02		so i suppose this one (,)
03		it’s the polymorphic region's deletion
04		but the other non-polymorphic region is not deleted (.) (1)
05		so is that showing just?
06		(5)
07		is it a polymorphism if it's causing (1) features?
08		(3)
09		i don’t know ((scribe writes topic on board))
10		((group laughter))
11	S7:	i it's (,) i can’t make sense out of it=
12	S9T1:	= wha- what does it mean if (1)
13		if you’ve got a polymorphism
14		what deletion that makes that probe not bind?
15		what does that (,) mean what does that mean?
16		does it mean anything?
17	S2:	i don’t actually know what it means at all (.)
18		((laughs))
19	S6:	[(?)]
20	S1:	[it could just be a cha- a base change
21	S9T1:	[what what's a (? polymorphism)
22		it’s probably just a random (,)
23		change you reckon it doesn’t (,) matter
24		so then you could use a probe that’s (1)
25		not got a polymorphic region in it (2)
26		if you find a sequencing
27		what non-polymorphic probe does bind (,)
28		what does that tell you, about the child? (2)
29	S7:	[that there is no (??)]
30	S6:	[(?)]
31	S9T1:	uh huh (2)
32		kind of (.)
33		(4)
34		the if the non-polymorphic
35		if you've got a region chromosome two (,)
36		the polym- the probe that's [?] polymorphic region
37		doesn’t bind properly (2)
38		but you take another probe that’s (,) specific
39		that’s non-polymorphic and it binds (,)
40	S1:	are they binding the same area? ((tutor nods))
41		ok (,)
42		that’s what’s confused me
43		cos i’m thinking they’re like (1)

44 [neighbouring areas
45 S9T1: [yeah
46 S7: [(? have a ?)] (PBL7.1)

In lines 1 to 3, S1 identifies a specific focus (deletions) but also their inability to solve the issue, ending with *i don't know* (line 9). Following group laughter possibly to mitigate the trouble spot (Glenn & Holt, 2013, p. 16), S7 agrees with S1 saying *i it's (,) i can't make sense out of it* (line 11). Just as the student finishes, the tutor immediately intervenes with a latched response, displaying their rights and access to knowledge and higher epistemic status. This is in the form of a complex, structured FPP question in lines 12 to 16, building up gradually from *what does it mean* in line 12, after a brief pause adding and stressing *polymorphism* in line 13, and then asking about *probe binding* in line 14. In this way the tutor provides information and direction and so scaffolds learning, going on to repeat part of the initial question again in lines 15 to 16. Throughout the building up of the question, the teacher takes very brief pauses, which may be to emphasise the questions and also allow processing time for the students.

In line 17, S2 provides a SPP response but while conforming to the normative requirement to respond, this is essentially a non-answer or dispreferred response as they verbally claim no knowledge, followed by group laughter again. S1 then makes a tentative suggestion, hedged with *it could just be* (line 20). At the same time the tutor starts another question. This overlaps with the end of S1's suggestion (*base change)*. Although no direct evaluative acceptance is provided, the tutor builds on the idea of change that had been mentioned in line 20 initially having started with a question (*what what's a*, at line 21). In acknowledging the suggestion of a *change*, the tutor then goes on to break down the question, starting with *probe*, emphasising *<u>not</u> got a polymorphic region* (line 25) and then expanding by inputting information and asking two questions at line 26 to 28. S7 starts to tentatively offer an answer as does S6. Eventually the tutor acknowledges with *uh huh* and an evaluation hedged with *kind of* (lines 31 to 32). Following a brief pause during which no other contributions are forthcoming, the tutor then elaborates further, adding more information piece by piece in lines 34 to 39. Although line 39 appears incomplete (said with slightly rising intonation), the short pause (,) allows S1 to ask a question to confirm her

own understanding. The transcriber notes indicate a non-verbal agreement nod by the tutor and S1 then acknowledges understanding with a change of state receipt, *ok* (line 41) and then elaborates on the source of their confusion (lines 42 to 44). S9T1 then provides a positive evaluation, *yeah*, at line 45. This extract then continues with other students asking their own check questions to the tutor and continues with a series of questions and answers primarily between the students as they continue to check and clarify their understanding.

In this example, we see the tutor building up a question incorporating cues to direct student answers. If the answer is not satisfactory or incomplete in some way, then a further question is asked. The student response may be acknowledged fully as correct (*yes*) or partially (*kind of*), or used as a building block (*base change*). In this way, we see the classic $IR^{(n)}$ F/E format emerging. When an answer is incomplete but the students ask questions, the tutor clarifies following a simple FPP-SPP question and answer format.

The tutor-question format is quite long and complex, directing the students by building on what they know but at the same time breaking up the points into manageable chunks and so scaffolding learning (Walsh & O'Keeffe, 2010). We also see here a concern for ensuring accurate knowledge (Dippold, 2014).

3.7.5 Clarification Checks

In a *simple information exchange*, the questioner positions themselves as less knowledgeable (i.e. K−) and is seeking a K+ response from anyone who can help. A successful exchange will involve a K+ response. In *instructional talk* a tutor (K+) asks a question, builds on the answer with prompts and cues and finally confirms or provides an answer. The turns-at-talk indicate that the tutor has a specific intention or learning point in mind, has themselves a higher epistemic stance and status, and we assume and later see that the recipient either doesn't know or needs help in arriving at the answer. In contrast, although similar in the format (question, answer (evaluation)), *confirmation checks* are subtly different in that the starting point is to check the understanding of content the speaker has previously made, i.e. the recipient of the question is at least to some extent already regarded as K+. The questioner may be a student checking their own

understanding of points the recipient has introduced. They do this by either checking or building on their knowledge, as a repetition, reformulation or as an additional question. In the case of the tutor, they are checking the accuracy of information the speaker holds. In either case the information request comes from a speaker who also has at least some knowledge. The questions ask for *further* demonstration of knowledge (Koole, 2010). That is to say, the person asking the question is asking *for* information and checking either that their own information is correct (from the point of view of a student[14]) or that the speaker understands a point (from the point of view of the tutor or at times students). While both seek to redress knowledge imbalances, this checking is in contrast to instructional talk, which aims to guide students *to* unknown information. As we shall see in PBL2s the responses appear to involve fuller answers as students frequently have more to say.

Example 18 illustrates a tutor-led clarification check. This excerpt occurs in the follow-up discussion stage to a presentation in PBL12.2. One of the tutors asks the student to clarify the difference between *syndromic* and *nonsyndromic* cleft. In this example, the tutor (both more knowledgeable and with more rights to knowledge) asks confirmation check questions about a specific aspect that the recipient of the question has researched (i.e. and so who should at least to some degree be K+). Questions continue until the tutor receives a satisfactory response.

(18)

01	S38T6:	can (,) can i ask you ehm (,)
02		you spoke about syndromic cleft versus nonsyndromic cleft
03		what what's the difference?
04	S7:	syndromic clefting is where um (1)
05		it is it has a particular inheritance pattern (1)
06		in the family (.)
07		when in nonsyndromic occurs sporadic most of the times (1)
08	S38T6:	but it wouldn't be (1)
09		sporadic if there's a genetic cause (2)
10	S7:	uhm (,) a mu- a mutation in these loci could be sporadic
11		which causes nonsyndromic clefts (.)

[14] Students may in fact be very knowledgeable in an area, for example having expertise from previous studies, or having researched the area themselves.

12		(3)
13		but (2) ((S27 whispers *it's not inherited* behind S7))
14		um it's not inherited yeah
15	S38T6:	ok ehm (,) (2) right (.) [next question removed] (PBL12.2)

The tutor (S38T6) starts off with a pre-opening, signalling they will ask a question (*can i ask*). Given that the topic is immediately introduced in line 2 followed by the specific question in line 3, *what is the difference*?, this opener acts as a token of politeness. Having researched the area, S7 is in a supposed K+ position and gives their response (essentially that syndromic cleft is inherited, and nonsyndromic cleft is sporadic) interspersed with brief pauses (although not interrupted) in lines 4 to 7.

In lines 8 to 9, the tutor contests this answer indicated by the disjunct *but* and by contesting the student's point on *sporadic* and *inheritance* (i.e. that it can't be sporadic if it is genetic/inherited). This is responding to the immediately preceding point asserting their higher epistemic stance and knowledge of the subject area (nonsyndromic is not sporadic if it has a genetic cause). This is followed by a pause of 2 seconds. The student then starts to repeat the point (lines 10 to 11) about sporadic occurrence. The pause may indicate online planning or indicate uncertainty at this point; the latter appears to be the case when we then see S27 whisper behind S7 *it's not inherited*, which is then repeated by S7. This confirms the tutor's point and is then accepted with the *ok* acceptance token. The tutor then continues with another follow up question (not shown here). The tutor's concern here is with checking that the information provided by the student is correct. In order to satisfy the tutor, the student needs to demonstrate, rather than just claim, understanding of the content.

In summary, the recipient of the information request in these checks is regarded to some extent as knowledgeable and it is expected that they can provide extended responses. This is particularly noticeable in the PBL2s after a period of study. However, the requester is also in these examples building on some previous knowledge and could be said to be somewhere on the epistemic stance gradient (Heritage, 2012a, 2012b); epistemic stance is relative. When it is the tutor asking, they are likely to be regarded by all as having both higher epistemic stance and status (e.g., the tutor in Example 18). The requester also

assumes the right to continue questioning until satisfied or other students start to join in (observed in other examples of similar episodes). There are often chained question-answer sequences of turns and may include (but not always) a final evaluation. In this type of talk the responses are also frequently long demonstrations of knowledge.

Clarification checks then indicate at their starting point some shared access to "epistemic territory" (Heritage, 2012a, p. 4), indicated by repetition of questions and responses in their attempts to redress an imbalance. These checks again demonstrate the collaborative nature of the development of content knowledge.

The types of talk discussed so far have shown how PBLs are managed and how knowledge imbalances are dealt with through highly interactive turns-at-talk. The next section looks at the interactive aspects of the longer student report or presentation, another form of talk related to developing subject knowledge.

3.7.6 Presentational Talk

CA typically concentrates on the links between turns in highly interactive talk. However, Jefferson (1978), in a consideration of story-telling, also considers aspects of talk that relate the bulk of a story to preceding and subsequent talk, thus making the contributions sequentially relevant. For the reports in PBL2s, I shall first of all consider the ways in which the long turns are signalled and embarked upon, when interruptions do occur, and how the closings link to subsequent talk. A fuller discussion of the content of the presentations, i.e. what is talked about, is provided in Section 3.8. (N.B. *student report* and *presentation* are used interchangeably here.)

3.7.6.1 Organisational Structure of Student Reports

In addition to the overall structural organisation of each PBL cycle, it is also possible to identify the overall structural organisation of the student report phase itself.

The reports include the following components:

- signalling the start
- introducing the topic
- reporting on findings
- (signalling ending)
- (referring to references)
- (inviting questions)

() indicates these components are not always present.

The turns in these elements largely serve to orientate the audience to the subsequent phase and, as we shall see, when it is 'permissible' to take the floor. The report on findings is necessarily a long individual turn but there was no clearly discernible internal shape common to all reports; the internal structure seemed to depend on the specific learning objectives. In most cases, the talks are uninterrupted and continue until the speaker specifically signals the ending in some way.

3.7.6.2 Openings in Student Reports

The signal to start a report is provided with a prompt by the Chair or in some tutorials the start is preceded by a long pause indicating the end of questions for the previous speaker. The topic of the report is then introduced in some way by the speaker themselves. When the speaker signals that they will start, this may be with a one-word or two-word transition signal, followed by the topic. This is illustrated in Example 19 below. After a six-second pause following the previous speaker's questions, the next speaker, S10, looks around the room and checks that questions are completed (*can I start? no more questions?*, line 2). They then signal the start of their specific talk in line 3 with an *ok*. *Ok* here is used as a pivot signalling a "change of matter" (Beach, 1993, p. 327), projecting forward into the topic. This is followed by the coordinator *so* (Biber et al. 1999, p. 80), used to "launch a new action trajectory" (Bolden, 2009, p. 990), in this case Spina Bifida.

(19)

01		(6)
02	S10:	can i start? no more questions? ((looks around the room))
03		ok (,) so i did em (,) Spina Bifida (PBL9.2)

After signalling the topic the students very quickly launch into their talks. This is in stark contrast to the advice given in many EAP and presentation study materials for students which suggest much longer introductions including introducing oneself, providing a summary of key components, ordering points to be covered (see, for example, Anderson et al., 2004; Rignall & Furneaux, 1998). The brevity is perhaps because the focus of the talk in the PBLs is known (as all have the learning objectives), or because of the informality of the group presentations. It may also be because they only, for example, talk about one area and do not feel the need to 'signpost' so much initially. No guidance on structuring the talk is provided in the PBL steps, but discussions with facilitators indicated that they do not expect long introductions or endings. Only in the more formal PBL12.2 talks is there a more formal introduction, which may on occasion also include an overview to the talk.

3.7.6.3 Closings in Student Reports

Following the main part of the report, the closings are generally very brief and while usually signalled, as in Example 20 below, they can end quite abruptly. There is also rarely any concluding summary. The speakers may themselves then call for questions, or these may be called by the Chair, after which the floor is turned over directly to the students.

In Example 20, we see the students closing the topic with *so that's it*, followed by a direction to references listed on the handout (line 2) and then a call for questions. Rather than launching the topics, *so* here serves to indicate the talk has finished (Biber et al., 1999, p. 877).

(20)

01	S5:	so that's it (.)
02		and that's the papers if you want to check it (2)
		((referring to references on PPT slide))
03		any questions? (PBL11.2)

Again, the abrupt ending, even in the more formal presentations, are in stark contrast to advice on student presentations (Anderson et al., 2004; Rignall & Furneaux, 1998). It should be noted that at no point did facilitators comment upon this as a problem.

3.7.6.4 Reporting on Findings

After the opening, no clearly discernible shape was identified across all the reports. However, within the individual reports more discreet stretches of talk, clearly related to the learning objective of the students and so contributing to the aims of sharing specific aspects of knowledge, were found. These include cause and effect, describing procedures, processes and conditions (and elaborating to clarify). Although not exhaustive, the most common academic functions identified are considered in Section 3.8.

In addition to a very clear orientation to the content of the talk, there is also clear evidence of orientation to the audience, reflecting the dual aims of sharing information and working as a group. This orientation to the group is most overtly seen in the reference to visuals, for example, deictic reference to a picture, diagram or figures illustrated on handouts or PPT slides. This aspect is commented upon in Chapters 5 and 6, relating to pronouns and lexical bundles.

3.7.6.5 Interruptions in Student Reports

Although for the most part the right to talk by anyone other than the presenter is suspended until the end is signalled, there are examples (14 in 43 talks) where the talk is interrupted or more accurately other participants contribute in some way for a short period. (These interruptions to the flow occurred for the most part after at least some pause or hesitation.) Four types were found. One example each of the two most common types is provided below. The first was an

offer of help in some way, for example, either in response to a student having difficulty with technical terms or hesitating over a word. In Example 21 below, we see the speaker struggle with the pronunciation of a term (line 1). After a pause indicating the student can't self-repair, S18T2 then provides the correct pronunciation, which the student repeats before moving on.

(21)

01	S17:	er (,) also amnioce-
02		(3)
03	S18T2:	amniocentesis
04	S17:	amniocentesis (.) (PBL9.2)

The second was with affiliative laughter, relating to some shared knowledge or common ground or humour. In Example 22, as part of the explanation of *pinna anomalies* S5 makes a funny gesture (line 4), to which the group responds with laughter. After this temporal interruption, S5 then resumes their presentation with *so* (Bolden, 2009).

(22)

01	S5:	pinna anomalies ah
02		which is ah the part of the ear
03		that projects like a little wing from the head
04		it's like that ((gestures))
05		((group laughs))
06	S5:	so (,) (PBL7.2)

Although less common, two other sets of interruptions were observed. One relates to confirmatory responses, where a point (usually a word) is repeated (e.g., *S10: twice tenfold, S12: twice* from PBL9.2*)*. The second example relates to direct interruptions, the least common type. Very rarely a tutor interjects with a direct challenge to the speaker about the accuracy of the information.

After each type of interruption, the speaker quickly resumes the report and, in all but two reports, continues to the end uninterrupted. Two PBL tutorials stand out: PBL9.2, which included interruptions of some sort to six of the eight

speakers. Once the first report had seen three interruptions, this seemed to lead to more interactive opportunities in five of the other reports. There were also no interruptions in the formal PBL12.2 presentations.

In the examples of interruptions or breaks in talk, we can see then where they appear allowable and how they are sequentially relevant. These are generally two part exchanges (with a verbal or non-verbal response), with *so* often used in the following subsequent turn to indicate they are moving forward. The interruptions during the reports are mainly positive, either in affiliative laughter, assistance or confirming understanding. It is unusual at this stage in the PBL for a student to be critiqued or challenged in some way mid-talk.

3.7.7 Summary of Types of Talk

The presentation of types of talk has identified and explicated the main types of talk found across the PBL tutorials. In line with many studies in educational contexts (Christie, 2000; Koschmann et al., 1997, 2000; Legg, 2007) two main frameworks were identified: one relating to the organisational management of PBLs, and one relating to content and knowledge building.

Organisational talk illustrates how the tutorials are managed with organisational roles taken on by both tutors and students. It also shows that certain prescribed and identified roles for students confer more rights and responsibilities to manage the learning event. Thus, while the management is more equally shared than in many learning events, role allocation continues to provide some hierarchy. In both cases, teacher and student-led organisational talk was primarily characterised by longer cues and instructions and short responses, and by short sequences of turns.

Students engaged in a variety of types of talk for dealing with knowledge imbalances and working towards a shared understanding of subject content. These also illustrate how students can teach and scaffold each other in their learning. These were characterised by distinct sequencing patterns. The simplest and shortest are the *simple information exchanges,* found as a standalone series of turns when there is a distinct unknown terminology phase. Although these

exchanges can occur within other types of talk, they were most common as ‘standalone’ sequences of an initiation and responses and possibly a final feedback in the unknown terminology phase. In this type of exchange, the questioner firmly positions themselves as less knowledgeable and any group participant can respond. When a final feedback turn is provided, this indicates a claim to understanding, as opposed to a demonstration of knowledge (Koole, 2010).

In contrast, *instructional talk* is initiated by a tutor, who is positioned as knowledgeable, even though the series of exchanges is also question and answer. In this type of talk we are more likely to see an IRF/E pattern emerging. Instructional talk was characterised by a series of questions and often short responses by students as the tutor directed students towards specific content knowledge. Tutors frequently scaffolded learning by breaking up complex points into a series of individual, cumulative questions, with a final third position being used to evaluate or add and confirm information. This building up of points also models reasoning skills. In contrast to the more democratic and open student-led interactions, these exchanges were firmly in the control of the tutor, evidencing their higher epistemic stance and status. This type of questioning was also found by Walsh et al. (2011) in instructional talk.

Problem-solving exploratory talk included many longer, more complex sequences, involving multiple participants and contributions. As students identified shared understanding and knowledge gaps, this pushed the sequences forward until some agreed conclusion was reached (e.g., agreement on a learning objective). Although a long inter-connected sequence of turns, these were not typically a sequence of sequences (e.g., a repeated question and answer series, Schegloff 2007, p. 207) but could include a statement, a question, an idea added, a challenge made and so on. There was a mix of knowledgeable and less knowledgeable participants and again all could contribute. A new section of this type of talk emerged when a new topic or term is introduced or students explicitly decide and signal the need to move on.

Clarification checks were included within the problem-solving exploratory talk but were also found to stand alone, particularly after the PBL2 reports. In this

type of talk, participants indicated more shared knowledge at the outset and the exchanges were characterised by more repetition and reformulation of points and questions. The clarification checks appeared particularly rich both in the amount of questioning and length of answers as students appeared to be building on points of knowledge that had either been presented in the reports or which the questioner, from the content of the questions, appeared to display at least some degree of certainty on the epistemic gradient (Heritage, 2012a, 2012b). Although initially they appeared similar in organisation to instructional talk, i.e. a series of questions and answers or a sequence of sequences, the tutor questions in clarification checks were notably different from tutor instructional talk (e.g., the question-answer series were shorter in length. Here the questioner and recipient both demonstrated some degree of knowledge, displayed in the interaction).

The study of aspects of the presentational talk identified the overall organisational shape and also yielded insights into how the student reports are opened and closed and the nature of interruptions.

The various types of talk indicate how participants are orientated to the different task requirements or 'steps' in the PBL cycle, specifically how knowledge gaps are handled in both student-to-student and tutor-led segments, and how the PBL cycle overall is managed. Although types of turns (e.g., initiation, response, question and answer) are useful starting points, it is by looking closely at the function of turns as they unfold that we can more clearly identify the purpose of the various types of talk and how they relate to the overall pedagogical aims. As Seedhouse (2004) found, the pedagogic focus and task orientation affected the turn-taking patterns and vice versa. For example, the discussion of issues led to long, complex multifaceted sequences of turns, in contrast to the tightly controlled, question and answer series in episodes of instructional talk. Although Walsh et al. (2011) found also empathic and argumentative talk in the interactions between students and seminars, these were not evident as distinct varieties in the PBLs, perhaps because the pedagogic aims are more focussed on a range of different tasks here.

The investigation of overall organisational structure and patterns of interaction has shed light on how PBLs are realised in talk. However, as we have seen, the bulk of the student presentation in PBL2s is a long uninterrupted turn-at-talk. While we can see how each talk links to previous and subsequent turns by other participants at the openings, closings, and interruptions, this provides only limited insights into the content of the presentations. The next section takes a closer look at the student presentations by considering *what* is being talked about in relation to academic functions.

3.8 Academic Functions in Presentational Talk

The content of the student presentation was closely tied to the learning objectives and on close inspection it was possible to identify common functions used on a regular basis. These have been called here *academic functions*, focussing on the communicative function of these segments of talk in academic contexts (i.e., what the student is doing in the specific context; Coffin, et al., 2009).[15] This section aims to provide a brief introduction to previous studies in the area of academic functions before going on to outline the main types found in the student reports.

Previous research on academic rhetorical and communicative functions has concentrated mainly on academic writing (Bruce, 2009, 2015; Hoey, 2001; Swales, 1990). In relation to spoken academic English, consideration of organisational and rhetorical patterns and functions has concentrated primarily on lectures. One notable exception of student-centred discourse is the analysis by Tanguay (2015) of phases and moves in student seminars from the BASE corpus. Tanguay identifies "Discussion Macro Genres" (DMGs), stretches of talk with shared functions and staging, for example, "responding" and "problem-solving". The stages in the DMGs consist of moves, e.g., "elicit". While this is useful in identifying rhetorical and functional organisation, it does not extend to address what students are talking about (e.g., describing a process), nor to longer turns.

[15] This does not mean that these functions do not occur in other contexts; merely that they were particularly common in the PBL corpus and clearly related to the pedagogic aims.

Studies of academic lectures include Young's (1994) investigation of macro and micro phases. These include "discourse structuring", "theory or content", and "evaluation" as macro-phases, and "relational process" as an example of a micro-phase found in conclusions. Steps in lecture introductions have also been analysed (Thompson, 1994; Shamsudin & Ebrahimi, 2013). Thompson identifies, for example, "setting up lecture framework" as a function, with "announce topic" as a sub-function (p. 176). Neither of these approaches provides clear insights into the content of the main phase.

More illuminating in terms of specific functions in content phases are studies by Dalton-Puffer (2007) and Deroey and Taverniers (2011). Dalton-Puffer discusses taxonomies of academic language functions in relation to CLIL teacher-student interactions in school lessons, concentrating on "defining", "explaining" and "hypothesising". Deroey and Taverniers (2011) identify a range of functions in lectures from the BASE corpus. They find, for example, an overall "informing" function, which includes "reporting", "describing" and "demonstrating", and an "elaboration" function, which includes "exemplifying and reformulating". Although it also extends beyond subject-related content to include discourse organisation and interactions, these aspects will not be considered here. (Additional linguistic aspects related to interaction and discourse organisation will also be discussed in the chapters on pronouns and on lexical bundles.) Deroey and Taverniers' work presents a more comprehensive taxonomy, looking as it does at short stretches of talk from across lectures in a range of disciplines. Where relevant, terms from this study have been applied or modified. Relevant descriptions are provided below.

3.8.1 Identifying Academic Functions

The approach to identifying academic functions followed that for identifying the overall organisational structure and types of talk, i.e. a bottom-up exploratory reading of each of the presentations, during which academic functions were ascribed to stretches of talk. This was done by looking for overt lexical signalling and semantic links. The categories were refined and developed during this process and examples grouped into collections. Following the analysis of 43 separate reports over the five PBL2 tutorials, the most common academic

functions identified are described and exemplified below, illustrated with examples taken from across the transcripts. It should be noted that these are broad categories only and less frequent functions (e.g., hypothesising, evaluating) which only occurred in a small minority of student presentations have not been included here. This analysis is also concerned primarily with identifying the academic functions, rather than an analysis of linguistic items, although salient points of interest are commented upon. An overview of the main academic functions and the number of reports in which they occurred is provided in Table 3.1 below.

Table 3.1: Academic Functions in PBLs

Academic Function	**No. of reports**	**No. of examples**
Describing/explaining		
• Characteristics/functions	64	123
• Procedures	20	39
• Processes	20	27
• Incidence/likelihood	22	26
Cause & Effect		
• Causes	26	46
• Effects	27	20
Defining	20	24
Exemplification/Elaboration	24	66

The functions reflected the independent learning objectives (LO) assigned to each speaker, but each main LO report typically involved more than one function. For example, a LO might have been to identify causes and effects of a condition with the presentation then including a series of cause-effect chains (i.e. one cause and related effect(s), followed by another cause etc.) but also a description of a condition. Alternatively, a student may have researched characteristics related to a medical condition, e.g., Spina Bifida. The presentation may then include a definition and a description with further exemplification and elaboration of the characteristics. These series of functions were typically separated and linked by informal linking phrases, for example *second one is* and *and they*.

Although I will go on to illustrate each main function, as indicated above one academic function commonly led into another. Before we move on to discrete examples of each specific function, Example 23 to follow provides an illustration of how this linking is realised. In the example, the overall purpose of the longer stretch here is to describe sub-categories of a condition, along with causes and effects. From lines 2 to 3 we see the *cause* clearly signalled lexically (*it is caused by*), followed by a *description of characteristics* in lines 5 to 8, and a comment on general *incidence* at line 10. The next type (Spina Bifida Melancae) is signalled in lines 11 to 12, followed by a description of characteristics in lines 13 to 15, a comment on *procedures* for treatment, and finally an indication of *effects* (or lack of), again in this case overtly signalled (*don't have any, long term effects,* lines 18 to 19*)*.

(23)

01	S16:	so it's it's the most common out the three (,) and em (,)
02		there is er it is *it is caused by*
03		like a small gap between the two (,)
04		vertebrae(.) vertebrae?
05		er but (,) because they they are so small
06		and they do not have any symptom at all
07		like you *don't get any symptom* at all
08		so they're are not that (,) dangerous
09		ehm they also are like (2)
10		[? found] like like i said they're quite common.
11		er second one is called uhm (,) Spina Bifida Men
12		i i can't yeah Melancae?
13		er it's a this one is ehm (,) it's the the gaps a lot bigger (,)
14		and the membranes are pushed (,) outside (,)
15		but em(,) but the nervous is still in side (,)
16		and they do not ca- like (,)
17		but you can actually remove it surgically (,)
18		and normally *they don't have any* (,)
19		*long term effects* or anything(.) (PBL9.2)

To follow are specific examples of each of the main functions that have been identified.

3.8.2 Academic Functions in PBL2s: Describing

In the reports, students frequently provided descriptions of conditions, procedures, processes and incidence and risk.

The description section has been separated into four sub-groups reflecting the different types that were most commonly found. Combined, these represented by far the most common academic function in the reports. *Describing characteristics* and features reflects the need to provide an account of, for example, a condition or the purpose of a technique or procedure. *Descriptions of procedures* which involved human agency included explanations for how something is done, e.g., techniques and stages involved in testing or screening for conditions. In explaining how genetic factors worked, this commonly involved *describing natural processes*, i.e. without human intervention. The final group added to the description category was *describing incidence and risk* in a population. This includes level of risk, likelihood of a condition in a certain population and numbers relating to screening tests, and is perhaps the most specific to medical genetics as a discipline. (The descriptions included examples and elaborations, but these specific aspects will be presented in 3.8.4 below.) Examples of each sub-function now follow.

A key part of many of the reports is to *describe characteristics* of conditions and procedures, and explain reasons for use. As such, it is quite a broad category. These frequently included additional elaborations, which have not been removed here as they illustrate how students work towards clarifying terms. In the examples, descriptive parts are shown in italics.

In Example (24), following a definition for IRT, we can see the description specifically starting with a comment on level (*higher in babies*) and its purpose (*a marker for screening*). The procedure (removed here) was then explained and followed again by reiterating the reason for use (*high in babies with CF*).

> (24) S23: right so we can use IRT is an enzyme produced in the pancreas and *is generally higher in, er the babies, with CF, yeah? than those without it. so this used as a neonatal marker for screening* [procedure

> removed] the IRT test is because, er for the babies with CF the, *the level of IRT will remain high*. (PBL10.2)

In (25) the student is talking about the methods used for detecting Cystic Fibrosis (CF), the advantages and disadvantages of tests (ARMS and RFLP). In describing features of the tests, relational processes, signalled by the copula *is* and with some material verbs, e.g., *detect*, are common.

> (25) S21: erm, for the detection of known mutations er ARMS reverse dot blot or RFLP, can be done. and erm, the advantage and disadvantage of each of them, er first of all ARMS um is it um *is* appropriate for large series and can *detect* up to twenty mutations by multiplex. but um, the main disadvantage of ARMS *is* uh, difficult to design the primer. er reverse dot blot, also it*'s* uh useful for large series but eh, and also can *detect* up to twenty mutations, by multiplex. (PBL10.2)

Students need to identify and explain *procedures* that medical geneticists or clinicians use for testing and treating genetic conditions. In Example 26 the student explains the steps involved in CGH procedures. We see in this example a range of active, again material, verb processes (*extract, use, co-hybridise*).

> (26) S3: this was for the traditional, comparative genomic hybridisation, there are more er improved, er CGH procedures. so basically they, *extract* the genomic DNA from patients' peripheral blood lymphocytes or from skin fibroblasts or any other available tissue. and if you can see on page, two diagram, basically this is the [?] DNA that i was telling you about and the test of DNA. in DNA labelling they usually mainly, they mainly usually *use*, er either Cy three, dye or Cy five dye. to differentiate between each reference or DNA tested er, test DNA er sequences. which by *they co hyb- co hybridise them* the to a slide (PBL7.2)

Although the professionals are referred to as *they*, the example appears more informal and conversational in style and we see the speaker actively engaging with the audience.

The sub-category *describing processes* reflects the need to identify the natural processes that occur in the genetic system. As we can see in (27) below, the process is closely linked to the cause and effect function and indeed in the example here, the process starts off with *what causes MAP*, followed by a description of the changes that occur. The material processes here are signalled with the active verbs *leads* and *happen* (Halliday & Matthiessen, 2013).

> (27) S28: ok er what *causes* MAP? er a mutation in this gene, and the oxidation, in the er proof reading er process, er *which leads the change* from GC to TA. and this diagram here will explain little bit the, (2) idea of this inheritance here in the middle, in the central pathway, you can see in general the change from GC, to TA, *happen* because the oxidation *here change the G to O*. (PBL11.2)

Describing the incidence or risk of a condition, or providing details of numbers, is not separated in the Deroey and Taverniers categorisation. However, this is quite a common requirement in the descriptions of genetically related conditions and has been added here. In the next example provided (28) we see a relational process and in addition to the presentation of percentages, we also see it linked to a consequence or effect to risk of getting carrier status.

> (28) S23: so er this test, because it's followed by the sweat test that's the good the good thing about *it is ninety per cent detection rate and zero point three per cent positive rate*. so this is, and here then when you do this there is no worry of getting a carrier status. (PBL10.2)

3.8.3 Academic Functions in PBL2s: Cause and Effect

Given that the subject area of medical genetics studies the causes and treatment of genetically related conditions, it is not surprising that a common function identified was related to cause and effect. Causes identify the results or consequences and in this field can be extended to include symptoms; the two are obviously closely linked to each other, and, as we have seen, also to processes. In the presentations, stretches of talk were found where either causes or effects were emphasised, or where the cause and effect appear more

closely intertwined. They were signalled lexically, e.g., as we have seen with *it is caused by*, or by a description linking either the causes or effects by a logical relationship. Causes in the examples are underlined, effects are shown in italics, with lexical signals in bold. The first example shows an emphasis on causes (29). Here the cause (of '*it*'/Spina Bifida Occulta) is clearly signalled:

> (29) S16: so it's it's the most common out the three, and em, there is er *it* **is it is caused by** like a small gap between the two, vertebrae. (PBL9.2)

In the next example (30), we see the *genes* highlighted as a cause of *clefting*; followed by smoking; this second cause is identified by the words *environmental factors/environmental uh, factor* (one of two main causes/factors relevant in genetics).

> (30) S7: there are three important genes on this locus, uh which, are found to be **cause** uh, i'm sorry [?i'm freaking] out uh **which are supposed to be causing** *clefting* but not syndromic clefting that's really important. **does not cause** *syndromic clefting* they cause *clefting*. maternal smoking, of all the **environmental factors** maternal smoking is supposed to be one of the most important uh**, environmental** uh, **factor** that, adds to the *deformities in the foetus*. (PBL12.2)

In Example (31), the effects (related to weight and lungs) can be understood from the cue *CF* (the condition and cause), with the final effect (a blockage) also signalled by *symptom*:

> (31) S2: em, people with CF, eh they just find it *hard to put on, em, find it hard to gain or lose weight*, and obviously just *have a very severe deterioration in the lung condition*. em, other **symptoms** include *link blockage of the bile ducts em* (PBL10.2)

In (32), the cause and effects stand more closely together and have equal prominence.

> (32) S15: NTDs caused er because of er deficiency in folate (PBL9.2)

3.8.4 Academic Functions in PBL2s: Defining, Elaborating and Exemplifying

Definitions should provide a precise explanation of a term. Flowerdew (1992) reviews the literature on definitions and in his analysis of science lectures provides the following classification system: formal definitions, which include the term, class and characteristics; semi-formal definitions, which omit the class; substitution (e.g., by an antonym or paraphrase); and finally, a minor type, ostentation (reference to some visual).

Although not common in quantity, definitions occurred in just under half the talks, frequently at the outset and framing the following discussion (e.g., of causes, effects or characteristics). After the definition is provided, students move on to elaborating or relating to specifics, for example causes or descriptions of conditions. Example (33) illustrates a formal definition with the term (IRT), the class (enzyme), followed by the characteristic.

> (33) S23: right so we can use *IRT is an enzyme produced in the pancreas and is generally higher in, er the babies, with CF,* yeah? (PBL10.2)

Example (34) shows a semi-formal definition, including the term (*multifactorial cleft lip*) plus a characteristic but no class (Flowerdew, 1992, p. 210); it is signalled lexically with *is ehm classified as*. The characteristic is also rephrased, i.e. elaborated (*on its own/no other features*).

> (34) S20: *so, mul- eh multifactorial cleft lip and palate is ehm classified as that when it occurs purely on its own*. there's *no other features* it just is the cleft lip and palate and ehm, yeah. (PBL12.2)

In the following example (35) the student signals a definition is to follow and provides synonyms, a type of substitution (*not normal; abnormal in form*), followed by an example to clarify further.

> (35) S5: ok er just a definition what is [?], ah it comes from the Greek, which is *means* ah, it's *not normal* but we can say here in this case er, it's something that *is abnormal in form*, and for example we can that it's ah *malformed ear or dysmorphic ear* (PBL7.2)

Deroey and Taverniers (2011) specifically separate elaborating and exemplifying. Following Hyland (2007), they regard them as not adding anything new and therefore not part of their informing function (which includes describing). However, in the examples from the PBLs, elaboration and exemplification are often intertwined with descriptions (and indeed definitions) and are not always quite so easily separated.

In Example (36), S16 is describing the cause of Spina Bifida. The failure of the spinal cord to close is described in relation to the way it closes (with a gesture), and a simplified comparison (flaps and rise up). Although it could be said that all elements after *fail to close* are an elaboration, they show the speaker attending to the group in trying to ensure he provides a full and clear description.

> (36) S16: Spina, Bifida, which is caused by the er, because of the spinal cords fail to close, during early development of the embryo, …. and the way it's em, it closes, it's like it's like this [speaker gestures] (? it's like) *it started from like uh, flaps, and then it just rise up and like you get a gap*. (PBL9.2)

More classic illustrations of examples overtly signalled with *for example* can be seen in (37):

> (37) S14: the the result will be entered in a software, in in MoM format, the software will give final result as a risk *for example* the final result will be, risk of one in five hundred or risk one in three hundred. (PBL9.2)

3.9 Summary of Academic Functions

The exploration of student presentations has identified the most common academic functions employed in the PBL2 presentations. These align closely to both the learning objectives and to the interests of the discipline, i.e. the description and investigation of the causes and treatment of genetically related conditions. In that PBL2s are student-led presentations (followed by discussion), this puts informing about the subject matter very much in the hands of the students. The most common academic functions, those relating to aspects of descriptions and cause and effects clearly address the informing, knowledge

building aspect of PBL, as does the less frequent but still prevalent function of defining. The descriptions category of characteristics, processes, procedures and incidence also reflects findings of lectures in the sciences by Deroey and Taverniers (2011, p. 6), who identify descriptions of things, processes and procedures. Closely linked to the overall informing function is the use of elaboration and exemplification. These demonstrate how the speakers attend to the listeners, working to ensure the information they present is clear.

Interestingly, although a few examples of hypothesising and evaluating were noted, these were not particularly prevalent. This may also reflect the largely absent stage of linking the reports back to the initial scenario (one of the steps). That is not to say, however, that no reasoning takes place. Although not perhaps overtly observed, the building blocks in terms underlying knowledge upon which later decisions can be made is being put in place and in some reports the conclusions reached do relate back to the original problem. The reports certainly lay the foundations for evidence-based decision making and by their nature with each member providing part of the input, the collaborative aspect of problem-based learning and the cumulative epistemology of the sciences is to some extent evident.

Although specific decontextualized examples have been provided for illustrative purposes, the different functions are closely related. Example 23 showed how one function leads into another, as do other examples. Examples of chains of functions include a description of one procedure followed by another (26), cause + effect + cause + effect (30), and (definition +) description + elaboration (35).

The main aim of this section was to identify what is being talked about, i.e. the content in relation to academic functions. Although these functions have not been analysed for lexico-grammatical features, a number of areas emerge that are worthy of further study. With a strong orientation to sharing subject-related content, the reports in PBL2s are characterised by a preponderance of technical academic vocabulary. This might be because in the student presentations information is more densely packed, but it is nevertheless much more noticeable than in the highly interactive sections of turns-at-talk. The examples also give some indication of how the speakers augment the informing function by

providing examples and elaborations. In this section on academic functions, other aspects of interaction with the audience emerge, for example, the use of personal pronouns and how the talks also appear to be linked with informal linking devices. These areas will be investigated in subsequent chapters.

3.10 Chapter Summary and Conclusion

This chapter has explored the different interaction patterns and academic functions of turns-at-talk in the PBL cycle and how problem-based learning is "talked into being".

By employing tools from CA to first map out the organisational structure, we can see how different phases may be signalled. Although specific steps are not always clearly delineated, we do find a clear orientation to the goals of the overall interactional project set out in the guidelines. While mapping out the overall organisation and considering how sequences of turns unfold, different types of talk emerged. Two main orientations were found: organisational talk and dealing with knowledge gaps.

The examples from the PBLs show how the goals of this particular variety of institutional talk are realised. Different from a more traditional classroom learning event, the PBLs in this corpus are quite student-centred. This is reflected in the high levels of student interaction in PBL1s and the student reports in PBL2s. However, as we have seen institutional roles and identities do come to the fore at points. This is notable in the ways tutors direct questions and answers in instructional talk and clarification checks, and in the management of PBLs, illustrated in organisational talk. However, we also see the distribution of roles amongst students, where the role of Chair or Scribe confers certain rights and responsibilities to direct proceedings. In this way, although student-centred, a hierarchical structure and expectations of roles can be seen in the organisation of the proceedings.

In dealing with knowledge gaps we see a variety of speech-exchange systems, each distinct in character and each contributing to the achievement of the overall aims of a PBL approach. These illustrated a number of ways to redress

knowledge imbalances. There are opportunities to specifically address epistemic imbalances and check information in simple knowledge exchanges and clarification checks, when tutors guide students to learning points in instructional talk, and through student discussions in problem-solving exploratory talk. The student reports provide a major opportunity to present subject-specific knowledge. These reports and the clarification checks that can follow are rich in opportunities to discuss specific aspects of content and demonstrate knowledge. The different types of talk may be open to all to contribute at any point (e.g., exploratory talk) or more tightly controlled as in instructional talk and presentational talk.

Thus we can see the PBL aims of collaborative development of skills and content knowledge are talked into being through a range of types of talk and academic functions. Although there are similarities, the two stages afford different opportunities for knowledge exchange. The PBL1s and follow-up discussions in PBL2s are highly interactive and dynamic explorations of issues, based on more limited knowledge. While the speakers do attend to the group by organising the discourse and with elaborations, the PBL2 presentations also have a more prominent informing function. This indicates that the linguistic resources used in each part of the cycle may vary.

The approaches employed so far have provided insights into how PBLs are structured and enacted and what types of things students talk about. They have also provided an initial indication of potential linguistic items that help to realise these interactions. The following chapters will look in more detail at a number of linguistic items and how they are used to realise the pedagogic aims and interactions in the PBL tutorials. Chapter 4 compares academic vocabulary use in PBL1s and PBL2s. Chapter 5 looks at personal pronouns as markers of interaction and engagement. Chapter 6 concludes the research into specific linguistic features in a study of structural and functional categories of lexical bundles and their use in discourse organisation, representing stance, and as referential expressions.

Chapter 4 Keywords and Academic Vocabulary in PBLs

"Genre specific reference corpora identify rather different keywords." (Scott, 2010, p. 11)

4.1 Introduction

Chapter 3 explored interaction patterns and academic functions in the PBL cycle. The different stages in the tutorials highlighted the prominence of subject specific vocabulary, particularly in the PBL2 student presentations. This suggests that vocabulary use may also vary between the stages. This chapter will employ keyword analysis in order to explore distinctive features of vocabulary choice in the PBL corpus in relation to a reference corpus. As this corpus is of academic speech, there is a specific focus on academic vocabulary representation, i.e. core academic vocabulary commonly found in academic registers and technical academic vocabulary which is more common in specific disciplines.

The overall aim is to illustrate how vocabulary as illustrated by keywords is used to represent the discipline, the epistemology of PBL, and how the varying requirements of each stage in the cycle are reflected in vocabulary choices. In order to do this, a reference corpus of lectures and seminars is used as a means of comparison to identify keywords in each stage of the PBL cycle. This shows words that are unusually frequent in a specific corpus. The keywords are then categorised according to frequency and academic type (academic core or technical), grouped semantically and finally considered in relation to process types represented in the transitivity system, drawing on concepts from Systemic Functional Linguistics (SFL).

This chapter will first outline the significance of vocabulary in academic settings. This is followed by a review of previous studies concerning academic vocabulary, and those specifically relating to spoken academic contexts and to process types. I will then discuss the concept of keyness in texts and the methodology used to identify keywords. A discussion of the different types of keywords found in each stage of the PBL cycle, of semantic groupings, and finally of process types then follows.

4.2 Vocabulary: An Indicator of Register and Genre Variation

Vocabulary and lexico-grammatical features have long been identified as a contributing factor to register and genre variation (e.g., Biber, 1988; Biber et al., 2002; McCarthy, 1998; Stubbs & Barth, 2003). This includes variations evident between spoken and written registers from conversation to academic registers, and between academic disciplines. McCarthy (1998), for example, identifies lower lexical density and different high frequency words used in informal spoken language. In contrast, in a series of studies on university registers, Biber and colleagues identify vocabulary as one of the features which differentiate registers shown in the distribution of types and tokens, in word class and word density (e.g., Biber et al., 2002; Biber et al., 2004; Biber, 2006b).

In addition to identifying the distinctiveness of vocabulary choice and use specifically in academic settings, an increasing number of studies are now starting to show how vocabulary varies between academic disciplines. Indeed, although Biber et al. (2004) state that mode and register overall are more likely to be predictors of linguistic variation, they also suggest that further study of vocabulary "will reveal important differences across disciplines" (p. 81). Disciplinary variation has been identified, for example, in the use of reporting verbs (Hyland, 2000; Hyland & Tse, 2007), collocations (Durrant, 2009), process types (Holmes & Nesi, 2009), and as we shall see in the growing number of studies into academic vocabulary lists which identify different frequencies and use across disciplines and sub-disciplines. These studies have not only found differences in frequency and type but have also related the items to the epistemological nature of the disciplines. For example, Thompson (2006) links the use of nouns to abstraction in economics lectures, while Ward (2009, p. 177) suggests that the use of non-technical words in a basic engineering list reflects the focus on processes and functions involved in the development of knowledge. One particular focus has been on illustrating how disciplinary concerns and aims of genres can be illustrated by studying meaning at the clause level through the transitivity system (e.g., Holmes & Nesi, 2009; Love, 1993; Martinez, 2001).

This interest in academic contexts and disciplinary variation has emerged from and has further motivated a wide range of studies of academic vocabulary. These include studies relating to coverage or representation, in academic texts, and to the level or size of vocabulary needed for comprehension of these texts. This is of course motivated by the interest of EAP practitioners in facilitating text comprehension for students studying in another language. The next section provides an overview of these issues.

4.3 Vocabulary: Coverage, Size and Comprehension

This section will outline the importance of vocabulary in academic settings and the related issues of vocabulary coverage in academic texts and the vocabulary size required to enable comprehension. It will then go on to discuss specific descriptions of academic vocabulary.

In order to be successful in academic settings, students obviously need to engage receptively and productively with a range of spoken and written academic texts. Vocabulary knowledge is recognised as playing an important part in the comprehension and production of academic discourse (Nation, 2006; Schmitt, 2008). Although this is true for both first and second language speakers (see Hu & Nation, 2000), much of the research in the field of second language acquisition has been motivated by an interest in determining the amount of vocabulary required to ensure a reasonable or adequate comprehension of texts (connected to lexical coverage), along with a related strand investigating the size of the vocabulary a learner needs in order to do this (lexical size), i.e., how many words or word families in English, based on frequency lists, are required to facilitate this comprehension. These studies have provided insights into the variations in vocabulary that are evident in different types of discourse and in different disciplines.

4.3.1 Vocabulary Coverage and Comprehension

A number of studies have worked to establish the percentage of words in a text that a reader needs to know in order to understand it (Hu & Nation, 2000; Laufer, 1989). These suggest that knowledge of up to 98% of the words in a text

is necessary for good comprehension, a figure which now seems widely accepted, certainly for written texts. Stæhr (2009) and Van Zeeland and Schmitt (2013) both point out that levels of comprehension and vocabulary size should not be assumed from studies of written texts alone as listening is a different process. They go on to suggest that the amount of vocabulary required for listening may be slightly lower at 95%, but again going up to 98% if a very high level of comprehension is necessary. Whether spoken or written, the research suggests that readers and listeners need to be able to understand a substantial number of items to successfully comprehend texts. This then raises the question of what size of a vocabulary is required to reach these levels of comprehension and the extent to which a general vocabulary of high frequency words is sufficient for different contexts.

Research into vocabulary size has been based around knowledge of frequency-based word lists, i.e. knowledge of the number of words, or in many cases word families on a word list (which can include inflections and also derivations, depending on the list) that would be necessary to understand the text. This is particularly relevant for language teachers and learners, although research in this area now extends to wider educational contexts, for example for NSs in secondary schools (e.g., Coxhead et al., 2015).

West's (1953) General Service List, the GSL, has been one of the most often cited and utilised lists of the most frequent words in English. From this list, the 2000 most frequent word families provide, it has been suggested, a coverage of approximately 80% of written texts. Although it is still often used as basis of word lists, the British National Corpus (BNC) word lists are now also commonly used. It is suggested that while the first two to three thousand word families provide extensive coverage of many texts, after this the increment is lower, i.e. the next 1000 words covers a much lower percentage of a text and so on (Nation, 2006). A number of studies have tried to determine the number of word families required to understand a range of written and spoken texts. Depending on the corpus, these can range from 8000 to 9000 word families for written and spoken texts in the BNC (Nation, 2006) to 3000 to 5000 for the CANCODE spoken corpus (Adolphs & Schmitt, 2003).

Adolphs and Schmitt (2004, p. 45) also found "differences in vocabulary coverage according to spoken discourse context". Interestingly, they found the most private and familiar contexts had low coverage but also that the pedagogic context had the least coverage, i.e. knowledge of more word forms to get a comparable coverage to other contexts was necessary.

This, then, indicates that not only can vocabulary be a differentiator of text type and context, but also that more vocabulary may be required for comprehension in certain contexts. If we can assume that academic texts are more specialised and lexically demanding and that there may be vocabulary specifically related to academic contexts, then we need to be able to clarify what constitutes academic vocabulary. This is the focus of the next section.

4.3.2 Academic Vocabulary

Baker (1988) and Farrell (1990) both identify three levels of vocabulary. These can be summarised as general *high frequency vocabulary*; a second group of *specialised* or *technical vocabulary*, which either is not evenly distributed or is significantly different in frequency between specialised and general texts; and finally *sub-technical*, context independent vocabulary used for rhetorical/organisational purposes in specialised genres. This third type can be equated with what is now regarded as *general academic vocabulary*, which Coxhead (2000) defines as follows:

> Academic words (e.g., *substitute, underlie, establish, inherent*) are not highly salient in academic texts, as they are supportive of but not central to the topics of the texts in which they occur. (p. 214)

As these words are relatively infrequent, they are less likely to have been learned; they are also less easily recognised compared to technical terms in a field. Coxhead and Nation (2001) also note that technical vocabulary has degrees of 'technicalness' depending on an item's use within a specific field. They add a fourth category of *low frequency vocabulary*, the preserve of technical specialists. However, the division between technical vocabulary and academic or

sub-technical vocabulary is not always distinct (Chung & Nation, 2004; Mudraya, 2006).

4.3.3 Academic Word Lists

The development of academic word lists in order to support teaching and learning in academic contexts has followed this differentiation of vocabulary. The most widely known is still Coxhead's (2000) academic word lists (AWL). This list was based on a corpus of written academic texts from a range of disciplines and excludes words from West's GSL. The resulting word list comprises 570 word families, including inflections and transparent derivations. Coxhead's work suggested that in addition to the GSL, the AWL could provide a further 10% coverage of a text, thus leaving a much smaller vocabulary to be accounted for.

The development of the AWL then instigated a range of studies into the extent to which such a list can actually be representative of all academic disciplines. Hyland and Tse (2007) question whether there actually *is* an academic vocabulary, in terms of one that is sufficiently useful to learners from a range of different disciplines. They also note that while the AWL can cover about 10% of a corpus, "individual lexical items on the list often occur and behave in different ways across disciplines in terms of range, frequency, collocation, and meaning" (p. 235). They investigate the extent to which the items on the AWL are represented in disciplinary specific novice and expert writing, finding not only uneven distributions, but also differences in meaning and collocations in different disciplines. They conclude among other things that, compared to the social sciences, "writing in the sciences demands more specialized and technical vocabulary" (p. 240).

Such criticisms are reflected in additional studies comparing and contrasting the AWL within specific registers and genres. These aim to identify the extent to which the AWL is sufficient, appropriate and representative of specialist disciplines. A number have been concerned specifically with medically related areas and have indicated that the AWL provides insufficient coverage and that academic words have differences in meaning in different contexts (Chen & Ge, 2007; Fraser, 2009; Wang et al., 2008; Yang, 2015). This strand of research

suggests that specific disciplines will be characterised at least in one respect by specialised vocabulary.

4.3.4 Academic Vocabulary and Spoken Academic English

In addition to studies of frequency and lexical size in general spoken corpora, Nesi (2002), Thompson (2006) and Dang and Webb (2014) have considered vocabulary in spoken academic corpora. In a preliminary study of academic words in the BASE corpus, Nesi (2002) found both general high frequency words and also words thought to be common to academic written texts (based on the AWL). She identified a further set of high frequency words in BASE which "reflect the interactive and interpersonal nature of spoken academic discourse" (p. 354), including expressions of politeness and vague words. In another study of academic vocabulary use in the BASE corpus, Thompson (2006) considered coverage of the AWL in lectures, with a specific focus on keywords in economics lectures. He suggests a smaller set of word families may be required for lecture comprehension than written language and also illustrates how keywords reflect different research concerns across the disciplines. In a third study of the BASE corpus, Dang and Webb (2014) investigated the vocabulary size required to understand lectures and seminars. They found not only a variation in the number of words required in different disciplines but also that medical sciences appeared more "lexically demanding" (p. 70).

Thus, in addition to variation according to academic disciplines, these studies suggest that vocabulary choices vary also according to mode (i.e. spoken) and genre.

4.3.5 The Academic Vocabulary List

While acknowledging disciplinary variations, Gardner and Davies (2014) have made a compelling case for a new Academic Vocabulary List (AVL), based on the much more comprehensive Corpus of Contemporary American English (COCA) corpus. The AVL is derived from a sub-corpus, COCA Academic (COCA-A) of 120 million words. Based on a much larger and more representative corpus (albeit written and American English), it categorises lemmas rather than word families,

noting different parts of speech and derivations may have different meanings. The selection is based on more rigorous frequency and dispersion criteria: to be included in the core AVL, lemmas should be 50% more frequent and found in at least seven of the nine disciplines in COCA-A than in the whole corpus, and to be included in the technical AVL component it cannot be more than "three times the expected frequency in any of the nine disciplines" (Gardner & Davies, 2014, p. 316). A specific online interface for academic words and phrases identifies lemmas as core academic, technical/sub-genre, and also high frequency non-academic and lower frequency non-academic lemmas if desired. This online resource has much potential in determining academic vocabulary use in texts although its applications have yet to be widely reported on or made use of in conjunction with spoken discourse.

While it is not the purpose of this research to develop an academic word list for the PBL corpus, the development of specialised corpora and word lists provide insights into aspects of vocabulary selection within disciplines, registers and genres. They also provide a point of reference when studying vocabulary in a specialised corpus such as the PBL collection.

In summary, academic vocabulary can be identified and categorised according to its prevalence in academic contexts and the extent to which items are found with more frequency in specific disciplines than others (academic technical). Research indicates that not only are different words found in different disciplines, but also that they can indicate and reflect differences in the interests and epistemologies of the disciplines. The next section provides an overview of such studies, including approaches to analysis.

4.4 Vocabulary Use in Academic Disciplines

In addition to attempts to identify disciplinary specific word lists, there have been a number of specialist corpus studies investigating aspects of vocabulary use in specific genres and academic disciplines, including one specifically on PBL tutorials.

Gledhill (2000) uses the WordSmith KeyWords programme (Scott, 1996) to identify internal salience. He investigates collocations of verbs in cancer research articles, finding differences in use in different parts of articles. These may be related to the "semantics of the specialist domain", or "the dominant discourse strategies in the research article" (Gledhill, 2000, p. 130). Groom (2010, p. 66) takes a "closed-set first" approach, again using keyword software. He identifies a number of patterns found in academic journal articles in art history. These include, for example, *property* + *of* + *phenomenon* for detailed analysis and *authority* + *of* + *domain* related to social power.

Research is now being extended to include cross-disciplinary collocations. Chen and Ackermann (2013) have developed an academic collocations list, identifying lexical collocations and highlighting the importance of collocational frameworks for EAP learners. Durrant's (2009) study of collocations in academic writing found mostly grammatical collocations, noting the potential for words to be linked to grammatical patterns. Drawing on Hunston and Francis' work on pattern grammar (2000), and Hoey's concept of lexical priming (2005), Durrant (2009, p.163) notes "productive patterns which are tied to specific lexis".

Both Leone (2010) and Malavasi and Mazzi (2010) investigate keywords in relation to specific academic disciplines. Leone considers keywords and discourse patterns in Italian history books, comparing general spoken language and school language. The keywords in history indicate, for example the importance of "narration, interpretation and explanation … and causality" (p. 243). Malavasi and Mazzi compare keywords in history and marketing and show how they bring insights into disciplinary epistemology, for example relating to shared empirical concerns but also differences. This includes forms of self-representation and co-occurrence with reporting verbs (with *historians,* or *we* in marketing).

In the only work that I know of related to vocabulary use in PBL sessions, Da Silva and Dennick (2010) use sophisticated corpus software, Wmatrix (Rayson, 2009) to tag parts of speech and semantic categories of vocabulary items in a PBL cycle of three PBL sessions. They illustrated how use of technical vocabulary and items related to pedagogic skills changed across PBL sessions, depending on

the focus of the particular part of the cycle. This included technical vocabulary related to probability in the first stage compared with words connected to treatments in the later stages.

These studies show that vocabulary use in disciplines can reflect the disciplinary epistemologies. This has also been found in studies of process types, discussed in the next section.

4.4.1 Vocabulary and Process Types

A number of studies have illustrated how the vocabulary choices exemplified in the transitivity system at the clause level reflect the concerns of specific disciplines or parts of genres. This section starts with an overview of the theoretical framework used.

The studies that will be discussed here draw on SFL, based on the underlying principle that language is a "resource for making meaning" (Halliday, 1994, p. xxvi). Briefly, language is used to represent how we experience and interact with the world and others. It allows us to articulate the inner world of thoughts, reflections and consciousness and the experience of the outer, external world, that of actions and events. Linguistically, Halliday goes on to explain how the transitivity system at the clause level "construes the world of experience into a manageable set of process types" (p. 106). There are three broad groups which reflect our inner and outer experiences and how these experiences are related to each other. Material processes reflect actions and events in our representation of outer experiences, mental processes are those that reflect thoughts and feelings, i.e. the inner consciousness, and the third group, relational processes, allows for classification and identification. In addition to the three main process types, there are those that sit on the boundaries. Between the mental and relational are verbal processes, between relational and material are existential processes, and between mental and material there are behavioural processes. Each process type is realised in the grammar of the clause by participants in the nominal group, the process in the verb group, and circumstances by adverbial groups or prepositional phrases (Halliday, 1994, pp. 107-109). For the three main process types, the participants within material

processes are the actor and goal or range, in mental processes the senser and phenomenon, and in relational processes the carrier and attribute, or token and value. As processes provide a frame of reference for interpreting "patterns of experience" and what is "going on" (Halliday, 1994, p. 106), studying how aspects of the transitivity system are used can give an indication of "the way disciplinary knowledge is conceptualised and expressed" by those involved (Holmes & Nesi, 2009, p. 58).

4.4.2 Process Types in Academic Disciplines

Studies of the processes evident in the sciences include the seminal work by Halliday and Martin (1993), with studies related specifically to academic disciplines or genres including work by Love (1993), Martinez (2001), and Holmes and Nesi (2009). Love (1993) examines process types in geology textbooks to illustrate the relationship between general (natural) processes in geology and the resulting specific geological products (e.g., weathering and the production of land formations). She finds that process types are frequently incongruent, drawing on Halliday's concept of grammatical metaphor where one grammatical form is substituted by another. This is achieved most commonly by nominalisation of verb forms, allowing for comparisons, identification, quantification etc., a common feature in writing in the sciences (Halliday, 1994; Halliday & Martin, 1993). This nominalisation, Love finds, results in a very high proportion of relational process types used to classify and identify characteristics. She thus makes a very clear link between the process types and the specific concerns of the academic field.

Martinez (2001) uses keyword analysis to investigate processes realised by finite verbs evident in different sections of research articles from physical, biological and social sciences. She finds different distributions of processes between different parts of the articles, including for example, higher numbers of mental and relational process types in the Results and Discussion sections, reflecting the importance of "deliberation and reflection" (p. 239). This is in contrast to the use of agentless passives in the Methods, regarded as the "action section" (p. 241).

Holmes and Nesi (2009) consider verbal and mental process types across four disciplines in student writing from the British Academic Written English corpus. Moving beyond process types realised in the verbal group to include other clausal elements (thus including inflected forms and derivations, e.g., verbs nominalised as participants), they illustrate how students represent themselves and their disciplines. Again using keyword analysis, they find important distinctions between hard, soft, pure and applied domains, suggesting "fundamental differences in disciplinary cultures" (p. 70) in how knowledge is developed and displayed. This includes claiming and counter claiming in history, establishing facts and causal and logical relationships in physics, and diagnosing properties and assessing outcomes in the applied disciplines (pp. 65-70).

4.4.3 Summary and Research Questions for Vocabulary

These studies indicate that vocabulary choice, use and frequency can vary depending on genre and also discipline. Specific studies have also indicated that items may be used in different ways, reflecting the specific concerns of the discipline and the type of discourse. However, as we have seen, much of the work to date has been based on written corpora with only a small number of studies starting to consider general and specific academic vocabulary in spoken contexts.

One of the aims of this study is to investigate the extent to which PBLs reflect general findings of academic, and in particular spoken academic, English. As the two stages differ in interaction patterns and purpose, differences in keywords in relation to the aims of each stage of the PBL cycle will be studied.

The specific research questions relevant to this chapter are as follows:

- What do the keywords indicate in terms of the 'aboutness' of PBLs?
- To what extent are academic core (AC) and academic technical (AT) keywords represented in the keyword lists?
- To what extent are keywords in PBL1s similar or different to keywords in PBL2s?

These three questions should provide an indication of the discipline and the type of communicative event, an indication of the academic and disciplinary specific nature of the PBLs, and of the different nature of each part of the cycle. In order to identify the extent to which vocabulary reflects the discipline and the specific type of communicative event and pedagogic aims through a manageable data set, vocabulary items will be selected by identifying keywords in relation to a reference corpus. The next section will expand on this approach.

4.5 Methodology: Keywords

This section first explains the concept of keywords and the methodological approach for identifying such items in the PBL corpus.

While raw frequency lists obviously identify the most common words in a corpus, function words are generally the most frequent. An alternative approach, also noted in some of the studies above, is to identify keywords in a specific study corpus in comparison to a reference corpus. Scott (1997) defines keywords as follows:

> A keyword 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. (p. 236)

These keywords give an indication of the 'aboutness' of a text. Keywords can be positive, with unusually high frequency in relation to a reference corpus, or negative, occurring much less often than in the reference corpus. The keyness is a measure of statistical significance and is measured either by Log-likelihood or Chi-square. The statistical significance is measured by a p-value, e.g., 0.000001, indicating a one in million "risk of being wrong". A low p-value also limits the number of keywords that will be identified (Scott, 1996).

The unusual frequency in comparison to a reference corpus will obviously be influenced by the nature of the reference corpus used. The reference corpus is usually larger, although comparisons between similar size corpora and even sub-corpora for internal salience have been made (e.g., Gledhill, 2000; Wilkinson,

2014). The question of the size and composition of a reference corpus has been the subject of some interest. Berber-Sardinha (2000) discusses the question of what size a reference corpus should be. By varying the size of the reference corpus according to the size of the study corpora (two to one hundred times), he concludes that a reference corpus five times the size of the study corpus was an optimum size. Any smaller may miss potentially important keywords but more than five times in size is not necessarily more reliable. However, Scott (2009) carries out a number of tests in an attempt to identify features of a bad reference corpus by varying a number of aspects: the size of the reference corpus, using a "deliberately strange" (p. 82) reference corpus, and the relevance (with genre as a variable). He suggests that size and strangeness do not affect the quality of the keywords greatly, although he is inconclusive about restricting genre. Perhaps not surprisingly, he concludes that the selection of the reference corpus will result in different aboutnesses being identified. Following this work, Goh (2011) points out that while Berber-Sardinha (2000) considered the size and quantity of keywords identified and Scott (2009) the quality, neither looked in sufficient detail at genre variations nor at corpus composition. His work considers both size and composition effects of the reference corpus, finding that size is not as significant a factor as genre and diachrony. Scott (2010, p. 11) later notes "genre specific reference corpora identify rather different keywords".

Once keywords have been identified, these can be studied for semantic and grammatical behaviour. As noted in Section 4.4 above, researchers often explore lexical keywords but Gledhill (2000) and Groom (2010) also make a strong case for considering closed-class keywords, suggesting that from these patterns can be identified, illustrating that this can provide insights into meaning and use within specific disciplines and genres. Keywords can also be studied in relation to the academic discipline to which they are connected (e.g., Holmes & Nesi, 2009; Love, 1993; Martinez, 2001; Thompson, 2006). Such studies have included investigations into semantic grouping, meaning and collocational patterns and process types again noted above (in Section 4.4).

Keywords, then, can provide an indication of the aboutness of texts in relation to a reference corpus. In the study of the PBL tutorials, a larger reference

corpus of academic seminars and lectures from a range of academic disciplines (the BASE corpus) will be used to identify distinctive features of this student-centred disciplinary specific discourse.

4.5.1 Selecting a Reference Corpus and Creating Keyword Lists

As one aim is to compare keywords in the PBLs with a wider range of academic disciplines, each sub-corpus, i.e. PBL1s and PBL2s, was compared with the readily available BASE holdings (over 1,640,000 tokens)[16]. As it is a specific academic corpus, it should indicate the extent to which each stage is distinct from the spoken academic registers represented in BASE. Although it is technically possible to select only the seminar section of the BASE corpus, the PBL2s include extended presentations which, in terms of interaction patterns, are more similar in format to a lecture than more interactive seminars. In addition, the seminars at around 430,000 words, are less than five times the size of the PBL corpus. Because of this, and because the aim is to show how student speaking in one discipline and in a specific type of academic speech event is distinct, the whole BASE corpus was selected.

In order to identify keywords, the first stage was to make word lists for the study corpus, i.e. each stage of the PBL cycle, and for the selected reference corpus. To do this, the BASE files for both lectures and seminars across the four disciplinary groups were converted to plain text and a word list generated using the WordList tool in WordSmith. To make a keyword list, the word lists from the study corpus and the reference corpus are then loaded into WordSmith and the KeyWords tool employed. The settings required a minimum of three instances, with the p-value set at 0.000001. The p-value was the Log-likelihood procedure used to calculate statistical significance.

Once the keyword lists for PBL1s and PBL2s were created, items which were clearly part of transcription notes (e.g., *coughs, laughter, inaudible, nods,*

[16] Initial explorations with the BNC World corpus, one section of BASE (Sciences), and comparing PBL1s and then PBL2s to the whole PBL corpus indicated that the BNC would provide too large a number of keywords for reasonable study, and comparing each stage with the whole PBL corpus too small a number, as was the case when using one sub-section of the BASE holdings.

aside, refer) and initials indicating names were removed. A small number of items that were a combination of transcriber notes and words that were part of the PBL discussion (e.g., *board, read*) were checked at this stage and removed if they did not occur in at least three texts. The word lists were then saved as Excel worksheets in order to facilitate further sorting and analysis. As I am interested in items that are key across the majority of PBL1s and PBL2s, not in only a minority of texts, the remaining keyword lists were then sorted to eliminate any items that did not occur in at least three of the five texts in each cycle (e.g., *polymorphic* occurs 31 times but only in one text). Negative keywords were also not considered.

The total number of positive keywords after cleaning and sorting found in a minimum of three texts in each cycle was 95 in PBL1s and 153 in PBL2s. Somewhat surprisingly, although pronouns, conjunctions, and auxiliary modals were noted, no closed-class preposition items were identified in the lists, thus precluding the possibility of following Gledhill (2000) and Groom's (2010) work mentioned above. It was, however, possible to categorise keywords into those which are academic core or technical according to the AWL using the COCA-A online resource. This is expanded upon in the next section.

4.5.2 Working with Keywords: Tools and Categories

As one of the aims was to determine the extent to which academic words are a distinguishing feature of the PBLs, the first step was to identify academic core (AC) and academic technical (AT) words in relation to categories set out in COCA-A. Although there is the obvious limitation that the COCA-A corpus does not include a spoken component, it was used as the main resource for a number of reasons. The first is that it allows the researcher to categorise words by group (e.g., AC, AT), and specifically highlights in which disciplines the AT lemmas are most commonly found. It also identifies other words that are either high or low frequency in the COCA-A corpus. The interface is quite easy to use and also provides information on common collocations and meanings. While there is now also a New Academic Word List (Browne et al., 2013), based on the CEC and incorporating MICASE and BASE, a check of the keywords from the PBL corpus found a particular low coverage (only 11 in PBL1s and 20 in PBL2s). The interface

is also less flexible compared to COCA-A as it does not distinguish academic core or technical words, nor does it provide information on frequency in different academic disciplines. The word lists suggested by both Nesi (2002) and Thompson (2006) developed from the BASE corpus were not readily available. Finally, while the BASE holdings are searchable with Sketch Engine, it does not identify the most common meanings nor does it categorise words as AC or AT.

In order to identify keywords which could be considered academic and to confirm part of speech (PoS), necessary for the categorisation, three tools were used. As mentioned, using COCA-A it was possible to identify which keywords are included in the AC and AT groups. The interface provides information on the most frequent lemmas in each category. While the PBL keyword lists only identify the words, not the part of speech (PoS), because the corpus is quite small it was possible to manually check and assign a PoS for each of the PBL keywords; this was done with the help of the Concordance tool in WordSmith. As the information is readily available, high- and low-frequency items (HF and LF) in relation to the COCA-A lists were also noted. N.B., HF and LF relate specifically to whether they are high or low frequency within the COCA-A corpus, and not whether they are HF or LF in relation to general words lists. Items that did not fall under these categories were grouped as Other (O), i.e. not in the top 20,000 in the AVL, or as Semi-lexical (SL), for example *ah*, *uhum*.[17]

Although the use of COCA-A does enable the identification of academic words found on the PBL keyword lists, as noted the obvious limitation is that COCA-A does not include a spoken component. In order to confirm the most common PoS of the academic keywords in the PBLs, as a final check, Sketch Engine was used to search and check lemmas of these words in the BASE corpus i.e. to identify the most common PoS for academic words in a spoken corpus (only *testing* was found to differ, being more common as a verb in BASE but mainly as a noun in the PBLs). A final check of whether the HF, LF and O words were included in the

[17] A range of terms are used for referring to items such as *uhum* including *non-lexical conversational sounds* (Ward, 2006), and *filled pauses, backchannel cues, exclamations*, MICASE (Ädel et al., 2007). This work follows the term used in the SCOTS corpus, which uses the term *semi-lexical* (http://www.scottishcorpus.ac.uk/corpus-details/).

top 1000, 2000, and 3000 word lists in the BNC was also made (Leech et al., 2001).

As noted in the previous section, 95 positive keywords were identified in PBL1s and 153 in PBL2s. However, the number of keywords represented in each group here is slightly higher than the final number of keywords identified. This is explained as follows. As we have noted, the COCA-A list presents the lemma of a word and its inflections, i.e. differentiating parts of speech (PoS). This means that *test, tests,* and *testing*, all keywords in the PBL corpus, would all be categorised as academic core in the verb form but academic technical in the noun form. Following the detailed analysis of PoS, a small number of words were found to be used in different forms and so are categorised into different groups: this comprised three of the 95 keywords in PBL1s (*test, separate* and *inherited*) and two of the 153 keywords in PBL2s (*affected* and *test*). This allows for the actual use of the keywords to be illustrated. The presentation of the keywords in this way therefore means that 98 words are included in the keyword list for PBL1s and 155 for PBL2s. However, it should also be noted that the keyword lists present word forms, not lemmas. This means that inflections (e.g., *test, tested*) are counted individually. The keyword analysis also considers types rather than tokens. As such, the results presented here of the keywords are not a comment on the overall coverage of types nor of the frequencies of academic words as a whole in the corpus.

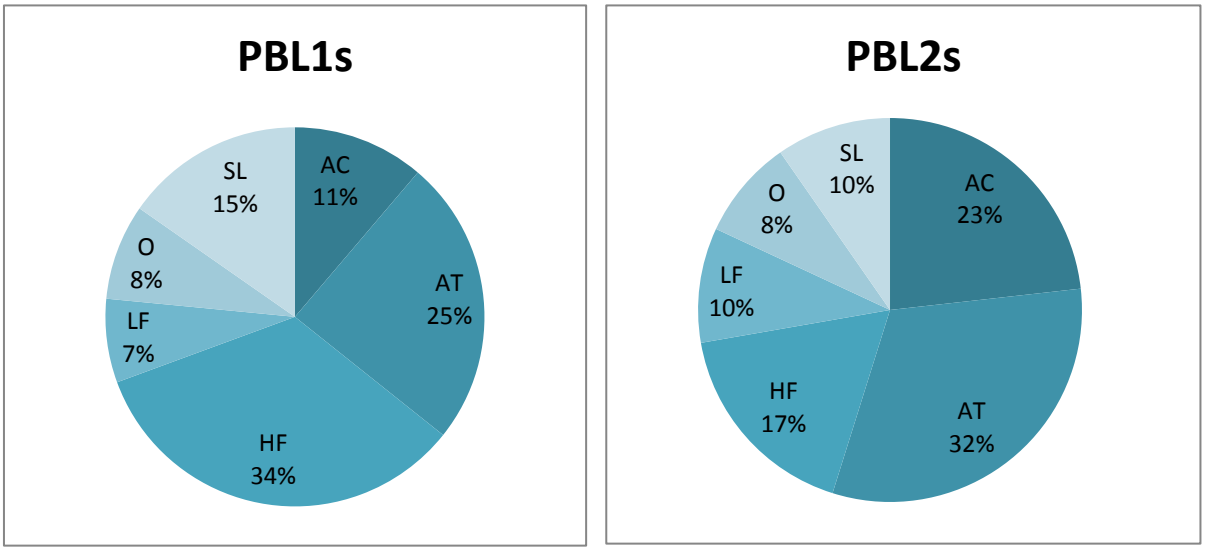
4.6 Findings

This section first provides an overview of the proportion of keywords found in the different groupings in PBL1s and PBL2s and comments on how the keywords indicate the general aboutness of the corpus.

4.6.1 Distribution of Keywords in PBL1s and PBL2s

Comparing the groupings of the keywords in PBL1s and PBL2s does indeed show different aboutnesses in the two parts of the cycle. The proportion of keywords found in each of the six categories is illustrated in Figure 4.1.

Figure 4.1: Keywords in PBL1s and PBL2s



Academic Core	AC	11	11%
Academic Technical	AT	24	25%
High Frequency	HF	33	34%
Low Frequency	LF	7	7%
Other	O	8	8%
Semi-lexical	SL	15	15%
		98	

Academic Core	AC	36	23%
Academic Technical	AT	49	32%
High Frequency	HF	27	17%
Low Frequency	LF	15	10%
Other	O	13	8%
Semi lexical	SL	15	10%
		155	

In PBL1s, 11% of the keywords are identified as AC and 25% as AT, i.e. 36% combined. A much higher proportion of keywords in PBL1s are HF words (34%). In contrast to PBL1s, PBL2s have a higher proportion of both academic core (23%) and academic technical keywords (32%), 55% combined, and a correspondingly lower proportion of high frequency words, only 17%. This may reflect the more interactive and exploratory nature of PBL1s compared to the much greater concentration of time and talk devoted to conveying pre-prepared, subject-specialist information in PBL2s (also indicated in Chapter 3). Notwithstanding the fact that the PBL2 sub-corpus is of course larger, in addition to different proportions, there are around twice the number of distinct lemmas in PBL2s, indicating a wider variety of items.

The keywords in each category are provided in Appendix 3. The keywords were sorted first by keyness, then alphabetically and then by frequency in the PBL sub-corpus. (Keywords found in PBL1s and PBL2s in the same part of speech are also indicated.)

That PBL1s are also more informal and conversational is reflected in the much higher level of high frequency words. These HF words are identified by Gardner and Davies (2014) as high frequency in the COCA-A, but an additional check also showed that they are almost all included in the top 1000 most frequent words in the BNC list. The PBLs words in this HF group include pronouns (e.g., *she, they, it*) and modal verbs (*could* and *should* in PBL1s). Even the high frequency words can often be seen to be related to the specific interests of medical genetics, for example, *family, baby, causes* and *inherited*.

In his academic lecture corpus, Thompson (2006, p. 259) finds overall a higher proportion of general high frequency words compared to the Coxhead AWL items. Although his study was based on the older AWL and compared the overall coverage, the prevalence of high frequency items in the spoken corpus is also reflected here, certainly in PBL1s.

Mauranen (2006, p. 279) points out that "it is more than obvious that disciplines differ in their specialist terminology" and because of this, she does not concentrate on these items in her study of a corpus of English as a lingua franca. However, the particular interests of the discipline can be seen in the academic technical keywords in both parts of the PBL cycles. When we consider the AT words in PBLs, they are indeed predominantly nouns, but we can also see that the majority can be clearly identified with natural processes, for example *genetic, syndrome, mutations,* and with the procedures in which medical geneticists are professionally involved, for example *counselling* and *testing*. These words occur on both the PBL1 and PBL2 keyword lists. We can also see the subject area and its interests reflected in the AC words. For example, there is a focus on identifying and exploring specific conditions with *testing, factors,* and *results*. These groupings will be explored in more detail in the next section.

In both stages, there is a smaller percentage of LF words (7% in PBL1s and 10% in PBL2s), and words grouped in the O category (8% in each stage). The LF group includes very specific terms (e.g., *pedigree* and *chromosomal* in PBL1s; *marker, pedigree* and *carrier* in PBL2s), but also words which can be clearly identified as common in spoken interactions (e.g., *ok, yeah,* and *oh* in both stages). This mix is also found in the O category, with keyword lists including *PBL* and *FISH* (in

PBL1s) and *MLPA* (in PBL2s), the latter two items relating to specific types of testing, and *cos* and *I*. The limitation of the lack of the spoken component in COCA-A is most evident in these groups.

The semi-lexical group indicates the online processing and highly interactive nature of PBL1s with back-channelling and semi-lexical responses such as *uhum*, *mhum* and *ah* and markers of hesitations such as *er, uhm*. Although the proportions seem high (15% in PBL1s and 10% in PBL2s), this is because each item is counted and stands out because the PBL corpus is not tagged, unlike BASE which tags semi-lexical items (i.e. these items do not show as words in a search of BASE).

The keywords can be seen to reflect the subject area and the communicative event, with, not surprisingly, the academic core and academic technical reflecting the concerns of the subject area. The following section discusses the items in more detail, concentrating primarily on the AC, AT and HF words, as together these account for 70% or more of the keywords in each stage.

4.6.2 Semantic Groupings of Keywords

In addition to grouping words by frequency and academic 'aboutness', the keywords can also be grouped semantically to highlight the common subject or topic areas. This grouping into related semantic fields also reflects to a large extent the academic functions identified in Chapter 3. Table 4.1 to follow illustrates the keywords for each of the following broad groupings: those related to describing genetically related syndromes and conditions, a group for people and patients, one relating to natural processes in the genetic system, one for genetic procedures and protocols, and a small group related to PBLs and study. The final column is a general grouping for the remaining keywords.

Table 4.1: Semantic Groupings of Keywords

	Medically related conditions	People	Natural processes and genetic system	Procedures/ Tests	PBLs and Study	General
PBL1 **AC**	factors, features	population	inherited, inheritance, affected	prenatal, results, screening, test, testing	database, scenario, scribe, separate	
AT	abnormality/ies, syndrome	age, geneticist, risk/s, parents,	genetic, gene/s, molecular, mutation/s, pregnancy	counselling, diagnostic, diagnosis, screening, test/s		
HF		baby, family	causes, inherited		board, chair, check, report, separate	could, do, else, everyone, it, just, know, like, maybe, mention, need, no, shall, she, should, someone, stuff, sure, think, we, what, wondering, would, yes
LF		pedigree	chromosomal			yeah, ok, oh, suppose, tomorrow, whoever's
O				fish, cytogenetic	pbl/s	cos, Glasgow, i, non, pre

PBL2 AC	features, factors, common	population, maternal	affect/ed, deletion/s, inherited, pathway, transcription, unaffected	accurate, detect/ed, identified, negative, positive, result, sensitivity, sequence/ing, specific, specificity, technique, test/ed/ing, prenatal	group, overlap, scenario	associated, mainly
PBL2 AT	abnormalities, cancer, cystic, normal, severe, syndrome/s, tube, tumour	child, foetus, incidence, parent/s, patient, pregnancy, risk	affected, allele, autosomal, chromosome/s, DNA, gene/s, genetic, molecular, mutant, mutation/s, normal, phenotype, polymorphism, protein	biochemical, criteria, counselling, diagnostic, diagnose/ed/is, kit, probes, screening, surgery, test/s, ultrasound	slide	weeks
PBL2 HF		age, baby/ies, cases, family/ies	high, causes/ed	follow, found, offer	paper/s, questions, mentioned	also, five, it's, like, no, other, she, six, so, thank, they
PBL2 LF	Delay	carrier, pregnant, pedigree		detection, false, marker/s, terminate		maybe, mostly, ok, oh, sixty, yeah
PBL2		percent	recessive, penetrance	MLPA	pbl	cannot, cos, de, don't, i
O						wee, UK non

Initially a group relating to incidence was included but this only included 3 items and so words were subsumed into one of the broader categories (e.g., *risk* and *population* referring to patients and people). A number of items from the HF and LF COCA-A AVL can also be added to these groups and have been categorised in to the relevant section (e.g., *pedigree* related to people, *terminate* related to procedures). There is, of course, some overlap, as boundaries are not always clear-cut in this essentially subjective type of exercise. For example, *affected* indicates the outcome of a natural process but also can relate to either the patient or a part of the genetic system; *population* obviously relates to groups of people in general but in this context was also used at times to talk about incidence and testing procedures; *risk* can relate to the condition, to incidence, or to specific groups of people; *prenatal* is obviously a condition and relates to a person, but in this context is most commonly used with *testing* and so has been grouped under procedures. In order to determine the most appropriate and common grouping at this stage, the concordancing tool was used to explore the surrounding co-text. Personal pronouns, although referring to patients or fellow students, have been put into the general category. Where a word occurs in different categories because of PoS, the word will occur again in the relevant grouping (e.g., *separate* as AC and HF).

From Table 4.1 above, we can see that in both stages of the cycle, apart from the General category in PBL1s the majority of keywords are connected to natural processes and to the procedures geneticists employ. This is particularly noticeable in PBL2s, where we see many more words related to both processes and procedures. This latter group includes words for describing specific types and characteristics of tests, e.g., *technique, sequencing*, *prenatal, accurate* and *specificity*, and words relating to treatments, for example c*ounselling* and *screening*. Both stages include words related to describing conditions, e.g., *features, syndrome, severe, abnormalities,* and to the groups of people directly involved in some way, e.g., *parents, family, babies,* and also *geneticists*. Thus, the keywords appear to reflect the concerns of medical genetics but also indicate a change in focus across the two stages, with increasing specificity in PBL2s.

In PBL1s a higher number of items are related to PBLs as a mode of academic study, for example *separate, scribe, board* and *database*, reflecting the concern with checking work for the next part of the cycle. In PBL2s, we see *papers, questions* and *slide*, reflecting the focus on presentations and responses to the content (*questions, mentioned*).

Within the general grouping, i.e. the remaining words that were initially categorised as either HF, LF or O, we can also see an indication of the mode of communication and the difference in orientation between PBL1s and PBL2s. We have already noted the more informal nature of the PBL1s, exemplified here with *like, cos* and high personal pronoun use, and the inclusion of response tokens (e.g., *yeah, oh, ok*, in both stages), but we can also see more high frequency verb forms in PBL1s (e.g., *think, suppose*), suggesting more of a focus on reflection and hypothesising.

The groupings so far have indicated the academic nature of PBLs and topic areas but aspects of use have yet to be considered in detail. As discussed in Section 4.4, various approaches have been used to illustrate vocabulary use and behaviour. The next section will consider one approach: how the process types that are represented at the clause level can give an indication of how the vocabulary is used to represent the subject area, mode and epistemology of PBL.

4.6.3 Keywords and Process Types in PBLs

This section presents the main process types identified in the keyword lists. The analysis here draws on Halliday's categorisation of process types, described in Section 4.4.1 above. As the focus of each stage in the PBL is quite different both in format and aim, we will see that the process types identified also vary.

The discussion of the process types will concentrate on keywords in the AC, AT and HF categories. This includes finite verbs, but also considers nouns derived from the verbal group as it is clear from the keyword lists that a significant number were nominalised verbs. Table 4.2 which follows presents the keywords grouped by process type, with nominalised verbs indicated in italics. The

keywords have then been further grouped according to natural processes in genetics, procedures, and a final 'Other' group.

As we shall see, a consideration of the process types evident in the keyword lists further emphasises the findings from the semantic grouping exercise. There is a very clear focus on natural processes and procedures, reflecting the concerns of this applied discipline, with the majority of keywords falling under material process types. This is true for both PBL1s and PBL2s. The keywords in both stages also include a number of nominalised processes, which again refer to natural processes and procedures. The main differences between the stages are that, as we have seen in the semantic grouping, within the material process verbs in PBL2s there is not only more variety in type of words, but there is a stronger focus on procedures. However, the study of process types also highlights that the keywords in PBL1s include a number of mental process verbs. Inspection of the nominalised verbs indicates that relational processes are evident, but do not predominate and that these nominalised forms frequently co-occur with other material verbs.

Table 4.2: Process Types in PBL1s and PBL2s

	Natural processes	**Procedures**	**Other**
PBL1s **Material**	affected, causes, inherited *causes, inheritance, mutation/s*	test/ing *counselling, diagnosis, results, screening, test/ing*	check, cover, do, separate
PBL2s **Material**	affect/ed, causes/ed, inherited *causes, deletion, deletions, transcription, mutation/s*	detect/ed, diagnose/d, found, follow, identified, offer, sequence, terminate, test/ed/ing *counselling, diagnosis, detection, result, screening, sequence/ing, test/s*	*risk/s*
PBL1s **Mental**			think, suppose, wondering, know, mean
PBL2s **Mental**			associated
PBL 1 Other **(Existential/ Verbal/ Behavioural)**			mention
PBL 2 Other **(Existential/Verbal/ Behavioural)**			mentioned, thank

In addition to categorising by process type and semantic area, the main verbs were also categorised by voice. Of the keywords which are verbs, the majority in PBL1s and PB2s are used in the active voice. In PBL1s of the 15 verbs from the AC, AT and HF groups, 12 were always used in the active voice, with only two used in the passive (*inherited* and *affected*). In PBL2s, of the 22 verbs in total, thirteen were always used in the active voice; only three were used solely in the passive (*inherited, affected* and *associated*); and six were shared (*caused, detected, tested, identified, diagnosed, found*, with the majority in all cases used in the passive). That keywords occur much more frequently in active constructions is reflective of spoken discourse. The higher proportion in PBL2s is likely to be due to the stronger focus on prepared content and reporting on procedures and causes. These points will be highlighted where relevant in the examples that follow.

4.6.3.1 Material Processes in PBLs

The majority of keywords under consideration in both stages are related to material processes, but becoming more precise and specific in use in PBL2s. These material process verbs are used to indicate causal relationships, for example with natural processes showing how elements of the genetic system interact and how patients are affected by diseases, for describing technical procedures, and with a small sub-group that are related to study.

In PBL1s we see examples of material process types used when speculating and asking questions about what or how something is *inherited*; who, or how someone, is *affected*; or what *causes* a condition. Identifying specific causes is unsurprisingly less common at this exploratory stage. The actor is frequently implicit but may be animate, e.g., the generic patient who inherits, with an inanimate goal in the subject position (*disease* and *mutation*), as in Examples 1 and 2.

> (1) S5: no but you mean that, they can how they those diseases *are inherited* (PBL11.1)

(2) S5: and why is going to talk about you can say about the you know the mutation and maybe *is* it *inherited* (PBL11.1)

In the examples of *affected*, the goal is almost always human (*baby, foetus, people*), with the actor implicit (e.g., *affected* by the disease in Example 3).

(3) S16: genetic counselling like genetic counselling so if the baby*'s affected* then what would you do? (PBL9.1)

Although *causes* is much more common as a noun, when it is used as a verb in PBL1s, it is used in six out of seven cases in a question, as in Example 4.

(4) S27: just, explain what's cancer what *causes* it what (PBL11.1)

In PBL1s, all examples of *inherited* and *affected* are passive and all examples of *causes* are active.

In PBL2s, having carried out their research and now ready to provide a more informed report, students are more specific about *causes* and who or what is *affected*. While both active and passive constructions are used, passives are more common for these verbs. In Example (5) below, the student is responding to a question about nonsyndromic cleft lip and palate.

(5) S20: it's not *caused* by a mutation it'*s caused* by variation and both environmental and genetic factors. (PBL12.2)

Affect is also used to talk about links between inanimate agents, for example drugs and mutations, and processes (as opposed to people) in the genetic system:

(6) S19: then you have in class three the mutations *affect* the chloride channel regulations and gating (PBL10.2)

(7) S15: another hypothesis is that, these drugs *affect* folic metabolism (PBL 9.2)

In contrast and similar to PBL1s, a*ffected* is most commonly used to talk about people, rather than a part of the anatomy:

> (8) S12: the diagnosis confirmed and J is *affected* with Van der Woude syndrome (PBL12.2)

These process verbs are used when students are making use of their independent research to make conclusions about conditions. The participants are more precisely defined here and identify specific factors and people involved.

Material process types related to procedures are evident in both stages. Not surprisingly, the larger PBL2 corpus with a greater focus on informational content also provides a wider variety of keywords. This focus on procedures also reflects the purpose of the report stage, which is not only to identify the genetic causes of conditions but also to establish the protocols medical geneticists should employ, from testing procedures to diagnosing conditions through to counselling of patients who may be affected. This focus on "practical ends" and "techniques" is common in applied fields (Becher, 1989, p. 15).

Along with the applied practical nature of this field, we find the role of the geneticist evident either explicitly in active constructions (9, 10, and the second instance in 11) or implicitly (12). By using a generic personal pronoun, this appears to allow the students to align their identity with their (future) professional work.

> (9) S27: the mutation because *we* cannot *test* if *we* do not know the types of mutation (PBL11.1)

> (10) S27: fertilisation, no implantation nothing. whereas when *we terminate* a pregnancy there is a foetus that has already begun (PBL12.2**)**

> (11) S14: our test could *detect* ninety percent of affected cases. that's sensitivity of the test. and *we* could *detect* ninety nine per cent of (PBL10.2)

(12) S32: but still not sure because the same mutations have also been *identified* in VWS (PBL 12.2)

We can add to the material process group examples where a dummy (delexicalized) verb is used and the nominalised (keyword) form can be described as the range of a process. Here there is a close relationship between process and participant "where the verb is emptied of its content, and the meaning expressed though the nominal Range constituent" (Eggins, 2004, p. 219). In Examples 13, 14, 15, *test, sequence* and *risk* are all nominal range constituents of *do* (13, 14) and *have* (*15)*. Interestingly, in the example with *sequence* it is used both in a nominalised and main verb form (this pattern of mixing is not uncommon in the keywords studied).

(13) S13: like if you're *doing* a screening *test* and you found that they had, they were positive for Spina Bifida (PBL9.1)

(14) S26: you'd then *do* a *sequence* of that gene you wouldn't *sequence* these two (PBL11.2)

(15) S15: so you're saying the, babies going to *have* a high *risk* of a lump? (PBL9.1)

In addition to being range constituents, a significant number of these nominalised verb forms in the role of participant co-occur with other nominalised nouns (16) or other material process verbs (17).

(16) S23: but is it's a positive result the first you got to, report as a high risk and give *genetic counselling* (PBL10.2)

(17) S12: Protein karyotype will *detect* large *deletions* and translocations (PBL12.2)

The related verbs can be subdivided into those related to the actions/changes occurring in natural processes and those bringing in the geneticist, focussing on research and procedures. For example, *mutations* can *cause or affect* (a change/condition), but they can also be *found* and *identified (*in Examples 19 and 20 the passive form is used with an implicit agent).

(18) S2: this is the kind of hot spot, for *mutations* that *cause* severe FAP (PBL7.2)

(19) S32: because the same *mutations* have also been *identified* in VWS (PBL12.2)

(20) S23: if one *mutation* is *found* it will also be followed by the second IRT test (PBL10.2)

As a participant in the clause structure and in subject position, this makes the entities the main focus and as actors in the clause structure highlights their agency.

In both stages, but most notably in PBL1s, we find a group of material process keywords related to study. In PBL1s, the focus is on essentially sharing or exploring informational content and on establishing what they will need to do. Examples of keywords include *check* and *separate* (a learning objective). Here the actors are the students themselves and again commonly used with personal pronouns (*I* and *you)* (21).

(21) S1: and then if you want like if *you* want to *check* the database, then then *check* it and then you know if *we*'re interested *i*'ll probably *check* it too (PBL12.1)

In PBL2s, overt reference to study is illustrated with *found*. In addition to being used as part of a procedure (e.g., as in found in the results of a test), it is also used to introduce research findings from research experiments (i.e. active scientific discovery and so described as a material process):

(22) S12: er *it's been found* that there's an increase er in the level of AFP in pregnancies with er open NTD (PBL 9.2)

These study related processes are all about specific actions. The other processes related to study were found in the mental process group.

4.6.3.2 Mental Processes in PBLs

One of the most notable differences in the keyword lists between PBL1s and PBL2s is the set of mental process verbs evident in PBL1s**.** The keywords representing these mental processes reflect the student concerns with hypothesising and presenting their own position in relation to the specific scenario. That they are tentatively expressing possible issues to explore is shown with *think,* which has the highest keyness of the verbs in the HF group, *wondering*, *suppose*, and *know*. These mental process verbs nearly always follow a personal pronoun, most notably and overwhelmingly *I*, followed by *you*: the students' inner thoughts are clearly 'present' in the PBL1 discussions, as is the involved nature of the interactions with reference to other participants (Examples 23 to 26 below).

(23) S18: i *think* it's e- OMIM i don't *think* it's a search for, fea- (PBL12.1)

(24) S7: i was just *wondering* whether we are sure that (PBL7.1)

(25) S1: so i *suppose* if it is a de novo mutate (PBL7.1)

(26) S12: do you *know* which abnormality causes Spina (PBL9.1)

In contrast, the PBL2 keyword list has only two mental process types, *associated* (27) and *found* (28). *Associated* is used here to indicate a mental connection of cognition (Fontaine, 2012, p. 75) and is presented in a passive form with an implicit senser. *Found* is used to talk about what the student learned in their own independent study.

(27) S22: also cardio abnormalities [?], or em, problems with thumb and so it's *associated* with syndrome. (PBL12.2)

(28) S8: but i *found* that oligo microarrays is the best one, for detecting er for the resolution, it's the most improved one, i don't know if you can see (PBL 7.2)

We also see one verbal process, the process which lies between relational and mental, with *mention* used to refer to earlier points a student has made in their talk. The speakers use mention with a metadiscoursal role of reminding listeners and guiding them through the talk, particularly important in the longer reports in PBL2s.

> (29) S5: again those cancers i *mentioned*, endometrium ovarian pancreas, and should be synchronous means have the same time, or metachronous that means er, can happen before (PBL11.2)

> (30) S38T6: uh yeah i have a question, uhm you you *mentioned* empiric risk so how is that one out of fifty generated? (PBL12.2)

The mental processes reflect the ongoing thought processes, interactions and construction of knowledge in PBLs and the independent study that becomes evident in PBL2s. The mental processes are most commonly found in active constructions with personal pronouns, i.e. the students as the active agent.

4.6.3.3 Relational Processes in PBLs

Relational processes have been shown to be common in scientific writing (Halliday & Martin, 1993; Love, 1993). Although they are not apparent in the process types in the PBL keyword lists of the main lexical verbs[18], relational processes are found to co-occur with keywords that are nominalised verb forms. Where they occur, they serve to identify and relate specific features and characteristics.

Identifying relational processes enables the students to specify types, as in the examples below of *mutation* (31) used with *showed* (i.e. replaceable by *to be* with no change in meaning) and the *test* type in Example (32).

> (31) S8: and they found that sixty-eight percent of the cases *showed* an IRF six *mutation* (PBL12.2)

[18] *It's* is included in the high frequency list for PBL2 but as part of a contraction; it has not been considered here as the focus is on the main lexical verbs.

(32) S19: that's our [?] yeah that's already diagnostic because, the *screening test is* just the the IRT (PBL10.1)

Attributive and possessive relational processes are also found, allowing for comment on significance (33) and encoding possession (34) (Eggins, 2004, p. 247).

(33) S1: i suppose the *counselling is* quite important in this family (PBL12.1)

(34) S20: so they they won't *have the* same *mutations* and stuff so you won't be able to it's just (PBL10.1)

4.6.3.4 Summary of Process Types in PBLs

The consideration of the process types as exemplified in the keywords further emphasises the focus on procedures and natural processes found in the semantic groupings. They indicate not only the concerns of this applied discipline, i.e. to find causes of genetic conditions and the treatment and the protocols associated with these but also the specific pedagogy underlying problem-based learning. This latter point is most evident in PBL1s with the mental process types but also to an extent in PBL2s with reference to findings from independent study. The change in focus from exploration and questioning in PBL1s to more specific procedures and identifying casual connections in PBL2s also reflects the specific aims of the staging in a PBL cycle.

In PBL1s, the students' representation of their inner states and thoughts are also clearly reflected in the mental process type verbs. This type of process is much less prominent in PBL2s. This may be because students are concentrating more on conveying the subject matter, or because the focus of the second stage is not on hypothesising and sharing potential knowledge, but more on demonstrating learning from background reading through discussion of specific processes and procedures. In this way, the pedagogical focus and aims align neatly with the focus of the discipline (i.e. the aim being to identify and discuss relevant subject content).

Students represent natural processes of the human system and procedures used by geneticists mainly through material process types. Where verbs have been nominalised, these nominalised participants also quite commonly co-occur with other keywords and also material processes. This again reflects also the applied nature of medical genetics: activity is prominent and emphasised. Relational processes are found, although at least in relation to keywords are less frequent; they are used to identify specific features and characteristics.

The results here reflect previous findings of studies that have found that the process types reflect specific concerns of the genre and subject matter. Da Silva and Dennick (2010) also found changes in focus, for example students talking about diseases in the early stage of a PBL cycle, moving on to more specific treatments in later stages. While Love (1993) identified a schematic structure related to processes and products in geology reflected in the transitivity system, in PBLs we find a concern with natural processes and procedures. Where Martinez (2001) found different processes reflecting different aims of stages in research articles, including a focus on action in the methods section, in PBLs we also find a move from the more general and exploratory questioning in PBL1s to the more specific focus on practical procedures in PBL2s. In both stages, actions and active constructions are common. These points reflect the specific spoken register and also the different aims of each stage.

4.7 Chapter Summary

The study of the keywords aimed to identify the distinct 'aboutness' of the PBLs in relation to a reference corpus of academic seminars and lectures from a range of domains. The proportion of keywords that can be identified as academic, either core or technical, the semantic grouping exercise, and study of process types all contribute to this picture of aboutness.

We can see here that both the subject matter and the type of interaction, particularly the informal and interactive nature of PBL1s, are clearly illustrated in the keywords, not least in the higher proportion of high frequency words, including those which specifically indicate oral interaction (e.g., response

words) and semi-lexical items. We can also see the specific focus and link in content between each stage in the PBL cycle illustrated in the keywords.

The academic technical words clearly identify the specialist subject area. Here we also see a preponderance of nouns, common in academic discourse and highlighting the centrality of the subject matter. These words also give insights into the specific concerns and focus of this field. In both stages we find keywords related to natural processes and procedures, becoming more specialised and varied in the second stage. In PBL2s we see a higher number of keywords, to be expected in a larger corpus. This not only reflects the subject area but also the purpose of the communicative event: sharing their findings from research and thus extending their knowledge. The AC keywords were also seen to closely align to the discipline, although no significant differences in terms of part of speech or meaning were identified in the explorations carried out here.

The semantic grouping and the exploration of processes reflected many of the academic functions identified in Chapter 3, i.e. describing conditions, natural processes and procedures, illustrating the disciplinary concerns and focus in medical genetics.

The study of process types also reflected the focus of this applied field/discipline, the epistemology of PBL and the specific type of communicative event. In the study of process types found in the AT, AC and HF keyword lexical verbs and nominalised forms, the most notable difference between PBL1s and PBL2s is the inclusion of a number of mental processes, with the stance and internal concerns of the speakers more prominent in PBL1s. The students are more focussed on reflections and suggestions in PBL1s compared with a more concentrated focus on reporting on precise content in PBL2s. This reflects both the epistemology and also the type of interaction in each stage of the cycle.

This chapter has only considered positive keywords in PBLs and does not look at total word frequency or coverage in relation to academic word lists. However, the keyword approach has provided a distinct set of items that inform us about the concerns of the subject matter and the mode of communication. They also

indicate the way in which students start to position themselves as an active member of the profession through their use of active constructions and particularly personal pronouns. The use of personal pronouns is the subject of the next chapter.

Chapter 5 Personal Pronoun Use in PBLs

"No no no *i* know what *you*'re saying but shall *we* see if, *we* can get two questions if can't then *we*'ll split that into two." (S26, PBL11.1)

5.1 Introduction

As we have seen, PBLs are fundamentally about interaction and engagement both with the propositional content and fellow participants. First and second person pronouns are one recognised marker of involvement, engagement and orientation to the audience (Ädel, 2010, 2012; Biber et.al., 1999; Hyland, 2005a). This chapter considers the use of *I, you* and *we* in the PBL corpus. The first section provides an introduction to their use and a summary of relevant studies in spoken academic English. Their frequency and aspects of use in PBLs are then presented.

5.2 First and Second Person Pronouns

First and second person pronouns and related forms are common not only in everyday conversation but also in spoken academic English, indicative of the interactive and involved nature of many forms of spoken discourse (Biber et al., 1999; Biber, 2006b; Carter & McCarthy, 2006). These personal pronouns are also very often interpretable with reference to the context in which they are used, particularly in the case of *I* and *you* (Carter & McCarthy, 2006; Halliday & Hasan, 1976, p. 48). Their frequency and aspects of use have been studied in large scale corpora, for example the Longman Spoken and Written English Corpus, Cambridge English Corpus (CEC), and the T2K-SWAL; within academic writing (Harwood, 2005; Hyland, 2001; Kuo,1999; Thompson, 2001), and to an extent within spoken academic English, most notably lectures, conferences and colloquia (e.g., Ädel, 2010, 2012; Fortanet, 2004, 2005; Morell, 2004; O'Boyle, 2014; Rounds, 1987a, 1987b; Webber, 2005;). The next section will outline the use and frequency of personal pronouns and summarise related research focussing on spoken academic English to date.

5.2.1 First and Second Person Pronoun Use

The prototypical use of first and second person pronouns presented in general and student grammar reference books of English identifies the first person singular as referring typically to the speaker, with the first person plural also including the addressee(s); the second person singular refers to an addressee, with the plural (in the same form) to addressees. (Carter & McCarthy, 2006; Quirk et al., 1985; Yule, 1985). In addition to referring to specific addressees, the second person singular and plural (*you*) can refer to any potential addressee (i.e. not necessarily present), and can also include the speaker (generic /general use) (Carter & McCarthy, 2006). Descriptions of the first person plural *we* also commonly differentiate between inclusive and exclusive uses, the former referring to the speaker and the addressee and others not necessarily present, the latter referring to the speaker and (an)other, potentially non-present, member(s) of a group, but not including the addressee (Carter & McCarthy, 2006; Halliday & Hasan, 1976; Kamio, 2001; Levinson, 1983; Wales, 1996). In all cases, the referent, used here to mean the person the pronoun actually refers to (as opposed to a presupposed item in the text, an antecedent), may be identified as a mental representation (e.g., Emmott, 1997) or directly evident from the situational context (Carter & McCarthy, 2006; Halliday & Hasan, 1976; Levinson, 1983; Lyons, 1977).

In explaining how pronouns are interpreted, Lyons (1977, p. 638) points out that "the canonical situation of utterance is ego-centric", i.e. the centre of the utterance is the speaker and referents are viewed from their standpoint. Kamio (2001), in a discussion of the generic uses of *we, you* and *they* in English and Japanese, also notes the speaker as the centre, discussing pronoun use in terms of "speaker territories". He concentrates on the proximal and distal spheres, similar to Lyons, with *I* and *we* in the (psychologically) proximal domain of the speaker (and where the speaker regards the addressee as part of some group or formed alliance), and *you* in the distal domain of speaker (but proximal domain of the hearer). He suggests that *I* and *we* indicate (psychological) closeness, *you* less so. He does note, however, that this is not always the case. In terms of the use of *you* he identifies two groups: those where there is a clear contrast (to *we*) and those where the use of *you* and *we* is comparable (a near, but not exact,

synonym, when the speaker is not aware of a difference). In such cases he suggests the territories of the speaker and of the hearer can almost merge (Kamio, 2001, p. 1120). For Kamio, my understanding is that the primary concern is with differentiation and contrast: the extent to which terms can be substituted and the closeness of reference.

In addition to consideration of referents, *I* has been identified as an indicator of subjectivity and of stance (Biber et al., 1999; Hyland, 2005b). The second person *you* in its generalised form may be used as a way to move from specific to general, to expand the scope and authority of statement (Scheibman, 2007) and as a means to invoke membership categories (e.g., Stirling & Manderson, 2011). The use of *we* has been connected with inclusiveness (e.g., Rounds, 1987a, 1987b). These all indicate how the pronouns can be used to position the speaker in relation to the content and the addressees.

5.2.2 Frequency and Use in Academic English

In the LGSWE, [19] Biber et al. (1999) states that the first person singular and second person pronouns are much more frequent in conversation than in other registers (*we* is more evenly distributed), with subject pronouns repeated more, reflecting 'real-time' production, and the associated false starts and disfluencies (pp. 333-335). In terms of academic English, Biber (2006b, p. 51) shows that spoken university registers rely on pronouns much more than written registers, with first and second person pronouns found in all the spoken registers they researched in the academy.

In a study of university language, Biber (2006b, p. 4) notes the pronouns *I, we* and *you* as one of the "features that directly acknowledge and engage the audience". Carter and McCarthy (2006, p. 284) state that in academic English, both written and spoken *we* is used to create "a sense of an academic community shared by all participants in the discourse". In addition to general large-scale corpus findings which have identified general patterns of frequency and use, the research into the use of pronouns within specific academic speech events has also included identification of referents, discourse functions and

[19] Conversation, fiction, news and academic registers are compared in Biber et al. (1999).

common linguistic environments (clusters/lexical bundles and collocations) in which they occur. These are discussed in the next section.

5.2.3 Frequency and Referents in Spoken Academic English

Rounds' (1987a, 1987b) study of a small corpus of five mathematics lectures delivered by native and non-native speaker teaching assistants is perhaps still the best-known research into personal pronoun use in spoken academic English. She points out the importance of pronouns and the deictic system in linking language use to context, going on to identify various referents of pronouns. She comments on their role for the lecturer not only in imparting knowledge of the subject but also in *establishing roles, power relations* and *solidarity*. In identifying the different roles the speaker takes on as represented in their pronoun use, she expands the traditional semantic mapping to offer a more fine-grained analysis.

Rounds (1987a) identifies examples where *I* = *they* ("I'm going to define", but where the speaker is not the originator of the definition, i.e., it should be *they define*) and where *I* or *you* = one ("if you take zero", i.e. not exclusively the speaker or the addressee, but a general use of *you* noted above). In addition to the traditional mappings of *we* = *I* + *others* (exclusive) or *we* = *I* + *you* (inclusive), she identifies the following uses:

> *We* = *I* ("we said", but actually the speaker said, not the audience)
>
> *We* = *you* ("some of the problems we had", but not actually problems for the speaker)
>
> *We* = *anyone doing calculus,* with a possibly substitute of 'one' ("*we* (mathematicians) call that number")
>
> Rounds 1987a, (pp. 17-19; my italics)

This final use appears to me to be the same as the traditional, more generic inclusive use (*we* = *I* + *you* and possibly others) because the lecturer can be part of the group as can the students, if not in an expert capacity. I find the

discussion of this final use in Rounds and later by Fortanet (2004) unclear. Rounds states it is a use where an indefinite *one* could be considered "since the scope of the potential addressee goes beyond simply the copresent participants" (Rounds 1987a, p. 19), and also suggests in her semantic mapping *anyone who does calculus* = pronoun *I, you, we*. This seems to me to suggest an inclusive use. Fortanet (2004, p. 48) in contrast, describes this use of 'anyone' as an example of 'exclusive *we*', "since the hearers, the students, are not included in the reference scope". Perhaps the first inclusive *we* needs to be differentiated and specified to include only *we* + *you* and *I*. This still provides 5 semantic mappings, but differentiates between a very specific *we* = *I* + *you* (only) and a more encompassing *we* = *I* + *you* + anyone else involved in this field.

Rounds' (1987a, p. 21) focus was on identifying the features of lectures by successful teaching assistants.[20] She found that assistants who were identified as being more successful used *we* (around 62% to 65% of the time) significantly more than *I* or *you* (only around 20% to 15% of the time). The use of *we*, she suggests, is to increase a sense of inclusiveness and engagement (substituting *we* for *you* being seen as more "distancing"). It also allows the speaker to change roles easily from the exclusive *we* (*we=I/self* and other mathematicians) to the inclusive (*we* = *self* + *you/audience*), thus being the more "egalitarian choice", allowing the lecturer to show solidarity with the different groups (Rounds, 1987b, p. 649). She notes the use of pronouns is not only related to politeness and the need to find common ground but also the need to develop an atmosphere of consensuality for educational purposes, the context and genre clearly shaping the choice and use of language (Rounds,1987a, 1987b). However, this work focussed on teaching assistant lecturers (as opposed to experienced lecturers) and was carried out over 25 years ago. Most importantly in relation to my corpus, it did not investigate student-centred discussions.

More recently, Fortanet (2004, 2005) also has researched the use of these personal pronouns in a larger academic corpus. She compared the frequency of *we* with other pronouns in university lectures and colloquia in the MICASE corpus, and then, in order to compare with Rounds' mathematics lectures, in an even smaller still sub-corpus of only mathematics related speech events drawn

[20] Rated by the mathematics department supervisors.

from MICASE. In her research, she studies the linguistic contexts (clusters), referents, and discourse functions. Adjusting Rounds' results to allow for a comparison of frequency counts, she finds that in stark contrast to Rounds, the frequencies for *I* and *you* in her sub-corpus of MICASE were almost double the use of *we*. In the smaller sub-corpus of maths related classroom speech events she noted a much higher use of *I* (than *you* and *we*). Her results do not support those of Rounds. She concludes the use does not seem related to disciplinary (maths) instructional concerns but is perhaps related to the increased acceptance of *I* in research articles (Fortanet, 2004, p. 93). She does not, however, expand on this point.

Fortanet also uses the frequency data to compare the relative use of inclusive and exclusive *we*. Her results suggest that overall the inclusive *we* is more often used in academic speech than the exclusive *we*, and, like Rounds, she suggests that when it is used it is done to involve the audience. The referents of *we* and their discourse functions are considered and although she points out difficulties in ascribing functions to many instances (for example, after a pause or a change of direction, and with reported speech), she separates these into two categories: "meta discourse" (to guide the audience through the talks) and "representations of larger groups of people" (the latter being the most common function of *we).* Although I appreciate the difficulties of categorising all instances, this seems a rather generalised discourse function.

Having found more examples of *you* and *I* than *we*, Fortanet (2005) reports on an extended study to investigate the discourse functions of *I* and *you*. Again in contrast to Rounds her findings indicated interaction in lectures is related to a higher frequency in the use of the pronouns *I* and *you* (interaction here referring to actual levels of involvement). The use of *I* was also identified as being more common in interactional lectures, *you* more so in the more monologic lectures, and that *I* was often used to signal opinions and attitudes of the speaker.

In considering the semantic mappings, Fortanet (2005, pp. 44-45) found the impersonal use of *you* higher than specific personal reference. She also found that the impersonal *you* can have four "hidden" referents: *they* (identified in the context), *people* (indefinite), *we,* or *I* and concluded that when *you* = *they*

or *people,* this is used to engender a greater sense of involvement and so reduce the distance between the speaker and hearer (the alternative *they* being more depersonalising and distant). Fortanet also identified the occurrence of *I mean* and *you know* as pragmatic markers and did not include these in her categorisation of referents.

In summary, Fortanet finds higher frequencies of *I* and *you* compared to *we*. The impersonal *you* is used to either create distance or to create a sense of involvement, with *we* used mainly to organise the discourse and to represent a larger group. As the speaker needs to identify themselves and their stance, this explains the high frequency of *I*.

Following Fortanet's investigations into use of *we* in MICASE, Chapman and Wulff (2010) in a kibbitzer look a little further into the question of whether the frequency and use of *we* is related at all to academic disciplines.[21] Compiling a sub-corpus of a variety of speech events across subjects in the physical science disciplines from MICASE, they found similar (although slightly lower) frequencies in the use of *we* to Fortanet. Although they acknowledge this was a small-scale study and without an in-depth analysis of the discourse functions and referents, they do find that the one speech event with a higher than average use of *we* was a group study session. They suggest the higher frequency in this event "may indicate a possible positive correlation of the frequency of *we* with informal registers" (Chapman & Wulff, 2010, p. 3).

In another academic speech event, Webber (2005) compared the use of a number of linguistic features in scientific conference monologues to research articles and reviews on the same topic. The aspects considered in the study included personal deictics specifically chosen as they are audience-orientated. Although the monologues are an academic speech event, she highlights important differences from a lecture, most notably the composition of audience and the potentially different power relations from a 'regular' university lecture (a conference being more likely to be expert-to-expert); that the talks on works-in-progress while partly planned may include more discussion, and are by their

[21] A kibitzer is a discussion page of a language problem which allows others to observe and comment on a problem.

very nature often an unfinished product open to discussion, unlike an article (this reflects in some ways a PBL, albeit the latter without the same level of expertise). Webber found extensive use of personal pronouns (*I, you* and *we*) and states this "is typical of non-scientific language and closer to spontaneous spoken discourse" (p. 159). She suggests that they are used to "create empathy and bond the group" (p. 173).

Although Webber found a much higher use of *we* than Fortanet, she suggests the inclusive *we* may be used for "reasons of deference or affiliation", to indicate shared knowledge, e.g., "as we know", (Webber, 2005, p. 163), and the use of exclusive *we* when speakers may be discussing and presenting research carried out as part of a team. In terms of reference, she notes that personal reference was much more frequent than the impersonal use for *you*, and used to refer to potential patients or the researcher (and for the latter being inclusive). This personal use of *you* directly referring to the audience was used, for example, to refer to visuals (particularly important in these talks) or to emphasise specific aspects of the data. The impersonal use of *you* to refer to patients or fellow professionals is, she suggests, used to identify with the audience as part of a specific group (p. 163). With regard to the first person singular pronoun, she found "many instances of personal reference in the data, particularly *I think*, used either for purposes of hedging or for declarations of stance" (p. 159). She specifically excludes markers such as *I mean* and *you know* as not at all common in her data but more common in casual conversation, although noting the high collocation with *know*. Broadly speaking, Webber identifies uses for structuring the discourse and for engendering a sense of involvement and group identity.

Morell (2004) includes personal pronouns in her study of the textual and interpersonal discursive aspects of three interactive and three non-interactive lectures.[22] Primarily concerned with identifying features that make lectures more interactive (and thus facilitating comprehension for non-native speakers), she found interactive lecturers made more use of pronouns overall and particularly of the second person pronoun *you* (singular and plural) and *we* compared to the non-interactive ones, which had a much higher frequency of *I*. This appears partly to support the findings of Rounds in the use of *we* in

[22] Lectures in an English-medium programme in Spain.

interactive lectures and the higher use of *I* in non-interactive lectures but not in the use of *you*. Morell suggests that the use of these pronouns reflects a more inclusive style, reducing the distance between the speaker and students, between the students and the content, and "can be interpreted as an effect of personalization" (Morell, 2004, p. 335).

Okamura (2009) in a comparison of undergraduate lectures and public lectures, again from the MICASE corpus, finds the second person plural *you* the most common pronoun, compared with *I* which was more frequent in public lectures. These differences in frequency, he suggests, reflect the purpose of each event: the student lecture passes knowledge on, compared with a public lecture where the aim is to present the speaker's view. He suggests *you* is used to engage students in the talk.

The use of personal pronouns has also featured in more specific studies related to metadiscourse. Ädel (2010) selects personal pronouns (*I, we* and *you*) as a marker of involvement in her study of reflexive metadiscourse in both writing and spoken academic discourse (specifically lectures and essays), and *you* as an indication of audience orientation. In both studies she notes *you* is used in a range of functions in the lectures and that a number of these, not surprisingly, were clearly related to the mode of delivery (i.e. lack of time to plan and the presence of the audience, repair, marking of asides and contextualising were only found in the lectures).

O'Boyle (2014) compares the use of *you* and *I* in a corpus of university seminars with a smaller corpus of English language learner talk. Noting their use to mark stance and shared knowledge, she finds that the learner talk varies in use from NNS and lecturer talk. Overall, she found *you* most common in both corpora, particularly in tutor talk. Cluster analysis shows high use of *I think* in language learner talk along with first person pronoun repetition, and *I don't know* in NS student talk. This study however, concentrates on comparing language learner talk with native speakers and does not focus on specific disciplines.

These studies have shown that pronoun use can vary depending on the type of interaction, that semantic references are more complex than prototypical

mappings and can indicate for example, group solidarity, attempts at inclusive engagement, involvement and attention to the audience. Some also indicate pronoun use in relation to clusters, the use of which is the focus of the next section.

5.2.4 First and Second Person Pronouns: Collocations and Clusters

While the following studies into personal pronoun use are not specifically academic speech events, they are included here as they provide additional insights into a range of linguistic contexts in which they may appear.

Baumgarten and House (2010) investigate the use of the collocations (their term) *I think* and *I don't know* as stance markers (following Biber, 1999) in a corpus of native and non-native speaker conversations using English as a lingua franca.[23] They argue that the use of first person pronoun is very clearly an indicator of the subjectivity and the speaker's stance.

In a cluster analysis of the spoken corpus of the CEC, Carter and McCarthy (2006) find two of the most common discourse markers are *you know* and *you see*. They see this as evidence of the speaker attending to the addressee, either to check the state of shared knowledge (*you see*), or with *you know* because it "projects the assumption that knowledge is shared or that assertions are uncontroversial … or checks that the listener is following" (p. 221). Erman (2001, p. 1339) includes *you know* and *I mean* as pragmatic markers which "have little or no meaning in themselves and can only be understood either through clues in the context and/or situation". In her review, the function of *you know* in discourse she suggests that the most important function of pragmatic markers is that of textual monitors, for example to change the direction of discourse in order to make it coherent and to guide the listener, thus showing audience awareness. They also provide a text encoding function, reflecting on-line processing as the speaker selects the right linguistic resources, commenting on the use of *I mean* in this role also (p. 1340).

[23] Selected as being the most frequent *I* + verb (*I think*) and *I* + negative collocation (*I don't know*) in the BNC and COCA corpora.

Macaulay (2002) also investigates the use of *you know* as a discourse marker in two data sets of conversations in Scotland. He says that although much of the work on *you know* as a discourse marker assumes that the core meaning is retained and is based on an assumption of shared knowledge, his own findings do not support this. He claims that essentially most of the meaning has been "bleached out" (p. 755). However, while they do act as discourse markers, they also indicate the presence of others and, I would tentatively suggest, this vagueness is similar to many of the generic impersonal examples of *you*. Scheibman (2002, p. 76) also comments that a "fixed expression … is always slightly referential as it tacitly refers to the addressee".

These studies demonstrate some of the uses of personal pronouns, in particular *I,* as an element of a stance marker, as components of pragmatic markers and in structuring discourse, e.g., with *you know*. They can indicate the speaker's engagement with the content, and their orientation to the audience and can be a reflection of the mode.

5.2.5 Summary of First and Second Person Pronouns

First and second person pronouns are acknowledged as markers of engagement and interaction, and also orientation to the addressee. A range of studies of both general and academic speech events have indicated their significance, although variations in frequency and use have been identified. We can see that they have a number of potential referents beyond the prototypical. These may be impersonal and, in the case of *you* and *we* in particular, can be seen to have "hidden" referents. In a brief consideration of their linguistic environments, *I* has been noted as a stance maker and as part of pragmatic markers with *I mean* and *I see*, as is the second person in *you know*. All can be viewed as a reflection of the engaged, involved and interactive nature of spoken discourse.

However, while these personal pronouns have been considered in larger scale and general corpora and in certain types of academic speech event, there does not appear to be any significant work on student-centred, disciplinary-specific academic speech. As PBLs are quite different in purpose, levels of interaction and in the power-relations between the main contributors from lectures, it is

reasonable to assume that there may be some differences in use. The next section goes on to present the investigation into frequency and aspects of use in the PBL corpus. It attempts to address the following questions:

- What is the frequency of *we*/*you*/*I* and related forms in PBLs overall, and within PBL1s and PBL2s?
- Is there a significant difference between PBL1s and PBL2s?
- What referents can be found for *you* and *we* in PBL1s and PBL2s?
- In what linguistic context does each one appear?

5.3 Methodology

A qualitative and quantitative approach is taken in the study of the corpus. Both the whole PBL corpus (AllPBLs) and the two sub-corpora (PBL1s and PBL2s) were analysed using WordSmith Tools (Scott, 2004). The quantitative analysis identifies three sets of data.

- The overall frequency of each item, *I, you* and *we,* and related forms. [24]
- The most common collocates immediately to left and right of each pronoun.

Frequency of use (per thousand words) is compared with the estimated figures for both the Fortanet (2004) sub-corpus of MICASE and Rounds' corpus (reported on in 1987a, 1987b). The qualitative analysis considers how the pronouns are used in terms of reference (generic, impersonal or person(s) specific, where relevant) and in relation to the aims of PBL sessions.

5.3.1 Coding and Categorisation

The concordancing tool in WordSmith, which allows for examination of the fuller co-text when a concordance line is insufficient or potentially ambiguous, was used to facilitate the subsequent qualitative analysis. This was primarily

[24] *You* and related forms (*you, your, yours, yourself, yourselves*), *I* and related forms (*I, me, my, mine, myself*) and *we* and related forms (*we, us, our, ours, ourselves, let's)*.

concerned with the semantic mapping exercise of *you* and *we* to identify the referents, i.e., either a context-specific deictic referring to a specific individual or specific groups (e.g., students in these sessions, staff on the programme), or an impersonal general use. A brief explanation of the coding system that was developed is now provided, and then presented in Table 5.1.

I refers to the speaker, apart from when used in a quoting function (see disregarded data below).

Each occurrence of *you* was categorised as referring to either a specific individual *YS1* (i.e. you + specific; group 1), *YGS2* referring to this (present) specific group, or *YG* as a generic, impersonal *you*, referring to anyone (in this area). There are also numerous examples of *you know*, which I have included as a sub-group of *G(YK)*, i.e. a general *you know*.

At times the distinction between the specific group in the room (*YGS2*) and a wider of students on the course (which came under *YG*), was not always clear. Rather than disregard these or be left with a large group categorised as 'unclear', the following practical solution was applied. Each item was coded with the two alternatives on the first reading. This was followed by a second reading a week later. The category assigned this stage was used (where two categories were still thought possible, the recordings were checked again). This approach was decided upon for two reasons. One was the practical difficulty of finding a second person to categorise the data[25]. The second was that while on first reading there appeared quite a large number that fell into the 'unclear' category, I found I could in fact categorise most on the second reading. This approach of reading and returning to the data at a later stage for ambiguous cases was taken for other ambiguous items throughout the coding process.

An example is provided in (1). While *you* does sometimes refer to all students on the programme, I would suggest in this instance it was directed at the specific group and has therefore been coded *YGS2*.

[25] This would have required a considerable commitment of time along with practical issues involved in accessing the data.

(1) S24T3: ok, don't *you* have a a sheet that gives *you* the location of *your* (PBL10.1)

In Example (2) the question was related to the preceding presentation. This could be specifically to the student who spoke but could also be for the group.

(2) ST24: picture of um IR six expression in a in an embryo, can you tell me what what are *you* actually looking at there expressions of (PBL9.2)

On the second reading and viewing, this was coded as *YGS2* as the facilitator appeared to be directing this to this group at this specific time.

Referents for *we* were categorised into one of five groups, two 'exclusive' uses, *we* = I and others (exclusive of addressee), *we* = I (exclusive of addressee), and three 'inclusive' uses: *we* (including audience/addressee) where *we* = you and I (this group); *we* = you (actually excluding the speaker essentially); *we* = anyone in this field.

A number of examples in each sub-corpus were disregarded at the semantic reference coding stage either because there was a lack of contextual information, even with consideration of the co-text (e.g., 'so *you*'); because they were false starts ('this *i* already, three types of'); included repetition of the item ('putting, *we we* will say that'); when they were used in a quoting function and thus not the actual speaker or addressee(s) in the PBLs (e.g., 'going to be like, *we* tested P and'); three instances in the *I* concordance regarded as part of a fixed phrase ('oh *my* God'); or were not actually a pronoun but instead an initial ('then you can say that J and I don't have').

The categories for *you* and *we* are presented with examples in Table 5.1 to follow.

Table 5.1: Semantic Categories for *You* and *We*

Category	Use/Refers to	Example
Semantic Categories for *You*		
1. YS1	Specific individual	sorry what do *you* mean by big?
2. YGS2	The present specific group	so if *you* look at the first figure
3a. YG	Generic, impersonal you/ anyone (in this area)	like if *you*'re doing a screening test and *you* found that
3b. YG(YK)	Part of *you know*	for CF *you know* if someone, say you met someone
Semantic Categories for *We*		
1. We	I and others (exclusive of addressee)	i think that's the one that *we* requested
2. We	I (exclusive of addressee)	(No examples were found in the data)
3. We	You and I (this group; inclusive)	so do you think, do *we* have do you think *we*'ll have to do like
4. We	You (excluding the speaker)	treatments *we*'ve got, (2) i think the the testing could be
5. We	Anyone in this field	or *we* can just go for the colonoscopy if she wants to be tested

Unfinished turns and disfluencies reflect the nature of spoken English and the fast moving highly interactive nature of the PBL1s and on-line processing in the PBL2s. However, the items were not included in this part of the analysis in order not to unduly inflate the numbers. While I have excluded 'oh *my* God' I have chosen to retain such phrases as *I mean* and *you know*.

5.4 Results

This section considers first of all the frequency of the pronouns, followed by the results for the semantic mappings. It finishes with a consideration of the most frequent collocates.

5.4.1 Overall Frequency

In consideration of the frequency word list for the whole corpus and each sub-corpus, we can see in Table 5.2 on the next page that the use of the personal pronouns *I, you* and *we* is high. Overall, *you* is the fourth most frequent word, *I* the fifth, and while a little less frequent, *we* is still the sixteenth most frequent word in the whole corpus (AllPBLs). This pattern is largely reflected in each sub-

corpus but with minor variations. In PBL1s *I* is the fourth most frequent word, *you* the fifth, and *we* is the tenth. In PBL2s, this time *you* is the fifth most frequent word and *I* is the seventh. Again, somewhat less frequent is *we,* at twenty-seventh. The biggest differences are in the frequency and ranking of *we*.

Table 5.2: Ranking and Raw Frequency of *You, I* and *We*

	Ranking	Frequency
AllPBLs		
You	4	2361
I	5	2297
We	16	1177
PBL1s		
I	4	1095
You	5	1031
We	10	716
PBL2s		
You	5	1330
I	7	1202
We	27	461

Note. Raw figures before disfluencies removed and contractions added.

In order to compare with previous studies more easily, the frequency per 1000 words was calculated for PBL1s, PBL2s and for the whole corpus. Fortanet (2004, p. 51) provides the frequency per thousand words for her sub-corpus of MICASE and calculates also for the Rounds corpus; as the corpus size is provided by Webber (2005), I have similarly calculated the frequency per thousand words for the pronouns she studied. It should be noted that while these do not allow for an exact comparison (e.g., there are more related forms), they can provide an indication of similarities and differences in use.

Table 5.3 to follow provides an overview of all *I* and related forms, *you* and related forms and *we* and related forms (henceforth referred to as *I*-forms, *you*-forms and *we*-forms*)* from across the different corpora. It provides the frequency along with normalised frequencies (per thousand words).[26] In terms of overall frequency, *I*-forms and *you*-forms are significantly more frequent in the whole corpus (AllPBLs) and within each sub-corpus than *we* (*I*-forms have a frequency

[26] The full breakdown of all related forms is provided in Appendix 4.

of 23.7 per 1000 words in AllPBLs, *you* forms are slightly lower at 22.4, with *we* forms standing at only 12.1). These overall figures reflect the patterns found in the Fortanet (2004) corpus, MICASE (F), (*I* at 21.03, *you* slightly higher at 22.8, and *we* also much lower at 12.3). In contrast, Rounds' figures overall are much lower for *I* and *you* (both 12.6) but with a noticeably much higher frequency for *we* (39.1). Webber (2005) similarly found *we* more common than *I* and *you* (15.0 to 10.2 and 10.9), although her results are obviously not quite as striking as the figures for Rounds.

Table 5.3: Pronoun Frequencies Across the PBL Corpora

	***I*-forms**		***You*-forms**		***We*-forms**	
	No.	Per 1000	No.	Per 1000	No.	Per 1000
AllPBLs	2723.0	23.7	2581	22.4	1392.0	12.1
PBL1s	1237.0	28.0	1126.0	25.5	814.0	18.4
PBL2s	1486.0	20.8	1458.0	20.5	578.0	8.1
MICASE (F)	16251.0	21.03	17664.0	22.8	9489.0	12.3
Rounds (1987a)	329.0	12.6	338.0	12.6	1052.0	39.1
Webber (2005)	355.0	10.2	379.0	10.9	523.0	15.0

Note. MICASE (F) refers to Fortanet (2004).

When we look at the PBL sub-corpora, we can see the frequency for all three pronouns is higher in PBL1s than in PBL2s, with similar patterns in use for both *I* and *you* within each sub-corpus (in PBL2s, it is almost identical). For both of these the frequency is lower in PBL2s, but more noticeably so for *I* which goes from a frequency of 28 per thousand words in PBL1s to 20.8 in PBL2s. *You* falls from 25.5 to 20.5. What is particularly interesting in my data is that while *we* overall is used significantly less then *I* or *you* (with an overall frequency of 12.1), the breakdown also shows that *we* is used considerably more often in PBL1s than in PBL2s (a frequency of 18.4 compared with 8.1). This lower figure in PBL2s is in contrast to both Fortanet for lectures (12.3), and Webber (2005) who found a relatively higher frequency of *we* (15.0) than for *you* and *I* in conference talks. This is interesting as the format of conferences and lectures, albeit with a different power relationship and audience, are more similar to PBL2 presentations than PBL1s.

This noticeable change in the use of *we* may be an indication of the extremely involved collaborative nature of PBL1s in comparison to the stage 2 tutorials, which all involve longer presentations and less turn-taking. It should be noted that in four of the five PBL2s, there is much more of an emphasis on presenting individual work, along with follow-up questions primarily to individuals (hence perhaps a reason for the higher use of *I* and *you*).

PBL1s are about working together, sharing ideas, negotiating what to do and coming to a consensus. In the following examples, we see direct attention to co-participants:

> (3) S12: so, let*'s* start with the main issue. what do *you* think the main issues here? (PBL9.1)

> (4) S20: well if *we*'ll see how what *we*'ve got but yeah if *we* don't have enough stuff *we* could put it in but, *i* don't like that but *i* don't know if that's (PBL12.1)

The discussion of the scenario and consideration of potential issues and learning objectives is also reflected in the use of *I*. Students tentatively give opinions, make suggestions, hedge statements and show uncertainty:

> (5) S14: say you have to memorise, er this diagram and this way of this calculation, so *i* think it's maybe impossible or at least it's difficult to memorise very complicate (PBL9.2)

> (6) S26: *i* think though if you look it up you're going to end up, oh what bowel cancers (PBL11.1)

> (7) S28: endometrial or the three of them? *i* don't know (PBL11.1)

Also on occasion, we can see students guiding the group through their talks in PBL2s, for example with *i'll:*

> (8) S7: we are the group three. uh these are our members. (2) *i'll* be presenting about the multifactorial causes and inheritance pattern (PBL12.2)

They also clarify and check, often prefaced with *i mean:*

> (9) S26: so
> S5: so. no *i* mean that you know the pedigree for example is autosomal recessive (PBL11.1)

Although it is not possible with WordSmith Tools to distinguish between subject and object use of *you*, it is possible to clearly distinguish subject and object forms for *we* and *I* and related forms of *you* (*your*/*yours*). The next section goes on to comment on the use of items that stand out: *me*/*my, our* and *your*/*yours*.

5.4.2 *I-, We-* and *You-* Related Forms in PBL1s and PBL2s

My has a slightly higher frequency in PBL2s than PBL1s (1.2 compared with 0.7). In the PBL1s, *my* is frequently about personal non study-related matters, including recounting personal stories or anecdotes, as in Example 10:

> (10) S1: so *my* dad's away so she's coming down to do Christmas shopping, cos *my* sister's working most of Christmas (PBL 7.2)

In contrast, in the PBL2 sub corpus, nearly half of the instances of *my* (42 of the total 85) are specifically related to individual work and reports:

> (11) S9: to be degraded by the pancreas or something, so that's why it's high. this is *my* assumption, this is *my* sus- suspicion, i'm not hundred per cent sure, but (PBL11.2)

> (12) S23: okay. (2) right i'll start by giving you a brief overview of *my* talk today. ehm, first of all i'll talk about OFC (PBL10.2)

In both stages of the PBLs, the use of *me* and *my* is frequently related to classroom management, for example checking the order of speakers or concerning the distribution of hand-outs and clarification checks.

(13) S1: all *my* copies

S2 that's [? how] i've got one (7) do you want *me* to go at the start? (PBL7.2)

Although perhaps not of note when comparing across the different studies, *our* stands out relatively within the PBL corpus, as we can see higher frequency in PBL2s (49 instances or a frequency of 0.7 in PBL2s compared with 15 or a frequency of 0.3 in PBL1s).[27] In PBL2s it is often used to refer to the scenario in hand (and so an example of the inclusive *we*). The participants often appear to be using this form to generalise their statements and perhaps to bring the scenario into the classroom.

(14) S1: fourteen and half days um in the wild type the palate has em come together but in *our* compound heterozygote the palate has still got a cleft down the middle. (2) (PBL12.2)

It is also used when students are discussing the findings of the group:

(15) S9T1: the target DNA, and step D is the annealing step where the probe will anneal to *our* to *our* specific er, DNA target (PBL7.2)

Similarly, *us* is used to talk inclusively about the work of a specific sub-group of students (identifiable from the context):

(16) S7: handouts and in a few minutes. so like how we ask S38, for an MSI, she told *us* that MSI was positive like how S27 told MSI is done for, er the genes the (PBL11.2)

The use of *your* in PBL1s and 2s is very similar in frequency (1.0 and 0.9). It is used in reference to descriptions of cases and conditions and general processes (rather than to a specific individual in the group):

[27] See Appendix 4 for figures.

(17) S16: sphincter muscle which help you to, well it contracts and relax to, to store *your* urine and stools, but sometimes it not well controlled so, when a baby has (PBL9.2)

These examples have provided an indication of the levels of engagement and orientation the group. They have also hinted at an orientation to disciplinary identity. The specific semantic mappings of *you* and *we* are discussed in the next section.[28]

5.4.3 Semantic Mapping of *You* and *We*

This section reports on the uses of *you* and *we* in relation to semantic groupings in order to explore in more detail how they relate to the pedagogic aims of PBLs, the type of communicative event, and to disciplinary identity, either as a student or novice medical geneticist.

A close reading of each item within its surrounding co-text was undertaken in order to establish the 'referent', either specific and identifiable, i.e. referring to a specific individual in this group or the group as a whole, or used as a generic, impersonal reference. The specific categories were provided in Section 5.3.1.

5.4.3.1 Semantic Mapping of *You*

As we have seen, *you* is a high-frequency item in both PBL1s and PBL2s, although slightly higher in frequency in the first stage. The specific semantic mappings give an indication of the different purposes of each part of the cycle. Table 5.4 shows the semantic mapping for both stages.

[28] As *I* clearly refers to the speaker (and so excludes the addressee and others) in all but a very small number of cases where it was used with a quoting function, it will not be discussed in terms of semantic reference.

Table 5.4: Referents of *You* in PBL1s and PBL2s

Personal pronoun	**1. YS1**	**2. YGS2**	**3a. YG (3b. YG(YK))**	**Total**
PBL1s				
You	187	372	370 + (87) = 457	1016
Your	13	15	16	44
Yours	0	0	0	0
Yourself	0	0	0	0
Yourselves	0	1	0	1
Total	200 (19%)	388 (36%)	473 (45%)	1061
PBL2s				
You	243	251	686 + (125) = 811	1305
Your	11	0	36	47
Yours	10	0	0	10
Yourself	1	0	2	3
Yourselves	0	1	0	1
Total	265 (19%)	252 (19%)	849 (62%)	1366

Note. YS1= a specific individual; YGS2= this specific group; YG= generic reference; YG(YK)= generic *you* with *you know*.

PBL1s are much more orientated to the group (*YGS2* at 36% compared with 19% in PBL2s), while PBL2s involve much more impersonal reference to processes and procedures.

We can see from Table 5.4 that in PBL1s of the three categories the generic, impersonal use of *you* (*YG*) is the most common (45%). However, the combined use of singular (*YS1*) and plural specific (*YGS2*) *you* is higher overall (55%). This reflects the highly interactive group work of the PBL1s. There are frequent references to specific individuals and questions addressed to the whole group.

In PBL1s we can see questions directed to the individual:

> (18) S4: so do *you* think it might be, quite rare maybe what the, the wee boy had then? (PBL7.1)

More frequent appears to be the comment or question to the whole group:

(19) S9: how would *you* look at it? how what type of screening (PBL7.1)

We can also see a speaker directing the group through their talk in PBL2s:

(20) S7: now what is a dominant negative mutation. here as i already told *you* that in PPS the mutations occur in the DNA binding domain (PBL 12.2)

In both stages, individuals may also be identified further, for example to emphasise a specific contribution, as shown below in Example 21:

(21) S25: well why don't the person who's doing, *S29* see i think *you*'re right (PBL11.1)

In both stages we find a substantial number of the instances of *you* used as a general/generic reference, although with a much higher percentage of this use in PBL2s (62% in PBL2s compared to 45% in PBL1s). In PBL2s, this appears to reflect the format of the presentations, characterised by less interaction. This generic *you* is commonly used to talk about processes and procedures medical professionals would carry out.

(22) S25: uhuh, exactly in the family. then how can *you* test other family members for (PBL11.1)

Within this generic grouping we see the imprecise nature of *you* and the hidden referents exemplified in specific contexts. For example, y*ou* may represent a carrier or patient with a condition (23), the student in the role as a professional geneticist or another professional in the field (24), or as a very general (anyone) *you (25)*:

(23) S28: like bowel or colon but a lot of cancers, if *you* have like a gene for it, *you* pass on the gene (PBL11.1)

(24) 24T3S: is going to be ill but it's all about because *you*'ve identified them i suppose *you*'ve got a responsibility (PBL10.1)

(25) S26:i think i've a random question [laughs]. see how in America *you* have to pay for *your* health care (PBL10.2)

On occasion, what appears as a very general reference is actually limited to a specific group in some way. The following extract was from a discussion about changing names at marriage, talking about a specific country and gender, where *you* actually referred only to women:

(26) S22: but when *you* get married *you* can *you* can change it or *you* can decide not to change (PBL10.2)

The varied uses noted above appear to allow the students to take on a multiplicity of roles (some of which will never be realised), allows them to align with their disciplinary community, make generalisations and overall reduces the distancing effect of *they*/*them* or *one*.

The generic impersonal uses also include the numerous examples of *you know*. It is indicative of interaction and audience orientation and although not specific it is, I would say, as generally vague and imprecise as some of the uses of *you* above.

In the next examples, in terms of the discourse functions of *you know*, these appear to be buying time to hold the floor (27), emphasising information (28), or checking comprehension (29):

(27) S40: in some er hospitals they give er free er, surgery, erm *you know* er in some it it is like er, trial or something, it is there but without (PBL12.2)

(28) S30: forty forty nine and then thirty four years, and then the other factor is that *you know* er there are, patterns of primary cancer segregation *you know* like er, (PBL11.2)

(29) S25: cos someone else's going to understand, *you know*? (PBL11.1)

A further exploration of three- and four-word clusters, many of which include *you know*, will be presented in Chapter 6.

5.4.3.2 Semantic Mapping of *We*

We have seen that while much lower in frequency than *you* and *I*, *we* is still of note as the sixteenth most frequent word in the corpus. Although overall frequency is similar to Fortanet (2004), with PBLs at 12.1 compared with 12.3 for Fortanet, in my corpus the frequency of *we* is significantly lower in PBL2s than in PBL1s *(*8.1 compared with 18.4). Rounds and Fortanet both identified a range of referents, finding lecturers used *we* in a variety of ways, permitting changes in footing and moves between differing roles as specialists and lecturer. Both also found uses of *we* = *I* and *we* = *you*. However, the PBL corpus presents a different profile of use (see Table 5.5 to follow).

Table 5.5: Semantic Mapping of *We*

	***We* 1**	***We* 2**	***We* 3**	***We* 4**	***We* 5**	**Disregarded**	**Totals**
We							
PBL1s	6	0	620	4	38	106	774
PBL2s	76	0	172	2	238	3	491
Us							
PBL1s	4	0	10	0	0	1	15
PBL2s	21	0	8	0	2	0	31
Our							
PBL1s	1	0	10	0	0	4	15
PBL2s	14	0	22	0	7	6	49
Ours							
PBL1s	0	0	0	0	0	0	0
PBL2s	0	0	1	0	0	0	1
Let's							
PBL1s	0	0	9	0	0	1	10
PBL2s	2	0	4	0	0	0	6

Note. *We* 1 = speaker + others (exclusive); *we* 2 = *I* (speaker only; exclusive); *we* 3 = speaker and listener(s) (inclusive); *we* 4 = *you* (listeners only, not speaker); *we* 5 = anyone working in this area.

Although we can see both inclusive and exclusive uses of *we*, with students also moving between 'roles', they also appear to be staking claims to groups, and notably groups that are very specific to the context. In PBL1s there is a much higher use of the traditionally inclusive *we* (*I*, the speaker plus *you* the addressees), strikingly so when compared to PBL2s (620 instances to 172 for *we*

alone). This reflects the purpose of PBL1s, i.e. to share knowledge with peers (30), to work as a group to negotiate content (31), and to identify outcomes and objectives for further study (32).

(30) S15: ok but whatever diagnostic tests *we* are using like CVS or AS
S16: no it wouldn't be CVS because
S8: sorry?
S14: *we* can say what's em, what's what is the disease (PBL9.1)

(31) S26: didn't *we* didn't kind of, decide
S25: yeah
S26: shall *we* see what other questions *we* have and then (PBL11.1)

(32) S12: any more issue?
S14 Down syndrome i think. *we* haven't mention Down Syndrome. (PBL9.1)

In PBL2s by contrast, there is a much higher use of *we* referring to 'anyone in the field' of genetics. Here again students take on the role of the professional. In the example below, as they describe a condition, they indicate shared knowledge:

(33) S10: and any chromo- chromosomal rearrangements as well, em, give rise to NTDs. so *we* know they must be lying in there somewhere, in some of the genetic material (PBL9.2)

In the examples above *we* could be substitute by *you*. The speaker perhaps selects *we* to personalise the process more. In doing so it appears that the students are frequently taking on the role of novice researchers and again aligning themselves to this discourse community. Other alternatives would be *they* or *geneticists,* which would exclude them from that group.

In terms of the traditional exclusive use of *we*, Rounds found it was often used to refer to the lecturer and other professionals (who more often than not, were not present). However, in these PBLs *we* is used primarily as a marker of specific

group in their programme. Students may identify themselves as members of a specific student group, for example referring to work they have done for a presentation (34):

> (34) S22: so just to conclude that the da- databases were very useful, ehm they helped *us* narrow down possible causes, in mice IRF six is highly expressed in tissues (PBL12.2)

We also see on occasion the facilitators identifying themselves (*us*) as a separate, authoritative group from the students (*you*):

> (35) S24T3: in to the office with it if *you* want *us* to listen to you (PBL10.2)

In the PBL corpus I could only identify 6 examples of *we* = *you*. As with Fortanet and Rounds, these examples were used by the facilitator and appear to be employed to make the statement more inclusive. In Example (36) below, the facilitator means 'that way *you* don't go down the wrong path', as she will clearly not be involved in the individual and group interim study.

> (36) S9: seen a bilateral? there is right? make sure it's bilateral cleft lip. that way *we* don't go down the wrong path by accident (PBL12.1)

There were no clearly identifiable examples of *we*= *I*.

In summary, while *we* is less frequent than *you* and *I*, it is interesting in two ways. The first is in the difference in frequency of the item between PBL1s and PBL2s, and the second is in how it is used compared to other studies which have focussed on either expert-to-novice or expert-to-expert groups. In the PBLs, students use *we* to show solidarity and indicate membership of their student group, notably in PBL1s and to an extent in PBL2s, and the professional community as novice researchers (PBL2s), including themselves in this group.

5.4.4 Collocates of Personal Pronouns

This section will look briefly at the collocates retrieved for each of the pronoun groups in each corpus. This provides an indication of the most common linguistic environments each pronoun is found in.

The collocate search functions in WordSmith Tools retrieves "patterns of repeated phraseology" (Scott, 2004), e.g., frequent collocations (words that co-occur to a designated span left or right of a keyword). The examples given in Tables 5.6-5.8 to follow are of the top ten collocates one word to the left and one word to the right of the relevant pronoun. (A discussion of longer clusters or bundles follows in Chapter 6.)

From the collocate search, we see in both stages *I think* and *I don't* as the most common partnerships for *I,* closely followed by *I mean.* The collocates shown in Table 5.6 on the next page indicate that *I think* while frequent in both stages, is much more so in PBL1s (415 cf. 272). This shows student attitudes, stance and most likely the tentativeness of their contributions: students have to state their level of certainty or otherwise when discussing specific ideas and concepts and since the aim is to generate ideas and identify what they do and do not know as a group, the PBL1s provide a 'safe' environment for this. More commitment and more public objective presentation of their independent study is required in the PBL2 report, possibly reducing the use of *I.*

Table 5.6: Top 10 Collocates for *I* in PBL1s and PBL2s

PBL1s	**No.**	**PBL2s**	**No.**
i think	415	i think	272
i don't	151	i don't	89
i mean	89	but i	86
but i	72	i mean	79
yeah i	67	so i	64
so i	39	i was	62
i just	34	yeah i	50
and i	27	i just	44
what i	31	and i	43
i know	22	i have	40

I mean is used to show students highlighting and checking points (37) or reiterating and elaborating (38).

(37) S6: *i mean* is it is it useful to do a ca- a carrier carrier population (PBL10.1)

(38) S20: but we experiment mostly on mouse so i think uh, it's mostly *i mean* well all experiments we use mouse instead of [?] so, i think that must be (PBL12.12)

I is frequently used with *but, and,* and *so* which may indicate linking in either the speaker's contribution or responding to another participant (adding, contrasting and developing points) in both parts of the cycle.

Table 5.7 indicates that *we* is commonly used with full modals, e.g., *we can, we should;* with the semi-modals *need* and *have* (to), and in questioning (*do we, shall we* in PBL1s). In PBL1s we see, for example, *we need* and *we should* are used when students are discussing what is involved in researching a scenario, prioritising and making decisions about what they will need to do in the independent study stage. In PBL2s, however, we see a slightly different pattern of use, referring to information (*we know*), and providing reasons (*because we*).

Table 5.7: Top 10 Collocates for *We* in PBL1s and PBL2s

PBL1s	**No.**	**PBL2s**	**No.**
we have	72	we can	62
we can	68	so we	39
we need	62	we have	34
think we	51	that we	26
so we	44	and we	24
do we	43	we know	24
we should	43	we are	23
if we	41	because we	22
we could	32	if we	22
shall we	29	we could	20

Table 5.8 which follows shows the most frequent collocates for *you*. *You know* is the most frequent in both stages. In PBL1s we can also see *do you,* and *you think,* reflecting the higher level of questioning and speculation. In PBL2s, there is a much higher instance of *you can* and *if you* in PBL2s and there are more collocates which appear to indicate that the speaker is directing the listeners through the talks and elaborating points (*so you, and you*).

Table 5.8: Top 10 Collocates for *You* in PBL1s and PBL2s

PBL1s	**No.**	**PBL2s**	**No.**
you know	157	you know	195
do you	147	if you	162
if you	85	you can	157
you can	66	you have	92
you think	66	do you	73
you want	46	so you	68
so you	42	and you	57
you mean	39	thank you	56
want you	36	you get	51
you could	39	that you	50

This section has taken a very brief look at immediate collocations for the first and second person pronouns. As with the frequency results, the items retrieved from the collocate findings appear to reflect the pedagogic purposes and

procedures of PBLs as a whole and for each sub-corpus (i.e. novice researchers investigating a field and the questioning and speculating in PBL1s, and the guiding of the group through their presentation of findings in PBL2s).

5.5 Discussion

The results show the high frequency of first and second person pronouns, particularly of *you* and *I*, with a correspondingly much lower use of *we*. Their use overall and within stages of the PBL sessions appears indicative of three (overlapping) areas: their use indicates a high level of involvement, engagement and orientation to other participants; this also appears to reflect the nature and purpose of PBL sessions as a genre (which obviously requires interaction); and finally I believe they function as a clear indicator of group membership and group solidarity, and in doing so reflect this specific community of practice (Wenger, 2006) and the context of the sessions (PBLs necessitating interaction and group work).These three points are discussed below.

The overall preponderance reflects the student-centred nature of PBLs, particularly in PBL1s which are more interactive and evidence an extremely high level of active participant involvement. In this respect they are more similar to casual conversation, where personal pronoun use is also very common, particularly for *I* and *you* (Biber et al., 1999). Although the PBL2 presentations necessarily involve less group discussion, the use of personal pronouns also suggests that participants are very orientated towards their audience, both by involving them in topics (as novice geneticist researchers) and guiding them through the presentations.

The level of involvement is of course also a feature of the PBL as a learning event. What is of particular interest are the specific uses of each pronoun group, as the evidence available suggests that these uses also closely reflect the purpose and nature of the PBLs and so also differentiate them from lecturer use (cf. Rounds, 1987a and, to an extent, Fortanet, 2004). The sessions are dominated by peer group discussions and presentations, which helps to explain, for example, the quite different profile of *we* found here compared with that of Rounds. In the PBL1s, the overwhelming proportion of the instances of *we* are

inclusive, used when students share ideas and determine their group learning objectives (this may be similar to the "collective" nature of business meetings that McCarthy and Handford (2004, p. 178) find in the CANBEC corpus). In contrast to the PBL1s, in PBL2s we find a much higher use of *we* to indicate 'anyone in the field', as students bring the scenario and role as professionals into the classroom. Here they take on the role of a novice researcher as they develop their skills. We also find the exclusive *we* when students are clearly signalling group membership (primarily as students, as opposed to *we* + others not present). In the PBLs we find very few examples of *we* = *you* and none of *we* = *I*, both of which were found by Fortanet and Rounds in their studies of lecturer talk.

In PBL1s, *you* is commonly used to refer to the study group and to a lesser extent to specific individuals. These uses are a clear indication of the co-presence of addressees and of the group discussion and questioning required for investigating the scenarios. However, of the three categories, *you* is most commonly found in its generic use. The alternative here is most often *they,* which would sound more distancing. The constant use of *you* when focussing on the content and discussing the procedures and processes involved in any given scenario again seems to allow students to become part of a group of novice researchers. Stirling and Manderson suggest "that generalized uses of *you* ... can be viewed as context-delimited invokers of membership categories to which the speaker and, depending on context, the addressee, are seen to belong" (2011, p. 1587). This generalised use may also be a way of widening the authority and power of the point of the statement (Scheibman, 2007).

I is the most common of the personal pronouns studied here, overall and within each stage. Higher in PBL1s, it illustrates not only the greater level of interaction (and thus moves between speakers) but also indicates attitude and opinions as the students present ideas along with their individual stance towards the content.

The pedagogic purpose of the PBLs necessarily requires a high level of involvement and team work and we can see that the uses of *you,* and *we* indicate very strong group identities, particularly as members of this specific

group of students on this programme. Students direct questions to the group as they search for answers, and identify themselves strongly as part of a whole group working towards specific aims (in PBL1s), guiding others through talks (in PBL2s) and as members of specific sub-groups, for example when they present their work. This very strong group orientation may be a way of invoking camaraderie (rather than individual effort) when asking questions and preparing a plan of action; it also frequently delineates staff (as a separate, authoritative group). Even the impersonal *you* for talking about the processes and procedures involved in medical genetics is used, I would say, to invoke a sense of group and of being members of a specific community and allows them to approximate closeness to professional groups.

5.6 Chapter Summary

This chapter has considered the frequency and aspects of use of first and second person pronouns in the small corpus of PBL recordings, concentrating on the use of *you, I,* and *we* and their related forms. In contrast to studies of lectures (Rounds, 1987a, 1987b) we can see quite different levels of frequency and use. While overall frequency results are similar to the larger academic corpus study by Fortanet (2004) variations in the purpose of the two PBL stages indicate different patterns of use. This would suggest that the purpose and type of interaction has a notable effect on frequency and use. Although only considered in brief, the collocation search also provides preliminary indication that they also can be tied to purpose and levels of interaction. How these pronouns are used to indicate stance, organise discourse and refer to subject matter will be explored in more depth as part of the following chapter on lexical bundles, a type of multi-word sequence.

Chapter 6 Multi-Word Sequences: The Use of Lexical Bundles in PBL1s

"Much language use consists of repeated expressions."
(Conrad & Biber, 2004, p. 56)

6.1 Introduction: Multi-Word Sequences

In the previous chapter, I introduced a number of contiguous collocates, a type of multi-word sequence, of personal pronouns along with an initial discussion of their potential discourse functions. This chapter extends that work by identifying the related structural categories and discourse functions of the three- and four-word clusters, a type of multi-word sequence also known as lexical bundles, found in each stage of the PBL cycle. Again, there is a consideration of their role as markers of involvement and interaction and the extent to which these items reflect the disciplinary specific characteristics of this PBL corpus. The chapter starts with a brief discussion of multi-word items before moving on to summarise previous research on lexical bundles specifically. The methodology and findings for the study of structural and discourse functions of lexical bundles in the PBL corpus are then presented.

The prevalence and importance of multi-word sequences is reflected in the considerable research in this area. This includes investigations into aspects of syntactic structure, functional use, degrees of compositionality, issues related to storage and retrieval, and register variation (e.g., Altenberg, 1998; Biber et al., 1999; Biber & Barbieri, 2007; Cowie, 1998; Durrant, 2015; Durrant & Mathews-Aydınlı, 2011; Ellis, 1996a, 1996b; Hyland, 2008a, 2008b; Moon, 1998; Pawley & Syder, 1983; Simpson, 2004; Wray, 1999; Wray & Perkins, 2001). Depending on the focus of the research, the definition and approach to identification vary.

Much of the interest from a language user's point of view derives from Pawley and Syder's (1983, p. 191) seminal work which discusses "two puzzles … nativelike selection and nativelike fluency". This raises the issue of how the NS, and indeed the NNS, select appropriate linguistic items from the wide range of language available and so communicate effectively across stretches of discourse, particularly given the limitations of processing capacity in the memory. In

considering naturally occurring conversational data they note the relatively limited range of linguistic items actually used and conclude that the ability to select, encode and also decode appropriate language "rests to a considerable extent on the knowledge of a body of 'sentence stems' which are 'institutionalised' or 'lexicalised'" (Pawley & Syder, 1983, p. 191).[29] Sinclair (1991) echoes the conundrum in his discussion of the idiom versus open choice principle. Here he highlights the limitations of a generative grammar model of language use (which allows for the construction of an infinite number of phrases and clauses) in accounting for the prevalence of what appears to be formulaic language in actual language use. In order to explain why we regularly make use of a relatively limited set of choices which cannot be fully explained by context and register restrictions, he maintains that we have "a large number of semi-preconstructed (or pre-fabricated) phrases that constitute single choices, even though they may be analysable into segments" (Sinclair, 1991, p. 110). In other words, it is suggested that we rely in part on memorised chunks when constructing, and possibly decoding, discourse.

Studies of multi-word units include work on formulaic language, which is understood to be stored and retrieved as holistic units of meaning. Such units are thought to provide a psycholinguistic processing advantage, by-passing the need to create and construct discourse from scratch each time and so contributing to more native-like production for language learners (Wray, 1999). As Wray (1999) notes, multi-word sequences have been described and categorised in a number of ways. They can include idioms (semantically opaque, at least to some degree), sentence stems and frames (which may be described as lexico-grammatical patterns and may include 'slots' for variability) and multi-word units of varying degrees of semantic transparency and fixedness of structure.

Yorio (1980, p. 433) uses the term "conventionalised language forms" and identifies a number of formal properties and functions. His main point is that this conventionalised language has socio-linguistic functions, is regulatory in nature (e.g., organising discourse) and is group identifying (e.g., for establishing

[29] Stems are units of meaning wholly or largely fixed in structure and which are culturally recognisable.

membership and rapport). In a similar vein, Moon (1998, p. 8) discusses what she terms *fixed expressions*, noting the importance of context for fully understanding their use. Within a specific academic context, DeCarrico and Nattinger (1988) and Nattinger and DeCarrico (1992) also identify "lexical phrases" in a collection of different styles of academic lectures. These lexical phrases, identified by perceptual salience rather than frequency alone, are described as 'chunks' of language which can be of varying length and flexibility but which are viewed as "conventionalized structures that occur more frequently and have more idiomatically determined meaning than language that is put together each time" (DeCarrico & Nattinger, 1988, pp. 91-92). They also highlight the importance of considering their pragmatic function and note that a lexical phrase may have more than one function.

While there may be variations in degrees of compositionality, fixedness in form and syntactic completeness, common themes that emerge are a sense of unity, that items may be stored and retrieved as units of meaning, that they can be assigned pragmatic functions and uses and, implicitly, that these may vary across genres and registers (e.g., Nattinger & DeCarrico, 1992). The studies mentioned so far have all, however, identified items with at least an element of intuition and reliance on perceptual salience.

The next section goes on to look at lexical bundles, a particular type of multi-word unit which are prevalent in many registers but which are not always perceptually salient.

6.2 Lexical Bundles

The study of multi-word sequences now routinely extends to and includes lexical bundles, described by Biber et al. (1999, p. 989) as "extended collocations: bundles of words that show a statistical tendency to co-occur", comprising three or more recurrent, fixed, sequences of words. In contrast to intuitive approaches to the identification of potentially significant multi-word items, the study of lexical bundles starts from an automated empirical search, items being identified by frequency alone. They are sometimes referred to also as clusters

(e.g., Hyland, 2008a; Scott, 2004) and recurrent word combinations (Altenberg, 1998; De Cock, 2000).

These recurrent strings of words are fixed (i.e. do not vary in compositionality), may be structurally incomplete (e.g., *a lot of*), can cross syntactic boundaries (e.g., *i think it's*), may not always be semantically transparent and tend not to be perceptually salient. This is an important point: as they are not generally perceptually salient, they may not normally be identified intuitively as significant. While bundles may not all be viewed as complete semantic or holistic units, there is also increasing research to suggest that at least some lexical bundles are processed and retrieved in a similar way to other types of formulaic language (i.e. a holistic, meaning-based unit) (Conklin & Schmitt, 2008; Ellis, 1996a; Jiang & Nekrasova, 2007; Nekrasova, 2009).

Notwithstanding the question of whether or not lexical bundles are processed in the same way as other types of formulaic expressions, as Biber and Conrad (2004, p. 56) note with the expansion of corpus based research "it has become impossible to ignore their importance for describing the lexicon of a language". Bundles have been identified as a significant component of a wide range of registers, spoken and written (e.g., Ädel & Erman, 2012; Altenberg, 1998; Biber et al., 1999; Biber et al., 2004; Byrd & Coxhead, 2010; Chen & Baker, 2010; Cortes, 2004; Durrant, 2015). They have been said to account for anywhere between 20% (Conrad & Biber, 2004) to 80% (Altenberg, 1998) of a text, depending on the corpus, frequency and length of bundle.

6.2.1 Overview of Structural and Functional Categorisations of Lexical Bundles

The extensive work by Biber et al. (1999), which investigated a large corpus of four registers culminating in the *Longman Grammar of Spoken and Written English* (LGSWE), includes a whole chapter devoted to the description and categorisation of lexical bundles. This work includes an identification of fourteen main structural groupings of bundles in conversation and twelve in academic prose and the identification of two overall groupings and three sub-groups. The two main groups separate bundles which are described as *oral*,

including a main or dependent verb phrase fragment (sub-groups or Types 1 and 2) and found to be more common in conversation, and the second group classified as *literate,* including items that include a noun or prepositional phrase fragment (sub-group/Type 3), more common in academic prose. However, as Biber notes, these categories not mutually exclusive: for example, *I think we should* could come under *personal pronoun + lexical verb* but also includes *verb+ that clause fragment* (albeit omitted) (1999, p. 1001). Their analysis concentrated on the use of four-word LBs, noting "three-word lexical bundles are too numerous to list in a work of this scope" (1999, pp. 1000-1001). The taxonomy was then developed by Conrad and Biber (2004) and Biber et al., (2004) and now offers a framework for categorising bundles by structural patterns and also discourse functions.

In addition to structural groupings, the main functional categories in Biber et al. (2004, p. 384) comprise four groups, *stance expressions, discourse organisers, referential expressions* and a fourth smaller group of *special conversational functions. Stance expressions* are sub divided into *epistemic stance,* indicating degrees of commitment, and *attitudinal/modality stance* expressing attitudes. *Discourse organisers* link ideas and sections of the discourse, and *referential expressions*, referring to entities or particular attributes (Biber et al., 2004, p. 394). There is also a category of *special conversational functions* including markers of politeness.

A fuller discussion of the structural and functional categories, along with modifications made for this study of PBLs will be provided in Sections 6.3 and 6.4 below.

6.2.2 Lexical Bundles in Academic Discourse

There have been a number of studies of both the structure and functions of lexical bundles in large-scale academic corpora, most notably by Biber et al. (2004), Biber and Barbieri (2007), related work by Cortes (2004, 2006), Hyland (2005b, 2008a, 2008b) and Salazar (2011, 2014). Both Hyland and Salazar use a framework adapted to suit specific registers. Research has focussed on NS use, NNS use, or provides comparisons between the two. Some studies concentrate

only on written or spoken registers and genres, some compare written and spoken discourse and some look at specific student learning events. Studies of lexical bundle use in written registers include Hyland's (2008b) investigation of disciplinary variation in theses and published writing and Salazar on published writing in the health sciences (2011) and on native and non-native use (2014). Chen and Baker (2010) and Cortes (2004) both compared published and student academic writing and Durrant and Mathews-Aydınlı (2011) and Durrant (2015) consider disciplinary variation in student academic writing. These studies indicated variations across disciplines and also between expert and novice writers.

Those comparing lexical bundle use in spoken and written registers include Biber et al. (2004) who analyse bundles in university teaching and text books, Conrad and Biber (2004) comparing conversation and academic prose and Biber and Barbieri (2007) studying university spoken and written registers. In their analysis of bundles in an academic corpus, the T2K-SWAL corpus, Biber et al. (2004) found classroom teaching uses more bundles than text books and that structural use varies across written and spoken academic registers. Biber and Barbieri also provide an analysis of lexical bundles across spoken and written university registers and claim that bundles "serve important discourse functions in both spoken and written texts" and "these word sequences turn out to be consistently functional, indicating that high frequency is a reflection of pre-fabricated or formulaic status" (2007, p. 265).

Focussing solely on spoken registers, Nesi and Basturkmen (2006) studied the cohesive role of four-word bundles found in lectures in the BASE corpus. They find that lectures contain a mix of 'oral' and 'literate' bundles. However, while the lecture corpus does include interactive lectures involving student contributions, they are primarily lecturer-led and represent an expert-to-novice interaction. Because of the extensive number of three-word bundles, as with Biber (1999), again they are not considered by Nesi and Basturkmen. In studies of more student-centred work from smaller corpora, Hernández (2013) considers bundles in three oral corpora of undergraduate NS and NNS and Sahin-Kızıl and Kilimci (2014) compare spoken general interviews between NS and NNS. There are to my knowledge as yet no studies of the structural categories or discourse

functions of postgraduate student-centred disciplinary-specific learning events, large or small.

6.2.3 Research Questions for Lexical Bundles

Although both stages of the PBL cycle are relatively informal and very much student-led, the strong sense of the group and interaction in the first stage of the cycle contrasts with opportunities for more 'formal' presentations of densely packed informational content in the second part of the cycle, which is to an extent more similar to a lecture. As a result of the distinct purpose and nature of each stage, the use of bundles might be expected to differ. This chapter, then, sets out to identify first of all the most common structural categories used in the PBL corpus as a whole (AllPBLs) and within each sub-corpus, PBL1s and PBL2s. This is then followed by a further investigation into the discourse functions of three- and four-word items in each sub-corpus.

The specific research questions are as follows:

- What are the most common structural patterns and discourse functions of the three- and four-word lexical bundles in PBL1s and in PBL2s?
- To what extent do the bundles reflect the aims of the PBL pedagogy?
- To what extent do they reflect the uses identified in other spoken academic registers?

6.3 Methodological Issues

In this section the operational requirements for identifying bundles will be introduced. In setting parameters for a study of bundles in any corpus, a number of factors need to be taken into consideration: the length of bundles, the frequency cut-off and whether the number is raw or normalised, the range or distribution requirements and the issue of overlaps. Along with detailed discussion of the operationalisation of structural categories and discourse functions in this thesis, these are now discussed.

6.3.1 Frequency and Range Requirements

In addition to setting the length of bundles to be studied, e.g., sequences of three, four, or five words, the frequency cut-off, or threshold, i.e. minimum frequency, needs to be set. This can be either a raw score or normalised per million or thousand words.[30] Previous studies have ranged from as low as a raw frequency of two with a small corpus (Altenberg,1998; De Cock, 2000), to 20 to 40 per million words in larger corpora. Nesi and Basturkmen (2006) and Conrad and Biber (2004) set a minimum of 10 per million, with Cortes (2004) then suggesting a minimum of 20 per million. However, as Conrad and Biber (2004, p. 59) note, the cut-off point is "somewhat arbitrary" and indeed can be dictated by pragmatic considerations in terms of working with a manageable data set (Biber et al., 1999; Byrd & Coxhead, 2010; Nesi & Basturkmen, 2006). In addition, while normalised frequencies are usually employed in large corpus studies and are said to allow comparison across corpora, if corpora are different sizes, using a normalised frequency will in effect mean quite different raw scores (Biber et al., 2004, p. 268; Chen & Baker, 2010, p. 32). Both convincingly show how 'normalised rates' can result in very different raw frequency scores depending on the size of the corpus. For example, in a small corpus of 100,000 an item occurring only three times would have a normalised rate of 30 per million (compared to a normalised score of 3 per million in a corpus of 1,000,000), giving the impression of much higher frequency and significance if compared across corpora. As the PBL collection is a small, specialised spoken corpus, the frequency rate for analysis was set at a raw frequency of ten in the first instance (the size of the corpus means this still allows for a manageable number of three-word items to be considered).

While a frequency threshold of ten occurrences in the corpus does provide manageable numbers for three-word bundle analysis, this also has a significant effect on four-word, and indeed longer bundles, drastically reducing the numbers available for analysis: a threshold of ten instances over six texts for the whole corpus (AllPBLs) would leave only nine items. Reducing the cut-off to five instances across five texts provides 42 four-word bundles. In comparison,

[30] This is done by dividing the raw frequency of the item by the total number of words in a corpus and multiplying by 1,000 or 1,000,000. The figure of 1,000 has been selected for this corpus as the PBL corpus is quite small.

reducing the frequency for three-word bundles to five would make the study much less manageable. Therefore, in order to bring into consideration a larger number of items of four-word items with potentially more structures and functions (Hyland, 2012), different cut-offs were applied. The differing thresholds obviously means that comparisons between three- and four-word bundles are problematic but it does mean that a wider range of items are considered and provides a fuller picture of use. This approach is not inconsistent with other studies: Biber et al. (1999, p. 1001) also reduce the frequency requirement for longer less common sequences, as does De Cock (2000).

In order to eliminate the effects of individual speaker idiosyncrasies and of localised topic related repetition, bundles should also be evident across a range of texts (Biber et al., 1999). For the purposes of this study, when analysis involved looking at the whole corpus (AllPBLs), items should be found in a minimum of six texts (of ten) and for each sub-corpus occur in three of the five texts for that stage in the cycle (the comparison of the whole corpus and the two sub-corpora was used when considering the structural categories only).

At the analysis stage, it is also useful to note not only the frequency of each type of bundle but also of the number of occurrences of each item in order to achieve a fuller description. A register may have a limited number of types but with repeated occurrence or vice versa (Biber et al., 2004; Chen & Baker, 2010).

6.3.2 Overlaps

In addition to frequency and range considerations, there is also the question of complete or partial overlaps. These should at least be acknowledged because of the potential impact on frequency counts. For example, the three-word bundle *i don't know* can be totally subsumed into four-word bundles, *i don't know what*, or *i don't know if*. Partial overlaps include *but i don't* overlapping with *i don't know* (*i don't* being the shared part). Chen and Baker (2010) again note that this can inflate quantitative findings. However, possibly due to the painstaking nature of checking and categorising each potential item and the difficulty of deciding what to do with partial overlaps (i.e. which group to categorise them within), studies do not always appear to have manually excluded such items. In

the study of structural categories in the PBL corpus, items that completely overlap have been noted (as are partial overlaps) but are not excluded from initial frequency counts. In the analysis of the four-word and three-word discourse functions, however, the much closer reading required made it more practical and manageable to delete items which were completely subsumed within a four-word bundle (this was also the case in the very small number of five-word bundles that were identified in the initial exploratory stage). For the study of the discourse functions then, three-word items that are completely subsumed in a four-word bundle have been removed from the three-word lists. This means that we can differentiate clearly between discourse functions associated only with three-word bundles or only with four-word bundles.

6.3.3 Comparing Corpora: A Note of Caution

Although we can compare three-word bundles in PBL1s and PBL2 and similarly four-word bundles in PBL1s and PBL2s, as noted above there are obvious limitations to comparisons when the frequency cut-offs for three- and four-word bundles differ. The note of caution must extend to any comparisons made with other corpora for a number of reasons. For example, corpora obviously vary in a number of ways, from their different composition, size, to distributional and frequency requirements and approaches to overlaps. In relation to lexical bundles, while the broad categories may appear to be the same, the definition and boundaries of specific sub-categories may vary. Hyland (2012) also warns against potential reliability issues of comparing bundles found in small corpora to those from large corpora. However, it could also be argued that the more comprehensive qualitative work possible with a small corpus helps to address this. Last but not least, the majority of studies have focussed on four-word bundles (three-word bundles being too numerous to analyse). As such only very general comparisons can be made to other corpora. This study only notes tendencies in use which might suggest potential similarities or differences.

6.3.4 Structural Categorisation in the PBL Corpus

The work of Biber et al. (1999), further developed by Biber et al. (2004) and Biber and Barbieri (2007), was used as the basis for the structural categorisation.

This section first discusses the modifications to these broad structural categories made in this study and then goes on to present the results of the structural classification of three and four-word bundles in the whole corpus and then in each sub-corpus.

6.3.4.1 Modifications to Structural Categories

Building on the work of Biber et al. (1999), Biber et al. (2004) identify three major structural types and related sub-types: the first two types include verb phrases and are categorised in the *oral* grouping. Those with nouns and prepositional phrases are categorised as part of the *literate* group. The finalised list of categories used in this study, adapted from Biber et al. (2004, p. 381) and with examples of both three- and four-word bundles from the PBL corpus, is presented in Table 6.1 on the following page. To follow is a further explanation of the sub-categories.

The categories are not mutually exclusive and at times overlap. For example, Biber et al.'s first category for conversation (Type 1; oral), *personal pronoun + lexical verb phrase fragment* (1a), can include a complement-clause fragment, for example, *i'm going to do* and *you want to do* (four-word bundles taken from Biber et al., 1999, p. 1002). This overlaps with the first sub-category of the *dependent clause fragment grouping*, Type 2a, *1st/ 2nd person pronoun + dependent clause fragment* (Quirk et al. (1985) name a complement clause as a type of dependent clause). For the purposes of this study, I have included *to be (is/are/was/were)* and *going to* as auxiliaries to main lexical verbs; similarly modals and semi-modals are categorised as auxiliaries (e.g., *we need to*) in Type 1a. I have taken *want to do* as including a dependent clause fragment (*want* as the main lexical verb followed by dependent clause fragment; Type 2d).

The first two types, *(connector) + 1st/2nd person personal pronoun + lexical VP fragment* and *(connector) + 3rd person pronoun + VP fragment*, were broadened to also include *pronoun/noun + be* to give the current Types 1a and 1b. In addition, *Yes-no question fragments* now also include an *initial connector+ yes-no*, (Type 1f). For Type 1 bundles the *Other* grouping was added to include, for example, *and then you* and *and then they:* a closer examination of the

concordance lines indicated that these lexical bundles were largely followed by a verb. Similarly, after some consideration *because of the* was allocated to the *Other noun phrase (NP) expressions* (Type 3e), in the noun phrase Type 3 section. While this does indicate a dependent clause, it is also followed by a noun or noun phrase. One other item, *no no no* was classified in an overall *Other* category, now 3f.

Table 6.1: Structural Types and Sub-Categories of Lexical Bundles

Type 1: Lexical Bundles with *verb phrase fragments*		
1a	(Connector+) 1st/2nd person personal pronoun + lexical VP fragment	*you don't have; i don't think*
1b	(Connector) + 3rd person pronoun + VP fragment (includes also pronoun/noun + be)	*it's going to; so this is*
1c	Discourse marker +VP fragment	*you know it was*
1d	VP (with non-passive verb (active))	*going to talk*
1e	VP (with passive Vb)	*[can be used]*[31]
1f	(Connector) Yes-no question fragments	*but do you*
1g	Wh-question fragments	*what do you*
1h	Other	*and then you; yeah but I*
Type 2: Lexical Bundles with *dependent clause fragments*		
2a	1st/2nd person pronoun +dependent clause fragment	*you know (that) it's; you know what*
2b	Wh-clause fragments	*don't know what*
2c	If-clause fragments /Connector +*if* + *PN*	*don't know if; so if you*
2d	(Verb/adjective+) to clause fragment	*to do with*
2e	That-clause fragments	*is that the*
2f	Verb + that clause fragment (*that* omitted)	*think we should*
Type 3: Lexical Bundles with *noun phrase and prepositional phrase fragments*		
3a	(Connector+) NP with *of*-phrase fragment	*a lot of*
3b	NP with other post-modifier fragment	*mutations in the*
3c	Other noun phrase expressions	*the screening test; the first one*
3d	Prepositional phrase expressions	*of them are*
3e	Other NP expressions	*because of the*
3f	Other expressions	*no no no*

Note: Types 1 and 2 are oral, Type 3 literate; examples taken from the PBL corpus with the exception of 1e.

[31] Example from Biber et al. (2004, p. 381).

While the categorisation of many bundles was straightforward, as with *and then they* above, it was necessary on occasion to look more carefully at the concordance lines and at times the fuller surrounding co-text to identify the use of some items. For example, the concordance lines for bundles starting with *it, that* and *what* were checked in order to clarify use, e.g., whether *that* was used in a bundle for anaphoric reference or as part of a *that-clause fragment* and *what* to confirm whether an instance was a dependent clause fragment or a question.

Perhaps because the three-word bundles are too short to provide a fuller picture of use, no examples of the sub-category 1c (with *i mean you* and *you know it (was)* as a discourse marker), were identified at the structural classification stage.

6.3.5 Discourse Function Categorisation

This section looks specifically at the discourse functions used to categorise the three- and four-word bundles in PBL1s and PBL2s. The categorisation follows the broad categories first identified by Conrad and Biber (2004) and outlined above in Section 6.2.1. As we have noted, the main functional categories presented and described in Biber et al. (2004, p. 384) comprise four groups, *stance expressions, discourse organisers, referential expressions* and *special conversational functions*.

Stance expressions are sub-divided into *epistemic stance* expressions, which indicate the degree of commitment or certainty to "the knowledge status of the following proposition" (Biber et al., 2004, p. 389), and *attitudinal/modality stance* expressions, which indicate the speaker's attitudes to actions or events in the following proposition. The latter is then sub-divided into desire, obligation/directive, ability and intention/prediction. The final sub-category of intention/prediction is worthy of further discussion and I shall return to this in Section 6.3.5.2.

The second major group is *discourse organisers*, which serve to "reflect relationships between prior and coming discourse" (Biber et al., 2004, p. 393).

The third category, *referential expressions*, "make direct reference to physical or abstract entities, or to the textual context, either to identify the entity or to single out some particular attribute of the entity as especially important" (Biber et al., 2004, p. 394). This group originally included topic introduction/focus and also topic elaboration or clarification bundles.

There is also a final category of *special conversational functions,* which appears to be largely related to politeness. For example, it includes expressions such as *thank you very much*.

As we have noted, this taxonomy has been used in a number of studies of academic discourse and was used as the basis for this work. However, as part of the initial work of categorising progressed, a number of modifications were required. These are discussed below.

6.3.5.1 Modification of Functional Categories

While the functional categories initially appear straightforward, it should be noted that Biber et al. (2004) present their taxonomy as preliminary and indeed it has been developed further, for example by Cortes (2004, 2006) and by Biber and Barbieri (2007).[32] While the main categories have remained largely the same, some modifications have been made. This has included extending, re-naming or adapting sub-categories to better reflect the specific genres being studied. For example, Cortes' (2004, p. 408) study of student writing in history and biology includes within discourse organisers *focus bundles* and *framing bundles* (previously in *referential expressions*). In 2007, Biber and Barbieri developed the 2004 taxonomy by re-categorising many *referential expressions* previously termed *identification/focus* as discourse organisers, noting that they are "often used to preview a major topic by stating the main point first" (Biber & Barbieri, 2007, p. 271). In the current study, specific sub-categories also proved problematic. These issues and the resulting modifications are discussed below.

[32] This preliminary nature may account for issues encountered in categorising by Ädel and Erman (2012), who also found some confusion or lack of clarity in the use of terms.

6.3.5.2 Modification of the Stance Category

The first issue encountered was related to the *epistemic stance* category and the apparent overlap with the sub-category of *intention/prediction* under *attitudinal stance*. The concept of stance and its linguistic realisation (for example, through adverbials, modality and lexical bundles) has been the subject of much theoretical discussion and research. The area is wide ranging and research is addressed under a variety of headings from appraisal (Martin, 2000) and evaluation (Hunston & Thompson, 2000) to stance and engagement (Hyland, 2005b) and can include considerations of intersubjectivity (e.g., Du Bois, 2007; Kärkkäinen, 2006). A common theme, however, is the distinction between *affect* or *attitudinal* stance on the one hand and that of *epistemic stance*. The concept of affect or attitudinal stance draws on the work of Ochs and Schieffelin (1989, p. 9) and is said to "convey and assess feelings, moods, dispositions and attitudes". This aspect is also summarised more recently by Gray and Biber (2012, p. 17) as the "*personal feelings, emotions and attitudes* rather than evaluations of knowledge" (my italics). They also note that epistemic stance, in contrast, is concerned with the speaker's attitude to the *level of certainty* of the propositional content. Biber and Finegan (1988) and Biber et al. (1999) draw on the work of Palmer (1986), Chafe (1986) and Kärkkäinen (2003) in their discussion of epistemic stance in modality. In this respect, the concept of evidentiality concerning the *source* of one's understanding arises (i.e. what evidence is available that may influence one's stance). While there is discussion of the degree to which evidence and the assessment of certainty can be separated or otherwise (Dendale & Tasmowski, 2001; Mushin, 2001), the area of evidentials is often subsumed or discussed in relation to epistemic stance. Although using different categories, both Palmer and Chafe, and then Biber, discuss likelihood and evidence under epistemic stance. In discussions on modality specifically, Biber and Finegan (1988) and Biber et al. (1999, p. 485) note the influence of the work of Palmer (1986), Chafe (1986) and Quirk et al. (1985) and align intrinsic and extrinsic modality to deontic and epistemic modality respectively, clearly putting prediction under extrinsic, epistemic meanings. Biber et al. (1999, p. 485; my italics) note:

> Each modal can have two different types of meaning, which can be labelled intrinsic and extrinsic (also referred to as 'deontic' and 'epistemic' meanings). Intrinsic modality refers to actions and events humans (or other agents) directly control: meanings related to *permission, obligation, or volition (or intention*). Extrinsic modality refers to the logical status of events or states, usually relating to assessments of likelihood: *possibility, necessity, or prediction.*

For this reason, it is surprising that within the sub-category for attitudinal/modality stance (note attitude and modality are grouped together), the intention/prediction impersonal bundles include those "expressing predictions of future events that do not entail the volition of the speaker… usually used when explaining a logical or mathematical process that involved several steps" (Biber et al., 2004, p. 391). This seems at odds with the definition in Biber et al. (1999) above which has volition in the intrinsic, deontic group. Following this pattern, many of the bundles that, in my view, comment on the likelihood of outcomes would be categorised under attitudinal stance. However, this, I feel, makes a false distinction between the bundle *i think it's* included in 'the karyotyping would take only about probably ah, *i think it's* ten *days'*, which clearly comes under epistemic stance, with others where the speaker is stating their view on the likelihood or certainty of a result which is related to their background knowledge. For example, with *it's going to be*, while impersonal, in the following example the speaker is giving their view based on logical likelihood: *if it's familial it's going to be mostly genetic*. This could be replaced with *it'll be* or *i think it will be*. For this reason, in this study bundles that within the context indicate the speaker is speculating on the outcomes of tests, for example, have been categorised under epistemic stance.

6.3.5.3 Modification of the Discourse Organisers Category

Discourse organisers which show the links between prior and coming discourse also required further consideration and development. If we take a broad view of discourse organisers to include bundles that help the speaker organise and order their discourse logically, direct the receiver to relevant information and help them to negotiate content, the category expands from the original topic

introduction/focus and elaboration/clarification bundles to include identification/focus. Biber and Barbieri (2007) themselves move indication/focus bundles to discourse organisers, previously classified as referential expressions. Similarly, although studying academic writing, Cortes (2004) appears to rename the *discourse organisers* category as text organisers and includes focus and framing, previously again in *referential expressions*. While the original taxonomy includes time/place/text deixis under *referential expressions*, the bundles in this corpus that direct participants to, for example, visuals and handouts (text deixis, which is more suited perhaps to written discourse) seem here to be to the speaker directing students through their talks. For this reason, such expressions are included under *discourse organisers* bundles. Finally, within the corpus a small number of items were found such as *you know the* and *the kind of* which appear in certain contexts to be floor-holding. These are primarily a reflection of the nature of spoken discourse, characterised by real-time processing and production, and allow the speaker to continue their turn. A sub-group of *discourse organisers floor-holding* (DOFH) was therefore added to this group.

This broadening of the functions under discourse organisers is not inconsistent with other discussions concerning phrases used in text organisation, for example in the study of metadiscourse (e.g., Ädel, 2010; Hyland, 2005a).

6.3.5.4 Referential and Other Expressions

The *referential expressions* category was necessarily reduced somewhat as expressions that primarily direct participants through the discourse (identification/focus and text deixis) have moved to the *discourse organisers* category. Referential expressions here focus on the specification of attribute, imprecision and quantification.

Other expressions include a "*special conversational/speech function*" to allow for an emphatic no (*no no no*) and *subject-specific bundles in the discipline* to account for additional items that occurred across a number of texts (i.e. that did not seem restricted to the focus of one specific scenario).

6.3.6 Procedure for Identifying Lexical Bundles

This section explains the procedure for identifying three-and four-word lexical bundles. The first stage was to compile word lists of three- and four-word bundles for the whole PBL corpus and for each sub-corpus (i.e. AllPBLSs, PBL1s and PBL2s) using the WordList function of WordSmith 6.0 (Scott, 2004). (As part of the initial explorations, five and six-word bundles were included, but as there were few examples of five-word bundles and none of six-word bundles, these are not discussed here). Relevant texts were selected and an index file created for each. From this it is possible to compute for bundles, generating lists where frequency minimums and length of bundle can be set. Separate lists for three- and four-word bundles were then automatically generated. In this study, contractions were counted as one word. Since the only punctuation in the corpus is for pauses of two seconds or more (marked by a comma) and question marks, this determined the boundary of the search, i.e. bundles across such boundaries are not included. Items with no clear meaning were deleted (*the the the* and *the er the*). These lists were then stored as WordLists and on Excel spreadsheets. The latter allowed for easier manipulation of the data, for example in tracking changes to lists, for selecting and sorting only by range and/or frequency and, at a later stage, for adding additional notes and clarifications. The lists were then refined to ensure they met the range requirements. This went some way towards excluding both idiosyncrasies of individual speakers and highly topic dependent, localised use of items, the importance of which is noted by Biber, et al. (1999). For example, *cleft lip and* occurs 61 times but in only two texts; *Alpha Feto Protein* occurs 26 times but in one text only.

Once the word lists of bundles for each corpus was generated, items were categorised for structural patterns and later for discourse functions. The Concordance tool of WordSmith was used in order to generate individual lists of items, for example all instances of *we need to*. The Concordance function allows items to be analysed within their surrounding co-text, necessary for the precise clarification of some items, and for the more fine-grained analysis of discourse functions. The categories compiled by Biber et al. (1999) and Biber et al.,

(2004), outlined above were used as a starting point for both the structural classification and identification of discourse functions.

6.4 Results

This section is in three parts. The first looks briefly at the types and frequency of the bundles identified. This is followed by a discussion of structural categories and finally discourse functions.

6.4.1 Types and Frequency in the PBL Corpora

The lists presented and analysed here have not been further modified to take into account overlaps (complete and partial), but items were marked if they occurred within a longer bundle. Table 6.2 below presents the total types and frequency of three- and four-word bundles in the whole corpus (AllPBLs). The numbers presented are before any modifications (e.g., for dysfluencies and disregarded items)

Table 6.2: Types and Frequency of Three- and Four-Word Bundles

	PBL1s		**PBL2s**		**AllPBLs**	
	Types	Frequency	Types	Frequency	Types	Frequency
3-word	58	1088	77	1195	171	3354
4-word	5	56	3	40	9	125

We can see that in the PBL1 corpus there are approximately ten times as many types of three-word bundles as four-word bundles. The finding is consistent with that of Biber et al. (1999, p. 993). However, in PBL2s there are over 25 times as many three-word bundle types as four-word bundles and for the whole AllPBL corpus, approximately 20 times as many three-word bundles as four-word ones. These two findings are in contrast to Biber et al.'s (1999) comparison between conversation and academic prose (ten times as many three-word to four-word and four-word to five-word).

The overview of types and the number of specific bundles suggests that PBL1s and PBLs vary in their use of bundles, both in types and in frequency. Potential

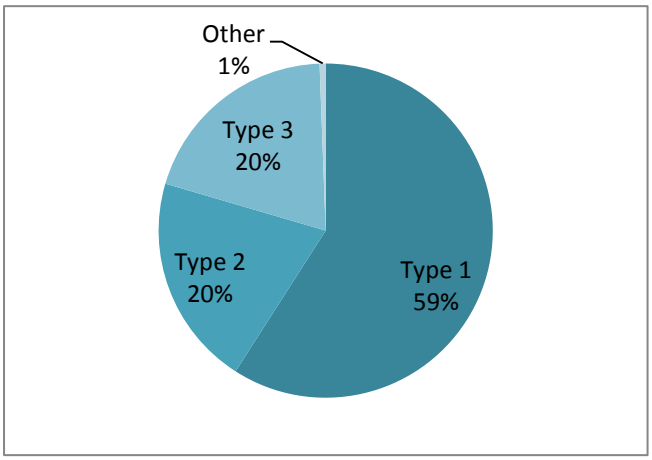
differences will now be explored through a consideration of structural categories and the discourse functions assigned to the items.

6.4.2 Structural Categories of Three-Word Bundles (AllPBLs)

The presentation of the results of the structural classification covers first a detailed discussion of the three-word bundles and then an overview of four-word bundles.

When we consider the overall distribution of the structural categories of three-word lexical bundles in the whole corpus (AllPBLs as shown in Figure 6.1), we can see the predominance of *oral* types (79%), but with a not insignificant *literate* element (20%). Of the oral types, 59% include A lexical verb phrase (Type 1), with 20% including a dependent clause fragment (Type 2). Although I have commented on difficulties comparing across corpora and of course between results for different lengths of bundles, Biber et al. (1999) found a much higher number of four-word bundles including part of a verb phrase in conversation (90%); the overall pattern in AllPBLs is more similar to classroom teaching, which has a more even spread of the three types (Biber et al., 2004, p. 382). The 20% that make up the Type 3 bundles (including a noun or prepositional phrase) in the AllPBL corpus possibly indicates more of an orientation to the informational content.

Figure 6.1: Structural Categories of Three-Word Bundles in the Whole Corpus (AllPBLs)



Type 1: bundles with verb phrase fragments (101 types)
Type 2: bundles with dependent clause fragments (35 types)
Type 3: bundles with noun/prepositional phrase fragments (34 types)
Other: (1 type)

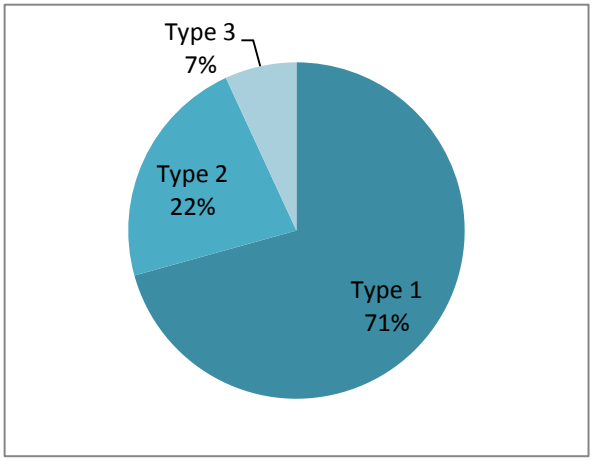
Note. Types 1 and 2 are oral, Type 3 literate.

6.4.3 Structural Categories of Three-Word Bundles in PBL1s

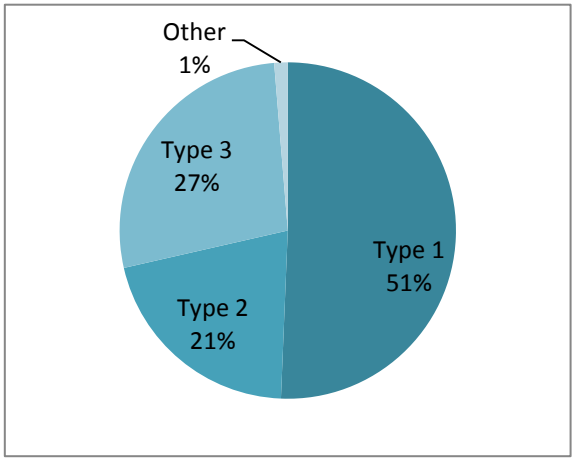
While the findings above for AllPBLs appear to indicate that overall the PBLs include more of a mix of structural categories than conversation (and so to some extent appear more similar to lectures and other classroom teaching), when we look at each sub-corpus we can see a striking difference between the two stages. Figure 6.2 shows that the structural categories in PBL1s are overwhelmingly oral in nature (93% in contrast to 72% in PBL2s). In contrast, PBL2s include a higher percentage of the literate Type 3 lexical bundles, i.e. those including NP or prepositional phrases (27% in PBL2s compared to only 7% in PBL1s).

Figure 6.2: Structural Categories of Three-Word Bundles in PBL1s and PBL2s

PBL1s



PBL2s



Type 1: bundles with verb phrase fragments (41 types, PBL1s; 39 types, PBL2s)
Type 2: bundles with dependent clause fragments (13 types, PBL1s; 16 types, PBL2s)
Type 3: bundles with noun/prepositional phrase fragments (4 types, PBL1s; 21 types, PBL2s)
Other: (1 type, PBL2s)

Note. Types 1 and 2 are oral, Type 3 literate.

In PBL1s, of the overall 93% categorised as oral (Types 1 or 2), we can also see a higher proportion including main lexical verb fragments (Type 1 account for 71%). Type 2 is still significant at 22% although part of the reason for a relatively high proportion having a dependent clause fragment can be attributed to a number of LBs including a *think* + clause, *e.g., i think it's* (with *that* omitted).

From the initial categorisation it was noted that a high proportion of bundles appear to indicate personal stance, e.g., *i think it's* and *i don't know*. There was also more questioning with the use of '*Yes-no*' and '*Wh-Q*' fragments. These points are investigated in more detail in the analysis of discourse functions.

6.4.4 Structural Categories of Three-Word Bundles in PBL2s

The structural categories found in PBL2s are also still largely oral at 72%. Again, there is a predominance of items in the Type 1 group, with bundles including a verb phrase fragment accounting for 51%. This is noticeably lower than the 71% found in PBL1s, but we can see a broadly similar percentage of bundles with a dependent clause fragment (21% here, compared to 22% in PBL1s). The main

difference, however, is that there is a more notable literate element in PBL2s at 27%, i.e. we see a change in distribution from bundles with a main verb element in PBL1s to more items with noun/prepositional phrase elements in PBL2s (see Figure 6.2 above). This might be accounted for by the longer presentations with a focus on informational content and based on prepared written reports. For example, instances of the phrasal literate items include *a lot of, one of the, in this case*, all appearing to refer to aspects of the academic subject. The uses will be discussed in the exploration of discourse functions in Sections 6.4.9-6.4.15. The next section looks in more detail at the sub-categories of each structural type.

6.4.5 Distribution of Three-Word Structural Sub-Categories

This section takes a closer look at the distribution of sub-categories of each type of bundle and again highlights the differences in use between PBL1s and PBL2s. Table 6.3 on the next page shows the number and percentage of the different three-word lexical bundles by structural type.

The description that follows the table compares the percentage of bundles within the three different types and sub-categories. Numbers in brackets following the example bundles in italics refer to the number of bundles found in the relevant sub-corpus, unless otherwise explained. The list of the thirty most frequent bundles is provided in Table 6.4.

Table 6.3: Distribution of Structural Sub-Categories of Three-Word Bundles

		Example	AllPBLs	%	PBL1s	%	PBL2s	%
Type 1: Number and % of Three-Word Bundle Types with *verb phrase fragments*								
1a	(Connector +) 1st/2nd person personal pronoun + lexical VP fragment	*you don't have* *i'm not going*	41	24%	16	28%	16	21%
1b	(Connector +) 3rd person pronoun + VP fragment	it's going to so this is	33	19%	7	12%	15	19%
1c	Discourse marker + VP fragment	*you know it (was)*	0	0	0	0	0	0
1d	VP (with non-passive verb (active))	*going to talk*	10	6%	8	14%	2	3%
1e	VP (with passive Vb)	*[can be used]*	0	0%	0	0%	0	0%
1f	Yes-no question fragments (including also connector + yes-no Q)	*but do you*	8	5%	5	9%	2	3%
1g	Wh-question fragments	*what do you*	7	4%	5	9%	1	1%
1h	Other	*and then you*	2	1%	0	0%	3	4%
Type 2: Number and % of Three-Word Bundle Types with *dependent clause fragments*								
2a	1st/2nd person pronoun + dependent clause fragment	*you know (that) it's* *you know what*	11	6%	8	14%	6	8%
2b	Wh-clause fragments	*don't know what*	2	1%	1	2%	4	5%
2c	If-clause fragments Connector + if + PN	*don't know if* *so if you*	9	5%	1	2%	4	5%
2d	(Verb/adjective +) to clause fragment	*to do with*	10	6%	2	3%	2	3%
2e	That-clause fragments	*is that the*	2	1%	0	0%	0	0%
2f	Verb + that clause fragment	*think we should*	1	1%	1	2%	0	0%
Type 3: Number and % of Three-Word Bundles Types with *noun phrase and prepositional phrase fragments*								
3a	(Connector +) NP with *of*-phrase fragment	*a lot of*	16	9%	3	5%	9	12%
3b	NP with other post-modifier fragment	*mutations in the*	3	2%	0	0%	1	1%
3c	Other noun phrase expressions	*the screening test;* *the first one*	11	6%	1	2%	6	8%
3d	Prepositional phrase expressions	*of them are*	4	2%	0	0%	4	5%
3e	Other NP expressions	*because of the*	0	0%	0	0%	1	1%
3f	Other expressions	*no no no*	1	1%	0	0%	1	1%

PBL1s have a higher proportion (28%) of 1a (*Connector+) 1st/2nd person personal pronoun + lexical VP* than PBL2s (21%). For example, in PBL1s the three most common items in this group are *i don't know* (73), *we need to* (49) and *yeah i think* (31); the three most frequent three-word items in PBL2s in this category include *i don't know* (72), but also *you can see* (38) and *i don't think* (28). In contrast PBL2s evidence a higher proportion of 1b (*Connector) + 3rd person pronoun + VP fragment*, 19% compared with 12% in PBL1s. For example, *there is a* occurs 27 times and *this is the* 23 times in PBL2s (as we shall see later, neither occurs in the thirty most frequent items in PBL1s).

In PBL1s, the category 1d, *VP (with non-passive verb*) is much higher proportionally than in PBL2s (14% to 3%). Examples in PBL1s include *going to be* with 39 occurrences and (not shown) *need to do* (12) and *going to do* (10). The two examples in PBL2s are *going to talk* (17), clearly linked to the main purpose of reporting in a short presentation, and *going to be* (11). Interestingly, although passives are generally thought to be common in academic discourse, there were no examples of 1e, *VP (with passive Vb)* in either sub-corpus. PBL1s also, unsurprisingly, include more 1f, *Yes-no* and 1g, *Wh-question fragments*: 9% each, compared to only 3% and 1% respectively in PBL2s.

In the Type 2 group, 2a, *1st/2nd person pronoun + dependent clause fragment* are also much more frequent in PBL1s (14% to 8%), clearly represented by *i think it's* (33), *i think we* (33) (both of which have the *that* element omitted) and *i think that* (32); all three items are included in the ten most frequent three-word bundles in PBL1s. While lower proportionally at only 8%, this 2a category is also dominated by *think (that)* items, e.g., *i think it's* (30) and *i think that* (15).

PBL2s are, as noted, characterised by a higher proportion of the literate, Type 3 phrasal lexical bundles. For example, 3a bundles *(Connector+) NP with of-phrase fragment,* account for 12% compared to 5% and only three items in PBL1s. In PBL2s, this type of bundle includes *a lot of* (49) and *one of the* (33), compared with lower numbers in PBL1s.

The structural types reflect the nature of the two aspects of the genre: the highly interactive, tentative and questioning language used in PBL1s which

involves individuals sharing and putting forward ideas, compared to the stronger focus on informational content in PBL2s. In this way, we can see in PBL1s some parallels to conversation (characterised by a much higher proportion of declarative clauses and questions) and for PBL2s to classroom teaching and lectures, which include a higher proportion of literate lexical bundles (Conrad & Biber, 2004; Nesi & Basturkmen, 2006).

The charts above only indicate the distribution of the types within the structural categories of the three-word bundles and although some examples have been included for illustrative purposes, I have yet to identify the most frequent individual bundles. For this we need to look at lists of actual types and the number of occurrences of each one. Again, we can identify differences when comparing the whole corpus and the two sub-corpora.

6.4.6 Types and Frequency: The Most Frequent Three-Word Bundles

Table 6.4 which follows lists the 30 most frequent three-word lexical bundles in PBL1s and PBL2s. Shared items (i.e. in PBL1s, PBL2s and the AllPBLs) are shaded and items that have complete overlaps with four-word bundles are asterisked. It should be noted that in relation to occurrences, the total number of bundle types and the thirty most frequent items for the whole corpus include items not necessarily found in both of the sub-corpora. This is because in order to be included in the whole AllPBL corpus list, items need to occur ten times and in six texts. However, one item might occur very frequently in PBL1s, for example, *i think we* occurs 33 times in PBL1s and in five texts but only five times in PBL2s. It still reaches the minimum frequency of ten for AllPBLs, but although evident in PBL2s it is not sufficiently frequent in the second stage to be included in the PBL2 list. This is a good example of an item very clearly linked to the specific aim of the PBL1 stage (i.e., collaborative group work, sharing knowledge and identifying agreed learning objectives). It also demonstrates the importance of studying sub-corpora in detail, particularly when component parts may be very different. If we had considered only the corpus as a whole, aspects clearly related to the specific communicative purpose of each stage might be overlooked.

Table 6.4: 30 Most Frequent Three-Word Bundles (Structural Types)

AllPBLs	Freq.	Texts	PBL1s	Freq.	Texts	PBL2s	Freq.	Texts
*i don't know	145	10	*i don't know	73	5	*i don't know	72	5
*a lot of	75	10	*we need to	49	5	*a lot of	49	5
i think it's	63	10	*do you think	47	5	*you can see	38	5
*do you think	61	9	*going to be	39	5	one of the	33	5
we need to	58	8	i think it's	33	5	i think it's	30	5
one of the	56	10	*i think we	33	5	i don't think	28	5
going to be	50	10	*i think that	32	5	there is a	27	5
*i don't think	49	10	it would be	32	5	this is the	23	5
i think that	47	10	yeah i think	31	5	*i'm going to	22	5
*it would be	47	9	*you want to	31	5	if you have	22	4
there is a	42	10	a lot of	26	5	the level of	21	3
you can see	42	7	that would be	24	5	*going to talk	17	5
*you want to	40	9	one of the	23	5	*the first one	17	5
i think we	38	9	*i don't think	21	5	in this case	17	4
that would be	37	9	i think the	21	5	so if you	17	4
yeah i think	36	8	*what do you	21	5	*don't know if	16	5
but i think	34	9	but i think	20	5	you don't have	16	5
this is the	33	7	do you want	20	5	first of all	16	4
i think the	32	9	it could be	20	5	it will be	16	4
*what do you	32	8	do you know	19	5	i think that	15	5
do you know	31	8	*want to do	18	5	so this is	15	5
*do you want	29	8	i think it	17	5	it would be	15	4
it could be	29	8	i'm not sure	17	5	so i think	14	5

AllPBLs	Freq.	Texts	PBL1s	Freq.	Texts	PBL2s	Freq.	Texts
but i don't	28	10	it should be	17	5	there is no	14	5
so i think	28	10	we don't know	16	4	but i think	14	4
i'm going to	28	9	i think you	16		do you think	14	4
it should be	28	9	*but i don't	15	5	you have a	14	4
i'm not sure	26	10	there is a	15	5	you know the	14	4
*don't know if	26	8	i think that's	14	5	so that's why	14	3
i think it	25	10	so i think	14	5	but i don't	13	5

Note. Items in PBL1s and PBL2s and the whole corpus are shaded; an asterisk denotes a bundle which is part of a larger bundle or overlap.

When we consider in more detail the proportion of types to occurrences, the thirty most frequent three-word lexical bundles are disproportionately representative of the overall totals. In the whole corpus (AllPBLs) the thirty most frequent lexical bundle types (of 171) represent 38% of the total occurrences. The thirty most frequent items in PBL1s (of 58) account for 71% of actual occurrences and in the PBL2 sub-corpus, the thirty most frequent items (of 79) account for 55% (after this, there appears to be a more even distribution of frequency). This would again suggest more repetition of a smaller range of items in PBL1s.

As noted, the shaded items in Table 6.4 are found in the whole AllPBL corpus and in each sub-corpus. We can see that although some lexical bundles occur in both the PBL1s and PBL2s, for many the frequency is different. Only 12 lexical bundles occur in the thirty most frequent items in both PBL1s and PBL2s. While *i don't know* and *i think it's* are very frequent in both, there is a difference in the number of occurrences and relative frequency of other items within each sub-corpus. For example, *a lot of* occurs only 26 times in PBL1s but 49 times in PBL2s.

In PBL1s, although the numbers have not been adjusted for overlaps, the 71% coverage by approximately half the total number of types as I have noted suggests a high level of repetition. We can also see similarities in the composition, with bundles containing mental process verbs *think, know*, or a modal or semi-modal, e.g., *would* and *need to* being quite common. Examples (with number of bundles in brackets) include:

> *do you think* (47), *i think it's* (33), *i think we* (33), *yeah i think* (31), *i don't think* (21)
>
> *we need to* (49), *that would be* (24), *it could be* (20).

Although not as high proportionally, in PBL1s there is also evidence of a number of items starting with a connector, e.g., *but i think* (20), *so i think* (14), the first two possibly reflecting the collaborative nature of the PBLs with participants sharing, agreeing and challenging as they build knowledge.

In PBL2s we also see a slightly higher number of item types starting with a connector (6), with four bundles starting with *so* (e.g., *so if you* (17), *so that's why* (14)) and two with *but (but i think* (14), *but i don't* (13)). This appears to reflect the need for speakers to provide explanations and to link and organise ideas in the longer turns in PBL2s. In contrast, an initial preliminary survey of the examples of *but i think* in PBL1s shows this bundle appears to occur more often in response to previous interactions, while *but i think* instead contrasts points within a turn in PBL2s (e.g., 'the recurrence risk is low, *but i think* when it's in').

The structural categorisation indicates a higher proportion of *Yes-no* and *Wh-questions* in PBL1s, e.g., *do you think* (47), *what do you* (21), with only one of this type in the top thirty of PBL2s (*do you think*, (14)). In PBL2s there is much more use of declarative forms.

In terms of modals, the two stages include quite different examples. PBL1s include items with *would, could* and *should. It would be* occurs 32 times (compared with 15 in PBL2s), *that would be* 24, *it could be* 20, and *it should be* 17 times, none of which appear in the most 30 most frequent bundles in PBL2s. PBL2s in contrast include *you can see* (38), which appears to direct attention, and *it will be* (16) expressing possibility.

PBL2s also show a higher use of *going to*, e.g., *i'm going to* (22), *going to talk* (17) and *going to be* (11), but with only *going to be* (39) occurring in the thirty most frequent items in PBL1s. The first two examples above are clearly linked to the introduction of the presentation (*going to be* appears related to a predictive or resultative function of explaining medical conditions).

We can also see a higher use of first and second person pronouns in lexical bundles in PBL1s, as participants put themselves and their ideas forward, sometimes quite tentatively. Although the thirty most frequent bundles represent different percentages in each corpus, PBL1s appear to have a much higher orientation to the person and the group, reflecting the findings from the study of the individual pronouns in the previous chapter. Twenty-one of the thirty most frequent lexical bundles in PBL1s include either *i* (14), *we* (3) or *you*

(6), with two items including two pronouns *i think we* and *i think you*. This is compared to fifteen of the thirty most frequent bundles in PBL2s, with only eight instances of *i* and seven of *you* and no examples of *we*. In contrast to the higher personal pronoun use in PBL1s, we see more bundles with noun phrase or prepositional phrase fragments in PBL2: *a lot of* (49), *one of the* (33), *the level of* (21), *the first one* (17), *in this case* (17), *first of all* (16): six items (and 153 occurrences in total). This is compared with *a lot of* (26) and *one of the* (23), a total of only 49 occurrences in the top thirty bundles in PBL1s.

In considering the most common types and occurrences this again reflects the communicative purpose of each stage in the PBL cycle. In PBL1s, the orientation is more to the group and each other in contrast to PBL2s which focus more on communicating the informational content, reporting findings and guiding the listeners through a longer extended turn. The structural categories found here indicate that when we look more closely at the specific communicative purpose of each stage, the use is not exactly the same as in classroom teaching and lectures, noticeably so in PBL1s. This is very possibly a reflection of the much more student-centred interactive nature of the discourse, particularly in PBL1s and to the possibly more involved but also interactive nature of PBL2s, with students engaging in questioning at the end of reports possibly more often even than in interactive lectures.

6.4.7 Frequency of Four-Word Lexical Bundles

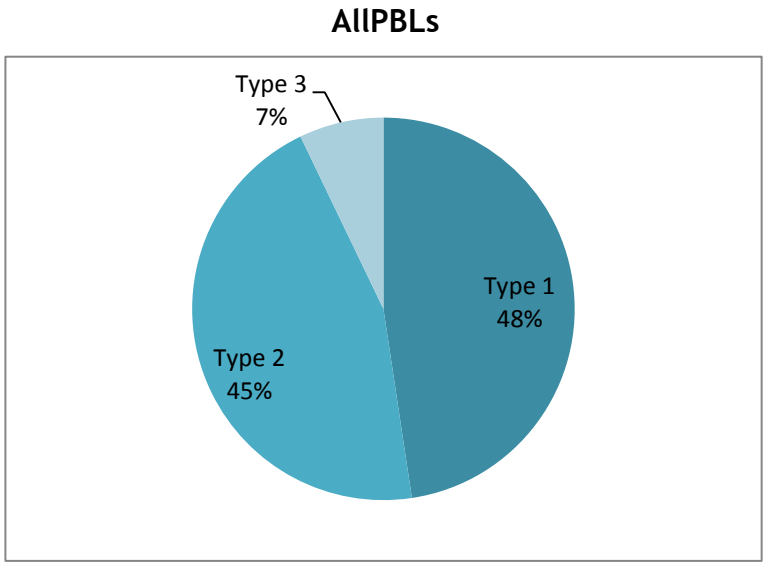
Although much less frequent in use in the PBL corpus, the four-word bundles have also been categorised by structure. Using the same parameters as three-word bundles (i.e. frequency ten; range three in each sub-corpus or six in AllPBLs), we only find a very small number of four-word bundles (five in PBL1s, three in PBL2s and nine in the whole corpus). By changing the raw frequency to five we find 32 four-word bundles in PBL1s, 23 in PBL2s and 42 in the whole corpus.

6.4.8 Structural Categories of Four-Word Lexical Bundles

Figure 6.3 illustrates the structural types of four-word bundles in the combined AllPBL corpus and the structural types in PBL1s and PBL2s.

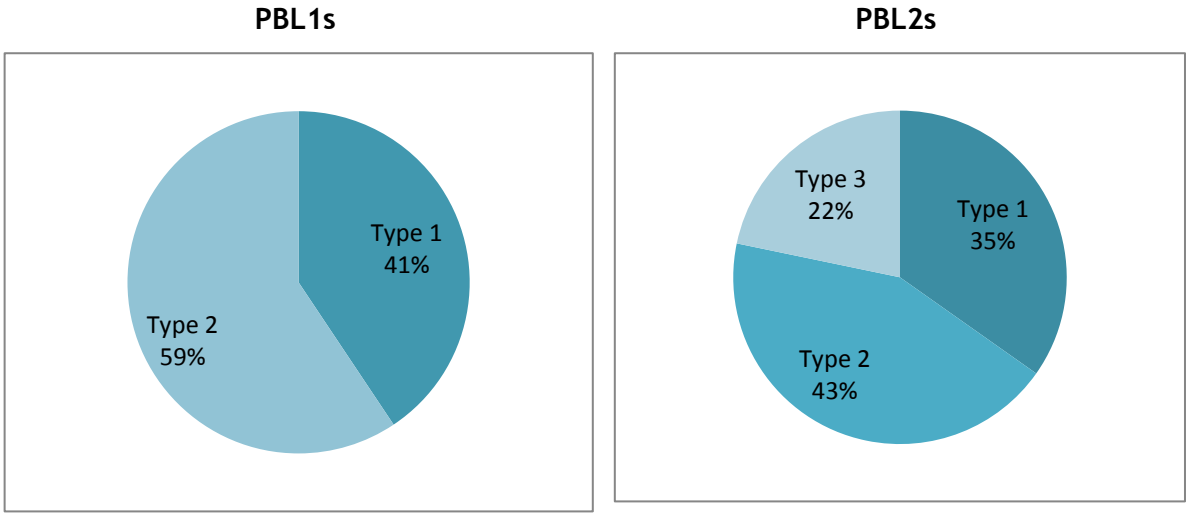
Figure 6.3: Structural Categories of Four-Word Bundles

Structural Categories of Four-Word Bundles (AllPBLs)



Type 1: bundles with verb phrase fragments (20 types)
Type 2: bundles with dependent clause fragments (19 types)
Type 3: bundles with noun/prepositional phrase fragments (3 types)

Structural Categories of Four-Word Bundles (PBL1s and PBL2s)



Type 1: bundles with verb phrase fragments (13 types in PBL1s; 8 types in PBL2s)
Type 2: bundles with dependent clause fragments (19 types in PBL1s; 10 types in PBL2s)
Type 3: bundles with noun/prepositional phrase fragments (5 types, in PBL2s)

As we can see, the initial overview from AllPBLs masks the different distributions found within the individual stages. The second row shows the clear distinction between the types used in PBL1s and those in PBL2s, with the second stage again making more use of the literate type bundles.

What is striking is that in PBL1s, none of the four-word bundles, even with a reduced cut-off, is of the literate type. In PBL1s, the most common type of bundle is the Type 2a *1st/2nd person pronoun + dependent clause* fragment (16 of the 19 Type 2s and half of all the types in this corpus). In PBL2s, the most common type is Type 1b *(Connector+) 1st/2nd person personal pronoun + lexical VP fragment*, accounting for five of the 23 types. While PBL2s present a more balanced and mixed picture, the pattern in PBL1s is in stark contrast to classroom teaching and lecture corpus findings where a mix is found (Biber & Barbieri, 2007; Nesi & Basturkmen, 2006). Table 6.5 on the next page presents the four-word bundles found in PBL1s and PBL2s. Shaded items occur in both PBL1s and PBL2s.

We can again see in PBL2s that common bundles reflect the nature of an oral presentation to a group, with a number of items directing the listeners through the information, including *you can see in* (7), *as you can see* (6) and *for example if you* (5). In contrast, PBL1s include more bundles which reflect the need to identify group objectives. They include *i think we should* (10) and *yeah i think we* (5) and bundles to share information, state uncertainty or question each other about the topic, for example *it's going to be* (6), *i don't think it's* (5), *what do you think* (7).

Only four four-word bundles are common to both PBL1s and PBL2s: *but i don't know* (9 and 6 occurrences respectively*)*; *i don't know if* (8 and 15), *i don't know what* (both 7) and *what do you mean* (6 and 5). Otherwise, the most frequent items in each stage of the cycle are different. PBL1s also show more variation: even without adjusting for overlaps, there are notably more types in PBL1s (32 compared with 23 in PBL2s).

Table 6.5: Types and Frequency of Four-Word Bundles (Structural Categories)

PBL1s			PBL2s		
Four-Word Bundles	**Freq.**	**Texts**	**Four-Word Bundles**	**Freq.**	**Texts**
do you want to	14	5	i don't know if	15	5
*know what i mean	11	3	*i'm going to talk	14	5
*you know what i	11	3	going to talk about	11	5
*do you know what	10	3	i don't know what	7	4
i think we should	10	5	or something like that	7	3
*but i don't know	9	5	you can see in	7	3
i don't know if	8	3	*as you can see	6	4
think we need to	8	3	but i don't know	6	5
we need to look	8	3	if there is a	6	4
*i don't know what	7	4	the first one is	6	3
we need to do	7	4	there's a lot of	6	3
what do you think	7	4	*you can see the	6	4
yeah that's what i	7	3	a lot of people	5	3
you want to do	7	4	a lot of the	5	3
do we need to	6	3	can see in the	5	3
do you think we	6	3	*for example if you	5	4
i don't think so	6	4	if you have a	5	3
i think we have	6	3	if you have the	5	3
if you want to	6	4	*if you look at	5	4
it's going to be	6	3	is one of the	5	3
not going to be	6	4	one of the most	5	3
what do you mean	6	3	that kind of thing	5	4
but i think that	5	3	what do you mean	5	3
going to have to	5	3			
i don't think it's	5	4			
i think it's a	5	3			
is going to be	5	3			
so it would be	5	3			
that's what i was	5	3			
what does it mean	5	3			
yeah i think that	5	3			
yeah i think we	5	4			
32 Types	222		23 Types	152	

Note. Shaded items occur in both PBL1s and PBL2s; asterisked items were found to be part of a longer 5-word bundle.

The increased number of types may be because of varied content in PBL2s and so less repetition of bundles. While PBL1s involve many instances of putting forward ideas, giving opinions and asking for clarifications, PBL2s concentrate more on reporting findings of medical conditions, procedures and processes, which vary obviously according to topic of the scenario. PBL1 four-word bundles are characterised by a much higher proportion including personal pronouns. More

than two thirds include *I, you* and/or *we*. There are also more modals/semi-modals and again more bundles including *think* and *know*. While over half of the four-word bundles in PBL2s include *I* or *you*, there are no examples of bundles with *we*.

Interestingly, only three bundles in the PBL corpora can be found in the twenty most frequent bundles in the corpus of lectures study presented by Nesi and Basturkmen (2006). In my corpus, two occur in PBL1s, *if you want to* (6) and *is going to be* (5), rather than PBL2s where one might expect more similarities as it includes the long report presentations. Instead we find only *if you look at* (5) in PBL2s, suggesting again that student-led classroom discourse is different from classroom teaching and lecture discourse (both of which are likely to be more teacher-led).

In order to obtain greater insights into specific uses in the PBLs and to confirm or otherwise initial impressions of use, the next stage of the analysis consisted of the identification of discourse functions.

6.4.9 Discourse Function Categories in PBLs

Section 6.3.5 discussed the discourse functions and modifications made for the analysis in this corpus. The final categories and sub-categories are presented in Table 6.6 and form the basis of discussion of the discourse functions. To facilitate coding, the main groups were abbreviated as follows: the *stance category* is separated into two groups, Stance A, for *epistemic stance* (SAE), with a further distinction between personal and impersonal (SAEp/i), and Stance B with sub-groups 1-4 (SBA1-SBA4) for *attitudinal stance*. *Discourse organiser* sub-groups are named DOA to DOC and DOFH for the additional d*iscourse organiser floor-holding* sub-category. Similarly, the *referential expressions* sub-category are named REA to RED, explained in Table 6.6.

Table 6.6: Discourse Function Categories

Stance Expressions		
Stance A: Epistemic (SAEp/i)		
Code	Sub-category	Example
SAEp	Epistemic personal e.g., uncertainty; hedging	*I don't know* *I think it's*
SAEi	Epistemic impersonal	*it would be*
Stance B: Attitudinal (SBA1-SBA4)		
SBA1	Desire	*want to do*
SBA2	Obligation/directive	*we need to*
SBA3	Intention/prediction within speaker's volition	*you're going to*
SBA4	Ability	*we can see*
Discourse Organisers (DOA-DOFH)		
DOA	Topic introduction/focus	*first of all*
DOB	Topic elaboration/clarification	*so if you* *do you think*
DOC	Concluding/summarising	*so that's why*
DOFH	Discourse organisers: floor-holding	*the kind of*
Referential Expressions (REA-RED)		
REA	Identification/reference to a specific attribute	*which is a* *there is a*
REB	Imprecision	*it's kind of*
REC	Quantification	*a lot of*
RED	Time/place	*in two thousand* (year)
Other Expressions		
Special conversational /speech functions		*no no no*
Subject specific bundles in the discipline		*mutations in the*

The next section explains the procedure for categorising bundles according to their discourse function.

6.4.10 Procedure for Categorising by Discourse Function

In order to study the discourse functions of three-word bundles within each sub-corpus, the same range and frequency parameters for the structural classification were used. In the much closer reading required to categorise discourse functions, it soon became clear that a number of quite high frequency three-word bundles were entirely subsumed within longer four-word bundles.

The detailed reading of all bundles required for the study of discourse functions also made it much more practical to identify and remove such items. In order to establish a clear distinction between the discourse functions of three-word (only) and four-word bundles and to make the consideration of discourse functions more manageable, any three-word bundles found to be entirely part of a four-word bundle were manually removed. As a result of this modification, while the three-word items discussed below might overlap with another three-word item, for example *so i think* and *i think that*, none of the bundles in the three-word list are part of any longer bundles. Any that were removed are now included in the discussion of four-word bundles.

A final point concerning the multifunctional nature of lexical bundles should also be made. Bundles may of course have more than one function, apparent in different contexts. For each bundle type the categorisation ensured that different function(s) for each occurrence were identified and quantified. However, even within one context, bundles may appear to have a dual function. Because including each item with a dual function in one context in both categories would unrealistically inflate the findings, a decision was made to categorise and count only one primary use. As with the categorisation of pronouns, this was done in two stages. In the first round of categorisation a function was assigned (primary and secondary noted), and checked one week later.

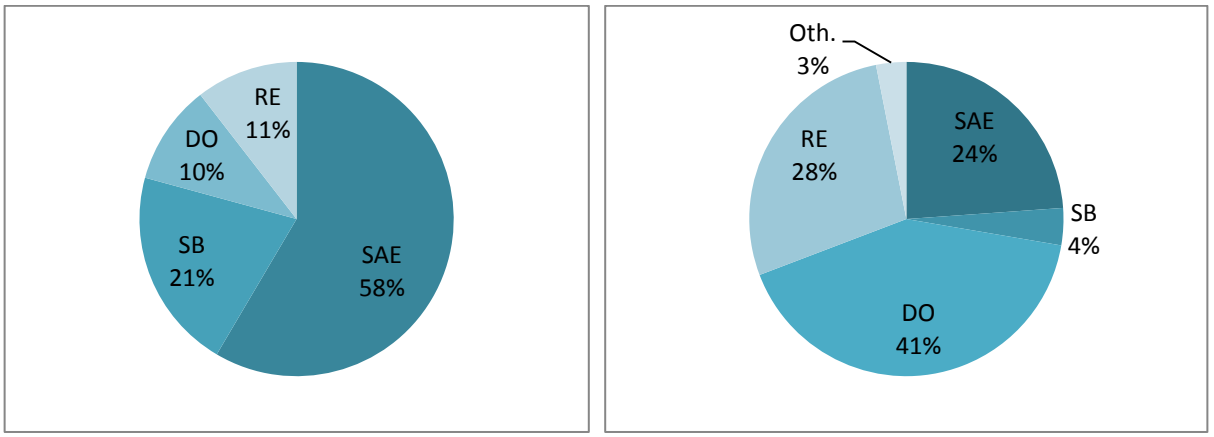
The following section will present the overall results of the common discourse functions that have been identified. Although general patterns in the larger scale studies will be commented upon in relation to findings in this study, the note of caution about such comparisons given in 6.3.3 above should be remembered.

6.4.11 Discourse Functions of Three-Word Lexical Bundles: Patterns of Use

In this section the frequency of three-word lexical bundles by overall category in PBL1s and PBL2s will be presented. This is followed by a more detailed consideration of the sub-categories for each group in PBL1s and then PBL2s.

The overall discourse functions of three-word lexical bundles in PBL1s and PBL2s are illustrated in Figure 6.4 below. As we can see in in the pie chart for PBL1s, the stance function overwhelmingly dominates, accounting for 80% of the bundles (59% epistemic and 21% attitudinal stance). Although stance is one main grouping, because the sub-categories account for such high proportions of use, they are sub-divided here. Referential expressions and discourse organisers each account for 10%.

Figure 6.4: Discourse Functions of Three-Word Lexical Bundles in PBL1s and PBL2s



SAE: Stance A: Epistemic stance SBA: Stance B: Attitudinal stance
DO: Discourse Organisers RE: Referential Expressions O: Other

This heavy reliance on epistemic stance in particular reflects the purpose of PBL1s, which involve hypothesising and sharing tentative knowledge. This stage is highly interactive and involved, with many short turns and requiring less obvious direct organisation of the message in the way of discourse organiser bundles. Stance expressions were also found to be dominant in all spoken registers by Biber and Barbieri (2007, p. 273) and in general terms, the patterns represented in PBL1s are similar to those in study groups where stance expressions far outstrip both discourse organisers and referential expressions.

The pattern is quite different in PBL2s. While stance bundles are not insignificant, accounting for 28% of the lexical bundles, the majority of three-word bundles (41%) are categorised as discourse organisers, followed by referential expressions (28%). The *Other* category includes a special speech function item (*no no no*) and a subject-specific bundle (*mutations in the*). These

results indicate a much more equal distribution of functions and again reflect the purpose of the sub-genre. In contrast to PBL1s, the presentation in PBL2s requires more organisation of the discourse and more work on the part of the speaker to orient the listener to the content. They also involve a much more in-depth discussion of the subject matter, thus involving more reference to concepts and conditions and resulting in a higher proportion of referential bundles. This more even distribution is similar to the patterns identified by Biber et al. (2004, p. 396) in classroom teaching, where relatively high proportions of stance and referential bundles were found (although they found lower uses of discourse organisers than we find in PBL2s). As Biber at al. (2004) note:

> Classroom teaching combines the functional and communicative priorities of involved spoken discourse (shown by the dense use of stance bundles) with the priorities of informational written discourse (shown by the dense use of referential bundles) …. This pattern apparently reflects the complex communicative demands of this register. (p. 397)

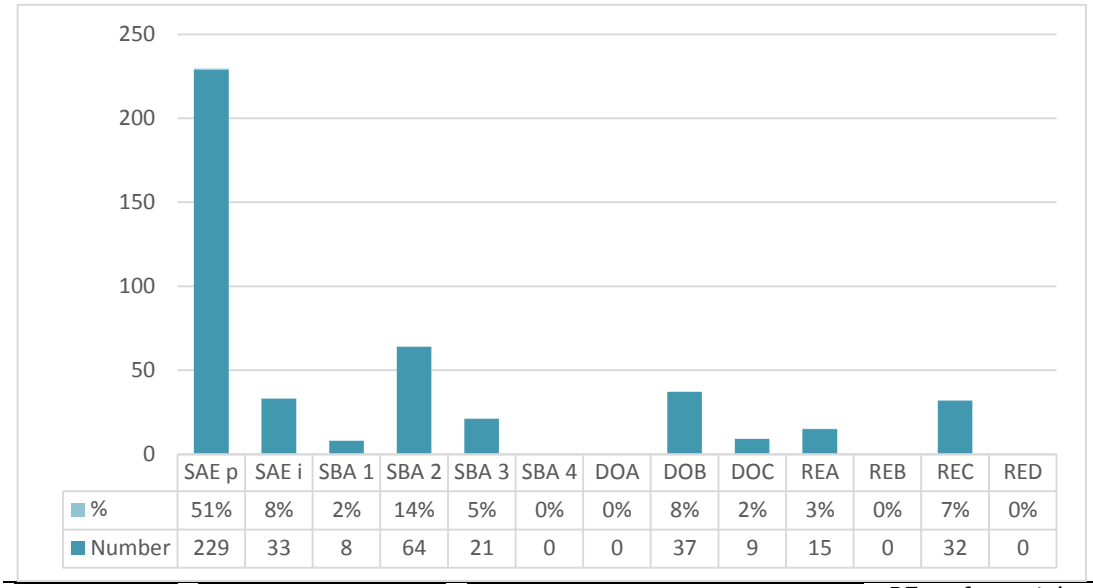
The similar pattern found in PBL2s may be because this stage in the cycle is more similar to a lecture style, teacher-fronted classroom teaching situation with one person in control of and leading the discourse at any one point in time.

Biber and Barbieri (2007, p. 274) in their study of spoken university registers also note that, in relation to stance bundles at least, "each register relies on different functional sub-categories, reflecting the particular communicative needs of the register". The next section will consider the specific sub-categories found in the PBL1 and then PBL2 corpus.

6.4.12 Discourse Functions of Three-Word Bundles in PBL1s

Figure 6.5 which follows provides a more detailed breakdown of the number and percentage of discourse functions of three-word bundles in PBL1s, illustrating the distribution across the subcategories. Table 6.7 then provides a breakdown of the number and sub-category of individual bundles. The results are discussed in detail in the following sections.

Figure 6.5: Discourse Functions by Sub-Category of Three-Word Bundles in PBL1s



	SAE p	SAE i	SBA 1	SBA 2	SBA 3	SBA 4	DOA	DOB	DOC	REA	REB	REC	RED
%	51%	8%	2%	14%	5%	0%	0%	8%	2%	3%	0%	7%	0%
Number	229	33	8	64	21	0	0	37	9	15	0	32	0

SAE: epistemic stance
SAEp: personal
SAEi: impersonal

SBA: attitudinal stance
SBA1: desire
SBA2: obligation/directive
SBA3: intention
SBA4: ability

DO: discourse organisers
DOA: identification/focus
DOB: elaboration/clarification
DOC: concluding/summarising

RE: referential expressions
REA: identification/attribute
REB: imprecision
REC: quantification
RED: time/deixis

In Table 6.7, where one type has been identified as having a different function in a different context, this is shown in column four. For example, *going to be* has been categorised on some occasions under epistemic stance (SAEi) and on others as part of the attitudinal stance sub-group of intention (SBA3).

Table 6.7: Types, Frequency and Discourse Functions of Three-Word Bundles in PBL1s

	Three-Word Types	Frequency	Discourse Function Sub-Categories
1.	a lot of	22	REC
2.	be part of	10	REC
3.	but i think	11	SAEp
4.	going to be	22	SAEi (13) SBA3 (9)
5.	how do you	10	DOB
6.	i'm not sure	17	SAEp
7.	i don't know	56	SAEp
8.	i think it	17	SAEp
9.	i think it's	26	SAEp
10.	i think so	12	SAEp
11.	i think that	21	SAEp
12.	i think the	17	SAEp
13.	i think we	12	SBA2 (5) SAEp (7)
14.	i think you	15	SAEp
15.	it could be	20	SAEi
16.	it should be	14	SBA2
17.	so i think	12	DOB (3) DOC (9)
18.	there is a	15	REA
19.	want to do	11	SBA1(8) SBA2 (2) SBA3 (1)
20.	we don't know	16	SAEp
21.	we need to	22	SBA2
22.	what are the	10	DOB
23.	what is the	14	DOB
24.	yeah i think	14	SAEp
25.	you have to	11	SBA2
26.	you need to	10	SBA2
27.	you're going to	11	SBA3
	27 types	448	

6.4.12.1 Stance Bundles in PBL1s

The stance grouping is further sub-categorised into epistemic stance and attitudinal stance.

As we have seen, the most striking finding is the extensive use of epistemic stance bundles, accounting for 59% of the bundles used, with 14 types and 262

individual occurrences. The numerous epistemic stance bundles in PBL1s reflect the requirements to speculate on conditions, to share knowledge and then to identify issues and gaps, which become learning objectives. Although there are 14 types, all but three are extended variants of *i think* or *don't know*. In sharing knowledge the students may appear tentative and so downplay their contributions.

There are 56 occurrences of *i don't know*, although when we consider the co-text we can see it can be used in different ways. It may be presenting the speaker's hesitation or lack of certainty or knowledge, i.e. the more prototypical function. In the next example, *i don't know* is used frequently and obviously to indicate lack of certainty or knowledge:

(1) S25: but *i don't know* i mean like, uteral. like something in the uterus and something in the colon, *i don't know* if they are, linked are they? (PBL11.1)

Quite a number are short answers (2) and near or at the end of utterances (3):

(2) S34: they could could they see it on ultrasound?
S32: *i don't know* (PBL12.1)

(3) S26: there's something about burning [?] *i don't know* (PBL11.1)

As noted above, depending on the context, the same bundle or type may be used in different ways and on occasion indicate a dual or at least primary and secondary function. *I don't know* can indicate uncertainty but it also appears to have another floor holding function, combining uncertainty but also allowing the speaker to maintain the floor:

(4) S5: but in those cases, they they have polyps in the early age. around, *i don't know*, i mean twenty or thirty (PBL11.1)

It may also in fact, indicate tentative knowledge, as in the next example (5).

(5) S26: it affects a lot of people it's *i don't know* exactly one in ten one in twenty five it affects, it's quite high it's like (PBL11.1)

This variety in functions for *i don't know* reflects the findings of Tsui (1991), who identifies a range of functions from “prefaces to disagreement” to “markers of uncertainty”. She suggests the common pragmatic function is “often a concern to save the face of self and others” (p. 607).

Nine of the bundles in the epistemic stance category include *i think* and again express degrees of certainty or uncertainty on the part of the speaker. By considering the items within their context, we can identify a number of more specific uses. Rather than a more direct pronouncement, *i think we* can be used in part to show a degree of certainty, acknowledging the need for agreement of other members of the group:

(6) S14: but we don't need to talk about Down Syndrome again
S13: yeah
S8: yeah *i think we* already know it (PBL9.1)

Rather than showing tentativeness, in the next example we can see *i think it’s* being used to indicate certainty politely, as here the speaker does in fact appear fairly confident in their knowledge, going on to repeat their assertion:

(7) S3: what does it mean hemizygosity?
S4: *i think it's* when there's only like one copy
S3: hemizygosity?
S4: *i think it’s* cos it's only one there's only one copy or something (PBL7.1)

The impersonal epistemic stance group also includes *going to be* and *it could be,* used when students hypothesise about causes or outcomes of procedures based on their current knowledge:

(8) S25: well if it's familial it's *going to be* mostly genetic if (PBL11.1)

(9) S1: dysmorphology is a spectrum so *it could be*, some in some people something that's normal, *it could be* a syndrome (PBL 12.1)

We can also see that these epistemic stance bundles are verb-phrase based falling under the oral category discussed in the structural classification section. They reflect the involved and highly interactive nature of PBL1s and may be used to express differing degree of certainty and a concern for politeness. Biber and Barbieri (2007, p. 276) found in study groups around "half of the stance bundles have an epistemic function (used) to hedge claims rather than asserting certainty". Conrad and Biber (2004, p. 67) note that conversation is predominated by personal stance expressions and have "a concern for politeness and not imposing on others". This orientation to the other participants also supports the suggestion by Kärkkäinen (2006) and Du Bois (2007) that stance is in fact intersubjective and not (solely) a reflection of the internal subjective position of the speaker.

The attitudinal stance category accounts for 21% of the bundles in PBL1s, with eight types and ninety-three occurrences. Within PBL1s students are required to identify individual learning outcomes to be allocated at the end of the session. When we look in detail at the sub-categories we see that the most common attitudinal stance bundles are the SBA2 group, i.e. those indicating an obligation or directive, used in relation to the identification of these specific learning objectives and frequently including *you* or *we*.

The more common items are immediately recognisable as obligation/directive bundles, often indirectly indicating requirements or generally for the group as a whole, for example:

(10) S16: well *you need to* think about the worst possibility which is the baby has a disease like (PBL9.1)

Similarly, the example below show *i think we* as primarily directing students in what they need to do:

(11) S12: i *think we* should, general like symptoms, tests diagnosis (PBL 9.1)

6.4.12.2 Discourse Organiser Bundles in PBL1s

Discourse organiser bundles account for 10%. The PBL1 cycle is as we have seen very interactive and characterised by short turns and as such there is less of a need for the speaker to orientate the audience through a long stretch of information. There were only four types in the discourse organiser category and forty-six occurrences, the majority of which served the elaboration (DOB) function and were primarily in the form of questions.

The bundles function in context to extend the exchanges with requests for clarification or elaboration, as the following example illustrates:

> (12) S7: *how do you* find that out? (PBL11.1)

A small number served to summarise information (DOC). In the example here we can also see a secondary epistemic stance in the tentativeness that is conveyed:

> (13) S1: *so i think*, we've had developmental delay what exactly is developmental delay, have you got a (2) (PBL7.1)

6.4.12.3 Referential Expression Bundles in PBL1s

Referential expressions also account for 10% of the bundles in PBL1s. Similar to discourse organisers, they are represented by a very small number of types (three) and occurrences (47), although two of these have a relatively high occurrence (22 examples of *a lot of* and 15 of *there is a*). These referential expressions serve two functions: one is quantifying in some way (*a lot of* and *be part of*), while the other introduces a specific entity or attribute (*there is a*). In the majority of the examples the participants are referring to the specific cases and conditions of the particular scenario, describing or identifying an attribute:

> (14) S19: it says here that she thinks, em with high IRT, *there is* a higher carrier frequency. (PBL10.1)

Students are quite vague in their quantification at this stage:

(15) S32: you have like *a lot of* causes (PBL12.1)

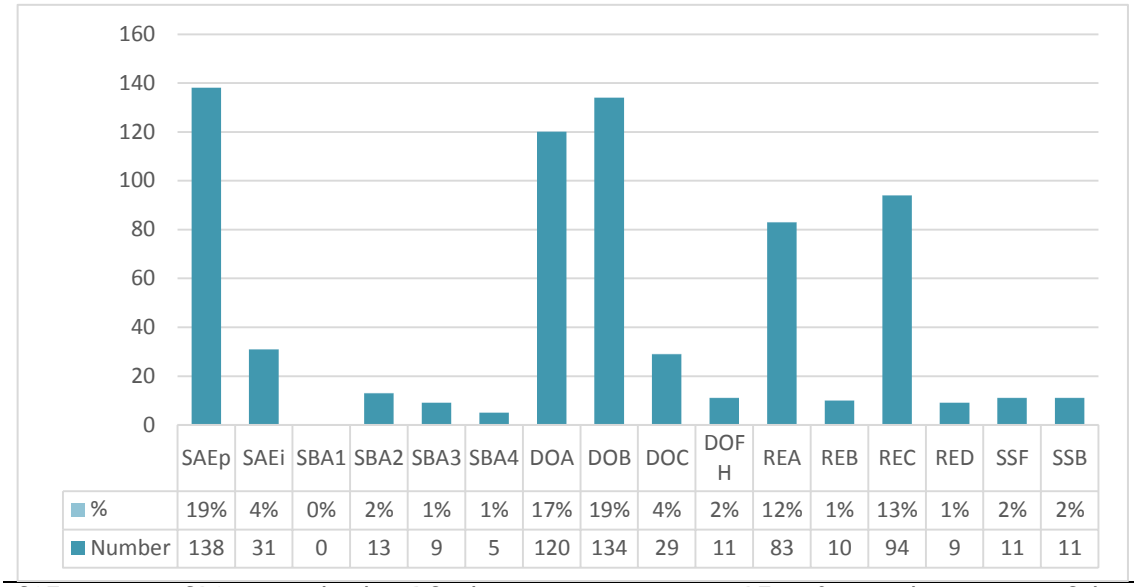
6.4.12.4 Concluding Comment on Three-Word Bundles in PBL1s

The discourse functions of lexical bundles found in PBLs are predominantly from the *epistemic stance* group with a notable number of *attitudinal stance* items. The functions indicate uncertainty or politeness and actively draw in other participants, particularly through the use of directives and questions, reflecting the communicative purposes of the PBL1 stage and the nature of the interaction. The majority are verb-phrase based and can be linked to the oral structural classification. The next section goes on to consider the use of three-word bundles in PBL2s.

6.4.13 Discourse Functions of Three-Word Bundles in PBL2s

PBL2s show a different picture, with a much more even distribution of discourse functions. This is, in fact, more similar to the classroom teaching category patterns found in Biber and Barbieri (2007, p. 273). In PBL2s, the majority of time is given over to individual students presenting the findings of their individual research. There is some follow-up discussion but this is relatively limited. What is different here from the Biber and Barbieri findings is the relative distribution. They found in classroom teaching a lower use of epistemic stance bundles (less than 25%) and a much higher percentage of obligation (around 35%) and intention (30%) bundles. This is the opposite to PBL2s. This high proportion may be due to the inclusion of the function “expressing prediction of future events that do not entail the volition of the speaker” (Biber & Barbieri 2007, p. 275), which I have included in epistemic stance. Although overall the number of stance bundles in PBL2s is much lower than the 80% in PBL1s, it is clear from the PBL2 bundle analysis that epistemic stance bundles are again the single most common discourse function at 24%. The detailed breakdown of the sub-categories for three-word functions in PBL2s is given in Figure 6.6.

Figure 6.6: Sub-Categories of Three-Word Discourse Functions in PBL2s



	SAEp	SAEi	SBA1	SBA2	SBA3	SBA4	DOA	DOB	DOC	DOFH	REA	REB	REC	RED	SSF	SSB
%	19%	4%	0%	2%	1%	1%	17%	19%	4%	2%	12%	1%	13%	1%	2%	2%
Number	138	31	0	13	9	5	120	134	29	11	83	10	94	9	11	11

SAE: epistemic stance SAEp: personal SAEi: impersonal	SBA: attitudinal stance SBA1: desire SBA2: obligation/directive SBA3: intention SBA4: ability	DO: discourse organisers DOA: identification/focus DOB: elaboration/clarification DOC: concluding/summarising DOFH: floor holding	RE: referential expressions REA: identification/ attribute REB: imprecision REC: quantification RED: time/deixis	Other: SSF: special speech functions SSB: subject specific

Note. Excel provides slightly different percentages when sub-categories are included in calculations (cf. Table 6.4; 1% lower here for SAE and RE and 1% higher for DO and SSF).

The PBL2 corpus is bigger and here we see more occurrences (708 compared with 448 in PBL1s) and many more types (49 to 27 in PBL1). While we would expect more due to the size, it is also interesting to note that there appears more variety in the types of bundles, as will be shown in Table 6.8. A similar pattern with classroom teaching (i.e. with longer turns and focussing on content information) having double the number of bundles to conversation was found by Biber at al. (2004, p. 282).

Table 6.8: Types and Frequency of Three-Word Bundles, PBL2s

Discourse Functions			
	3-word types	Frequency	Discourse function sub-categories
1.	this is the	19	DOA(6),DOB (1) DOC (3) REA (9)
2.	so i think	10	DOB (9), DOC (1)
3.	it's kind of	10	DOFH (1) REB (9)
4.	it would be	13	SAEi (11) REA (2)
5.	you can see	20	DOA (15), DOB (1) SBA4 (4)
6.	you don't have	16	SBA2 (3) REA (13)
7.	it should be	11	SBA2 (10), SBA3 (1)
8.	we can see	11	SBA4 (4) DOA (9), DOC (1)
9.	the kind of	10	DOFH (7) REA (2),REB (1)
10.	going to be	12	SAEi (4) SBA3(8)
11.	these are the	12	DOA (4) REA (8)
12.	you know the	12	SAEp (9) DOFH (3)
13.	in the first	11	DOA (7) REC (4)
14.	this is a	12	REA (7) DOA (3), DOC (2)
15.	in two thousand	10	REC (1) RED (9)
16.	first of all	16	DOA (16)
17.	so this is	14	DOA (4), DOC (10)
18.	the other one	10	DOA
19.	the first one	11	DOA (10) REC (1)
20.	in this case	16	DOA (10), DOB (3) DOC (3)
21.	and then you	12	DOB
22.	because of the	10	DOB
23.	but i think	12	DOB
24.	do you think	13	DOB
25.	if you have	12	DOB
26.	so if you	13	DOB
27.	to do with	10	DOB
28.	so that's why	14	DOB (5), DOC (9)
29.	it's the same	26	DOB
30.	that would be	12	REA
31.	there is a	20	REA
32.	which is a	10	REA

33.	one of them	11	DOB (7) REC (4)
34.	the level of	10	REC
35.	the most common	10	DOA
36.	there are two	10	REC
37.	there is no	14	REC
38.	a lot of	33	REC
39.	both of them	10	REC
40.	one of the	23	DOA (16) REC (7)
41.	i don't know	47	SAEp
42.	i don't think	21	SAEp
43.	i think it's	28	SAEp
44.	i think that	12	SAEp
45.	i think that's	10	SAEp
46.	i think the	11	SAEp
47.	it will be	16	SAEi
48.	mutations in the	11	SSB
49.	no no no	11	SSF
	49 types	**708**	

6.4.13.1 Three-Word Stance Bundles in PBL2s

The discussion of stance bundles in PBL2s is again separated into the epistemic and attitudinal categories.

There are ten different types of epistemic stance bundles and 169 occurrences in PBL2s. As with PBL1s nearly all include either *i think* or *i don't know,* with *going to be* and *it will be* being two of the exceptions. Stance bundles in PBL2s are primarily concerned with indicating a lack of knowledge, e.g., *i don't know*, or lack of certainty*, i don't think* (68 in total) or in contrast may in fact indicate a stronger degree of commitment with *it will be* and, on occasion, also with *i think*. In Example (16) below, *so i think* (which overlaps with *i think that)* demonstrates the dual function, partly showing epistemic stance but also acting as a summarising discourse organiser (DOC):

> (16) S5: so *i think that* the girl in, F and G well this is the same girl, has ah, hyperactivity. (PBL7.2)

We can also see here *going to be* to express likelihood based on their knowledge (this could be replaced by *it won't be*):

(17) S23: so because er i think i my-my suspi-suspicion is that IRT when it's produced it's not being degraded it's not being *going to be* degraded by the pancreas or something, so that's why it's high. (PBL10.2)

In the following example, *we know that* indicates the speaker's certainty; they also appear to have a secondary discourse organising function of highlighting specific information as important; here the speaker is introducing a topic:

(18) S12: so first of all how do *we know that* there's, there's a genetic component in the first place? em, there's a couple of factors firstly that *we know that* neural tube defects (PBL9.2)

While PBL1s saw a quite high proportion of attitudinal stance bundles (21%), PBL2s have a fraction of this figure at only 4%, with five types and only 30 occurrences. The majority relate to the SBA2 obligation/directive group and frequently in response to follow-up questions:

(19) S40: eh no no, it it has s-uh you know it has the surgery has some stages, i mean *it should be* done when they are small, yeah. (PBL12.2)

We also see a small number of speakers talking about their intention (SBA3):

(20) S16: right em so today topic is *going to be* about Spin-Spina Bifida (PBL9.2)

In fact, a number of other related bundles made use of *going to* in this way, but are either part of four-word bundles (e.g., *going to talk*), or fell just below the frequency threshold of ten. Only in PBL2s did we see a very small number of bundles relating to ability (SBA4), as in the example of *you can see* below (more usually this bundle was used as a discourse organiser):

(21) S26: so if you've got a mutation that, er like the protein's truncated or it's like not there are all when you do the immunohistochemistry *you can see* it's not there. so that's the kind of two tests that you'd start off with (PBL 11.2)

In summary, the relative proportions of epistemic and attitudinal stance bundles in PBL2s are not the same as in PBL1s, with far fewer attitudinal bundles in PBL2s.

6.4.13.2 Discourse Organiser Bundles in PBL2s

As discussed, this study takes a wider view of discourse organiser bundles than the original Biber et al. categories of 2004. Here I have included also bundles that direct the listeners' attention to visuals and figures, that help organise longer turns explaining logical links, for example reasons and causes, and those which enable the speaker to maintain their flow. All of these serve the larger function of enabling the speakers to engage with and show they are attending to other participants, aiming to elicit a level of involvement.

While the discourse organiser bundles in PBL1 were mainly requests for elaboration in the form of questions, in the second stage there are more types and much more variety in the bundles used. In PBL2s, we see a much higher percentage of discourse organiser bundles, 41% compared to only 10% in PBL1s, and also much higher numbers of identification/focus bundles (DOA) and clarification/elaboration bundles (DOB). This is in part due to the need to maintain and direct attention during a longer talk (compared to the highly interactive short turns in PBL1s), but also in order to link ideas, and to clarify and elaborate terms.

With identification/focus bundles (DOA; 13 types, 120 occurrences), students may be introducing a topic or part of their talk:

> (22) S39: right i'll start by giving you a brief overview of my talk today. ehm, *first of all* i'll talk about OFC, what the causes of it are (PBL12.2)

In other cases, they may be highlighting specific pieces of information. In the example below (23) the speaker is referring to the specific scenario that is being discussed:

> (23) S1: but it- well in this one and *in this case* it seemed that she just had the balanced trans balanced inversion (PBL7.2)

The discussion of the scenarios and explanation of their individual objectives requires elaboration and clarification of concepts. Seven of the fourteen elaboration/clarification bundles (DOB) include a personal pronoun, as the speakers seek to either ask for clarification or elaboration or draw the audience in and personalise the discussion.

The speaker may be providing an additional explanation of reasons/causes:

> (24) S4: i don't know if i've picked this up right it was mandatory to ehm the screening for PKUs so *because of the* benefit to the child who was ehm [? at risk]. (PBL10.2)

In the follow-up question and answer section of each talk, participants have the opportunity to ask for clarification:

> (25) S1: so *do you think* that uhm perhaps maybe a more thorough clinical examination would've produced the the need for testing (PBL12.2)

Although the students typically rounded up their presentations quickly (see Chapter 3, 3.7.6.3), there was a relatively small number of bundles which serve to summarise and conclude, either at the end of a stage in their talk or at the end of the presentation. These summarising bundles (DOC) are relatively limited within only seven types and 29 occurrences.

The example below show one such bundle, *so that's why,* being used to draw a conclusion from the preceding explanation:

> (26) S20: em, one of the main issues is that there's a lot of carriers of Cystic Fibrosis way more carriers than there's of frequency of like, em, actual occurrences of Cystic Fibrosis. *so that's why* you'd want to test you know try and limit the number, em. (PBL10.2)

Only three types were identified in the Discourse Organiser Floor-Holding category (*it's kind of*, *the kind of* and *you know the*). This is a reflection of the longer turns and online real-time production:

> (27) S1: as far as i could see it was, very similar at *the kind of*, cause you've got the whole palate looks very similar in total (PBL12.2)

6.4.13.3 Referential Expression Bundles in PBL2s

In this study, referential expressions are limited to those that frame or introduce a specific attribute of the following noun. They are used primarily to describe the content in terms of features and to quantify. Interestingly, there are examples from imprecision (REB) and time/deixis (RED) categories. The referential expressions group also includes an 'Other' section with *subject-specific bundles*. Again in stark contrast to PBL1s, we have many more types (20) and occurrences (196). The breakdown of the different sub-categories now follows.

Referential expressions that specify attributes (REA), accounted for nine types and 83 occurrences. The following example shows a bundle which frames a feature of the following noun/entity:

> (28) S20: it's not worthwhile because of the proportion of people *that would be* identified as being a carrier (PBL10.2)

Two types of bundles used to indicate imprecision/vagueness (REB) were identified, each occurring ten times in total (*it's kind of* and *the kind of*).

In the following example, the speaker uses *it's kind of* as a marker of imprecision or vagueness:

> (29) S25: again you know *it's kind of* similar to SSPC em, ehm i think it's how far it's running [? the gel] (PBL7.2)

This may be used either to hold the floor or to make an illustrative example, and so having also a secondary discourse organising role.

Much of the talk presents information on patients, features and characteristics of conditions. Two examples of quantifying expressions (REC), of which there were ten types and 94 occurrences, used in these descriptions are provided below. The use of *the level of* may also be an indication of a disciplinary specific bundle within this category:

(30) S12: and, *the level of* AFP increased, because other abnormalities like a cancer (PBL 9.2)

While expressions referring students to figures and visuals and the context have been included in discourse organiser expressions, the original Biber et al. (2004) category related to place/time deixis (RED). There is only one use referring to time (occurring nine times), specifically for (part of) the date of publication of relevant articles (*in two thousand*).

6.4.13.4 Other Bundles in PBL2s

Included in the *Other* group are *subject-specific bundles* in the discipline (SSB) and *special speech functions* (SSF). There was in fact only one of each, *mutations in the* and *no no no*.

The final point of note is the example I have included in special conversational/speech functions (SSF). This could, perhaps, be included in a discourse organiser category as it serves to disagree and emphasise a point but it seems a particular feature of real time processing. In disagreement or to emphasise a point, the speaker uses *no no no:*

(31) S7: yeah, so is it the HNPCC and it's not polyposis then [?]
S27: yeah
S5: *no no no* [? non] polyposis means that there are a lot (PBL11.2)

6.4.14 Summary of Discourse Functions of Three-Word Bundles

Although very little work has been carried out on the discourse functions of three-word bundles in either spoken or written texts, and because of the

different composition, size of corpora and categorisation of bundles, only tentative comparisons with other studies have been made. A number of patterns and differences between PBL1s and PBL2s, however, can be identified from the analysis presented here.

Lexical bundles are used in both stages of the cycle but with different distributional patterns, with PBL1s being dominated by epistemic stance bundles and PBL2s showing a more even distribution between the three major groupings. The bundle types in PBL1s are limited in variety in comparison to PBL2s. Although in PBL1s the epistemic stance category has the larger number of types, these are mostly variants of *i think* or accounted for by the high number of *i don't know*.

The discourse functions identified reflect the specific communicative purpose of each stage. In PBL1s the collaborative, involved and interactive nature of the cycle, requiring hypothesising and sharing of knowledge or lack of, results in the high use of epistemic stance bundles. We also see collaborative team work through the use of discourse organiser bundles for questioning and attitudinal stance markers to highlight group and individual requirements for the learning objectives. PBL2s involve students speaking at length about genetic conditions and procedures, referring to concepts and issues; they also require speakers to organise and elaborate their points in order to guide the listeners through the information. This results in higher use of referential expressions and more varied discourse organisers. Although using an older categorisation and looking specifically at the cohesive function of four-word bundles in lectures, the general uses of the discourse organisers here are similar to those found by Nesi and Basturkmen (2006), which is to link pieces of information within a lecture and to sequence and label stages in a talk.

While we can see an orientation to the audience and to an extent the subject matter (in that PBL2s focus more on concepts and propositional knowledge), there appears to be limited evidence of disciplinary specific use of bundles at least for those meeting the range requirements. Where it is evident is in PBL2s, with the very limited number of subject specific bundles and, perhaps, the quantification bundle *the level of*. Further work and then comparisons with

subject specific corpora would be required to really substantiate this point. The following section goes on to consider the use of four-word bundles.

6.4.15 Discourse Functions of Four-Word Lexical Bundles: Overall Comparison for PBL1s

The discourse functions of four-word bundles follow the same categorisation and exclusion criteria for three-word bundles but with adjusted frequency cut-offs.

Notwithstanding the change in the frequency thresholds, a brief overall comparison was made between three- and four-word bundles. The first point to note is that in comparison to three-word bundles, while the stance group is equally prominent in both three- and four-word bundles, there were more four-word discourse organiser bundles. There were no bundles in the referential category in four-word types. As we shall see, this is most likely due to the nature of the referential bundles and the variation in the 'fourth slot'; for example *a lot of (X)*, X being the fourth slot.

The summary of the overall comparison of PBL1 three and four-word bundles can be seen in Table 6.9 below:

Table 6.9: Distribution of Three- and Four-Word Bundles in PBL1s

	PBL1 Three-Word Bundles	**PBL1 Four-Word Bundles**
Stance	80%	82%
DO	10%	18%
Ref	10%	0%

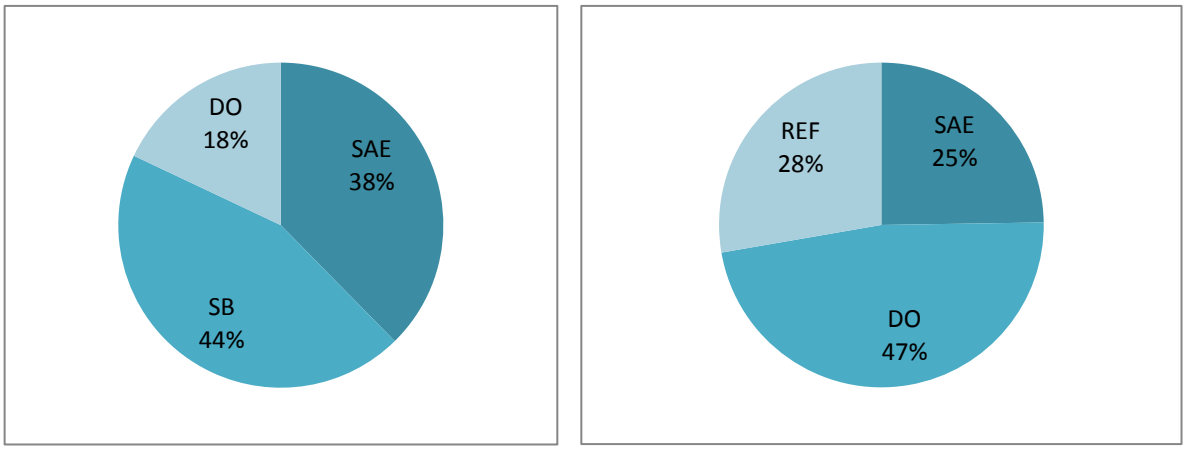
6.4.15.1 Overall Patterns of Use: Four-Word Bundles in PBL1s and PBL2s

This section considers the discourse functions of four-word lexical bundles in PBL1s and in PBL2s.

After all refinements were made, we find 28 types and 178 occurrences of four-word bundles in PBL1s but only 17 types and 101 occurrences in PBL2s, the larger corpus. The discourse functions by overall category are now presented in Figure 6.7.

Figure 6.7: Discourse Functions of Four-Word Lexical Bundles in PBL1s and PBL2s

PBL1s **PBL2s**



SAE: Stance A: Epistemic stance SBA: Stance B: Attitudinal stance
DO: Discourse Organisers RE: Referential Expressions

We can immediately see that in PBL1s stance categories continue to dominate, accounting for 82% overall. However, in contrast to three-word bundles in PBL1s, for four-word bundles within the stance group, it is attitudinal stance bundles that are the most common at 44%, followed by epistemic stance bundles at 38%. Discourse organiser bundles account for only 18% of the occurrences and, of particular note, we can see that in PBL1s there are no examples of four-word referential expressions. In comparison, the distribution of four-word bundles in PBL2s is relatively more evenly spread. Discourse organiser bundles predominate at a slightly higher 47%, with referential bundles accounting for 28% and epistemic stance much less frequent at 25%.

The overall patterns again reflect the overall aims of the stage in the PBL cycle but are more pronounced than the distribution of three-word bundles. In PBL1s we see the importance of attitudinal stance, as we shall see later with the majority being primarily the directives (part of the group decision making of the objectives), with again a large number of epistemic stance bundles indicating degrees of certainty. In contrast to PBL1s but reflecting the purpose of the second stage, in PBL2s while we still have 25% of the bundles categorised as epistemic stance, we see many more discourse organiser bundles, directing the audience through the message/longer turn and also allowing for questioning at the end. In PBL2s we also find four-word referential bundles, accounting for 28%.

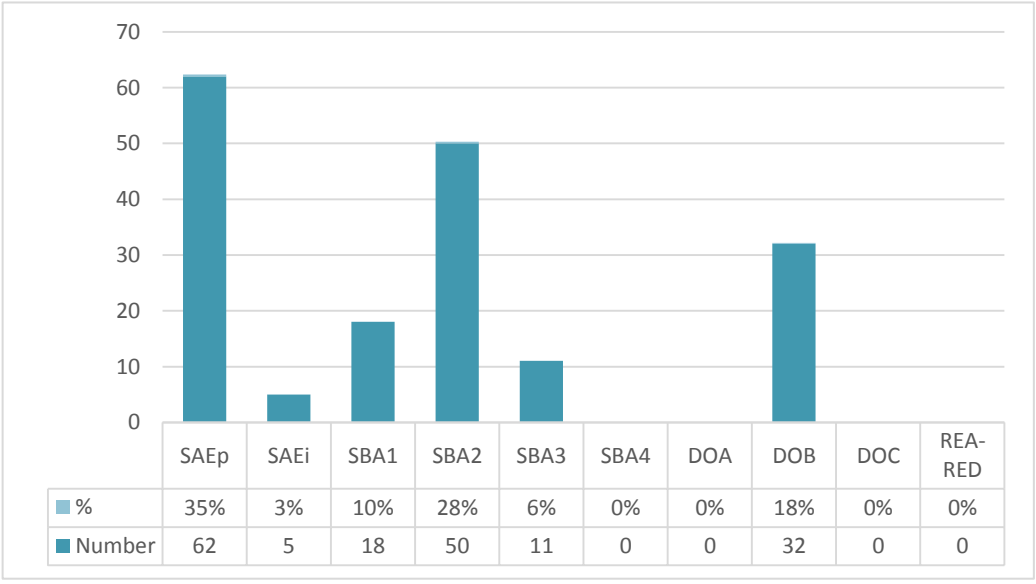
Although the frequency cut-offs are different, this general pattern of a more even distribution in PBL2s was also found in the three-word bundles in PBL2s, as were the patterns found by Biber et al. (2004, p. 396) in classroom teaching, noted previously.

Following the same format as the analysis for three-word lexical bundles, the next sections will now consider the main groupings in more detail, discussing prominent findings within the sub-categories.

6.4.15.2 Discourse Functions of Four-Word Bundles in PBL1s

The breakdown of the sub-categories of four-word bundles in PBL1s is given below in Figure 6.8. We can see that while personal epistemic stance bundles are the most common of each of the sub-types, this is quite closely followed by obligation/directive (SBA2) attitudinal stance bundles and overall attitudinal stance bundles are in the majority. The discourse organiser bundles identified are used for elaboration or clarification, the DOB sub-category.

Figure 6.8: Discourse Functions by Sub-Category: Four-Word Bundles in PBL1s



SAE: epistemic stance
SAEp: personal
SAEi: impersonal

SBA: attitudinal stance
SBA1: desire
SBA2: obligation/directive
SB3A: intention
SBA4: ability

DO: discourse organisers
DOA: identification/focus
DOB: elaboration/clarification
DOC: concluding/summarising

REA-RED: referential expressions

Table 6.10 to follow provides a list of types and frequency of four-word bundles in PBL1s. Where one item has been identified as having a different discourse function in different contexts, this has been noted and items categorised accordingly. Items which appear to have a dual function were noted in the analysis and comments included in parts of the discussion. While initially it appears that there is a wide variety of four-word types in PBL1 (28), closer inspection also shows that eleven of the bundles include *think*, eight of which have *i* and *think*, three include *know* and four each include *going to* and *need to*. Three of the six discourse organiser bundles include *what* used in a question. In addition, 22 of the 28 bundles include a pronoun (12 with *i*, eight with *we* and six with *you*), suggesting, not surprisingly, a strong personal, involved and interpersonal orientation.

Table 6.10: Types and Frequency of Four-Word Bundles in PBL1s

	Four-Word Types	Frequency	Discourse Function Sub-Categories
1.	but i don't know	8	SAEp
2.	but i think that	5	SAEp (4) SBA2 (1)
3.	do we need to	6	SBA2
4.	do you think we	6	SAEp
5.	do you want to	12	SBA1 (6) SBA2 (6)
6.	going to have to	5	SBA2
7.	i don't know if	8	SAEp (6) DOB (2)
8.	i don't know what	6	SAEp
9.	i don't think it's	5	SAEp
10.	i don't think so	6	SAEp
11.	i think it's a	5	SAEp
12.	i think we have	6	SAEp
13.	i think we should	9	SBA2
14.	if you want to	5	SBA1
15.	is going to be	5	SAEi (2) SBA3 (3)
16.	it's going to be	5	SAEi (2) SBA3 (3)
17.	not going to be	6	SAEi (1) SBA3 (5)
18.	so it would be	5	DOB
19.	think we need to	8	SBA2
20.	we need to do	7	SBA2
21.	we need to look	8	SBA2
22.	what do you mean	6	DOB
23.	what do you think	7	DOB
24.	what does it mean	5	DOB
25.	yeah i think that	5	SAEp
26.	yeah i think we	5	SAEp
27.	yeah that's what i	7	DOB
28.	you want to do	7	SBA1
	28 types	178	

6.4.15.3 Four-Word Stance Bundles in PBL1s

Stance bundles account for 82% of the four-word bundles: the majority of these are attitudinal stance at 44%, with 38% epistemic stance bundles.

All the epistemic stance bundles include either *i* or *you think*, *i (don't) know* (mental process verbs), or *going to* (logical certainty). As with PBL1s, they are mainly verb based and characterised by the inclusion of first and second person pronouns (in 11 of the 14 types).

As with PBL1s, the bundles indicate varying degrees of certainty or uncertainty, with relatively more of the latter (e.g., *but i don't know*). They also primarily include *i*, indicating the speaker's subjective position, but also potentially indicating attention or consideration to the receiver.

In the next example (32), we see *i don't think so* used for disagreement, further emphasised with *no*:

(32) S28: maybe eh, it's related to a the diagnosis method
S25: [? no]
S27: no *i don't think* so (PBL11.1)

We also find examples of *it's going to be* to comment on predicted outcomes based on logical conclusions from background knowledge:

(33) S25: well if it's familial *it's going to be* mostly genetic if (PBL11.1)

Within the attitudinal stance category, the most common by type and frequency of occurrences is the SBA2 category, obligation/directives. At 28% they account for eight of the 13 types in the attitudinal stance grouping and 50 of the 79 occurrences overall. Although there are 13 different types, four include *need to*.

The attitudinal stance bundles relate to one of the main purposes of the PBL1s: as a group to identify learning objectives that members will follow up in self-directed independent study and report back on in the next tutorial. We can see a strong sense of the collaborative nature of the group here with the inclusion of *we* in five of the thirteen bundles: *we need to look, do we need to, think we need to, i think we should*. There is a strong presence of modals and semi modals here, common in language associated with obligation and necessity. The

higher number may be connected to the structural type (*pronoun + semi-modal + verb*).

The collective responsibility to identify learning outcomes is evident in the following example of SBA2 bundles. *I think we should* in the example below (34) can be seen here to be a directive, indicating a necessity but with a secondary function of expressing a degree of certainty or opinion:

(34) S8: *i think we should* include conditions that cause the screening, like the (PBL9.1)

The bundles may also be embedded in a question, a less direct form and marker of politeness:

(35) S4: do you *think we need* to look at the reasons (PBL10.1)

6.4.15.4 Four-Word Discourse Organiser Bundles in PBL1s

The other category found in PBL1s is discourse organiser bundles, at 18%. All of these are within the clarification/elaboration group (DOB). There are 32 occurrences but only six types, half of which ask a 'Wh' question requiring further clarification or elaboration (e.g., *what do you mean; what do you think*). We can also see that this focus on clarifying ideas is reflected in the preponderance of mental process type verbs (*mean/think/know)*.

In the following example, we can see the bundle indicating agreement and elaborating the previous turn with additional support:

(36) S1: so that means the problem isn't a deletion so therefore what's causing, the dysmorphism in K
S7: *yeah that's what i* was telling you these scenarios (PBL7.1)

6.4.15.5 Four-Word Referential Expressions in PBL1s

Although no four-word referential expressions were identified (in comparison to 10% of three-word bundles in PBL1s), a possible explanation is offered. This lack

might be accounted for by the length of bundles included in this category in PBL1s. In PBL1s, the referential expressions included *a lot of, there is a, be part of,* all of which were then followed by a specific attribute or topic/content noun in the subsequent slot. As these are subject to variation depending on the scenario*, (e.g., a lot of + cancers/ syndromes/detail*) it is not surprising that they do not all occur as frequent four-word bundles, particularly in a small corpus.

6.4.15.6 Concluding Comment on Four-Word Bundles in PBL1s

The four-word bundles in PBL1s are broadly similar in the distribution of categories to three-word bundles in that stance bundles account for a much higher proportion than discourse organisers, with the obvious exception that there are no referential bundles. The epistemic stance bundles are again used to indicate degrees of certainty, or lack of, and the discourse organisers to question and seek clarification, a signal of students working to develop their knowledge. However, while the general overall categories indicate a similar pattern, overall the attitudinal stance bundles predominate, with bundles most notably being used to signal learning objective requirements and indicating with a strong sense of the collaborative nature of PBLs in inclusion of *we* in a significant number of the bundles.

Again we see a clear orientation to the self and group in the use of first and second person pronouns (in 22 of the 28 four-word bundles). While many bundles appear to indicate at first a subjective orientation to the self, in fact they could also be said to be intersubjective, acknowledging the group, often being used to play down and hedge contributions and so show a regard for 'the other'. At the very heart of the pedagogy behind PBLs is the aim to develop skills and knowledge through collaboration.

The bundles also appear to reflect, to an extent, the spoken interactive mode: we see items used in responses and questions, indicating the purpose of PBL1s to exchange and develop information and knowledge. The short exchanges may also account for the lack of summarising (DOC) bundles, as longer turns are rare. However, as with three-word bundles, while the communicative purpose and

mode are reflected in the bundles, the specific disciplinary academic content is not immediately evident.

6.4.15.7 Four-Word Bundles in PBL1s and PBL2s

If we consider three-word and four-word bundles in the PBL2 cycles, we see a similar distribution in three-word and four-word items (different from PBL1s). In the PBL2s, discourse organisers predominate, with stance and referential bundles making up just over half, in roughly equal proportions as shown in Table 6.11:

Table 6.11: Distribution of Three- and Four-Word Bundles in PBL2s

	PBL2s Three-Word	**PBL2s Four-Word**
Stance	28%	25%
DO	41%	47%
Ref	28%	28%
Other	3%	0%

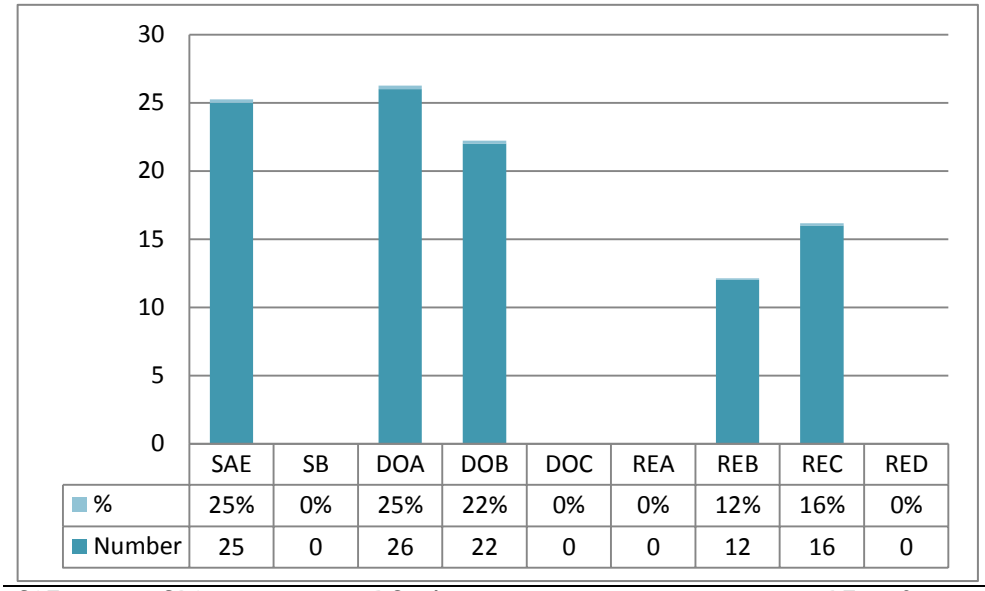
Note. Comparison for PBL2 Three-Word based on overall groups Table 6.4.

6.4.15.8 Discourse Functions of Four-Word Bundles in PBL2s

We have already observed that the overall patterns of use of four-word bundles varies between PBL1s and PBL2s, with PBL2s being characterised by a much higher proportion of discourse organiser bundles (47% to 18% in PBL1s; see Table 6.9 above) and by the presence of referential bundles at 28%. This obviously means that epistemic stance bundles are much reduced, accounting for only 25%, all of which are epistemic.

The detailed breakdown of the sub-categories of discourse functions for four-word in PBL2s is given in Figure 6.9.

Figure 6.9: Discourse Functions by Sub-Category: Four-Word Functions in PBL2s



	SAE	SB	DOA	DOB	DOC	REA	REB	REC	RED
%	25%	0%	25%	22%	0%	0%	12%	16%	0%
Number	25	0	26	22	0	0	12	16	0

SAE: epistemic stance (SAEp)
SBA: attitudinal stance
DO: discourse organisers
DOA: identification/focus
DOB: elaboration/clarification
DOC: concluding/summarising
RE: referential expressions
REA: identification/ attribute
REB: imprecision
REC: quantification
RED: time/deixis

In PBL2s, discourse organiser bundles are by far the most frequent, accounting for 47% of all the occurrences. The stance (in this stage of the cycle only epistemic stance) and the referential bundles provide 25% and 28% of the remaining items. Broadly speaking this would appear to reflect the nature of the second cycle, certainly in the use of discourse organisers and referential expressions. There is an increased importance of guiding the audience through the informational content, showing how parts of the talk relate to each other and referring to concepts and issues related to the scenario. The specific functions and examples of each grouping will be discussed in the following sections.

The types, frequency and discourse functions of the four-word bundles in PBL2s are shown in Table 6.12 to follow. With 101 occurrences and 17 types (compared with 178 and 28 in PBL1s), initially there appears to be much less variety despite PBL2s being a larger corpus. However, we can also see that while the majority of types in PBL1s were variants of particular verb forms, including many of the same verb forms (e.g., *think/know/going to/need to*), in PBL2s we can see more variety in the types, a wider range of verb forms and the presence of referential

expressions (e.g., bundles including *a lot of the, one of the most, the first one is*).

Table 6.12: Types and Frequency of Four-Word Bundles in PBL2s (Discourse Functions)

	Four-Word Type	Frequency	Discourse Functions Sub-Categories
1.	a lot of people	5	REC
2.	a lot of the	5	REC
3.	as you can see	6	DOB
4.	but i don't know	5	SAEp
5.	i don’t know if	14	SAEp
6.	i don't know what	6	SAEp
7.	if there is a	6	DOB
8.	if you have the	5	DOB
9.	if you look at	5	DOA
10.	is one of the	5	DOA
11.	one of the most	5	DOA
12.	or something like that	7	REB
13.	that kind of thing	5	REB
14.	the first one is	5	DOA
15.	there's a lot of	6	REC
16.	what do you mean	5	DOB
17.	you can see the	6	DOA
	17 types	101	

6.4.15.9 Four-Word Stance Bundles in PBL2s

The discussion of stance bundles in PBL2s covers only epistemic stance bundles, as no attitudinal items were identified above the frequency threshold. Stance bundles are not insignificant, at 25%, but are far fewer than four-word stance bundles in PBL1s (where epistemic stance alone accounted for 38%). Again, the requirements of the task and specific communicative functions are reflected but this time in the absence of certain bundles.

The epistemic stance bundles in PBL2s comprise only three types, all including *i don’t know*, the most common one being *i don’t know if*, used to indicate uncertainty, or a lack of knowledge. At other times the speaker appears to be hedging and playing down a point (at the end the speaker does in fact conclude that they think they are correct), as in Example (37):

(37) S4: em, so the right to choose, from what i read it seems to be a bit kind of different in different countries, and in America, eh *i don't know*

if i've picked this up right it was mandatory to ehm the screening for PKUs so because of the benefit to the child who was ehm [? at risk]. *but i don't know* i think that's right anyway (PBL10.2)

6.4.15.10 Four-Word Discourse Organiser Bundles in PBL2s

The discourse organiser bundles comprise the largest proportion of four-word bundles in the PBL2 corpus at 47%, with nine types and forty-eight occurrences. While in PBL1s the four-word bundles were in the DOB category mainly when asking for clarification or elaboration, the PBL2 cycle also includes identification and focus bundles (DOA). The identification/focus (DOA) bundles (five types and twenty-six occurrences) serve to introduce and single out new topics, direct the listeners' attention to specific information, for example in a visual or handout, and provide emphasis. These all attend to the listener, the speaker working to maintain their attention and guide them through the talk.

In the first example below, the speaker highlights a specific point (38) and in the second example (39) directs the group to the supporting visual (PBL10.2):

(38) S7: oral facial clefting is supposed to be *one of the most* common birth defects, and they have been ranged between one and ten (PBL12.2)

(39) S23: so *if you look at* the first figure, er *the first one is* definitely to the initial IRT test and there's, looking at the, level of IRT. (PBL10.2)

The elaboration/clarification bundles account for just under half the discourse organisers with four types and 22 occurrences. *As you can see* has a secondary function of focus, although in the specific context studied here it was seen as part of an elaboration.

(40) S3: (2) so expressivity of is variable in V-VWS, *as you can see*, eh no- all the clinical features such as lip pits cleft lips are eh [? a cleft] palate is seen in each case (PBL12.2)

Where there was a preponderance of discourse organisers used when asking for clarification and elaboration (DOB) in PBL1s, in PBL2s this is less significant. The longer reports are frequently followed by only a very short question and answer stage. Although in theory the students should continue the discussion of the information presented in relation to the scenario, in fact in these particular PBLs this is not a prominent feature.

There were no summarising, four-word bundles (DOC). This is interesting in that this might be expected at the end of an extended explanation. However, if we consider the three-word summarising bundles in PBL2s, for example *so this is, so that's why*, the following word or phrase may be open to more variation (similar to the three-word referential expressions).

Interestingly, given the extended nature of the turns, no discourse organiser floor holding (DOFH) bundles were found. These appear to occur only as shorter three-word items.

6.4.15.11 Four-Word Referential Bundles in PBL2s

As might be expected, there is a much higher number of referential expressions in the PBL2s given their focus on informational content. We find two bundles for imprecision (REB) and referring expressions for quantity (REC). However, surprisingly there are none in the REA category (referring to specific attributes or introducing concepts). Although not substantial, two referential expressions used to express imprecision (REB) are found, both quite syntactically and semantically 'complete' (e.g., *that kind of thing*). There are three types of referential expressions referring to quantity (REC), all including *a lot of*. Given that the presentation report is factual and based on background research, one might have expected more precision. However, in discussing the condition or issues related to the scenario, the speakers generalise, as in the next example (41).

(41) S27: six gene there is low penetrance and higher onset of age, but there not *a lot of people* who carry this mutation (PBL11.2)

As with PBL1s, there were no examples of four-word bundles for the first referring expression sub-group (specifying attributes or introducing concepts, the REA sub-group). No time/place deixis bundles were noted, nor were there any 'other' expressions.

6.4.16 Concluding Comments on Four-Word Bundles

In a similar way to the distribution patterns of three-word bundles in PBL1s and PBL2s, the types of four-word bundles vary between the first and second stage of the cycle. The most noticeable difference is the lack of four-word referential expressions in PBL1s, which appears to be related to the structural make-up of the bundles. Once again the interactive, collaborative work to build knowledge and identify learning objectives required in PBL1s is reflected in the high use of epistemic and attitudinal stance bundles. It is the four-word attitudinal stance bundles that dominate here though, perhaps because they are more phrasal in nature, compared to the epistemic bundles that typically frame propositions.

We also see more variety in types in the PBL2s, mirroring findings from the much larger studies of classroom teaching (Biber et al., 2004). This may be because students with different learning objectives are focussing on different things (relating to Biber et al.'s (2004) comment that classroom teaching covers a wider range of communicative tasks). In both stages, the functions reflect the communicative purpose of the stages. PBL1s are highly interactive and involved, PBL2s are involved though with less interaction (we find more discourse organisers in PBL2s). Although covering a wide range of linguistic features, Barbieri (2015) identifies involvement as a key feature of postgraduate and small group university settings.

In contrast to the large-scale studies of registers, even with reduced frequency thresholds, remarkably few occur on any four-word lists from studies of larger spoken academic corpora (cf. Biber et al., 2004; Nesi & Basturkmen, 2006).

6.5 Chapter Summary

The investigation into the structural and functional types of three- and four-word bundles in PBL1s and PBL2s has shown repeatedly that the types and distribution of bundles for both aspects reflect the communicative purpose, the level of interaction and the level of involvement expected in each stage. The specific communicative function is also reflected in the absence of certain bundles (e.g., the lack of attitudinal directives in PBL2s).

The categories developed by Biber and colleagues have been used with slight modifications to provide a framework for the categorisation of structural and discourse functions of both three- and four-word lexical bundles in the whole corpus and each sub-corpus.

The structural categories in PBL1s are overwhelmingly oral compared to PBL2s, which have a higher proportion of literate bundles. This reflects the more interactive nature of PBL1s, with much more turn-taking. PBL2s have more of an informational content.

The most common discourse functions are also different in the two stages. PBL1s include a very high number of epistemic and attitudinal stance bundles, with PBL2s displaying a more even spread and including discourse organisers and referential bundles. This again reflects the highly personal, subjective stance of speakers as they contribute to the discussion and the focus on identifying learning objectives (through attitudinal stance) in PBL1s. In PBL2s, the distribution reflects the requirement to direct participants through the talks along with the reference to content. Thus we see the pedagogic aims (collaborative sharing and building of knowledge) and the demands of the communicative events clearly reflected in the bundle types and number of occurrences of each bundle.

While the bundles overall highlight the communicative event type and pedagogy, it is in the referential bundles mainly in PBL2s that we see more of an indication of the specific discipline (e.g., in referring to attributes and quantities). This is in contrast to larger studies particularly of written registers which have

indicated disciplinary differences. This may well be because this is only a small corpus and could only be substantiated by comparing bundles from other spoken registers and specific disciplines. Although the PBL cycle as a whole could be seen as a type of study group interaction in contrast to a classroom teaching event typically dominated by a teacher, the individual stages within themselves still show variation in use. While some general comparisons to larger scale studies have been made and some patterns found to be similar in vein, the students in PBLs are still 'doing it differently'.

Overall the study of lexical bundles indicates that differences in expectations of the communicative event (collaborative discussion or presentation) and the underlying PBL pedagogy result in different types and proportions of bundles used.

Chapter 7 Conclusion

7.1 Introduction

This thesis set out to explore how problem-based learning, a student-centred speech event common in medicine, is realised linguistically. This type of event was selected not only because there is limited research into occluded student-centred dialogic genres, but because there is also very little on disciplinary-specific speaking. Yet speaking to learn is implicit in many approaches to education and explicit in PBL pedagogy. This pedagogy, which aims for students to develop professional skills and knowledge through collaborative problem-solving work, involves interacting and engaging both with peers and the disciplinary subject matter.

To explore this type of event, a bespoke corpus of five cycles of PBL sessions was built. The specific research questions aimed to find out how the pedagogic aims of PBL were reflected in talk. There were three broad areas of enquiry: how students engage with and orient to both their peers and the disciplinary content; how the PBL pedagogy is reflected in patterns of interaction and linguistic markers of engagement and interaction; and finally, how the mode is reflected in language choices, as this type of speech event contains different components in terms of format. In exploring these areas, I aimed to provide a more detailed description of disciplinary-specific speaking and engagement.

In order to explore how the participants interact and engage with peers and the disciplinary-specific subject matter within this pedagogy, a multifaceted approach was taken. This involved identifying patterns of interaction by applying a CA-informed analysis and also identifying academic functions. This was complemented by employing techniques from CL to explore a number of specific linguistic features which might exemplify interaction, engagement and disciplinary discourse. Comparisons were made between the two stages of the cycle.

In the next section, I will review each aspect of enquiry and provide a brief summary of each chapter. This will indicate the contributions of the research.

7.2 Review of Research Areas and Contribution

The first area of exploration was the organisation of the different stages in the PBL cycle. In order to explore interaction patterns and how PBLs are "talked into being", a CA-informed approach was employed. This aimed to uncover the different patterns of organisation and orientations to participants and the content. The findings in this section identified a general overall staging but also the fact that phases, particularly in the PBL1s, merged into each other. This reflects the highly interactive and contingent type of talk in this speech event. In orientating to the specific pedagogic goals, I identified six different types of talk that participants engage in as they work collaboratively to co-construct knowledge. These exemplified the different ways in which talk is organised to meet the pedagogical aims of PBL and to develop the skills and knowledge of the novice medical geneticist. The types of talk showed how students take on roles and responsibilities, how knowledge is exchanged and knowledge building is carried out and how knowledge imbalances are dealt with.

That the PBL is an academic speech event was also illustrated through these different types of talk. While the more traditional *instructional talk* by tutors was identified as they guided students to learning points, *simple knowledge exchanges, problem-solving exploratory talk* and *clarification check*s also demonstrated how students work together, speaking to learn. The complex series of exchanges in the latter two types were used in the first place to build up meaning and in the second to check and clarify learning. All served to redress epistemic imbalances. The *clarification checks* in particularly showed how participants on different parts of the epistemic cline exchanged and challenged information. In PBL2s, *presentational talk* provided students with an opportunity to demonstrate their learning, again consolidated in *clarification checks*. As CA is suited to the study of interaction and exchanges, in order to explore the longer stretches of talk in the PBL2 presentations, an alternative approach was taken. This involved reading and identifying the *academic functions* which predominate in this stage of the cycle and which reflect the specific concerns of the discipline.

The study of interaction patterns confirms Basturkmen's (1999, 2002) findings of the complexity of student interactions, Tanguay's (2015) findings of a variety of possible functions in student seminars and Walsh et al.'s (e.g., 2011) identification of different types of talk contributing to pedagogic aims. In the PBLs, we can see also a distinct set of types of talk related to dealing with knowledge gaps.

The identification of a range of types of talk and the study of interaction patterns have also demonstrated the different ways in which learning might occur in PBLs. In addition to clear examples of teacher scaffolding and checking of learning in *instructional talk* and *clarification checks*, we also see examples of peer-to-peer scaffolding. While research into learning events in HE has frequently highlighted how teachers guide and support learners in the extension of knowledge, less perhaps has been said about *how students* work together to support and scaffold each other in learning and so achieve the co-construction of knowledge. The types of talk shows that through *problem-solving exploratory talk, simple information exchanges* and *clarification checks*, learners have opportunities to teach and support each other. The *presentational talk* also affords an opportunity for students to report on their own learning and, for example by exemplifying and elaboration, it could be argued that in this way they also work towards scaffolding information for peers (although the extent to which this is then taken-up is not observed).

The study of orientation and interaction indicated how the differences in the aims of the PBL stages affect how students engage with each other and with the subject. To study these aspects in more detail, keywords, personal pronouns and lexical bundles were explored. These areas are now reviewed.

To discover how vocabulary, a recognised marker of disciplinary variation, is used in PBLs, I employed keyword analysis to identify how each PBL stage is distinct from the other and in relation to a reference corpus of seminars and lectures. The keyword analysis identified three main points. The first was that, as with the study of interaction patterns, the keywords for each PBL stage indicated differences from each other. PBL1 keywords included a set of mental process verbs (drawing on SFL transitivity analysis), indicating the cognitive

internal and subjective nature of the initial discussions. PBL1s also included more high frequency vocabulary, reflecting the more conversational nature of PBL1s. Finally, vocabulary in PBL2s included a strong focus on both natural processes and technical procedures in medical genetics, reflecting both the concentration on informational content but also the aim of the second stage in this context, which is to discuss in more detail conditions and procedures for treatment. There were similarities in some areas in terms of academic technical vocabulary and topics discussed. Thus by coding the keywords into general academic vocabulary and technical vocabulary, by SFL process type and by semantic groupings, the aims of the pedagogy and the concerns of the discipline and, to an extent, the format of the event were highlighted as distinct from the reference corpus.

Very little research has been concerned with academic vocabulary use in student-centred spoken interactions. This work contributes to studies exploring the use of academic vocabulary through keywords as indicators of aboutness and SFL process types, confirming how these approaches can be used to highlight the concerns of different disciplines.

Participants orient to each other through markers of involvement and engagement. Personal pronouns are one type of marker and were explored for frequency and semantic mapping of the referent. This part of the research builds on studies into pronoun use by more expert speakers in academic contexts. The semantic mapping of *you* and *we* highlighted differences from other academic registers. Through the study of personal pronouns, student identity was observed, both as students within the specific cohort and as novice professionals. The study of pronouns also included an initial exploration of contiguous collocates, which was followed up in the study of lexical bundles.

Lexical bundles have been shown to have important roles in displaying stance, in referential expressions and in organising discourse. This was the third linguistic aspect to be studied in detail. As the corpus is quite small it was possible to consider three-word bundles along with the smaller number of four-word bundles. The study of lexical bundles again showed differences between PBL1s and PBL2s. The differences in use and type were shown to relate to the specific

format and learning aims of each stage (e.g., tentatively sharing opinions in PBL1s compared to reporting on background reading and research in PBL2s). The communicative event and the pedagogy are prominent in the study of bundles.

Building on previous descriptions of spoken academic English, this study confirms as might be expected higher levels of markers of personal engagement than are typically found in written academic discourse. However, studying the two stages of the genre shows more clearly the very interactive and oral nature of PBL1s in contrast with the presentations in PBL2s, the latter more similar to lecturer talk in that they included more of a mix of oral and literate bundles. The detailed study of one specific type of speech event, one which is characterised by different stages with quite distinct aims, adds to the descriptions of spoken academic English.

In this context, the instantiation of PBL is through a highly interactive and collaborative cycle. Pronoun use, lexical bundles and interaction patterns all show how students orientate to their peers and to the aims of PBL of sharing and building knowledge. However, while the PBL is a distinct genre, the two stages also vary considerably in language use. The different aims of each stage are exemplified to an extent in different keywords but also in the types of pronouns and lexical bundles used. In investigating specific items in detail in both stages, this study provides additional insights into the use of these linguistic markers in characterising this particular type of speech event and pedagogy, one situated in a specific discipline.

A number of corpus linguistic techniques have been used to identify these features which were then explored using a range of approaches from qualitative analysis of semantic groups to SFL transitivity and a mapping of discourse functions and uses (pronouns and lexical bundles). In doing so, this study has also shown how combining a CA-informed approach with the detailed study of a variety of specific linguistic items from different perspectives can provide a rich description of a learning event. In this way, it also makes a contribution to increasing number of studies that employ a multidimensional approach to the analysis of a corpus of academic speech.

7.2.1 Lessons on Problem-based Learning

The primary aim of the research was to provide a clear and informative description of how PBLs are enacted in a specific programme and discipline, with a view to potentially informing EAP teaching and materials design. While it was not an analysis of the effectiveness of PBL nor of the extent to which claims about its potential to develop problem-solving, reasoning and collaborative co-construction of knowledge can be verified, some observations can be made based on the data available. The students were clearly working together and opportunities to add to each other's knowledge, question and scaffold peer learning were afforded. However, perhaps because of practical time constraints, although some discussion of individual learning objectives might take place after peer reports, the stage in which the students returned to the scenario and then discussed and applied their new knowledge as a group was not seen. To maximise opportunities for further discussion and to be able to ascertain more demonstrably if students can apply learning from independent study and peers, this is a stage that facilitators should ensure is routinely built in or an additional session incorporated. As the former is most likely to be more practical on a busy intensive programme, the time allocated for presentations could be slightly reduced to ensure the discussion stage at the end is incorporated (so also following departmental guidelines).

PBL scenarios are essentially about solving human issues. Through the identification of academic functions and keywords in each stage, we can see in this cohort a strong focus on describing natural processes and procedures for identifying and testing genetically related issues. While the students were very much 'on-task' in relation to content knowledge and the specific problems at hand, the data does not show much in the way of attention to personal attitudes and feelings. Only when the students make asides to their own personal experiences do we see more of their own opinions and attitudes emerge. While counselling (of patients and related family members) is invariably identified as a learning objective, in the PBL reports this area is addressed primarily as treatments and points to be raised. Although academic staff point out that the actual counselling would not be carried out by this group, an awareness of how such information might be communicated could be of use. For example, in one

scenario students were tasked with producing a leaflet for the potential patients. Following a discussion of what technical information to include, the facilitator had to intervene and ensure the content was not overly complex and attended to the feelings of the patients and family members who would read the leaflet. How such technical communication might be communicated, for example in empathy writing for the layperson (Nesi & Gardner, 2012), is a potential area for future research.

7.3 Contributions to the Field of EAP

The different aspects of the research have illustrated how students enact learning in this specific speech event and how they realise their disciplinary identities, both as students and novice researchers. This research has also shown how the different aims of the two stages of the cycle are reflected in different uses of specific language features. This of course may seem obvious but if teachers are to develop a more nuanced understanding of spoken academic and indeed disciplinary interactions, then a greater understanding of how this might be realised is important. By comparing the two stages, we can see that relying on general descriptions of academic speech is not in itself sufficient. By looking closely at each stage in the cycle, we see also that one student-centred learning genre, a PBL cycle, can have different linguistic profiles, clearly related to the aims of each stage. This in itself is not entirely surprising as studies of different written genres and parts of genres have similarly reported on disciplinary and genre variation (e.g., Nesi & Gardner, 2012; Flowerdew & Costley, 2016). However, this has not been much researched in speaking.

The study of the PBL interactions shows how students can engage with and take on multiple roles and collaborate to co-construct knowledge. This is important not only in general educational theory which highlights the importance of deep level learning and a constructivist approach, but also for language learning. Studies in second language acquisition show how a learner's language develops and is pushed forward by actively engaging with opportunities for pushed output, negotiation of meaning and receiving feedback (e.g., Samuda & Bygate, 2008;

Swain & Lapkin, 1995). Although the PBLs are not about language learning[33], they do show how collaborative tasks can provide opportunities for such engagement. The very rich explorations and development of ideas in this data can be seen particularly in the *problem-solving exploratory talk* and *clarification checks* which provide potential for task design. The study of the interactions also shows that when students are provided with authentic and motivating input, in this case academic content, they have multiple and varied opportunities to develop and share knowledge and to negotiate meaning. In an EAP context, this can provide students with opportunities to practise how to articulate this accurately and with an appropriate level of complexity, thus providing opportunities for academic language activation and recycling.

That students engage in collaborative talk and on an informal and very personal level is also important in setting expectations of student discourse in EAP teacher induction and in assessment tasks and criteria. Assessment criteria for EAP courses tend to be based on generalised expectations of academic language and the extent to which students can deliver an effective presentation. Yet PBL also involves highly interactive discussion and the sessions are quite informal, along with engaged follow-up to presentations.

There are a number of ways in which this in-depth description of how PBL is enacted within a specific discipline can inform EAP materials design. In PBL the focus is on speaking to learn and on articulating and communicating on academic content. In order to provide examples of appropriate linguistic content which could form the source of exemplars of language for noticing and practice activities, examples of the academic functions could be exploited, e.g. extracts of cause and effect chains and descriptions of natural processes and procedures. Similarly, how key words are used in specific contexts could provide a useful source of language work in an EAP context. This could include looking at collocations and colligational patterns to facilitate a focus on meaning and use. At a macro level, the PBL cycle of work highlights the importance in this approach of collaborative goal-setting to inform independent study and of dealing with and discussing academic content in both stages.

[33] There are parallels here with task-based learning, but in PBL the focus is firmly grounded in the academic content.

Although the focus in this research has been on one discipline and approach to study, the importance of dealing with academic content and subject matter is not unique to PBLs. Given that all students on EAP courses need to be prepared for academic study and have opportunities to engage with and activate language on academic subject matter, the approach may be useful on EAP courses. At the very least, the importance of providing students with opportunities to prepare for and discuss academic content should be incorporated more consistently into EAP task and materials design.

This in my view is an important point. EAP teaching and published materials to facilitate spoken academic discourse are still largely based around developing presentation skills and, to an extent, skills for taking part in seminars or discussions. The former may draw on a topic students have studied for extended writing, particularly on intensive pre-sessionals, but in coursebooks may be based on only limited pre-prepared academic content or none at all (e.g. Chazal & Moore, 2013; Hewings & McCarthy, 2012). Follow-up discussions to presentations in my experience are often limited in time and scope, as other students may be very unfamiliar with a topic and may have little to contribute. Seminar discussions are frequently on general topics and, with the exception of McCormack and Watkins (2012), do not as a general rule involve much in the way of preparation in relation to researching and reading around an area, nor require much discussion and negotiation of meaning in relation to academic subject content. Instruction tends to focus on how to organise and deliver a presentation and on peer and self-evaluation, with language work concentrating on functional exponents, for example signposting in a talk and, for seminars, phrases for agreeing or disagreeing. There is a lack of materials that focus on, for example, academic functions common to a discipline, nor that require students to meaningfully discuss academic content knowledge in the way that a PBL cycle of work, and I would suggest many academic seminars, would. In my view, more specific materials that reflect not only the typical interaction patterns and requirements of specific disciplines and for particular modes of learning, along with tasks and materials that require engagement with academic content are required.

On a wider note, we can see that the PBLs can be highly interactive and student-centred. This type of learning may be less familiar for students coming from different educational backgrounds (Remedios et al., 2008). By employing PBL on preparatory EAP courses, this may help develop student confidence by preparing them for the student-centred and autonomous learning contexts they may well engage in, as they begin their journey of socialisation into the academy (Duff, 2010).

7.4 Limitations of the Study and Directions for Future Research

While this study has provided a clear picture of disciplinary speaking in one type of speech event, there are limitations. The most obvious point is that the corpus is of only one discipline, of one type of speech event and at one level of study. A corpus of PBLs from different disciplines and of different types of student speaking would further enhance our understanding of disciplinary variation in spoken academic English.

The study of lexical bundles showed distinct differences in each stage of the cycle. Less obvious in this corpus are disciplinary differences (which have been identified in writing), which may be because of the small size of the corpus and low frequencies of occurrence. In order to explore, for example, lexical collocations and other lexico-grammatical features to inform EAP materials design, a larger corpus would be useful. The initial exploration of the academic functions in presentations is also an aspect that could be explored in further research.

This study focussed exclusively on student-centred speaking and has not considered the comparison with written reports that accompany the PBL2 presentations. As many of these were provided as part of the data collection, a future area of research would be to compare language use between the PBL oral reports and the written reports. By indicating potential ways forward, it is hoped that this research will motivate further studies into spoken academic English.

Although the benefits of a small, specialised corpus were identified in Chapter 2, the obvious limitation is that findings are not generalisable. However, as

Anderson (2006, p. 4) notes, “it is the nature of corpus research to be comparative”. This was most obvious in the exploration of keywords which makes explicit use of a reference corpus. Otherwise, results can only be compared in general terms and for general tendencies and patterns with findings from across registers, or with other similar corpora or sub-corpora.

Practical limitations also mean that after the quantitative identification of features, all qualitative interpretive work has been carried out by one researcher. While I have made every attempt to bring objectivity to the process, for example revisiting the data and categorisations on multiple occasions, and with notable gaps in between, no other researchers have been involved and so no inter-rater reliability methods or checks can be made.

The corpus comprises interactions of a multilingual international group studying in an English medium environment but I do not explore this from the perspective of English as a Lingua Franca.[34] Instead it merely takes the PBL as a prototypical postgraduate group in an English-medium environment. This corpus is also not tagged. As such, corpus searches for grammatical structures are not available. Similarly, features of pronunciation including intonation are not explored.

Finally, although motivated by EAP, the research does not at this point develop specific materials or tasks but rather makes suggestions for materials and draws conclusions on the implications for EAP course design. This is also a possible direction for future research.

7.5 Concluding Remarks

This research was conducted because I wanted to understand more fully what is involved in dialogic speaking in a specific discipline. At the outset I thought I might discover a neat way of describing spoken interactions. Instead I uncovered a range of complexities involved in only one type of speaking in one academic subject area. This study has highlighted the importance of considering in detail the different types of speech event students engage in if we, as EAP practitioners, are to provide appropriate models and opportunities for academic

[34] Much of this research has focussed on NNS-NNS interactions rather than NS-NNS groups.

language and skills development. While this work has focussed only on one type of learning event and discipline, it has shown how student interactions and their linguistic realisation vary not only depending on the aims of particular events, but from generalised findings of spoken academic English and lecturer talk. Students in this context ‘do PBLs differently’. As it is highly likely this will be found in other types of speech and subjects, it is thus important to continue exploring speaking in the disciplines.

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.

Ädel, A. (2012). 'What I want you to remember is…': Audience orientation in monologic academic genres. *English Text Construction,* 5(1), 98-124.

Ädel, A., & Erman, E. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31(2), 81–92.

Ädel, A., Simpson, R.C., Lee, D.Y.W., & Leicher, S. (2007). *MICASE manual the Michigan corpus of academic spoken English*. The Regents of the University of Michigan. Retrieved from http://www.lsa.umich.edu/eli/micase/index.htm.

Adolphs, S. (2006). *Introducing electronic text analysis: A practical guide for language and literary studies*. London: Routledge.

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38-52). London: Routledge.

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, 24(4), 425–438.

Adolphs, S., & Schmitt, N. (2004). Vocabulary coverage according to spoken discourse context. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 39–49). Amsterdam: John Benjamins.

Aguilar, M. (2016). Seminars. In K. Hyland & P. Shaw (Eds), *The Routledge handbook of English for academic purposes* (pp. 335-347). Abingdon, Oxon: Routledge.

Alavi, C. (1995). *Problem-based learning in a health sciences curriculum*. London: Routledge.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A.P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101-122). Oxford University Press, Oxford.

Anderson, K., Lynch, T., & Maclean, J. (2004). *Study speaking: A course in spoken English for academic purposes* (2nd ed.). Cambridge: Cambridge University Press.

Anderson, W.J. (2006). *The phraseology of administrative French: A corpus-based study*. Amsterdam: Rodopi.

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30(2). 141–161.

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*, 4(2), 91–105.

Barbieri, F. (2015). Involvement in university classroom discourse: Register variation and interactivity. *Applied Linguistics*, 36(2), 151-173.

Barnes, D. (2008). Exploratory talk for learning. In N. Mercer & S. Hodgkinson (Eds.), *Exploring talk in school* (pp. 1-16). London: Sage.

Barrows, H.S. (1986). A taxonomy of problem-based learning methods. *Medical Education*, 20(6), 481-486.

Barrows, H.S. (1994). *Practice-based learning: Problem-based learning applied to medical education*. Springfield, Illinois: Southern Illinois University of Medicine.

Barrows, H. (1999). A taxonomy of problem-based learning methods. In J. A. Rankin (Ed.), *Handbook on problem-based learning* (pp. 19–26). New York: Forbes.

Barrows, H.S. & Tamblyn, R.M. (1980). *Problem-based learning: An approach to medical education*. New York: Springer Publishing.

Basturkmen, H. (1999). Discourse in MBA seminars: Towards a description for pedagogical purposes. *English for Specific Purposes*, 18(1), 63-80.

Basturkmen, H. (2002). Negotiating meaning in seminar-type discussion and EAP. *English for Specific Purposes*, 21(3), 233–242.

Basturkmen, H. (2016). Dialogic interaction. In K. Hyland & P. Shaw (Eds), *The Routledge handbook of English for academic purposes* (pp. 15-164). Abingdon, Oxon: Routledge.

Baumgarten, N., & House, J. (2010). *I think* and *I don't know* in English as a lingua franca and native English discourse. *Journal of Pragmatics,* 42(5), 1184-1200.

Beach, W.A. (1993). Transitional regularities for 'casual' "Okay" usages. *Journal of Pragmatics*, 19(4), 325–352.

Becher, T. (1989). *Academic tribes and territories: Intellectual enquiry and the cultures of disciplines*. Buckingham: Society for Research into Higher Education & Open University Press.

Benwell, B., & Stokoe, E. (2002). Constructing discussion tasks in university tutorials: Shifting dynamics and identities. *Discourse Studies*, 4(4), 429–453.

Berber-Sardinha, T. (2000). Comparing corpora with WordSmith Tools: How large must the reference corpus be? *Proceedings of the Workshop on Comparing Corpora*. Annual Meeting of the ACL, 9 (pp. 7–13). Morristown, NJ: Association for Computational Linguistics.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.

Biber, D. (2006a). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, 5(2), 97–116.

Biber, D. (2006b). *University language: A corpus-study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(3), 263–286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181–190). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based approaches to issues in applied linguistics. *Applied Linguistics*, 15(2), 169-189.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36(1), 9-48.

Biber, D., Conrad, S., Reppen, R., Byrd, H., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. Princeton, New Jersey: TOEFL Monograph Series, Educational Testing Service.

Biber, D., & Finegan, E. (1988). Adverbial stance types in English. *Discourse Processes*, *11*(1), 1-34.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Bligh, J. (1999). Problem based learning in medicine. In J. Rankin (Ed.), *Handbook on problem based learning* (pp. 3-10). New York: Medical Library Association.

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer* [Computer program]. Version 5.3.51, Retrieved from http://www.praat.org/ (December, 2010).

Bolden, G.B. (2009). Implementing incipient actions: The discourse marker 'so' in English conversation. *Journal of Pragmatics*, 41(5), 974-998.

Bondi, M., & Scott, M. (2010). *Keyness in texts*. Amsterdam: John Benjamins.

Browne, C., Culligan, B., & Phillips, J. (2013). *The new academic word list*. Retrieved from http://www.newgeneralservicelist.org.

Bruce, I. (2009). Results sections in sociology and organic chemistry articles: A genre analysis. *English for Specific Purposes*, 28(2), 105-124.

Bruce, I. (2015). Use of cognitive genres as textual norms in academic English prose: University essays in English literature and sociology. *Bulletin Suisse de Linguistique Appliquée/Swiss Journal of Applied Linguistics*, 2(Special Issue, Summer 2015), 161-175.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney papers in TESOL*, 5(5), 31-64.

Cameron, D. (2001). *Working with spoken discourse*. London: Sage.

Camp, M.G. (1996). Problem based learning: A paradigm shift or a passing fad? *Medical Education Online* 1(2), 1-6.

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

Chafe, W. (1986). Evidentiality in English conversation and academic writing. In W. Chafe & J. Nichols (Eds), *Evidentiality: The linguistic coding of epistemology* (pp. 261-272). Norwood, NJ: Ablex.

Chafe, W. (1995). Adequacy, user-friendliness, and practicality in transcribing. In G. Leech, G. Myers & J. Thomas (Eds.), S*poken English on computer: Transcription, mark-up and application* (pp. 54-61). London: Longman.

Chapman, M., & Wulff, S. (2010). The frequency of *we* across academic disciplines and registers in MICASE. *MICASE online kibbitzer*. Retrieved from http://www.lingtechcomm.unt.edu/~swulff/research/Chapman%20&%20Wulff%20(2010).pdf (06/03/2017).

Chazal, E. d., & Moore, J. (2013). *Oxford EAP: A course in English for academic purposes*. Oxford: Oxford University Press.

Chen, Q., & Ge, G.C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles. *English for Specific Purposes*, 26(4), 502-514.

Chen, Y., & Ackermann, K. (2013). Developing the academic collocation list (ACL)-a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4), 235-247.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, 14(2), 30-49.

Christie, F. (2000). The Language of classroom interaction. In L. Unsworth (Ed.), *Researching language in schools and communities* (pp. 184-283). London and Washington: Cassell.

Chung, T.M., & Nation, P. (2004). Identifying technical vocabulary. *System*, 32(2), 251-263.

Coffin, C., Donohue, J., & North, S. (2009). *Exploring English grammar: From formal to functional*. London: Routledge.

Colliver, J.A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine: Journal of the Association of American Medical Colleges,* 75(3), 259–266.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, 29(1), 72-89.

Conrad, S., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, *20*, 56-71.

Cook, G. (1995). Theoretical issues: Transcribing the untranscribable. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on computer: transcription, mark-up and application* (pp. 35-53). London: Longman.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, 17(4), 391–406.

Cowie, A.P. (1998). *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.

Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. In J. Flowerdew & M. Peacock (Eds.*), Research perspectives on English for academic purposes* (pp. 252-267). Cambridge: Cambridge University Press.

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies*, 50(1), 121-135.

Crowdy, S. (1994). Spoken corpus transcription. *Literary and Linguistic Computing*, 9(1), 25-28.

Csomay, E. (2006). Academic talk in American university classrooms: Crossing the boundaries of oral-literate discourse? *Journal of English for Academic Purposes*, 5(2), 117-135.

Csomay, E. (2007). A corpus-based look at linguistic variation in classroom interaction: Teacher talk versus student talk in American university classes. *Journal of English for Academic Purposes*, 6(4), 336-355.

Da Silva, A.L., & Dennick, R. (2010). Corpus analysis of problem-based learning transcripts: An exploratory study. *Medical Education*, 44(3), 280-8.

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms*. Amsterdam: John Benjamins.

Dang, T.N.Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, 33(1), 66-76.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20),* (pp. 51-68). Freiburg im Breisgau 1999, Amsterdam: Rodopi.

DeCarrico, J., & Nattinger, J.R. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes*, 7(2), 91–102.

Dendale, P., & Tasmowski, L. (2001). Introduction: Evidentiality and related notions. *Journal of Pragmatics*, 33(3), 339-348.

Deroey, K.L.B., & Taverniers, M. (2011). A corpus-based study of lecture functions. *Moderna Sprak*, 105(2), 1–22.

Dippold, D. (2014). 'That's wrong': Repair and rapport in culturally diverse higher education classrooms. *The Modern Language Journal*, 98(1), 402-416.

Dolmans, D.H.J.M., Wolfhagen, H.A.P., Van der Vleuten, C.P.M., & Wijnen, W.H.F.W. (2001). Solving problems with group work in problem-based learning: Hold on to the philosophy. *Medical Education*, 35(9), 884–9.

Douglas, F. (2003). The Scottish corpus of texts and speech: Problems of corpus design. *Literary and Linguistic Computing*, 18(1), 23-37.

Downing, K., Kwon, T., Chan, S., Lam, T., & Downing, W. (2009). Problem-based learning and the development of metacognition. *Higher Education*, 57(5), 609-621.

Drew, P., & Heritage, J. (1992). Analyzing talk at work: An introduction. In P. Drew & J. Heritage (Eds.), *Talk at work* (pp. 3-65). Cambridge: Cambridge University Press.

Du Bois, J.W. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (pp. 139–182). Amsterdam: John Benjamins.

Duff, P. A. (2010). Language socialization into academic discourse communities. *Annual Review of Applied Linguistics*, 30, 169–192.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.

Durrant, P. (2015). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*. doi:10.1093/applin/amv011.

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

Edwards, J.A. (1995). Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up and application* (pp. 19-34). London: Longman.

Eggins, S. (2004). *An introduction to systemic functional linguistics* (2nd ed.). London: Continuum.

Ellis, N.C. (1996a). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18(1), 91–126.

Ellis, N.C. (1996b). Analyzing language sequence in the sequence of language acquisition: Some comments on Major and Loup. *Studies in Second Language Acquisition,* 18(3), 361-368.

Emmott, C. (1997). *Narrative comprehension: A discourse perspective*. Oxford: Oxford University Press.

Erman, B. (2001). Pragmatic markers revisited with a focus on *you know* in adult and adolescent talk. *Journal of Pragmatics,* 33 (9), 1337–1359.

Farr, F. (2003). Engaged listenership in spoken academic discourse: The case of student-tutor meetings, *Journal of English for Academic Purposes*, 2(1), 67–85.

Farrell, P. (1990). *Vocabulary in ESP*: *A lexical analysis of the English of electronics and a study of semi-technical vocabulary*. CLCS Occasional Paper, No. 25. Dublin: Trinity College.

Ferguson, G. (2001). If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes,* 20(1), 6-82.

Firth, A., & Wagner, J. (2007). Second/foreign language learning as a social accomplishment: Elaborations on a reconceptualized SLA. *The Modern Language Journal*, 91(1), 800-819.

Flowerdew, J. (1992). Definitions in science lectures. *Applied Linguistics*, 13(2), 202-221.

Flowerdew, J. (2001). Concordancing as a tool in course design. In M. Ghadessy, M.A. Henry & R. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 71–92). Amsterdam: John Benjamins.

Flowerdew, J., & Costley, T. (2016). *Discipline-specific writing*. London: Routledge.

Flowerdew, L. (2004). The argument for using English specialised corpora to understand academic and professional language. In U. Connor & T.A. Upton (Eds.), *Discourse in the Professions* (pp. 11-33). Amsterdam: John Benjamins.

Fontaine, L. (2012). *Analysing English grammar: A systemic functional introduction*. Cambridge: Cambridge University Press.

Fortanet, I. (2004). The use of '*we*' in university lectures: Reference and function. *English for Specific Purposes*, 23 (1), 45–66.

Fortanet, I. (2005). Interaction in academic spoken English: The use of *'I'* and *'you'* in the MICASE. In E. Macià, A. Cervera & C. Ramos (Eds.), *Information technology in languages for specific purposes: Issues and prospects* (pp. 35-52). Springer: New York.

Fraser, S. (2009). Breaking down the divisions between general, academic and technical vocabulary: The establishment of a single, discipline-based word list for ESP learners. *Hiroshima Studies in Language and Language Education*, 12, 151–167.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, N.J: Prentice-Hall.

Genetics Home Reference. Retrieved from https://ghr.nlm.nih.gov/.

Ghadessy, M., Henry, A., & Roseberry, R. L. (2001). *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19(2), 115-135.

Glenn, P.J., Koschmann, T., & Conlee, M. (1999). Theory presentation and assessment in a problem-based learning group. *Discourse Processes*, 27(2), 119-133.

Glenn, P., & Holt, E. (2013). Introduction. In P. Glenn & E. Holt (Eds.), *Studies of laughter in interaction* (pp. 1-22). London: Bloomsbury.

Goffman, E. (1981). *Forms of talk*. Philadelphia: University of Pennsylvania Press.

Goffman, E. (1983). The interaction order. *American Sociological Review,* 48(1), 1-17.

Goh, G. (2011). Choosing a reference corpus for keyword calculation. *Linguistic Research*, 28(1), 239-256.

Gray, B., & Biber, D. (2012). Current conceptions of stance. In K. Hyland & C. Sancho Guinda (Eds.), *Stance and voice in written academic genres* (pp. 15-33). New York, NY: Palgrave Macmillan.

Green, J.L., & Dixon, C. N. (2002). Exploring differences in perspectives on microanalysis of classroom discourse: Contributions and concerns. *Applied Linguistics,* 23(3), 393-406.

Groom, N. (2010). Closed-class keywords and corpus-driven discourse analysis. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 59-68). Amsterdam: John Benjamins.

Halliday, M.A.K. (1994). *An introduction to functional grammar* (2nd ed.) London: Edward Arnold.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M.A.K., & Martin, J.R. (1993). *Writing science: Literacy and discursive power*. London: Falmer Press.

Halliday, M.A.K., & Matthiessen, C.M.I.M. (2013). *Halliday's introduction to functional grammar* (4th ed.). Abingdon, Oxon: Routledge.

Harwood, N. (2005). 'Nowhere has anyone attempted…In this article I aim to do just that'. A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines. *Journal of Pragmatics,* 37(8), 1207–1231.

Henry, A., & Roseberry, R.L. (2001). Using a small corpus to obtain data for teaching a genre. In M. Ghadessy, A. Henry & R.L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 93-133). Amsterdam: John Benjamins.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action* (pp. 299–345). Cambridge: Cambridge University Press.

Heritage, J. (1997). Conversation analysis and institutional talk: Analysing data. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (pp. 161–182). London: Sage.

Heritage, J. (1998). Conversation analysis and institutional talk: Analyzing distinctive turn-taking systems. In S. Cmejrková, J. Hoffmannová, O. Müllerová & J. Svetlá (Eds.), *Dialoganalyse VI (Volume 2) (Proceedings of the 6th International Congress of IADA - International Association for Dialog Analysis)* (pp. 3-17). Tubingen: Niemeyer.

Heritage, J. (2004). Conversation analysis and institutional talk. In R. Sanders & K. Fitch (Eds.), *Handbook of language and social interaction* (pp. 103-146). Mahwah, NJ: Erlbaum.

Heritage, J. (2011). Conversation analysis: Practices and methods. In D. Silverman (Ed.), *Qualitative research: Issues of theory, method and practice* (3rd ed., pp. 208–230). London: Sage.

Heritage, J. (2012a). Epistemics in action: Action formation and territories of knowledge. *Research on Language and Social Interaction*, 45(1), 1–29.

Heritage, J. (2012b). The epistemic engine: Sequence organisation and territories of knowledge. *Research on Language and Social Interaction*, 45(1), 30-52.

Heritage, J. (2013a). Epistemics in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 370-394). Chichester: Wiley-Blackwell.

Heritage, J. (2013b). Language and social institutions: The conversation analytic view. *Journal of Foreign Languages*, 26(4), 2-27.

Heritage, J., & Clayman, S. (2010). *Talk in action: Interactions, identities, and institutions*. Chichester: Wiley-Blackwell.

Hernández, P.S. (2013). Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies*, 12(1), 187–209.

Hewings, M., & McCarthy, M. (2012). *Cambridge academic English: Upper intermediate / an integrated skills course for EAP*. Cambridge: Cambridge University Press.

Hmelo, C.E. (1998). Problem-based learning: Effects on the early acquisition of cognitive skill in medicine. *The Journal of the Learning Sciences*, 7(2), 173-208.

Hmelo-Silver, C.E., & Barrows, H.S. (2006). Goals and strategies of a problem-based learning facilitator. *Interdisciplinary Journal of Problem-based Learning*, 1(1), 21-39.

Hmelo-Silver, C.E., & Barrows, H.S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction*, 26(1), 48-94.

Hoekje, B.J. (2007). Medical discourse and ESP courses for international medical graduates (IMGs). *English for Specific Purposes*, 26(3), 327–343.

Hoey, M. (2001). *Textual interaction: An introduction to written discourse analysis*. New York/London: Routledge.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Holmes, J., & Nesi, H. (2009). Verbal and mental processes in academic disciplines. In M. Charles, S. Hunston & D. Pecorari (Eds.), A*cademic writing: At the interface of corpus and discourse* (pp. 58-72). London: Continuum.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–30.

Hunston, S., & Thompson, G (Eds.), (2000). *Evaluation in text*. Oxford: Oxford University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hutchby, I., & Wooffitt, R. (2008). *Conversation analysis* (2nd ed.). Cambridge: Polity Press.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Harlow: Longman.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes,* 20(3), 207–226.

Hyland, K. (2005a). *Metadiscourse: Exploring interaction in writing*. London: Continuum.

Hyland, K. (2005b). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies,* 7(2), 173–91.

Hyland, K. (2007). Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics*, 28(2), 266-285.

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics,* 18, 41–62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1), 4–21.

Hyland, K. (2009). *Academic discourse*. London: Continuum.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics,* 32, 150–169.

Hyland, K., & Tse, P. (2007). Is there an 'academic vocabulary'? *TESOL Quarterly*, 41(2), 235-253.

Jakonen, T., & Morton, T. (2013). Epistemic search sequences in peer interaction in a content-based language classroom. *Applied Linguistics*, 36(1), 73-94.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed), *Conversation analysis: Studies from the first generation* (pp. 13-31). Amsterdam: John Benjamins.

Jefferson, J. (1978). Sequential aspects of storytelling in conversation. In J. Schenkein (Ed.), *Studies in the organization of conversational interaction* (pp. 219-248). New York: Academic Press.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, 91(3), 433–445.

Johansson, S. (1995). The approach of the text encoding initiative to the encoding of spoken discourse. In G. Leech, G. Myers & J. Thomas (Eds.*), Spoken English on computer: Transcription, mark-up and application* (pp. 82-98). London: Longman.

Kamio, A. (2001). English generic *we*, *you*, and *they*: An analysis in terms of territory of information. *Journal of Pragmatics*, 33(7), 1111–1124.

Kärkkäinen, E. (2003). *Epistemic stance in English conversation: A description of its interactional functions with a focus on I think*. Amsterdam: John Benjamins.

Kärkkäinen, E. (2006). Stance taking in conversation: From subjectivity to intersubjectivity. *Text and Talk,* 26(6), 699-731.

Kilgarriff, A, Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1–30.

Koester, A. (2010). Building small specialised corpora. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). London: Routledge.

Koole, T. (2010). Displays of epistemic access: Student responses to teacher explanations. *Research on Language and Social Interaction*, 43(2), 183–209.

Koole, T. (2012). The epistemics of student problems: Explaining mathematics in a multi-lingual class, *Journal of Pragmatics,* 44(13), 1902–16.

Koschmann, T., Glenn, P., & Conlee, M. (1997). Analyzing the emergence of a learning issue in a problem-based learning meeting. *Medical Education Online*, 2(1), 1-9.

Koschmann, T., Glenn, P., & Conlee, M. (2000). When is a problem-based tutorial not tutorial? Analyzing the tutor's role in the emergence of a learning issue. In D. H. Evensen & C. Hmelo (Eds.), *Problem-based learning: A research perspective on learning interaction* (pp. 53-74). London; Mahwah, N.J: Lawrence Erlbaum Publishers.

Kuo, C. (1999). The use of personal pronouns: Role relationships in scientific journal articles. *English for Specific Purposes,* 18(2), 121–138.

Laufer, B. (1989). What percentage of lexis is necessary for comprehension? In C. Lauren & M. Norman (Eds.), *Special language: From humans to thinking machines* (pp. 316–323). Clevedon: Multilingual Matters.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus.* London: Longman.

Legg, M. (2007). From question to answer: The genre of the problem-based learning tutorial at the university of Hong Kong. *English for Specific Purposes*, 26(3), 344-367.

Leone, P. (2010). General spoken language and school language: Key words and discourse patterns in history textbooks. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 235-248). Amsterdam: John Benjamins.

Levinson, S. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Limberg, H. (2010). *The interactional organization of academic talk: Office hour consultations*. Amsterdam: John Benjamins.

Lindemann, S., & 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(1), 459–475.

Lindström, A., & Sorjonen, M. (2013). Affiliation in conversation. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 350-369). London: Blackwell.

Love, A. (1993). Lexico-grammatical features of geology textbooks: Process and product revisited. *English for Specific Purposes*, 12(3), 197-218.

Lyons, J. (1977). *Semantics 2*. Cambridge: Cambridge University Press.

Macaulay, R. (2002). You know, it depends. *Journal of Pragmatics*, 34(6), 749–76.

Malavasi, D., & Mazzi, D. (2010). History v. marketing: Keywords as a clue to disciplinary epistemology. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 169-184). Amsterdam: John Benjamins.

Markee, N. (2007). Conversation analysis: Issues and problems. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 1017-1032). New York: Springer.

Martin, J. (2000). Beyond exchange: Appraisal systems in English. In S. Hunston & G. Thompson (Eds.), *Evaluation in text* (pp. 142-175). Oxford: Oxford University Press.

Martinez, I.A. (2001). Impersonality in the research article as revealed by analysis of the transitivity structure. *English for Specific Purposes*, 20(3), 227-247.

Mauranen, A. (2006). Speaking the discipline: Discourse and socialisation. In K. Hyland & M. Bondi (Eds.), *Academic discourse across disciplines* (pp. 271-294). Linguistic Insights, 42. Bern: Peter Lang.

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

McCarthy, M., & Handford, M. (2004). 'Invisible to us': A preliminary corpus-based study of spoken business English. In U. Connor & T. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics (pp. 167-202)*. Amsterdam: John Benjamins.

McCormack, J., & Watkins, S. (2012). *English for academic study: Speaking: Course book* (New ed.). Reading: Garnet Publishing.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

McHoul, A. (1978). The organization of turns at formal talk in the classroom. *Language in Society*, 7(2), 183-213.

Mehan, H. (1979). 'What time is it, Denise?': Asking known information questions in classroom discourse. *Theory into Practice*, 18(4), 285-294.

Mercer, N. (1995). *The guided construction of knowledge*. Clevedon: Multilingual Matters.

Mercer, N. (2000). *Words and minds: How we use language to think together*. London: Routledge.

Mercer, N. (2008). Changing our minds: A commentary on 'conceptual change: A discussion of theoretical, methodological and practical challenges for science education'. *Cultural Studies of Science Education*, 3(2), 351-362.

Meyer, C.F. (2002). *English corpus linguistics: An introduction*. Cambridge: Cambridge University Press.

Mok, C.K.F., Dodd, B.J., & Whitehill, T.L. (2009). Speech-language pathology students' approaches to learning in a problem-based learning curriculum. *International Journal of Speech-Language Pathology*, 11(6), 472-481.

Moon, R. (1998). *Fixed expressions and idioms in English*. Oxford: Oxford University Press.

Morell, T. (2004). Interactive lecture discourse for EFL students. *English for Specific Purposes*, 23(3), 325–338.

Mori, J. (2002). Task design, plan, and development of talk-in-interaction: An analysis of a small group activity in a Japanese language classroom. *Applied Linguistics*, 23(3), 323-347.

Morton, T. (2012). Classroom talk, conceptual change and teacher reflection in bilingual science teaching. *Teaching and Teacher Education*, 28(1), 101-110.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, 25(2), 235-256.

Mushin, I. (2001). *Evidentiality and epistemological stance: Narrative retelling*. Amsterdam: John Benjamins.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes,* 63(1), 59-81.

Nattinger, J.R., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nekrasova, T.M. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning*, 59(3), 647–686.

Nesi, H. (2002). An English spoken academic word list. In A. Braasch & C. Provlsen (Eds.), *Proceedings of the tenth EURALEX International Congress*, EURALEX 2002, Volume I (pp. 351-358). Copenhagen: Center for Sprogteknologi.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics,* 11(3), 147-168.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.

Nesi, H., & Thompson, P. (2006). *The British academic spoken English corpus manual*. Warwick: University of Warwick.

Nieminen, J., Sauri, P., & Lonka, K. (2006). On the relationship between group functioning and study success in problem-based learning. *Medical Education*, 40(1), 64-71.

O'Boyle, A. (2014). 'You' and 'I' in university seminars and spoken learner discourse. *Journal of English for Academic Purposes*, 16, 40-56.

O'Connell, D.C., & Kowal, S. (1999). Transcription and the issue of standardisation. *Journal of Psycholinguistic Research*, 28(2), 103–20.

O'Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly* 37(3), 389-418.

O'Keeffe, A., & McCarthy, M. (2010). *The Routledge handbook of corpus linguistics*. Abingdon: Routledge.

O'Keeffe, A., & Walsh, S. (2012). Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education. *Corpus Linguistics and Linguistic Theory*, 8(1), 159-181.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B.B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York: Academic Press.

Ochs, E., & Schieffelin, B. (1989). Language has a heart. *Text* 9(1), 7-25.

Okamura, A. (2009). Use of personal pronouns in two types of monologic academic speech. *The Economic Journal of Takasaki City University of Economics*, 52(1), 17-26. Retrieved from http://www1.tcue.ac.jp/home1/kgakkai/ronsyuu/ronsyuukeisai/52_1/okamura.pdf (12/3/2012).

Palmer, F.R. (1986). *Mood and modality*. Cambridge: Cambridge University Press.

Pawley, A., & Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards & R.W. Schmidt (Eds.), *Language and communication* (pp. 191–226). New York: Longman.

Pekarek Doehler, S. (2010). Conceptual changes and methodological challenges: on language, learning and documenting learning in conversation analytic SLA research. In P. Seedhouse, S. Walsh, & C. Jenks (Eds.). *Conceptualising learning in applied linguistics* (pp. 105-126), Basingstoke: Palgrave MacMillan.

Peräkylä, A. (1997). Validity and reliability in research based on tapes and transcripts. In D. Silverman (Ed.), *Qualitative analysis: Issues of theory and method* (pp. 201-220). London: Sage.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Rayson, P. (2009). *Wmatrix: A web-based corpus processing environment*. Computing department, Lancaster University. Retrieved from http://ucrel.lancs.ac.uk/wmatrix/.

Remedios, L., Clarke, D., & Hawthorne, L. (2008). Framing collaborative behaviors: listening and speaking in problem-based learning. *Interdisciplinary Journal of Problem-based Learning*, 2(1), 1-20.

Rignall, M. & Furneaux, C. (1998). *Speaking*. Hemel Hempstead: Prentice-Hall.

Rounds, P. (1987a). Multifunctional personal pronoun use in an educational setting. *English for Specific Purposes*, 6(1), 13–29.

Rounds, P. (1987b). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly*, 21(4), 643–671.

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

Sahin-Kızıl, S.A., & Kilimci, A. (2014). Recurrent phrases in Turkish EFL learners' spoken interlanguage: A corpus-driven structural and functional analysis. *Journal of Language and Linguistic Studies*, 10(1), 195-210.

Salazar, D. (2011). *Lexical bundles in scientific English: A corpus-based study of native and non-native writing*. Unpublished PhD thesis, University of Barcelona. Retrieved from http://diposit.ub.edu/dspace/bitstream/2445/35023/2/DJLS_DISSERTATION.pdf

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing*. Amsterdam: John Benjamins.

Samuda, V. & Bygate, M. (2008). *Tasks in second language learning*. Basingstoke: Palgrave Macmillan.

Sarangi, S., & Roberts, C. (1999). *Talk, work and institutional order*. Berlin: Mouton de Gruyter.

Savin-Baden, M., & Major, C.H. (2004). *Foundations of problem-based learning*. London: Open University Press.

Schegloff, E.A. (1980). Preliminaries to preliminaries: 'Can I ask you a question'. *Sociological Inquiry*, 50, (3-4), 104-152.

Schegloff, E.A. (2007). *Sequence organization in interaction: A primer in conversation analysis, vol. 1*. Cambridge: Cambridge University Press.

Scheibman, J. (2002). *Point of view in grammar: Structural patterns of subjectivity in American English conversation*. Amsterdam: John Benjamins.

Scheibman, J. (2007). Subjective and intersubjective uses of generalizations in English conversations. In R. Englebretson (Ed.), *Stancetaking in discourse: Subjectivity, evaluation, interaction* (pp. 111–138). Amsterdam: John Benjamins.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329-363.

SCOTS corpus details. Retrieved from http://www.scottishcorpus.ac.uk/corpus-details/.

Scott, M. (1996). *WordSmith Tools manual*. Oxford: Oxford University Press.

Scott, M. (1997). PC analysis of key words-and key key words. *System*, 25(2), 233-45.

Scott, M. (2004). *WordSmith Tools version 4*, Oxford: Oxford University Press.

Scott, M. (2010). Problems in investigating Keyness, or clearing the undergrowth and marking out trails. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 59-68). Amsterdam: John Benjamins.

Scott, M. (2009). In search of a bad reference corpus. In D. Archer (Ed.), *What's in a word-list? Investigating word frequency and keyword extraction* (pp. 79-92). Oxford: Ashgate.

Seedhouse, P. (2004). *The interactional architecture of the language classroom: A conversation analysis perspective*. Oxford: Blackwell.

Sert, O. (2013). Epistemic status check as an interactional phenomenon in instructed learning settings. *Journal of Pragmatics*, 45(1), 13–28.

Sert, O., & Walsh, S. (2013). The interactional management of claims of insufficient knowledge in English language classrooms. *Language and Education*, 27(6), 542-565.

Shamsudin, S., & Ebrahimi, S.J. (2013). Analysis of the moves of engineering lecture introductions. *Procedia-Social and Behavioral Sciences*, 70, 1303-1311.

Sidnell, J. (2010). *Conversation analysis: An introduction*. Chichester, UK: Wiley-Blackwell.

Simpson, R.C. (2004). Stylistic features of academic speech: The role of formulaic expressions. In U. Connor & T.A. Upton (Eds.), *Discourse in the professions.* (pp. 37-64). Amsterdam: John Benjamins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1995). From theory to practice. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up and application* (pp. 99-109). London: Longman.

Sinclair, J., & Coulthard, M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils.* London: Oxford University Press.

Smith, N., Hoffman, S., & Rayson, P. (2008). Corpus tools and methods, today and tomorrow: Incorporating linguists' manual annotations. *Literary and Linguistic Computing*, 23(2), 163-180.

Stæhr, L.S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31(4), 577-607.

Stirling, L., & Manderson, L. (2011). About *you*: Empathy, objectivity, and authority. *Journal of Pragmatics,* 43(6), 1581-1602.

Stokoe, E.H. (2000). Constructing topicality in university students' small-group discussion: A conversation analytic approach. *Language and Education*, 14(3), 184-203.

Stubbs, M. (2004). Language corpora. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 106-132). Oxford: Blackwell Publishing.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language*, 10(1), 61-104.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*,16(3), 371-391.

Swales, J.M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tanguay, E. (2015). *Disciplinary talk: A systemic functional exploration of university seminar discussions*. (Doctoral dissertation). Retrieved from http://eprints.aston.ac.uk/26748/.

Tapper, J. (1996). Exchange patterns in the oral discourse of international students in university classrooms. *Discourse Processes*, 22(1),25-55.

Ten Have, P. (2007). *Doing conversation analysis: A practical guide* (2nd ed.). London: Sage Publications.

The Chambers dictionary (2008). (11th ed.). Edinburgh: Chambers.

The talking glossary of genetic terms. Retrieved from https://www.genome.gov/glossary/.

Thompson, G. (2001). Interaction in academic writing: Learning to argue with the reader. *Applied Linguistics,* 22(1), 58-78.

Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 59–70). Oxford: Oxbow Books.

Thompson, P. (2006). A corpus perspective on the lexis of lectures, with a focus on economics lectures. In K. Hyland & M. Bondi (Eds.), *Academic discourse across disciplines* (pp. 253-270). Bern: Peter Lang.

Thompson, S. (1994). Frameworks and contexts: A genre-based approach to analysing lecture introductions. *English for Specific Purposes*, 13(2), 171-186.

Tsui, A.B.M. (1991). The pragmatic functions of *I don't know*. *Text*, 11(4), 607-622.

Upton, T., & Connor, U. (2001). Using computerized corpus analysis to investigate the text-linguistic discourse move of a genre. *English for Specific Purposes*, 20(4), 313-329.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, 34(4), 457-79.

Ventola, E. (2002). Why and what kind of focus on conference presentations? In E. Ventola, C. Shalom & S. Thompson (Eds.), *The language of conferencing* (pp. 15-50). Frankfurt am Main: Peter Lang.

Wales, K. (1996). *Personal pronouns in present-day English*. Cambridge: Cambridge University Press.

Walsh, S. (2006). *Investigating classroom discourse*. London: Routledge.

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

Walsh, S., & O'Keeffe, A. (2010). Investigating higher education seminar talk. *Novitas Research on Youth and Language*, 4(2), 141–158.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, 27(4), 442–458.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes*, 28(3), 170-182.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics and Cognition*, 14(1), 129–182.

Webber, P. (2005). Interactive features in medical conference monologue. *English for Specific Purposes,* 24(2), 157–181.

Weissberg, B. (1993). The graduate seminar: another research-process genre. *English for Specific Purposes* 12(1), 23-35.

Wenger, E. (2006). *Communities of practice*. Retrieved from http://wenger-trayner.com/introduction-to-communities-of-practice.

West, M. (1953). *A general service list of English words*. London: Longman, Green and Co.

Wilkinson, M. (2014). Using the Keyword tool to explore lexical differences between British and American English in specialised corpora. *CALL-EJ 15*(1), 21-38. Retrieved from http://callej.org/journal/15-1/Wilkinson_2014.pdf.

Woodward-Kron, R., & Remedios, L. (2007). Classroom discourse in problem-based learning classrooms in the health sciences. *Australian Review of Applied Linguistics*, 30(1), 9.1-9.18.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, 32(4), 213-231.

Wray, A., & Perkins, M.R. (2001). The functions of formulaic language: An integrated model. *Language and Communication*, 20(1), 1-28.

Wulff, S., Swales, J.M., & Keller, K. (2009). 'We have about seven minutes for questions': The discussion sessions from a specialized conference in *English for Specific Purposes* 28, 79–92.

Xiao, Z. & McEnery, A. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics*, 33(1), 62-82.

Yang, M. (2015). A nursing academic word list. *English for Specific Purposes*, 37(1), 27-38.

Yorio, C.A. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly*, 14(4), 433-442.

Young, L. (1994). University lectures—macro-structure and micro-features. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 159-176). Cambridge: Cambridge University Press.

Yule, G. (1985). *The study of language*. Cambridge: Cambridge University Press.

Zareva, A. (2009). Informational packaging, level of formality, and the use of circumstance adverbials in L1 and L2 student academic presentations. *Journal of English for Academic Purposes*, 8(1), 55-68.

Zimmerman, D.H. (1984). Talk and its occasion: The case of calling the police. In D. Schiffrin (Ed.), *Meaning, form and use in context: Linguistic applications* (pp. 201-228). Georgetown University Roundtable on Language and Linguistics. Washington, DC: Georgetown University Press.

Zimmerman, D.H. (1992). The interactional organization of calls for emergency assistance. In P. Drew & J. Heritage (Eds.), *Talk at work: Social interaction in institutional settings* (pp. 418-469). Cambridge: Cambridge University Press.

Appendices

Appendix 1: The Steps of the PBL Process

This guide is provided by the Medical Genetics Department to all students.

The steps* of *the PBL process

The first meeting

Having been given the scenario, each member of the group will read it carefully, and then the Chairperson will invite one (or more) members of the group to read it aloud. The Chairperson will then lead the group through the following stages:

1. Explain unknown wording, statements and concepts: Each group member identifies words / phrases that s/he does not understand (linguistically or scientifically). Other members of the group will then, if possible, provide explanations based on their own knowledge / experience.

Note that this is not a test, and there is no shame in admitting ignorance, only in allowing yourself to remain ignorant!

2. Define the problem and identify all the main issues: Group members identify the key issues; the Scribe makes a note of each issue as a list on the whiteboard.

3. Brainstorming: For each of the main issues the Chairperson will invite the members of the group to contribute whatever previous knowledge they have, that will help the group understand and explain what is going on in the scenario. The issues should be addressed in as logical an order as possible; the Chairperson may consult the group to decide the best order. The Scribe will write notes of what is said on the whiteboard. As far as possible, based on the current knowledge of the group, hypotheses or plausible explanations should be proposed during the brainstorming.

4. Plan your learning:

- Identify what is already known from the brainstorming

- Identify knowledge needed (in other words, your learning objectives) – these are the areas which you require to research in order to fully understand what is going on in the scenario.

- Identify appropriate learning resources – these resources will obviously include textbooks, the WWW, journals, etc.

- Identify the group's response to the problem and the means by which it should be communicated to others, e.g., written report, patient information leaflet, oral presentation, etc.

5. Assign tasks* to *members: So the group should allocate the tasks between the members, in an equitable fashion. Make sure that individual tasks have been clearly defined so that effort is focussed on the correct area(s). Be specific rather than general, using assignments like, "Find out about the molecular basis for galactosemia" or "How is galactosemia treated and managed?", not simply "Find out about galactosemia". Group members should ask for their tasks to be defined more precisely if they are in any doubt about their assignment.

Individual work

6. Research and completion* of *assigned tasks: Each member of the group will spend some time over the next day or two investigating their allocated task(s), remembering to stay focussed, keeping a note of findings and sources of information, so that these can be reported back to the group. It may be desirable in some cases to maintain contact with the group between meetings, so an exchange of e-mail addresses might be useful.

7. Preparation* of *summary: Each student prepares, **using their own words,** a summary of their findings, no longer than one A4 side of text and no longer than two A4 sides in total including all text, diagrams, tables & references. All information should be presented in the student's own words (not cut and pasted) and appropriately referenced and acknowledged. Copies of this summary must be made for each group member and the facilitator.

Second Meeting

8. Present and debate new knowledge:

- Critically evaluate the new knowledge and its source – the Chairperson will ask each group member to report back on their findings. Students should not refer to their

- written reports while they are presenting, but may have a short list of bullet points to remind them of the main issues.

- During the presentations, other members of the group should ask for further clarification on points they do not follow and be prepared to discuss issues which may be problematical.

- Select new knowledge to be applied to the problem

- Combine old and new information and gain understanding of its application to the problem – the group should revisit the scenario to see how they would handle the case with the benefit of their new knowledge.

9. Group review and evaluation:

- Summarize what was learned.

- Identify what was not learned – it may be necessary to carry forward some objectives for further research and consideration at a subsequent meeting of the group.

- Assess group working and discuss changes which may be required to group etiquette for subsequent meetings.

- Provide **constructive criticism** for each other's reports and presentations: this means highlighting good points as well as points that could be improved

- Group members should invite criticism and not be defensive – no-one is perfect!

Individual Work

10. Individual review and evaluation:

- Test understanding of knowledge by application to another problem.

- Reflect on and evaluate your own contribution to the group.

- Assess each individual's contribution to the group.

Appendix 2: The Transcription System

The following guide was developed and used for the transcription of the PBL corpus. Praat software transcription software was used to facilitate transcription (freely available at http://www.fon.hum.uva.nl/praat/)

	Rule/guide	***Examples***
General	• Punctuation kept to a minimum. • Possessives and contractions kept. • Question marks used when utterance appears to be a question.	
Capitalisation	• Capitals are not used at beginning of turns and utterances nor for personal pronoun "i". • Used for proper nouns (e.g., acronyms of tests, country/place names, names of conditions, abbreviations for technical lexis).	*MLPA* Amsterdam Downs *CF* (for cystic fibrosis)
Repetitions & false starts	• Repetitions and false starts transcribed.	
Hyphenation	• Truncated/unfinished words marked by a hyphen. • Otherwise words not hyphenated.	it's a *phy- <u>phy-</u>*
Pauses	• Commas used for noticeable pauses. • Full-stops mark end of utterance and noticeable gaps. • Pauses of 2 seconds or more noted in brackets.	S14: Spina Bifida, or maybe the ante-natal screening. *(2)*
Spelling & lexical items	For orthographic spelling, *Chambers* dictionary was used for general vocabulary. For other technical items: • *Genetics Home Reference* http://ghr.nlm.nih.gov/ and • http://www.genome.gov/glossary/index.cfm. [35]	

[35] Sites recommended by Medical Genetics departmental staff.

	Rule/guide	***Examples***
	• British English spelling, e.g., 'ise' instead of 'ize'; 'colour' instead of 'color'. • Because and cos: each transcribed. • The following are written in full: • Numbers and percent • 'Gonna' written in full: Going to.	
Semi-lexical items	• ah, ahh, ahm, eh, ehm, em, er, erm mhum, uh, uhm, uhuh, uhum, um, umm.	
Unclear	• Unclear utterances surrounded by square brackets e.g., [?]. • Where a 'best guess' is possible, e.g., from context, this is surrounded by square brackets with the word preceded by a question mark [? word]. • Square brackets [ ] also surround significant non-verbal events mid-utterance e.g., laughter; referring to handouts, gestures.	S6: [?] [?de] de novo S6: [mumbles]
Censored	• Speakers all referred to by speaker number, e.g., Student→ S1. • Tutor by speaker number and tutor number, e.g., →S2T1. • Names in scenarios replaced with capital letter e.g., (J).	S3: i just remember S38_T5 say ask this
Transcriber notes	• Transcriber notes: for additional events or in silences, e.g., door opening, multiple overlaps. • Where words or intonation are noticeable exaggerated, these are included in transcriber notes.	

Appendix 3: Keywords in PBL1s and PBL2s

PBL1s				PBL2s			
Keyword	**Freq.**	**Keyness**	**Set**	**Keyword**	**Freq.**	**Keyness**	**Set**
TEST	169	513.40	AC	TEST	30	723.24	AC
TESTING	95	431.58	AC	AFFECTED	46	330.34	AC
DATABASE	34	176.29	AC	DETECT	66	298.64	AC
POPULATION	76	169.85	AC	MATERNAL	48	257.96	AC
PRENATAL	20	134.26	AC	DELETION	35	199.78	AC
INHERITED	12	119.85	AC	SEQUENCING	29	163.62	AC
FEATURES	40	118.55	AC	DELETIONS	25	153.67	AC
SEPARATE	43	103.88	AC	POPULATION	79	142.61	AC
SCENARIO	19	63.34	AC	TESTING	40	115.14	AC
INHERITANCE	13	61.87	AC	SENSITIVITY	30	106.73	AC
RESULTS	27	40.44	AC	SPECIFICITY	21	104.10	AC
SCREENING	190	1056.18	AT	POSITIVE	66	103.69	AC
TEST	169	513.40	AT	POPULATIONS	27	97.72	AC
TESTS	100	414.93	AT	GROUP	136	95.40	AC
COUNSELLING	61	369.55	AT	DETECTED	17	66.30	AC
GENETIC	79	339.82	AT	TESTED	22	62.52	AC
SYNDROME	78	312.95	AT	UNAFFECTED	13	61.74	AC
MUTATION	44	279.07	AT	OVERLAP	18	61.18	AC
MUTATIONS	44	259.12	AT	PATHWAY	17	60.29	AC
RISK	99	247.09	AT	NEGATIVE	44	58.53	AC
SCRIBE	29	194.68	AT	FEATURES	29	57.51	AC
DIAGNOSTIC	34	141.90	AT	ASSOCIATED	35	45.87	AC
PREGNANCY	27	111.02	AT	SCENARIO	17	45.03	AC
DIAGNOSIS	38	96.42	AT	ACCURATE	21	42.06	AC
PARENTS	29	78.07	AT	SPECIFIC	45	41.73	AC
RISKS	17	57.33	AT	FACTORS	41	37.64	AC
GENETICIST	10	53.30	AT	PRENATAL	6	36.88	AC
MOLECULAR	17	50.88	AT	INHERITED	11	32.63	AC
AGE	36	49.59	AT	RESULT	41	32.57	AC
AFFECTED	18	46.33	AT	TRANSCRIPTION	10	31.21	AC
GENE	17	45.05	AT	MAINLY	19	27.55	AC
ABNORMALITY	8	33.47	AT	TECHNIQUE	22	27.53	AC
ABNORMALITIES	8	32.09	AT	COMMON	44	26.98	AC
GENES	12	31.26	AT	IDENTIFIED	17	26.58	AC
FACTORS	31	29.61	AT	SEQUENCE	18	25.06	AC
MAYBE	167	328.98	HF	AFFECT	20	24.28	AC
THINK	491	308.62	HF	MUTATION	144	864.42	AT
LIKE	445	251.22	HF	MUTATIONS	131	756.71	AT
IT	1159	193.00	HF	TEST	209	723.24	AT
BABY	36	155.94	HF	DNA	113	694.72	AT
NO	271	149.34	HF	SCREENING	112	509.59	AT
DO	464	133.35	HF	GENE	96	420.06	AT
CAUSES	60	129.95	HF	CANCER	105	396.89	AT
SHE	134	120.90	HF	SYNDROME	97	363.18	AT
INHERITED	14	119.85	HF	RISK	138	341.88	AT
SEPARATE	43	103.88	HF	GENES	75	333.54	AT
BOARD	48	102.37	HF	AFFECTED	35	330.34	AT
WE	718	100.62	HF	PREGNANCY	64	290.11	AT
REPORT	43	99.05	HF	PROTEIN	110	271.36	AT
SHOULD	142	95.80	HF	PROBES	36	191.36	AT
JUST	379	91.84	HF	NORMAL	97	156.79	AT

PBL1s				PBL2s			
Keyword	**Freq.**	**Keyness**	**Set**	**Keyword**	**Freq.**	**Keyness**	**Set**
COULD	173	83.32	HF	CHILD	68	155.92	AT
CHECK	28	59.74	HF	GENETIC	49	153.67	AT
FAMILY	34	49.70	HF	COUNSELLING	29	146.76	AT
SHALL	35	46.37	HF	PARENTS	45	126.66	AT
SOMEONE	46	44.78	HF	CRITERIA	40	125.90	AT
SURE	65	43.72	HF	CHROMOSOME	35	124.06	AT
WHAT	512	42.96	HF	FOETUS	23	115.58	AT
WOULD	221	42.18	HF	TUMOUR	28	108.81	AT
STUFF	40	41.92	HF	SYNDROMES	21	104.10	AT
EVERYONE	28	34.90	HF	TESTS	42	101.57	AT
KNOW	311	33.52	HF	CYSTIC	18	90.16	AT
WONDERING	14	33.48	HF	DIAGNOSTIC	26	85.32	AT
MENTION	22	32.45	HF	POLYMORPHISM	15	84.82	AT
NEED	115	31.25	HF	ABNORMALITIES	19	84.73	AT
YES	96	28.96	HF	INCIDENCE	25	84.22	AT
ELSE	53	28.87	HF	ULTRASOUND	23	74.94	AT
CHAIR	9	25.91	HF	CHROMOSOMES	15	73.02	AT
YEAH	1347	3577.20	LF	SURGERY	37	65.19	AT
OK	285	1704.77	LF	PHENOTYPE	17	61.68	AT
OH	142	198.31	LF	SEVERE	28	56.08	AT
PEDIGREE	21	127.52	LF	PATIENT	68	52.82	AT
SUPPOSE	43	48.14	LF	DIAGNOSED	17	52.13	AT
TOMORROW	18	40.37	LF	MUTANT	8	49.17	AT
CHROMOSOMAL	6	31.42	LF	BIOCHEMICAL	12	45.93	AT
COS	113	502.52	O	SLIDE	33	44.65	AT
I	1094	226.07	O	DIAGNOSE	13	40.47	AT
PBL	29	194.68	O	TUBE	18	37.31	AT
NON	64	81.14	O	KIT	9	34.05	AT
FISH	21	58.93	O	AUTOSOMAL	7	33.68	AT
PBLS	5	33.56	O	DIAGNOSIS	24	33.20	AT
GLASGOW	6	27.10	O	ALLELE	8	32.32	AT
CYTOGENETIC	4	26.85	O	PARENT	12	31.70	AT
ER	160	1029.98	SL	WEEKS	33	30.07	AT
UHUH	63	422.96	SL	MOLECULAR	13	27.15	AT
EM	69	396.14	SL	LIKE	615	360.94	HF
UHUM	55	369.24	SL	IT'S	866	218.43	HF
AH	112	295.15	SL	FAMILY	88	208.61	HF
EH	50	293.71	SL	THEY	725	151.21	HF
MHUM	32	214.82	SL	BABY	37	141.24	HF
EHM	27	181.25	SL	CASES	59	86.70	HF
UHM	25	167.83	SL	FOUND	75	77.45	HF
UMM	18	113.07	SL	NO	272	71.57	HF
UM	20	75.33	SL	SHE	129	64.61	HF
AHM	8	53.70	SL	ALSO	213	62.99	HF
ERM	8	53.70	SL	SIX	93	61.25	HF
UH	8	33.47	SL	HIGH	81	54.79	HF
AHH	4	26.85	SL	THANK	48	50.75	HF
				FAMILIES	20	47.61	HF
				CAUSES	40	44.93	HF
				AGE	38	38.98	HF
				CAUSED	22	38.05	HF
				FIVE	131	38.02	HF
				OTHER	262	37.68	HF
				PAPERS	19	33.98	HF

PBL1s				PBL2s			
Keyword	**Freq.**	**Keyness**	**Set**	**Keyword**	**Freq.**	**Keyness**	**Set**
				BABIES	12	33.32	HF
				SO	1006	32.30	HF
				PAPER	35	32.21	HF
				QUESTIONS	66	27.97	HF
				FOLLOW	27	26.94	HF
				OFFER	17	25.18	HF
				MENTIONED	37	24.45	HF
				OK	201	1048.69	LF
				YEAH	590	581.83	LF
				OH	143	140.40	LF
				CARRIER	24	98.84	LF
				PREGNANT	25	93.74	LF
				DETECTION	20	92.70	LF
				MAYBE	104	90.21	LF
				TERMINATE	11	56.64	LF
				SIXTY	48	51.93	LF
				FALSE	23	47.47	LF
				DELAY	13	43.84	LF
				MARKER	14	40.93	LF
				PEDIGREE	8	39.36	LF
				MOSTLY	17	31.00	LF
				MARKERS	8	26.52	LF
				PERCENT	95	388.32	O
				PBL	28	172.12	O
				MLPA	21	129.09	O
				COS	42	108.66	O
				RECESSIVE	13	72.80	O
				PENETRANCE	9	55.32	O
				DE	32	49.41	O
				I	1105	45.65	O
				DON'T	269	44.65	O
				WEE	12	40.38	O
				UK	10	31.21	O
				NON	49	26.81	O
				CANNOT	28	24.05	O
				ER	961	5855.86	SL
				EM	365	2139.90	SL
				EHM	233	1432.82	SL
				EH	222	1303.19	SL
				AH	216	661.99	SL
				UH	113	642.02	SL
				UM	117	592.94	SL
				UHM	78	479.51	SL
				UHUH	66	405.73	SL
				ERM	34	209.00	SL
				AHM	32	196.71	SL
				UHUM	24	147.53	SL
				UMM	14	78.80	SL
				MHUM	6	36.88	SL
				AHH	4	24.59	SL

Note. Keywords are sorted first by keyness, then alphabetically and then by frequency in the PBL sub-corpus. Keywords found in both PBL1s and PBL2s in the same part of speech are shaded in grey.

Appendix 4: Raw and Normalised Frequencies of *I-*, *We-* and *You-* Forms across Corpora

Raw and Normalised Frequency for all *I-*, *We-* and *You-*forms across PBLs

Pronouns	PBL1s		PBL2s		AllPBLs	
	No.	Per 1000	No.	Per 1000	No.	Per 1000
First person						
I	1167.0	26.4	1332.0	18.7	2499.0	21.7
me	37.0	0.8	59.0	0.8	96.0	0.8
my	29.0	0.7	85.0	1.2	114.0	1.0
mine	4.0	0.1	10.0	0.1	14.0	0.1
myself	0.0	0.0	0.0	0.0	0.0	0.0
Total	**1237.0**	**28.0**	**1486.0**	**20.8**	**2723.0**	**23.7**
we	774.0	17.5	491.0	6.9	1265.0	11.0
us	15.0	0.3	31.0	0.5	46.0	0.4
let's	10.0	0.2	6.0	0.1	16.0	0.14
our	15.0	0.3	49.0	0.7	64.0	0.5
ours	0.0	0.0	1.0	0.0	1.0	0
ourselves	0.0	0.0	0.0	0.0	0.0	0.1
Total	**814.0**	**18.4**	**578.0**	**8.1**	**1392.0**	**12.1**
Second person						
you	1079.0	24.4	1378.0	19.3	2457.0	21.4
your	46.0	1.0	67.0	0.9	113.0	1.6
yours	0.0	0.0	10	0.1	10	0.1
yourself	0.0	0.0	3.0	0.04	3.0	0.04
yourselves	1.0	0.02	0.0	0.0	1	0.02
Total	**1126.0**	**25.5**	**1458.0**	**20.5**	**2581**	**22.4**

Raw and Normalised Frequency for all *I-*, *We-* and *You*-forms across Rounds (1987a), Fortanet (2004) and Webber (2005) Corpora

Pronouns	MICASE (F)		Rounds		Webber	
	No.	Per 1000	No.	Per 1000	No.	Per 1000
First person						
I	13827.0	17.90	301.0	11.5	312	8.9
me	1180.0	1.53	26.0	1.0	16	0.5
my	1244.0	1.60	2.0	0.1	27	0.8
mine	0	0	0	0	0	0
Total	**16251.0**	**21.03**	**329.0**	**12.6**	**355**	**10.2**
we	7450.0	9.70	907.0	34.0	450	12.9
us	610.0	0.80	35.0	1.0	23	0.7
let's	644.0	0.80	92.0	3.5		
our	785.0	1.00	18.0	0.6	50	1.4
ours	0	0	0	0	0	0
Total	**9489.0**	**12.30**	**1052.0**	**39.1**	**523**	**15.0**
Second person						
you (sub & obj)	16000.0	20.70	335.0	12.5	366	10.5
your	1664.0	2.10	3.0	0.1	13	0.4
Total	**17664.0**	**22.80**	**338.0**	**12.6**	**379**	**10.9**

Note. Corpus size indicated in brackets: MICASE (F) = Fortanet (2005), (770,353 words); Rounds (1987) (26,743 words), based on Fortanet's (2004) figures; Webber (2005), (34,692 words).