'You' and 'I' in university seminars and spoken learner discourse

O'Boyle, A. (2014). 'You' and 'I' in university seminars and spoken learner discourse. *Journal of English for Academic Purposes*, *16*, 40-56. https://doi.org/10.1016/j.jeap.2014.08.003

Published in:
Journal of English for Academic Purposes

Document Version:
Peer reviewed version

Minor Revision of JEAP-D-12-00068

BLINDED MANUSCRIPT

'You' and 'I' in university seminars and spoken learner discourse

Keywords
Pronouns, *you*, *I*, clusters, university classroom, spoken academic learner discourse

Abstract
You and *I* may be little words but they do a great deal. In spoken discourse they reference shared knowledge and mark stance. In pedagogical contexts, they maintain relations in teacher-student discourse. However, language classrooms may rarely explore this array of pragmatic meanings. A lack of awareness of the variety of these functions may be problematic for learners when seeking to construct interpersonal relations and operate successfully in particular spoken contexts. This paper presents a study of *you* and *I* in two spoken corpora: a corpus of English language learner task talk and a corpus of university seminar talk. Findings illustrate different patterns of *I* and *you* between the two corpora: *I* and *you* have a higher rate of occurrence in learner discourse, and pronoun repetition is more frequent in learner discourse, though it does not account for the higher rate of *you* and *I*. These findings suggest that language learner task talk displays more features tied to speech production and self-regulation and fewer features associated with attempting to point to the informational space of others, a key feature of university classroom talk. This paper concludes by outlining pedagogical applications to overcome features perceived as disfluent.

1. Introduction

University seminars and tutorials are driven by an underlying belief in the educational value of face-to-face discussions of subject knowledge. It is often explicitly stated that university classrooms and seminars focus on the exchange of opinions and evaluation of knowledge through dialogue (e.g. Griffiths, 2009). Yet, as Kim (2006) reports, it is participating in whole-class discussion and small group discussions which is of significant concern for English for Academic Purpose students. Although *you* and *I* may seem to be little functional words, in the complex arrangement of linguistic and non-linguistic features which make up participation in university classrooms, they do a great deal. Personal pronouns reference a shared knowledge of people, objects and entities (Carter and McCarthy, 2006); they signal agentive action and mark the territories of information of the speaker and listener (Kamio, 1997). With particular reference to the evaluation of knowledge, pronouns index stance-taking and can signal the alignment or, indeed, disalignment, of speakers to evaluations in face-to-face contexts (Du Bois, 2007).

Although powerful, these items have received more attention in academic writing than in speech (e.g. Tang and John, 1999; Hyland, 2002; Harwood, 2005). Observed differences in the use of pronouns in novice-expert or NS-NNS written discourse are considered to be indicative of different levels of mastery in particular writing events (e.g. Hyland and Milton, 1997; Gilquin and Paquot, 2007).

The domination of research in written discourse may not solely account for the backgrounding of *you* and *I* in spoken discourse. It may be the case that the array of pragmatic meanings expressed in the use of *you* and *I* are rarely explored in EAP classrooms in relation to spoken academic genres. Personal pronouns may be reduced to core meanings: to signal who is the speaker, who is the listener and who/what is being talked about.

However, it is important for EAP students and teachers to be aware of the variety of subtle and incremental functions which *you* and *I* perform in conjunction with other features in order that English language learners can operate successfully in particular spoken academic genres. When used contrary to genre expectations, seemingly discrete functional words such as *I, you* and *we* can become a source of controversy (Hinkel, 2004). The omission and misuse of pronouns in learner discourse are argued to be the negative effects of first language (Muñoz, 1991), which are conceptualized as interference and therefore problematic. Despite being a regular feature of spoken discourse (Fung and Carter, 2007; Svartvik, 1980; Altenberg, 1990; Clark and Fox Tree, 2002), repetitions of personal pronouns are generally regarded as markers of disfluency (e.g. Clark and Wasow, 1998) and are considered to interrupt the naturalistic flow demanded in some assessment frameworks (e.g. Council of Europe, 2001; Brown, 2007). It seems, therefore, that although these little items may not be salient in the stream of speech, their absence, overuse, misuse, or underuse becomes entirely significant for listeners' perceptions of fluency and genre expectations.

It is also important to consider that in relation to EAP students, a speaker's language use may reflect the classroom practices, teaching approaches and materials they have experienced. Therefore, it is of significant consequence to understand how and why learners use language in and out of language classroom events, without the need to describe this use as misuse or deficient. Of equal importance is evidence of the "genre-specific purposes and discipline-specific practices" (Groom, 2005: 257) of the spoken academic discourse in which EAP students will be involved.

Learner discourse has been investigated through corpus-based studies, which often compare learner or novice discourse with native speaker or expert discourse (Hyland and Milton, 1997; Gilquin et al., 2007; Gilquin and Paquot, 2007; Gilquin, 2008; Luzon, 2009; Martinez, 2005). Such research can be embedded within a 'different' versus 'deficient' model of communication and can be discussed in relation to how, for example, English as a Lingua Franca is conceptualized (Seidlhofer, 2005). Some see comparisons of learner and expert discourse as a means to learner empowerment (Martinez, 2005), as a means to the development of important competencies (Hyland and Milton, 1997) and as a means of overcoming non-fatal infelicities and misuse (Gilquin et al., 2007). To take up a 'difference' model is to investigate which patterns of use occur, and why. As an alternative to scouring learner discourse for examples of error or incompetency, an investigation of learner discourse can reveal something about the process of language learning itself, just as any spontaneous speech provides clues to the process of speech production (Clark and Fox Tree, 1997; Chafe, 1994).

Indeed, corpus-based studies offer both a means of identifying particular features of learner discourse and providing possible explanations for observed differences. One explanation for the observed overuse of pronouns in learner written discourse, when compared to native speaker writing, is that learners transfer what they know about the use of pronouns in speaking to the production of written discourse (Gilquin et al., 2007). Less has been written about pronouns in learner spoken discourse than in learner written discourse, with the notable exception of Gilquin (2008). Parallels can be drawn between learner writing and novice writing in EAP contexts, although differences between these groups can also be noted (Gilquin et al., 2007). However, parallels may be less clearly drawn between learner speech and novice speech, or between learner speech and expert speech, whatever that may be. In an EAP context, an undergraduate seminar may constitute novice speech, which is novice to the discipline and event, whereas an inaugural lecture may constitute expert speech. It may be more difficult to consider where a conference presentation or a postgraduate seminar would lie along such a continuum. These are related, yet different, events to which

corpus studies can contribute an understanding. For example, Csomay (2007) observes changes in student and teacher academic discourse as the level of instruction increases. In her study of US university classroom talk, Csomay (2007) suggests that an increase in the use of personal pronouns and active voice demonstrates an increase in interactivity with level of instruction (Csomay, 2007). In relation to language variety, Rowley-Jolivet and Carter-Thomas (2005) observes differences in *you* and *I* in conference presentations. NNS scientists appear to avoid such interactional features preferred by their NS peers. Therefore, it would seem that particular linguistic features such as pronouns must be understood in relation to the (genre specific) purposes and practices of the events in which they occur.

If the goal for EAP students is successful participation in university seminars, it may prove useful to compare the learner discourse of an ELT classroom with the talk of university seminars.

As spoken academic discourse becomes increasingly internationalized, through English medium universities, internationalization at home strategies, and an increase of online learning, understanding the various functions that personal pronouns play and how to avoid the negative consequences of differential use is ever more relevant. Pronouns clearly do more than index a speaker and a listener; they contribute to the overall sense of what it means to participate fully, or be fluent, in a particular genre; and they can be used differently in particular contexts for particular purposes. Potential hazards exist for all language users in relation to the complexity of pronoun use described as the 'politics of pronouns' (Pennycook, 1994). Knowingly or otherwise, the choice of one particular pronoun over another can index more than a shared reference; it can signal inclusion or exclusion or result in contestable representations (Wortham, 1996).To some extent, therefore, these little functional words carry quite a heavy social burden.

In terms of preparing such information for teaching purposes, personal pronouns in academic contexts are items which are far from amenable to a simple description of language use unrelated to their associated complex social signals.

However, EAP educators can use descriptions of academic discourse derived from research using spoken corpora and complementary methods (Basturkmen, 2002). Research-based evidence can assist EAP teachers to simulate academic tasks and develop suitable materials which prepare students for active participation in seminars. Alongside corpus studies of native speaker discourse, corpus studies of learner discourse have illustrated differences between native and non-native discourse, and such comparisons of learner and expert discourse can be discussed as a way of overcoming the non-fatal infelicities and misuse which often lead to the impression of non-fluency (Gilquin et al., 2007).

This paper investigates the use of personal pronouns in two spoken corpora. The use of personal pronouns *you* and *I* in seminars and group discussions are examined in relation to their frequency and cluster patterns in two corpora: UNITALK (a spoken corpus of university classroom talk) and ELLTTALK (a spoken corpus of English Language Learner talk taken from a university setting). The conclusions drawn in this paper suggest that the use of *you* and *I* by learners may be indicative of a self-regulatory function employed by language learners when they are involved in thinking in spoken interaction and where priority is assumed to be given to the content over delivery. The pedagogical applications of this research on personal pronoun use are presented together with discussion on why an increased awareness of the power of these little items might be a good thing for EAP learners.

2. Literature Review

The use of *you* and *I*, together with other features, reflect "directly interactive situations" in university classrooms (Biber et al., 2002:14) and occur more frequently in contemporary spoken university discourse than they have done in the past (Fortanet, 2004). Typically, personal pronouns refer to someone/thing already mentioned and as such they largely mark knowledge that is shared, or assumed to be shared, in interaction (Carter and McCarthy, 2006). The meaning of all pronoun referents is not semantically present; it must be inferred from the information available in a shared context. For example, it cannot always be assumed that *I* is the speaker and *you* is the listener (Biber et al. 1999). Even when signalling thought-to-be-shared information, with the use of pronouns, a degree of active participation is required to recover and maintain the focus of attention and the meaning of pronoun referents.

It is useful here to draw on Kamio's pragmatic theory of pronouns to examine both the pragmatic and psychological aspects of pronouns. Kamio (1997; 2001) proposes that personal pronouns, *I*, *you* and *we* indicate territories of information, described as either proximal or distal. From a speaker's perspective, *I* and *we* are located in proximal conversational space, that is, the space of the speaker, and *you* is positioned in distal space; the territory of the listener. However, on certain occasions, the territorial boundary between *we* and *you* is diminished and *you* becomes "a near- synonym of *we*" (2001:1119). Kamio (2001) suggests that the preference for *we* over *you* in such cases may indicate a previous alliance with speakers and hearers to the same group, and the preference for *you* over *we* (where they can be interchangeable) suggests an absence of previous alignment to the same territory. In this respect, territorial boundaries are not necessarily signalled in a straightforward manner and the meaning of a pronoun is more labile than might be expected. This understanding highlights that the referent is not semantically or necessarily immediately obvious and, therefore, significant effort may be required to recover the meaning intended.

In lectures, Fortanet (2006) finds that *you* and *I* not only index speaker and listener, but are used to signal a range of referents beyond those present. Moreover, they perform a range of discourse functions, from expressing attitudes and organizing discourse to drawing or distancing discourse participants. These are crucial aspects to note for EAP: the comprehension of the particular discourse function of pronouns enacted at any one time in the ongoing discourse is only possible by understanding their context of use. The task is, therefore, a considerable one for NNS who may be more focussed on the semantic short-hand of *you* and *I*, rather than on the pragmatic information that needs to be inferred.

In order to be able to cope with the multifarious and time-constrained nature of university classroom talk, speakers rely on pre-fabricated elements, sequences of words or lexical bundles, to navigate their participation in university classrooms (Biber, et al., 2004). Approaches to collocation (e.g. Sinclair, 1991) can identify recurring sequences of words in a corpus using frequency measures and it is now recognized that such sequences are "essentially the building blocks of spoken and written discourse" (Lin and Adolphs, 2009:34).

Pronouns make their way into the formation of these blocks and they cluster with other items to form such bundles. In addition to signalling physical and discourse entities, proximal or distance informational space, pronouns also signal psychological, social and cultural territories through their deployment in stance bundles such as *I think* (Karkkainen, 2003). These frequently occurring sequences of words, e.g. *and I think that* and *I mean you know,* occur twice as often in university face-to-face classroom teaching than in conversation (Biber et al., 2004). In particular, stance bundles such as *I don't know I* and *I think that* are

extremely common in university classroom discourse and appear necessary in order to be able to operate in this particular complex spoken genre (Csomay, 2007).

Pronouns are constituent elements of markers of shared knowledge such as *I mean* and *you know,* which are used to signal cooperation across turns and achieve intersubjectivity (McCarthy, 2010). These interpersonal expressions, highly frequent in native speaker discourse, signal a joint focus of attention for speakers and listeners. It has been argued that the use of these markers contribute to a sense of fluency and coherence in face-to-face communication when they are used to expand on previous utterances, monitor shared knowledge or invite listener inferences (Carter and McCarthy, 2006; Fox Tree and Schrock, 2002). In learner discourse, Gilquin (2008) demonstrates a comparative underuse of such items as *I mean* or *you know* to signal joint attention and suggests that this underuse contributes to an impression of non-fluency (2008).

Kesckes (2007) argues that in learner interaction goals of cooperation and shared knowledge are somewhat secondary to the primary aim to deploy the linguistic means available to them to ensure that a communicative goal is reached. It may be the case that in language learning classroom contexts there is less focus on reaching interpersonal, intersubjective positions and connecting with the informational space of others and more focus on expressing the content of a message, or regulating the speaker's own position. Kesckes (2007) suggests that speakers rely more on a linguistic code than a shared or common knowledge. This might suggest that the tendency for pronoun overuse in learner writing described as a lack of awareness of the genre or register, is actually more specific in terms of difficulties surrounding the lack of facilitation of reader engagement (Hyland, 2005). It may be the case that this lack of engagement is equally observable in the absence of features of spoken learner discourse which signal shared knowledge and align with the positions of other interactants.

Notions of sharing knowledge and incorporating a previous speaker's utterance into one's own are not only features of intersubjectivity (Du Bois, 2007). Developments continue in dialogic approaches to the study of language use and such ideas are becoming incorporated in definitions of orality and fluency (O'Connell et al., 2004). Fluency reconceptualised as confluence by McCarthy (2010) is a "jointly produced artefact which constitutes an efficient and successful interaction" (2010: 7) and takes into account how responsibility for meaning is distributed between both speakers and listeners.

It is important to note the impact of time and production constraints on classroom face-to-face interaction (Csomay, 2007). Some pronoun clusters are indicative of a university speaker's involvement in formulating their ideas 'online'. Items such as *I mean* and *I think* are cognitive discourse markers used to signal a delay, hold a turn, search for a word or indicate a commitment to continue speaking (Fung and Carter, 2007). The repetition of initial words, such as *I I,* provides a means of dealing with the time limited response expected in the flow of speech (Clark and Wasow, 1998). Therefore, such repetition phenomena can afford an insight into the processes of speech production. It might be suggested that pronoun repetition in learner discourse may be a part of a planning strategy (e.g. Clark and Wasow, 1998), or it may be used to fulfil a regulatory function in terms of positioning and focusing the speaker to the task at hand (Frawley, 1997). Indeed, Gilquin (2008) demonstrates that the repetitions *I I* and *you you* are more frequent in learner discourse than native speaker discourse.

Current research using non-native speaker corpora seeks to redefine "fluency deficiencies as organizational tactics" (Ruhlemann, 2007; 161). The implications of this type of research for pedagogical practices lie in illustrating the important role that cognitive discourse markers play as speech management tools and the importance of teaching discourse

markers in typical turn positions, and highlighting their role as carriers of stance (Aijmer, 2011).

For students, participation in small group teaching contexts is a complex activity which is not just informational. It is not enough to know the right content and exchange an opinion. The skill of knowing how to present their perspectives in stance-taking acts which engage with and anticipate the responses of others, subject to real-time constraints, is vital. Subsequently, it may be useful to examine the ways in which *you* and *I* cluster with other items.

3. The study

This paper uses data from two specialized corpora: a corpus of university classroom talk (UNITALK) and a corpus of English language learner classroom task talk (ELLTTALK) to understand the ways in which the personal pronouns *you* and *I* in university classrooms and language learning classrooms are used. Data from both corpora were collected from the same university setting, but at different periods of time. There is no overlap in individual speakers, i.e. the same speaker does not appear in both corpora. In the collection of UNITALK, speakers (tutors or students) were not excluded on the basis of first language spoken or on the basis of EAP student status. Of the speakers in the UNITALK corpus, two are non-native speakers of English; one student and one tutor.

3.1 UNITALK

UNITALK is a modest-sized untagged synchronic specialized full-text corpus of spoken academic discourse collected from fifteen university classrooms across disciplines. UNITALK was designed to study the genre of small group teaching contexts across the range of academic divisions and subject disciplines, specifically those teaching events whose goal is to work on collaborative ideas or tasks through speaking (Author, 2010). As a full text corpus UNITALK is not designed in terms of the number of words but in terms of full recordings from particular speech events, such as university seminars, tutorials, workshops. A principled approach to corpus design was followed (Adolphs and Knight, 2010) to ensure that the contents included in the corpus were selected according to the communicative function language fulfils in a particular community.

In comparison to other spoken academic corpora UNITALK includes only small group speech events, defined as teaching events with less than 12 participants, and is just over one third of the comparable sections of BASE (Thompson and Nesi, 2001) and one third the size of comparable sections of MICASE (Simpson *et al.*, 2002). Although modestly small, UNITALK is comparable in size to other specialized corpora which focus on one type of speech event (e.g. Camiciottoli, 2008). Working with smaller specialized corpora allows both quantitative and qualitative analyses of patterns of language use in particular settings (e.g. Farr and O'Keeffe, 2002; Koester, 2006; Vaughan, 2008). The breakdown of the UNITALK corpus by academic division, teaching context, speakers and words is given below in Figure 1:

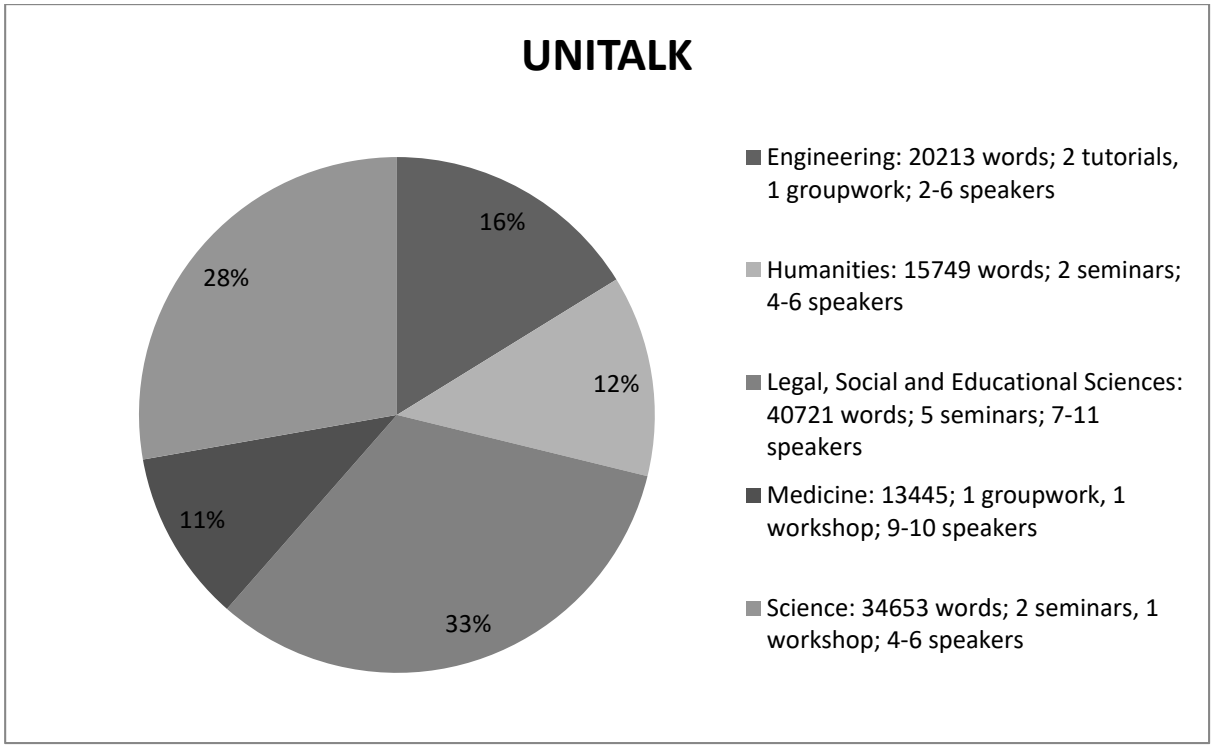

Figure 1. The composition of UNITALK by number of disciplines, texts and words

UNITALK speakers range in age from 18 to49. The activities in which speakers engage include discussions and group tasks. Topics include Irish politics, bridge and building design, contemporary composers, and learning styles. Full details on the activity, level of study, and topics of all UNITALK texts are provided in Appendix 1.

3.2 ELLTTALK

ELLTTALK is a small specialized corpus. The design of ELLTTALK was driven by a focus on speaking skills and a focus on collaborative speaking tasks. The eighteen texts which make up the corpus are recordings of speaking tasks carried out by adult English language learners in a classroom context.

Texts in ELLTTALK-General are taken from upper-intermediate to advanced level teaching contexts with a focus on speaking skills, conducted in a university setting. Speakers range in age from 19 to 45 years old and first languages spoken include Asian and European languages. Tasks are group discussions which are either convergent, i.e. speakers must come to some agreement on a topic, or divergent, i.e. students discuss a topic but are not required to reach agreement to complete the task. Texts in ELLTTALK-EAP are taken from upper-intermediate to advanced level university English language classrooms which have an explicit academic focus on seminar skills for undergraduate and postgraduate students. Speakers range in age from 19 to 40 years old and first languages spoken include Asian and European languages. Tasks are group discussions, which are either convergent or divergent.

In both ELLTTALK contexts a teacher is present during the tasks, but does not participate in the tasks unless requested by the speakers. The tasks take place as part of the context of a classroom lesson. In this respect, ELLTTALK is not a full-text corpus as only the student discussion tasks, not the full classroom lesson in which they occur, are included. Full details of the activities and topics of all ELLTTALK texts are provided in Appendix 2.

This corpus is dissimilar to learner corpora designed under criteria which control variables such as age, proficiency, and first language spoken (Granger, 2002). Although this information is available in relation to the speakers and a range of L1 backgrounds are represented in the corpus, it is not a feature of the corpus design and is not investigated as a variable here. However, it is important to note that corpus studies of learner discourse which examine first language as a variable suggest that, although features of learner discourse may vary according to first language spoken, there are also a number of features which are shared by a considerable number of learners. Those that are shared are understood to be developmental features (Gilquin et al., 2007).

The breakdown of the ELLTTALK corpus is given below in Figure 2:

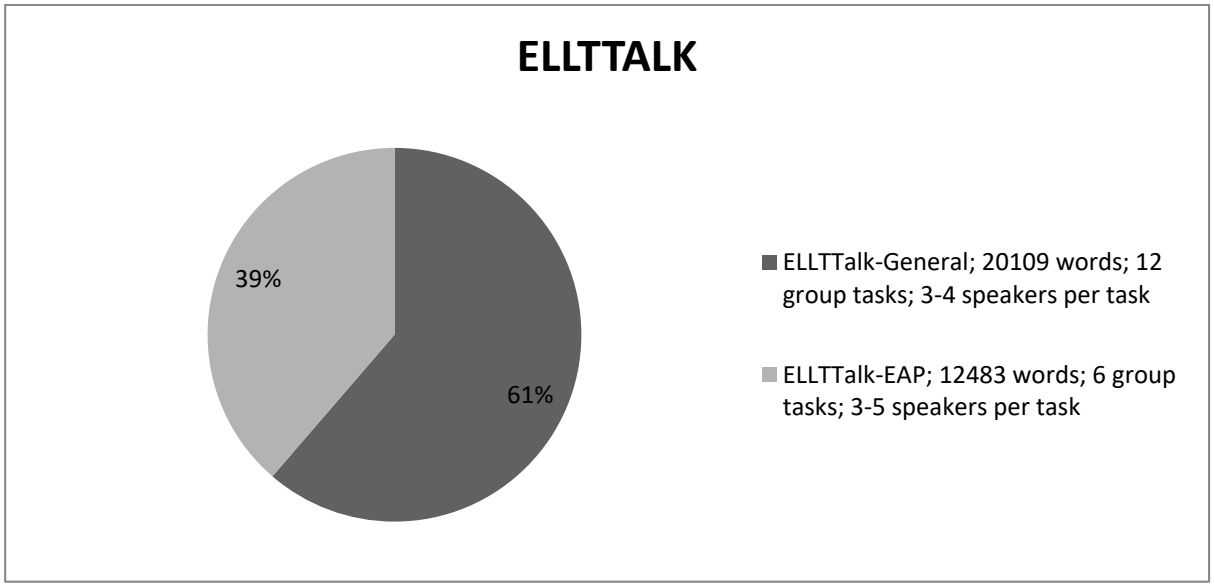

Figure 2. The composition of ELLTTALK by number of words, tasks, and speakers

3.3 Comparing corpora

As Kilgariff (2001) notes, identifying corpus similarity is complex. The purpose of this investigation is to describe the use of *you* and *I* in a sample of English language classrooms and to highlight any differences in observed use between such data and a corpus of university classrooms. Therefore, there are a number of dimensions which need to be taken into consideration when seeking to highlight different patterns of language use between the two corpora.

As illustrated in Fig. 1 and 2, the two corpora are not comparable in terms of size. In order to provide a descriptive comparison of the frequency of items and clusters across the two different sized corpora, normalized frequencies need to be calculated. Described by Evison (2010), this is obtained by dividing the raw frequency count of the item by the total word count for the corpus and multiplying by one thousand. This normalized frequency yields a rate of occurrence per 1000 words which can be used for comparative purposes. However, as specialized corpora, both are small-to-modest in size and, therefore, it is necessary to be cautious about overstating the basis (corpus size) on which the frequencies are made and to restrict analysis of normalized frequency patterns to descriptive rather than inferential statistics.

UNITALK and ELLTTALK both consist of multi-party talk in a classroom context and these two issues will be considered separately First, in relation to speaker roles in multi-

party talk, both corpora identify participant roles as 'student' and 'tutor'. However, ELLTTALK data consists of student discussion tasks. Although taken from within a classroom context where the tutor is drawn into the discussion at times by students (accounting for 10% of the corpus word count), the texts are largely comprised of student discourse. Therefore, any interpretations of observed trends must take into account that, although arising from a classroom context, ELLTTALK is primarily a corpus of learner data. By contrast, UNITALK classroom data is shared between those in student and tutor roles (Author, 2010). Second, in relation to the classroom context shared by both corpora, the purpose of both classroom genres is to achieve collaborative goals through talk. University seminars are designed to exchange opinions and evaluate knowledge through face-to-face interaction and speaking skills classrooms are designed to prepare and practice presenting and exchanging information through speaking. An examination of the types of discussion tasks demonstrates similarities. For example, in ELLTTALK, texts 07-09, the task is to make a decision on a hypothetical employee based on information presented. In UNITALK, text 03, the task is to decide which design is the most appropriate based on information previously presented. Both tasks require group discussion (presenting and exchanging information, responding, agreeing or disagreeing) and the convergence to one agreed outcome. Clearly, however, the focus is different, i.e. subject knowledge *vs.* language skills, and the type of knowledge being operated upon is different. Subject-specific knowledge is required in the engineers' discussion, whereas the knowledge being discussed in the ELLTTALK task may be drawn primarily from the speakers' social knowledge and personal opinion. It may be argued that the retrieval of disciplinary knowledge may be different from that of social knowledge, perhaps because it is less well-known or understood and social knowledge has been produced and reproduced many times. Although UNITALK speakers are already familiar with the content of the seminars from the associated lectures and received the topic and tasks or questions of seminars beforehand, ELLTTALK speakers are more often unaware of the topic and tasks prior to their participation.

The age ranges of participants are similar and both corpora are collected from the same institutional setting. The corpora most obviously differ in relation to the language variety they represent: ELLTTALK is a corpus of L2 speech, whereas UNITALK is primarily a corpus of native-speaker speech. It is this dimension which is the key locus of comparison for this paper. Ultimately, the rationale for such a comparison is to investigate the use of *you* and *I* in two spoken corpora, one of which might be termed a learner corpus and the other a native speaker corpus, in order to influence the achievement of language learning goals and create more effective participation in university classrooms for EAP students.

3.4 Data analysis

The recordings in both UNITALK and ELLTTALK were transcribed orthographically including repetitions (*I I I*) and fillers (*uh*, *um*). All texts were formatted for use by the software package Wordsmith Tools (5.0) (Scott, 2008). The Wordlist and Concordance features of Wordsmith Tools (Scott, 2008) were used to generate (i) the most frequently occurring words and (ii) two to five word *I*-clusters and *you*-clusters. The comparison of frequency lists can be used to indicate similarity or differences between established patterns (Kilgariff, 2001; Evison, 2010). The frequencies of *you* and *I*, and their associated clusters, are presented and compared within and between corpora using normalized frequencies. In order to avoid any inflation due to repetitions, which are discussed further in section 4.3, items such as *you you* and *I I* were removed from frequency scores for both corpora.

As data has the potential to be better understood through both quantitative and qualitative analyses (Conrad, 2002; Nesi, 2011) all concordance lines of pronouns and

pronoun clusters were examined qualitatively to identify repetitions of pronouns and to examine how they were being used.

4. Results and Discussion

4.1 Word lists and personal pronouns in top ten

The top ten most frequently occurring words in the corpora are presented in Table 1. The high ranking of the pronouns *you* and *I* across both corpora are indicative of the interactive nature of the face-to-face educational contexts from which the texts were gathered (Fortanet, 2004; Carter and McCarthy, 2006). Although spoken academic discourse is thought to be highly informational, *I* and *you* mark the conversational aspects of spoken academic discourse (Biber et al., 2002).). As evidenced in Table 1, there is frequent use of the hesitation marker '*uh*'. Found in speech to perform an anticipatory function and to indicate a commitment to future discourse (Clark and Fox Tree, 2002), this item occurs in the most frequent rankings of UNITALK (student only data) and ELLTTALK, and is not commonly used by university tutors. This is illustrative of the online constraints of speech production similarly experienced by students navigating either discipline-specific knowledge or language knowledge in these complex communicative events.

	UNITALK			**UNITALK student only**			**ELLTTALK**			**ELLTTALK-General**			**ELLTTALK-EAP**		
Rank	*Word*	*Raw Frequency*	*N Frequency per 1000 words*	*Word*	*Raw Frequency*	*N Frequency per 1000 words*	*Word*	*Raw Frequency*	*N Frequency per 1000 words*	*Word*	*Raw Frequency*	*N Frequency per 1000 words*	*Word*	*Raw Frequency*	*N Frequency per 1000 words*
1	The	4127	33	The	1243	37	The	1247	38	The	854	42	The	393	31
2	You	3709	30	It	787	23	You	821	25	You	525	26	You	296	24
3	And	2627	21	And	754	22	I	697	21	To	495	25	I	240	19
4	It	2585	21	You	735	22	To	677	20	I	457	23	To	182	15
5	That	2555	20	To	627	18	And	494	15	And	356	18	Is	172	14
6	To	2547	20	Of	619	18	Uh	455	14	A	329	16	Yeah	170	14
7	Of	2167	17	That	605	18	is	445	14	They	300	15	Uh	158	13
8	A	1798	14	I	556	16	Yeah	441	14	Uh	297	15	That	154	12
9	I	1666	13	Uh	519	15	A	439	13	That	281	14	And	138	11
10	Is	1587	13	A	472	14	That	435	13	But	276	14	Of	137	11

Table 1.Top ten most frequent words in UNITALK and ELLTALK including Student only talk in UNITALK and sub-corpora in ELLTTALK

Of further note is the higher ranking of *I* (3rd) in ELLTTALK than UNITALK (8th and 9th). This higher position of *I* may appear more similar to that of a ranking in a corpus of conversation (e.g. CIC, in which *I* is ranked 2nd, Carter and McCarthy, 2006). However, a closer inspection of normalized frequency lists and the use of *you* and *I* in context may indicate a more complex picture. Table 1 shows similar rankings of *you* and *I* in the top ten word frequency lists from ELLTTALK by sub-corpora. Eight out of ten items are ranked in the top ten of both sub-corpora, which shows a degree of similarity between the General English and the EAP contexts in ELLTTALK.

4.2 Pronoun distribution: you *and* I

Table 1 also shows that *you* is more frequent than *I* in both corpora. This finding from the UNITALK data supports previous research on the frequency of *you* and *I* found in American spoken academic corpora (Fortanet, 2004). *You* is understood to perform a range of discourse functions in academic contexts, from expressing attitudes and organizing discourse to drawing or distancing discourse participants (Fortanet, 2006). These discourse functions are discussed further in relation to *you*-clusters.

In relation to both *you* and *I*, UNITALK and ELLTTALK display significant statistical differences. Raw frequency counts were compared and used to calculate log-likelihood statistics. UNITALK has a significantly higher frequency of *you* than ELLTTALK (log-likelihood value= 19.06; $p<0.0001$). As noted earlier, the two corpora are different in relation to the amount of teacher talk included. The speaking tasks in ELLTTALK are student-led, however, in UNITALK discussions and tasks are tutor-led. In order to address this difference, frequency scores and rankings were calculated for student only data in UNITALK (see Table 1). The frequency scores of both *you* and *I* in UNITALK-student and UNITALK-tutor are compared with the ELLTTALK data. Figure 3 below shows that a larger proportion of *you* in UNITALK is attributable to tutors than to students. However, in ELLTTALK, where tutor discourse is minimal, the normalized frequency of *you* is significantly higher than in the student only data in UNITALK (log-likelihood value= 8.78; $p<0.01$). This shows that, although UNITALK tutors use *you* more frequently overall ELLTTALK speakers use *you* more frequently than UNITALK students.

To account for the greater frequency of *you* in ELLTTALK compared with the UNITALK-student only data, it could be suggested that ELLTTALK students may be employing discourse management functions of *you* within their student-led discussions in a manner similar to those employed by tutors in UNITALK. Alternatively, ELLTTALK speakers may be employing different uses of *you* to those displayed by both UNITALK teachers and students. To explore this frequency finding in depth, a comparison of the most frequent *you*-clusters is presented in section 4.3.

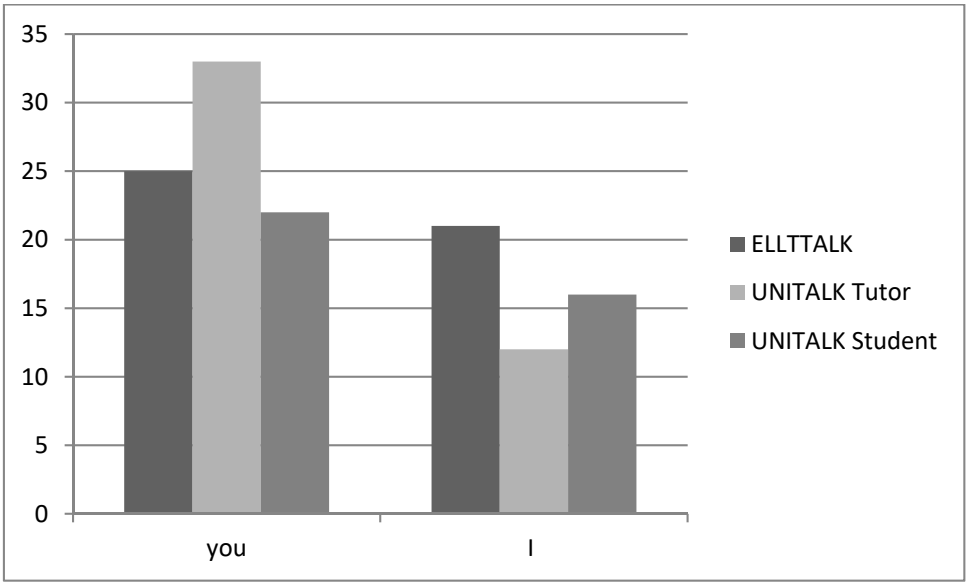

Figure 3. Graph showing the normalized frequency of *you* and *I* in ELLTTALK and UNITALK (student and tutor) per 1000 words

In relation to *I*, the comparison of the frequencies in the two corpora demonstrate that *I* is not just ranked higher in ELLTALK (Table 1) but occurs over 60% more often than in UNITALK.ELLTTALK has a significantly higher frequency of *I* than UNITALK (log-

likelihood value= 60.67; p<0.000001) (Rayson and Garside, 2000; Scott, 2008). Further comparison with the student only data in UNITALK shows that UNITALK students use *I* more than their tutors, but ELLTTALK speakers use *I* more frequently than either group of UNITALK speakers. In seeking to understand this difference it is important to recall the distinctions made in 3.3 between the speech events of the two corpora. The purpose of both UNITALK and ELLTTALK classroom genres is to achieve collaborative goals through talk. However, the corpora differ in relation to their respective focus on subject specific knowledge or language skills and on how aware or prepared speakers may be in advance of the topics under discussion. In an ELT context, speakers may rely more on a personal perspective to engage with content, whereas in university classrooms speakers may combine a personal perspective with reference to a disciplinary perspective gained though familiarity with lectures and readings. The ELLTTALK data clearly shows a preference for *I*, which has the effect of positioning the speaker at the centre of the informational space (Kamio, 2001) in face-to-face communication.

Previous studies demonstrated an overuse of *I* in student writing (e.g. Hyland and Milton, 1997). However, the present study finds that there is a statistically higher use of *I* in spoken learner discourse. Although the observed overuse of pronouns in learner written discourse has been explained as the erroneous transfer of what is known about the use of pronouns in speaking to writing (Gilquin et al., 2007) thereby creating an oral tone to learner writing, the present study shows that learners are already using *I* more frequently than native speakers in a spoken context.

In relation to *I*, it may be the case that learners are using *I* more than native speakers as a component of stance expressions, or it may be the case that *I* as a component in a category of discourse markers (Fung and Carter, 2006) are more prevalent as a cognitive tool in learner discourse than native speaker discourse. It is therefore necessary to examine cluster data.

4.3 Pronoun clusters

Using the personal pronouns *you* and *I* as search words in the Concordance feature of Wordsmith Tools, 2-word, 3-word, 4-word, and 5-word clusters were produced. Repetitions have been included for the analysis of all cluster data. The normalized rate per 1000 words is given in the figures and tables below. Of all the instances of *I*, 90% are accounted for in this cluster examination.

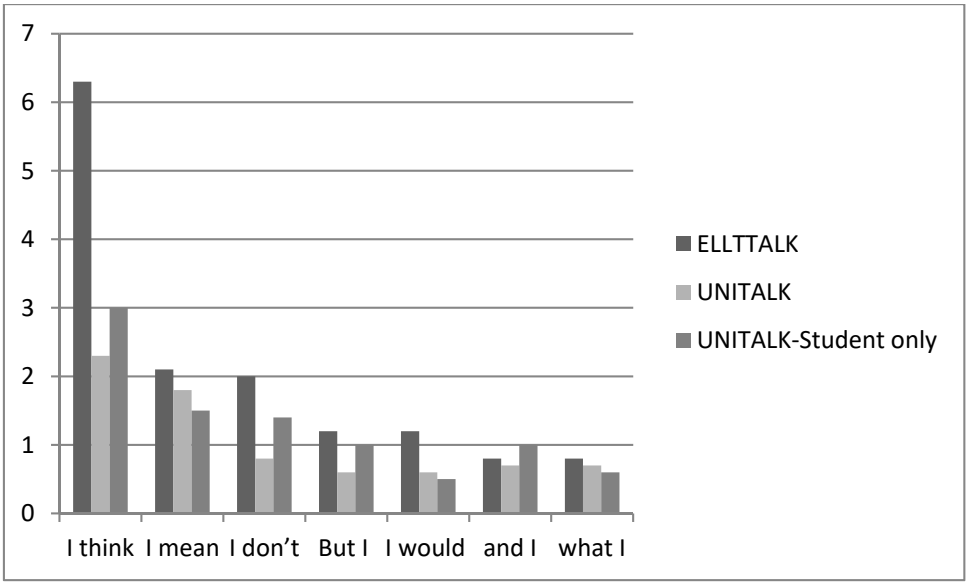

Figure 4. Graph showing the most frequent 2-word *I* clusters occurring in both ELLTTALK and UNITALK (including UNITALK student only for comparison) (per 1000 words)

Figure 4 shows that the most frequent 2-word cluster in UNITALK and ELLTTALK is the stance marker *I think.* It is unsurprising to find this stance marker in corpora of classroom discussions as this item establishes a speaker's stance or attitude and displays a certain orientation towards a proposition when individuals share a joint orientation or are seeking to establish something together (Karkkainen, 2003; Fung and Carter, 2007). Examples of *I think* from UNITALK and ELLTTALK are given below:

(1a) Student: actually **I think** the government are getting the money at the at the expense … of people's health **I think** it's just too bad … that's what **I think** (ELLTTALK)
(1b) Teacher: That's a good article…keep it **I think** it's very good (UNITALK-tutor)
(1c) Student: **I think** it's coming from the networks (UNITALK-student)

Although the verb *think* is polysemous (Aijmer, 1996), the semantic meaning of *I think* is in relation to degrees of certainty.

(2a) Student: uh **I think** it is because uh girls have you social social view it is very good I can't I can't think like you (ELLTTALK)
(2b) Student: **I think** maybe I'm wrong but it's sort of like they're not afraid of him (UNITALK)

As Karkkainen (2003) notes, in turn initial position, *I think* can guide participants' reaction and help signal that recipients are to align (or disalign) themselves to the upcoming discourse:

(3a) Student 1: I think most of the main parties just pay lip service to, like the only party in the down in the south that stands on the constitutional issue you know is Sinn Fein
Student 2: **I think** you had a good point there like (UNITALK)
(3b) Student 1: **I think** the purpose of the censorship is to …constitute some kind of...national…national…or collective… [S2: Hmm] consciousness [pause 4 seconds]
Student 2: Yeah I agree with the word collective and national but consciousness is … [S1: I don't know] it's quite vague (ELLTTALK)

The normalized rate of occurrence of *I think* shows a frequency in ELLTTALK which is almost three times that of the native speaker corpus. It may be the case that in ELLTTALK, *I think* takes the place of other or a variety of discourse markers. For example, at turn initial points *I think* may be used to ground the speaker's utterance in their own informational space first, rather than using other discourse markers which would connect with the previous utterances of others or relate to what has gone before. It may also be the case that learners are using one particular stance expression repeatedly.

A further difference between 2-word *I*-clusters in UNITALK and ELLTTALK is the frequency of pronoun repetition. Table 2 shows the top ten 2-word clusters with *I* which are not shared by both corpora. Figures in parentheses indicate the frequency of the item in the other corpus.

	ELLTTALK		UNITALK		UNITALK-Student Only
I I	4.2 (UNI: 0.5)	I'm	1.4 (ELLT: 0.5)	When I	0.9 (ELLT: 0.1)
No I	0.9 (UNI: 0.2)	I'll	1.0 (ELLT: 0.2)	I just	0.8 (ELLT: 0.2)
If I	0.7 (UNI: 0.4)	I was	0.7 (ELLT: 0.4)	Yeah I	0.8 (ELLT: (0.3)

Table 2. 2-word *I*-clusters which are not shared in the top ten lists of 2-word clusters

First person pronoun repetition is evident in both corpora. However, the normalized rate of occurrence per 1,000 words is 4.2 in ELLTTALK and 0.5 in UNITALK, which is more than eight times more frequent. This evidences the finding of Gilquin (2008) that pronoun repetition is more frequent in language learner discourse, but with data from a range of learners with different L1 backgrounds.

The concordance lines of these repetitions were examined in order to understand the immediate context of their use. Examples from UNITALK and ELLTTALK are presented below. In UNITALK, *I I* is observed next to a change of topic, hesitation marker, and phrases indicating disagreement.

Change topic

(4a) Teacher: but it'll help you to create a strong support if you use it to increase that and **I I** know I said that we were going to put some numbers to and uh I thought about doing it this morning I think I said I was going to do it but I thought well you've got this design to do so we'll start Monday's lecture looking at it

Uncertainty

(4b) Teacher: Yes I know it sounds very English uh **I I** haven't really heard of him I'm not really an expert on this but I guess it must be Korean

(4c) Student: and I when **I I I** uh I don't I'm sure those things have some significance you know

Pre-disagree

(4d) Teacher: Right well I understand that **I I** can see that point but creating something vertical doesn't mean you create blank walls
Student: Uhuh

In the student only component of UNITALK, only one *I* repetition occurs. However, in ELLTTALK this pronoun repetition occurs in 10 of the 18 texts of ELLTTALK and uttered by 13 different speakers. In the examples below, *I I* occurs at points of overlap, pauses, hesitations and attempts to seize a turn.

Repair-self and other

(5a) Student: **I I** didn't mean not to say what do you mean what do you think about it

Overlap/attempt to seize a turn

(5b) Student 1: She doesn't she doesn't insist the uh youngest one
Student 2: Yeah but **I I [**uh what I say is is to
Student 1: [to share the the rest of of of the sorry of the ice cream

Hesitations/pauses

(5c) Student: Me **I I** prefer I would say uh you …you go to some restaurants they accept uh accept … they accept uh … money or something else

In contrast to the amount of top ten 2-word *I*-clusters shared by UNITALK and ELLTTALK, only four 3-word *I*-clusters occur in both ELLTTALK and UNITALK. These are presented in Figure 5 and Table 3.

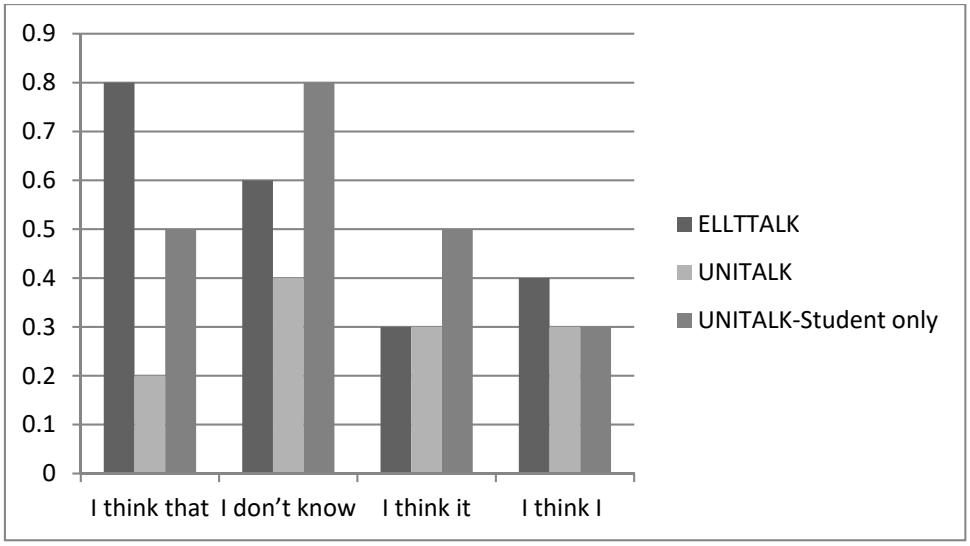

Figure 5. 3-word *I*-clusters common to ELLTTALK and UNITALK top ten clusters (per 1000)

It appears that as the size of *I*-cluster increases the degree to which items are common to the top ten clusters in both corpora decreases. This indicates that ELLTTALK speakers are clustering words differently to UNITALK speakers or are using them with different frequencies.

	ELLTTALK		**UNITALK**	**UNITALK-Student only**	
I I I	1.5 (UNI: 0.1)	I mean I	0.4 (ELLT: 0.05)	I'm just	0.3 (ELLT: 0)
I don't think	0.6 (UNI: 0.1)	I'm not	0.3 (ELLT: 0.2)	I'm not*	0.3
I I think	0.5 (UNI: 0.0)	I'm going	0.2 (ELLT: 0)	I'm going*	0.3
But I think	0.5 (UNI: 0.1)	I think you	0.2 (ELLT: 0.05)	When I was	0.3 (ELLT: 0.06)
I think the	0.4 (UNI: 0.1)	think I think	0.2 (ELLT: 0.02)	I when I	0.2 (ELLT: 0)
I would say	0.3 (UNI: 0.08)	what I mean	0.2 (ELLT: 0.2)	what I mean*	0.2

Table 3. 3-word *I*-clusters which are not shared in the top ten lists of 3-word clusters (* shared with UNITALK top ten)

As shown in Tables 3, 4 and 5, pronoun repetition is a more frequent feature of ELLTTALK than UNITALK. Although word repetition is often considered to be an error or signal of disfluency, Clark and Wasow (1998) demonstrate the degree to which it assists speakers to signal their commitment to ongoing discourse and overcome planning difficulties. When under pressure to speak, ELLTALK speakers, more so that UNITALK speakers, may

prematurely commit to an utterance with first word *I*, but are not yet ready to complete what comes next. Repeating the initial item therefore becomes a means of bridging the hiatus until continuity has been restored and what comes next has been formulated (Clark and Wasow, 1998). Such repetitions seem to be indicative of a need to hold a psychological predicate "in mind" as learners attempt to gain regulation over tasks and themselves in situations which are highly cognitively loaded (Frawley and Lantolf, 1985; McCafferty, 1992).

This difference in the frequency of pronoun repetition in ELLTTALK and UNITALK may be indicative of the differences in the management of spontaneous speech, such as planning (e.g. Clark and Wasow, 1998) and the strategies or processes of self-regulation (Frawley, 1997) employed by learners and native speakers. Again, it can be noted that as the size of the cluster increases, the similarity between the most frequent clusters in both corpora decreases. There are no 4-word clusters which are shared in the top ten lists of both corpora. There is only one 5-word cluster (*you know what I mean)* which is shared in both corpora. This item has a highly interpersonal function which is used to check listener understanding and to indicate a shared speaker-hearer position (Carter and McCarthy, 2006). The other five and four word clusters in ELLTTALK which occur with the same or greater frequency do not signal such an interpersonal function and are largely made up of repeated words.

	ELLTTALK		**UNITALK**	**UNITALK-Student only**	
I I I I	0.2 (0)	I think I think	0.08 (0)	You know what I*	0.1
I I I think	0.2 (0)	You know what I	0.08 (0.09)	Know what I mean*	0.1
This I I I	0.2 (0)	Know what I mean	0.07 (0.09)		

Table 4. 4-word *I*-clusters which are not shared in the top 4-word clusters (per 1000). (* shared with UNITALK)

	ELLTTALK		**UNITALK**	**UNITALK- Student only**	
This I I I I	0.09 (0)	You know what I mean	0.07 (0.09)	You know what I mean*	0.1
Yes uh uh I I	0.09 (0)	Do you know what I	0.05		
You know what I mean	0.09 (0.07)	Do you see what I	0.05		
Yeah yeah yeah I I	0.09 (0)				
No no no I I	0.09 (0)				

Table 5. 5-word *I*-clusters which are not shared in the top 5-word clusters (per 1000). (* shared with UNITALK top ten)

As noted by Gilquin (2008), learner discourse underuses interpersonal discourse markers such as *you know*, which are used to maintain the flow of speech in native speaker discourse. The 2-word cluster *you know* appears with greater frequency in UNITALK than in ELLTTALK. The comparison between 2-word *you*-clusters in ELLTTALK and UNITALK are shown below in Figure 6 and Table 6. 90% of instances of *you* are accounted for in this cluster examination.

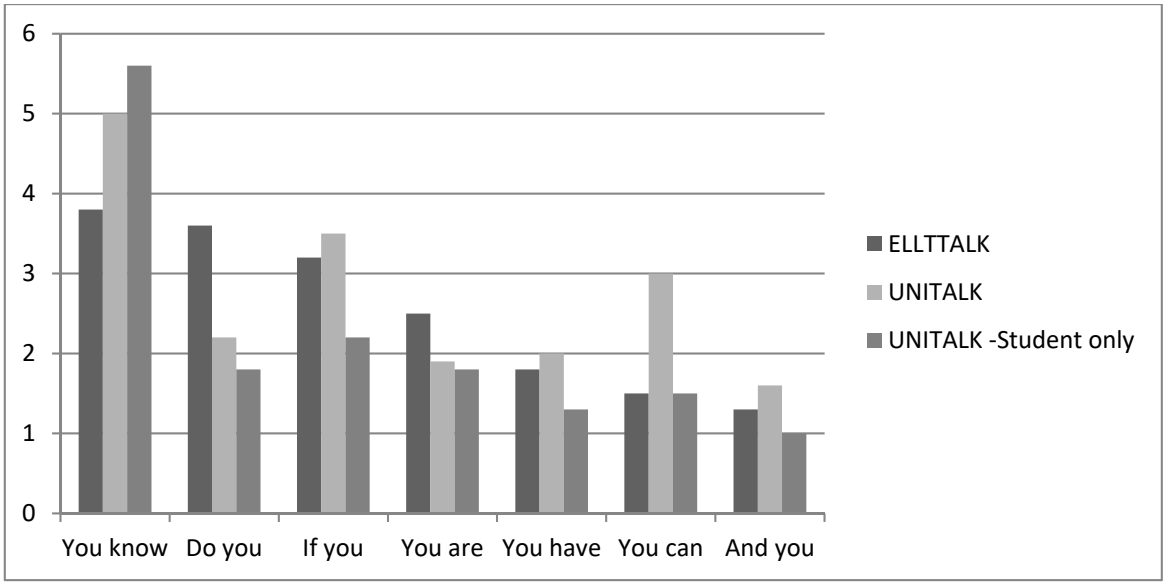

Figure 6. 2-word *you*-clusters common to ELLTTALK and UNITALK top ten clusters (per 1000)

	ELLTALK		UNITALK		UNITALK-Student only
you you	3.6 (0.8)	so you	1.7 (0.3)	so you*	0.8 (ELLT:0.3)
you think	2.5 (0.8)	that you	1.4 (0.5)	you see	1 (ELLT: 0.09)
you will	1.1 (0.2)	what you	1.1 (0.3)	you think^	0.7

Table 6. 2-word *you*-clusters which are not shared in the top 2-word clusters (per 1000)

(* shared with UNITALK top ten; ^ shared with ELLTTALK top ten).

Again, pronoun repetition is evident in 2-word *you*-clusters in both corpora, with different normalized rates of occurrence per 1,000 words: 3.6 in ELLTTALK and 0.8 in UNITALK. This pronoun repetition in ELLTTALK is produced by 10 different speakers, in 7 out of 18 texts. Pronoun repetitions are found more often in learner discourse, whether with *I* or *you,* than in native speaker discourse. Therefore, pronoun repetition may be a feature which marks learner discourse as non-native like. Indeed, for the items which are not shared, *you* appears to cluster with connectives in UNITALK data, e.g. *so you, that you*. Pronoun repetitions could be taking the place of these items in learner discourse, with the implication that more fluent speakers use connectives rather than pronoun repetitions.

The concordance lines of *you you* were examined in order to understand the immediate context of their use. In addition to occurring with hesitations and false starts, *you you* in UNITALK is observed next to other repetitions and it seems to be used to hold the floor.

False start

(6a) Student: So I would advise probably that literature and bring it in that helps s= **you you** score marks

(6b) Student: He was caught on CCTV jumping on a fella's head like jumping and you just …think oh my God and **you you** really take a step take a step back and think

Instances of *you you* in ELLTTALK occur with signals of hesitation (*uh* and pauses), false starts or repair. They seem to signal the overall process of thinking in speaking and do not

appear in relation to other discourse marking features which would signal a flow in interaction (Gilquin, 2008).

Hesitation

(7a) Student: It means that uh **you you** someone tells you not to do that anymore perhaps uh **you you** will have more problems to evacuate this this suffering

(7b) Student: … because because **you you** mention education level

False start/repair

(7c) Student: And if **you you** think..uh uh ..**you you** haven't get no you haven't got justice

(7d) Student: **You you** will you will feel you will feel…uh how do I mean to say you will feel that society is not is not fair wi=for to you

Hold floor

(7e) Student: I think yeah **you you** in my opinion it's a fair it's a right because every people should be responsible

The pattern of pronoun repetition is also observed in the 3-word clusters with *you*. The normalized rate of occurrence per 1,000 words is 1.1 and 0.1 in ELLTTALK and UNITALK respectively (Table 7). Again as found with the *I* clusters, many fewer of the highest ranking 3-word clusters are shared in both corpora.

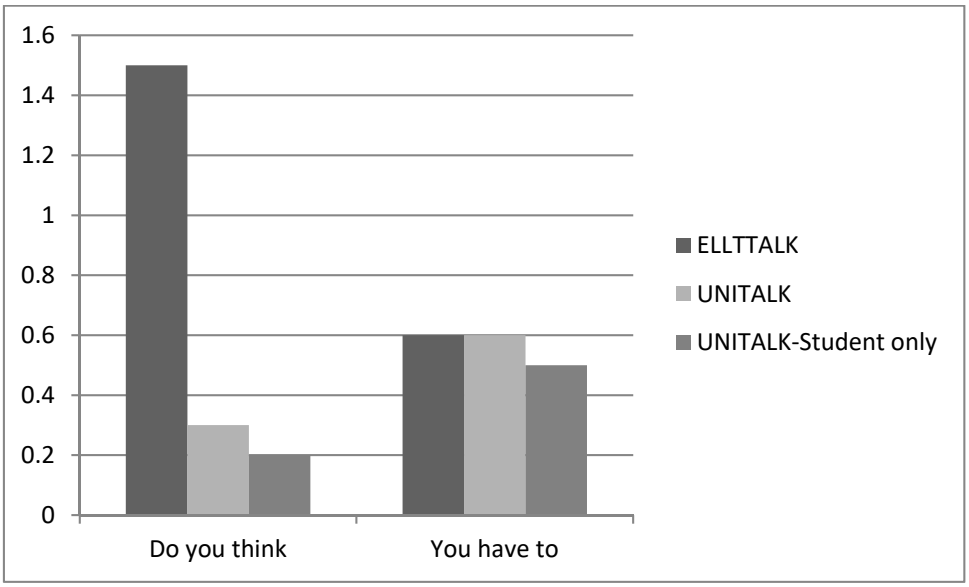

Figure 7. 3-word *you*-clusters common to ELLTTALK and UNITALK top ten clusters (per 1000) (including UNITALK student only for comparison)

	ELLTTALK		**UNITALK**		**UNITALK-Student only**
What do you	1.3(0.3)	You've got	0.8 (0)	You know the*	0.9
You you you	1.1 (0.1)	You want to	0.7 (0.1)	You look at*	0.4
If you are	0.6 (0.2)	You know you	0.5 (0.05)	You want to*	0.3
You think that	0.5 (0.07)	You need to	0.4 (0.02)	If you are^	0.3
You you will	0.5 (0)	You know the	0.4 (0.09)	Do you have	0.3 (ELLT: 0.09)
You if you	0.4 (0.2)	You can see	0.4 (0.09)	You are not	0.3 (ELLT: 0.2)
You think about	0.4 (0.06)	You look at	0.3 (0.02)	You can see	0.3 (ELLT: 0.09)

How do you	0.4 (0.2)	So if you	0.3 (0.07)	Do you want	0.3 (ELLT: 0.09)
				So you can	0.2 (ELLT: 0)

Table 7. 3-word *you*-clusters which are not shared in the top 3-word clusters (per 1000). (* shared with UNITALK top ten; ^ shared with ELLTTALK top ten).

As shown in Fig. 7, only two 3-word clusters are found in the top ten of both corpora: *you have to* and *do you think.* This shows, in combination with earlier findings, that as the size of the cluster increases similarities decrease between both corpora. This decrease in similarities is also borne out when ELLTALK is compared with the student only component of UNITALK (Figure 7 and Table 7). The differences in the frequency and type of *you*-clusters in ELLTALK and UNITALK show that ELLTTALK speakers are using *you* in a different manner than UNILTALK teachers and students. The larger clusters with *you* again show pronoun repetition (*you you you you)* in ELLTTALK absent from the UNITALK data presented in Figure 8 and Table 8.

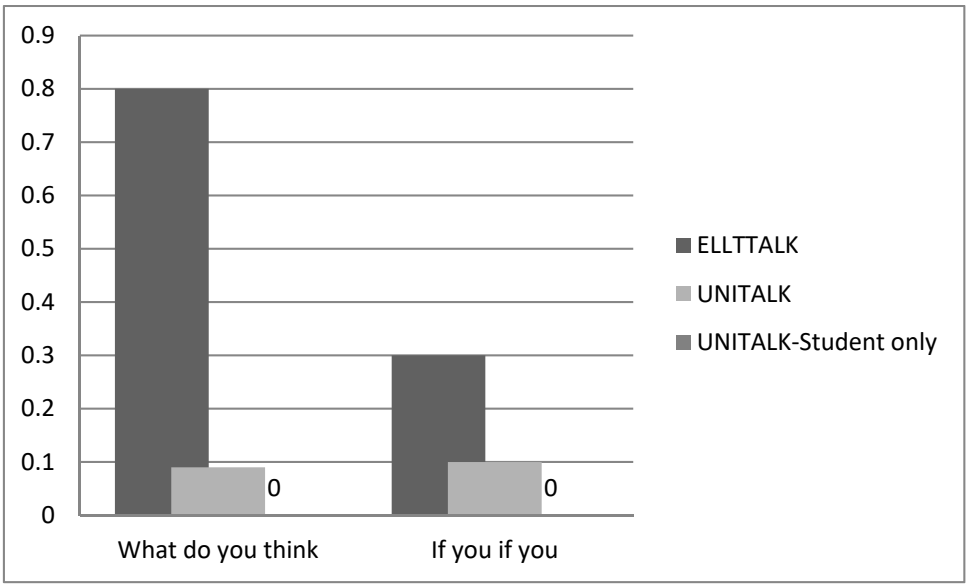

Figure 8. 4-word *you*-clusters common to ELLTTALK and UNITALK top ten clusters (per 1000) (including UNITALK student only for comparison)

	ELLTTALK		**UNITALK**		**UNITALK-Student only**
Do you think about	0.4 (.)	If you look at	0.2 (0.03)	If you look at*	0.2
Do you think that	0.3 (0.04)	As you can see	0.2 (0)	Do you have to	0.2 (ELLT: 0.03)
How do you say	0.2 (0)	If you want to	0.2 (0)		
Perhaps you you you	0.2 (0)	Do you want to	0.1 (0.09)		
Yeah you you you	0.2 (0)	So you have got	0.09 (0)		
You you you you	0.2 (0)	And as you can	0.09 (0)		
What do you mean	0.2 (.)	So you have to	0.07 (0)		
When you when you	0.2 (.)	If you have a	0.07 (0)		

Table 8. 4-word *you*-clusters which are not shared in the top 4-word clusters (per 1000) .=<0.0001(* shared with UNITALK top ten

	ELLTTALK
So what do you think	0.1 (.)
You what do you think	0.1 (0)
I think yeah you you	0.1 (0)
If you if you think	0.1 (0)

	UNITALK
So I think you you	0.04 (0)
I'd like you to	0.04 (0)
I'll tell you what	0.04 (0)

Table 9. 5-word *you*-clusters which are not shared in the top 5-word clusters (per 1000).=<0.0001

The comparison of the larger clusters with *you* indicates that, again, only two 4-word items are shared in ELLTALK and UNITALK. ELTTALK speakers use *what do you think* almost ten times more frequently than UNITALK teachers. The lexical verbs in the larger clusters in ELLTTALK appear cognitively focused (*think, mean*), further indicative of the thinking process during spoken interaction. In UNITALK, *you* clusters more with modal verbs. In the larger clusters with *you* in ELLTTALK shown in Table 9, question stems appear, e.g. *what do you think* and *how do you say* occurring at a rate of 0.2 per 1,000 words. These larger clusters in ELLTTALK seem to be focused on clarifying meaning or eliciting participation. However, in UNITALK the clusters with *you* used by tutors appear to signal obligation and direction (*I'd like you to*) and mark the introduction of a topic (*if you look at*) (Biber et al., 2004). As discussed in section 4.1, the data shows that it is not entirely the case that ELLTTALK speakers employ *you* clusters within their student-led discussions in a manner similar to those employed within the tutor-led discussions in UNITALK.
In relation to the student component of UNITALK as shown in Fig. 8, there is only one or fewer instances of the top four-word ELLTTALK *you* clusters in the student only component of UNITALK. For UNITALK students, the most frequent 4-word *you* clusters seem to be more similar to those of UNITALK teachers than ELLTALK speakers, showing some degree of obligation with the cluster: *do you have to* and the cluster common in teacher discourse *if you look at*. Overall, it is clear that ELLTTALK speakers use *you* differently to both UNITALK teachers and students. These results suggest that learner discourse displays more features tied to immediate constraints of online production of speech, self-regulation, and fewer features signalling joint attention and the conversational and informational space of others.

4.4 Summary of results and discussion

The analysis of *you* and *I* in these two corpora demonstrate that learners and native speakers use these pronouns with differing frequencies and that these differences may contribute to a perception of disfluency. These differences could be indicative of the suggestion that learners could have a psychological focus in language use which places them at the centre of informational space in interaction. By contrast, native speakers may not only point to their own informational space but they may relate and connect to that of others in interaction. Pronoun repetition may be a feature which marks learner discourse as non-native like. It may be the case that learners are repeating pronouns in place of the variety of discourse and stance markers used by native speakers. As cluster size increases similarity between the top ten most frequent clusters in both corpora decreases. This shows that ELLTTALK speakers are clustering words differently to UNITALK speakers or are using them with different frequencies.

5. Pedagogical Applications

Corpus studies have led to the development of EAP material designed to highlight the features of spoken and written discourse and genre and register variations, and to raise awareness of language in use (Gilquin *et al.*, 2007). Indeed, Pilcher (2009) finds that what students want are examples of successful and unsuccessful texts and a means of identifying what they need to do to produce appropriate and successful spoken and written discourse in a particular context.

The pedagogical applications of this research on *you* and *I* are concerned with a principled guided awareness of the functions of personal pronouns in spoken academic discourse. Data can be used to facilitate a language awareness approach to teaching and learning (Svalberg, 2007); a corpus provides a picture of language as a dynamic phenomenon which can be described and explored by both students and teachers; texts, features, and concordance lines of personal pronouns can be used in classrooms in order to discover and make explicit how a particular pronoun functions in relation to the array of possible pragmatic meanings; examples of real-life data can illustrate what is achieved through the successful (or unsuccessful) use of personal pronouns. Not only can the use of data and a language awareness approach be engaging for learners, but the depth of investigation may have a more lasting impact than just drawing a learner's attention to particular features alone (Thornbury, 2001).

This paper has discussed how pronoun repetition may be illustrative of the way in which thinking occurs in spoken interaction. While this may be a naturalistic feature of producing speech, it may have negative consequences when interpreted as a language learner's lack of control or disfluency. A text from ELLTTALK, marked by hesitations, false starts, pronoun repetition, and generic use of *you*, together with suggested activities of how they could be used in classrooms is given below:

Extract 8

S: if you you think… uh uh… you you haven't get no you haven't got justice you you will you will feel you will feel… mmm how do I mean to say… you will feel… that society is not is not fair wi- for to you because when… some somebody from your family or quite or very close related to you you you need to to… to be in some way recomfort for this uh loose or loss sorry loss and you you find that justice is not a very fair justice perhaps you you you you will have this sentiment to revenge

Students can be invited to examine ways in which they could make the text appear more fluent. In this respect, students 'clean up' real data themselves, and it could form one aspect of awareness raising training. Students could consider how they might reduce, omit or replace particular features to become more fluent (or less disfluent) in speaking contexts. This might be particularly relevant in preparation for speaking tests where such signals as hesitations, pauses and repetitions may be negatively perceived. Prior to any reformulation by students, the following types of questions could also be discussed:

- *Why do you think there are pauses?*
- *Why do you think there are lots of* you*?*
- *Who does* you *refer to?*
- *If you were listening to this speaker, what effect would the pauses and repetitions have on you?*

Using an extract from UNITALK, such as Extract 9, students could examine similar features, such as hesitations and self-repair, to notice that these are not just used by language learners and they are a feature of spontaneous spoken discourse.

Extract 9

T:	Yeah good so any other thoughts?

S:	Like the whole way through sort of it's been about **…** her father is a king so her identity should be from birth if **you know what I mean**
T:	Good
S:	She's very **…**she doesn't know who she is … and she's sort of waiting for **her for him** to tell her nearly
T:	Yeah
S:	**You know** who she is because she seems sort of rootless and **uh** even though it's sort of like symbolic because she's isolated on an island she's sort of isolated from herself
T:	Very good good yeah

By contrasting these examples, students' attention could be drawn to the absence of pronoun repetition and the presence of markers which the speaker uses to draw in the listener and connect with the informational space of other interactants. .

6. Conclusion

This investigation of the use of *you* and *I* in English Language Teaching classrooms shows that the speakers in the two corpora use pronouns differently. It appears that in language learner classroom discourse, repetitions of pronouns are used as hesitation devices and demonstrate something of the thinking process in which speakers are involved. It has been suggested that pronoun repetition is a means of regulating the thinking process for learners in the activity of speaking, allowing them a cognitive space (DiCamilla and Anton, 1997) to make future commitments to the discourse and to organize their thinking.

It may be the case that in native speaker discourse a variety of discourse markers occupies these functions (Gilquin, 2008). Items such as *I mean* and *you know* are linguistic features of online production constraints (Csomay, 2007), but they are multifunctional and can signal interpersonal and relational intentions. In effect, such discourse markers signal an attempt to reach a shared position, rather than a transactory exchange of information between speakers. Kesckes (2007) argues that NNS speakers do not search for shared common ground or knowledge and rely instead on repetitions, paraphrasing, and salient literal meaning (2007: 204). Kesckes (2007) suggests that the primacy of the linguistic code and literal meaning is more of a feature of NNS discourse than native speaker discourse, where the avoidance of formulaic language or conversational routines is evident. If this is indeed the case, then serious consideration needs to be given to the rationale and means of teaching the relational aspects of English language use explicitly in ELT classrooms. Further research could identify to what extent a lack of relational discourse is present in English Language Learner discourse together with further exploration of how learners express evaluation and intersubjectivity, so fundamental to communication in any language.

Hyland and Milton (1997) note that the greater frequency of *I* in L2 writing as compared with L1 writing decreases as proficiency increases. Further research examining the relationship between levels of proficiency, first language spoken and pronoun repetition may result in a clearer understanding of how learners regulate their speech and their language learning. Furthermore, investigations of targeted awareness-raising using oral communicative strategies to reduce or adapt hesitations, false starts and pronoun repetitions could contribute to this area. The data used in this study is taken from classroom contexts. Pronoun use could be further investigated in relation to teacher and student roles and the construction of discourse identity and knowledge positions and in non-formal contexts.

These differences between ELLTTALK and UNITALK show that learner discourse displays more features tied to immediate constraints of online production of speech, and

fewer features associated with attempting to point to the conversational and informational space of others. It may be that EAP teaching could include how learners can regulate their own speech in native-like ways and connect with the informational space of others, in order to appear more fluent. This could be achieved by highlighting the overuse of pronoun repetition, and by illustrating how it is possible to maintain their cognitive use, but including a native-like variety of discourse markers. Similarly, the over-reliance on one particular stance-marker can be highlighted and alternatives explored. Until such items are brought to the attention of learners, they are likely to remain and contribute to a sense of disfluency.

7. References

AUTHOR, (2010).

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O'Keeffe & M.J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*. Oxford: Routledge.

Aijmer, K., I think-an English modal particle, In:Swan, T. and Westvik, O.J., (Eds)., Modality in Germanic Languages: Historical and Comparative Perspectives. Mouton de Gruyter: Berlin

Aijmer, A. (2005). Evaluation and Pragmatic Markers. In E. Tognini-Bonelli & G. Camiciotti (Eds.), *Strategies in academic discourse*. Amsterdam: John Benjamins.

Aijmer, K. (2011). "Well I'm not sure I think…" The use of well by non-native speakers. Errors and Disfluencies in Spoken Corpora. *Special issue of International Journal of Corpus Linguistics* 16(2): 231-254.

Altenberg, B. (1990). Spoken English and the learner. In J. Svartvik (Ed.), *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund Studies in English 82. Lund University Press.

Artiga, M. (2006). The semantic-pragmatic interface of authorial presence in academic lecturing phraseology. *Iberica*, 12: 127-144.

Basturkmen, H. (2002). Learner observation of, and reflection on, spoken discourse: An approach for teaching academic speaking. *TESOL Journal* 11/2 26-30.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson Education

Biber, D., Conrad, S., Reppen, R., Byrd, P. & Helt, M. (2002). Speaking and writing in the university: A multi-dimensional comparison. *TESOL Quarterly*, 36, 9-48.

Biber, D., Conrad, S.& Cortes, V. (2004). 'If you look at...' Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25,3:371-405.

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In L.B. Taylor & P. Falvey (Eds.), *IELTS Collected papers: research in speaking and writing assessment*. Cambridge: Cambridge University Press.

Camiciottoli, B. (2008). Interaction in academic lectures vs. written text materials: the case of questions. *Journal of Pragmatics*, 40 (7), 1216-1231.

Carter, R.A. & McCarthy, M. J. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Chafe, W. (1994). *Discourse, consciousness and time*. Chicago: University of Chicago Press.

Clark, H. H. & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speech. *Cognition,* 84, 73-111.

Clark, H.H., Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, 37:201-242.

Council of Europe (2001) *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22, 75-95

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, 336–355.

DiCamilla, F. &Anton, M. (1997). Repetition in the collaborative discourse of L2 learners: A Vygotskian perspective. *Canadian Modern Language Review*, 53, 609-633.

Du Bois, J. (2007). The stance triangle. In R. Englebretson (Ed.), *Stancetaking in Discourse*. Amsterdam: John Benjamins.

Evison, J. (2008). *Turn-openers in academic talk: An exploration of discourse responsibility*. PhD Thesis, University of Nottingham.

Evison, J. (2010). What are the basics of analysing a corpus?. In A.O'Keeffe, M.J. McCarthy, (Eds.), *The Routledge Handbook of Corpus Linguistics*. Oxford: Routledge.

Farr, F. & O'Keeffe, A. (2002). Would as a hedging device in an Irish context: and intra-varietal comparison of institutionalised spoken interaction', in Reppen, R., Fitzmaurice, S. & Biber, D. (Eds) *Using corpora to explore linguistic variation*. Amsterdam: John Benjamins.

Fortanet, I. (2004).The use of *we* in university lectures: reference and function. *English for Specific Purposes*, 23, 45-66.

Fortanet, I. (2006). Interaction in academic spoken English: the use of 'I' and 'you' in the MICASE. In E. Arno Macia, A. Soler Cervera, & C. Rueda Ramos (Eds.), *Information Technology in Languages for Specific Purposes: Issues and Prospects*. New York: Springer

Fox Tree, J. E. & Schrock, J.C. (2002). Basic meanings of 'you know' and 'I mean'. *Journal of Pragmatics 34. 727–47.*

Frawley, W. & Lantolf, J. P. (1985). Second language discourse: a Vygotskian perspective. *Applied Linguistics*, 6, 19-44.

Frawley, W. (1997). *Vygotsky and Cognitive Science: Language and the Unification of the Social and Computational Mind*. Cambridge: Harvard University Press.

Fung, L.& Carter, R. (2007). Discourse Markers and Spoken English: Native and Learner Use in Pedagogic Settings. *Applied Linguistics 28/3: 410–439*

Gilquin, G. (2008). Hesitation markers among EFL learners. In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics: a mutalistic entente.* Berlin: Mouton de Gruyter.

Gilquin G. & Paquot M. (2007). Spoken features in learner academic writing: identification, explanation and solution. In *Proceedings of the Fourth Corpus Linguistics Conference, University of Birmingham, 27-30 July 2007*.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6: 319–335.

Goffman, E. (1981). *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching,* Amsterdam: Benjamins.

Granger, S., Hung, J. & Petch-Tyson , S. (2002). *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam & Philadelphia: John Benjamins.

Griffiths, S. (2009) Teaching and learning in small groups, in Fry, H., Ketteridge, S., Marshall, S. (Eds) *A handbook for teaching and learning in higher education* (2ed.) Oxon: Routledge.

Groom, N. (2005). Pattern and meaning across genres and disciplines: an exploratory study. *Journal of English for Academic Purposes 4/3*: 257-277

Harwood, N. (2005).'We Do Not Seem to Have a Theory . . . The Theory I Present Here Attempts to Fill This Gap': Inclusive and Exclusive Pronouns in Academic Writing. *Applied Linguistics* 26/3: 343–375.

Hinkel, E. (2004). *Teaching academic ESL writing: practical techniques in vocabulary and grammar*. Oxford: Routledge

Hyland, K. & Milton, J. (1997). Qualifications and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*, 183-205.

Hyland, K. (2001). Bringing in the Reader: Addressee Features in Academic Writing. *Written Communication* 18(4): 549–74.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics* 34: 1091–112.

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. *Discourse Studies* 7(2) 173-192.

Koester, A. (2006) *Investigating Workplace Discourse*. London, Routledge.

Kamio, A. (1997). *Territory of information.* Amsterdam: John Benjamins.

Kamio, A. (2001). English generic *we, you,* and *they:* An analysis in terms of territory of information. *Journal of Pragmatics 33 (2001) 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.

Kesckes, I. (2007). Formulaic language in English Lingua Franca. In I. Kesckes & L. Horn (Eds.), *Explorations in Pragmatics: Linguistic, Cognitive and Intercultural Aspects.* Berlin/New York: Mouton de Gruyter

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6 (1), 1-37.

Kim, S. (2006). Academic oral communication needs of East Asian international graduate students in non-science and non-engineering fields. *English for Specific Purposes,* 25: 479–489.

Lin, P. & Adolphs, S., (2009). Sound evidence: A multimodal corpus-based study into the notion of holistic processing of multiword units. *In A.* Barfield & H. Gyllstad (Ed.), *Collocating in Another Language: Multiple Interpretations* Palgrave Macmillan.
Luzon, M. (2009). The use of we in a learner corpus of reports written by EFL engineering students. *Journal of English for Academic Purposes*, *Volume 8, Issue 3*, *September 2009*, *Pages 192-206*
Martinez, I.A. (2005). Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. J*ournal of Second Language Writing*, *Volume 14, Issue 3*, *September 2005*, *Pages 174-190*
McCafferty, S. G. (1992). The use of private speech by adult second language learners: A cross-cultural study, *Modern Language Journal*, 76, 179-189.
McCarthy, M. J. (2010). Spoken Fluency Revisited. *English Profile Journal,* 1,1, 1-15.
Muñoz, C. (1991). Why are he and she a problem for Spanish learners of English?. *Revista Española de Lingüística Aplicada, 7: 129-136.*
Nakatani, Y. (2005). The Effects of Awareness-Raising Training on Oral Communication Strategy Use. *The Modern Language Journal, 89: 76-91.*
Nattinger, J. & DeCarrico, J. (1992). *Lexical Phrases in Language Teaching. Oxford: Oxford University Press.*
Nesi. H. (2011). Laughter in university lectures. Journal of English for Academic Purposes, 11,2,69-79
Pennycook, A. (1994). The politics of pronouns. *English Language Teaching Journal, 48,2: 173-178.*
Pilcher, N. (2009). What do EAP Students Think Make Good Materials. *Proceedings from the Different approaches to EAP workshop held at English Language Teaching University of St Andrews on 28th February 2009.*
Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In proceedings of the *workshop on Comparing Corpora held in conjunction with the 38th annual meeting of the Association for Computational Linguistics(ACL 2000)*. 1-8 October 2000, Hong Kong, pp. 1 - 6.
Rowley-Jolivet, E. & Carter-Thomas, S. (2005). Genre awareness and rhetorical appropriacy: manipulation of information structure by NS and NNS scientists in the international conference setting. *English for Specific Purposes, 24*(1), 41–64.
Ruhlemann, C. (2007). *Conversation in Context: a corpus driven approach.* London: Continuum.
Scott, M. (2008). *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.
Seidlhofer, B. (2005). English as a Lingua Franca. *English Language Teaching Journal, 59, 4, 339-341*.
Simpson, R. C., S. L. Briggs, J. Ovens, & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.
Sinclair, J. McH. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Stenström, A. (2004). What is going on between speakers?. In A. Partington, Morley, J. & Haarman, L. (Eds.). *Corpora and Discourse*. Bern: Peter Lang.
Svalberg, A. (2007). Language awareness and language learning. *Language Teaching, 40, 287–308.*
Tang, R. and S. John. 1999. 'The ''I'' in identity:Exploring writer identity in student academic writing through the first person pronoun,' English for Specific Purposes 18:23–39Thompson, P. and Nesi, H. (2001).The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research* 5 (3) 263-264.
Thornbury, S. (2001). *Uncovering grammar*. Oxford: Macmillan Heinemann.
Vaughan, E. (2008) 'Got a date or something?: an analysis of the role of humour and laughter in the workplace meetings of English language teachers' in Adel, A., &Reppen, R. (eds.) *Corpora and discourse: the challenges of different settings*. Amsterdam: John Benjamins.
Wortham, S.E.F. (1996) 'Mapping participant deictics: A technique for discovering speakers' footing', *Journal of Pragmatics,* 25, 331-348.

Appendix 1: UNITALK. Details of the activity and topic of all UNITALK texts

Text	Level of study	Subject	Topic	Activity	Academic Division
01	UG	Architecture	Building Design	Discussion of a design	Engineering
02	UG	Architecture	Building Design	Discussion of a design	Engineering
03	UG	Engineering	Building Design	Discussion and plan of a design	Engineering

04	UG	English	Literary Text	Discussion	Humanities
05	UG	Music	Contemporary Composers	Student presentations and discussion	Humanities
06	PG	Education	Learning styles and strategies	Discussion and tasks	LSES
07	PG	Education	Discourse and Education	Discussion and tasks	LSES
08	UG	Management	Motivation	Discussion	LSES
09	UG	Politics	Irish politics	Discussion	LSES
10	UG	Politics	Media	Student presentation and discussion	LSES
11	PG	Pharmacy	Development of Pharmacy	Discussion and tasks	Medicine
12	PG	Pharmacy	Practice of Pharmacy	Task and discussion	Medicine
13	UG	Chemistry	Molecular Chemistry	Task and discussion	Science
14	UG	Chemistry	Organic Chemistry	Task and discussion	Science
15	UG	Computers	Graphics	Task and discussion	Science

Appendix 2: ELLTTALK. Full details of the goal-type, activity, and topic of all ELLTTALK texts

Text	Subject	Topic	Activity
01-03	General English: Speaking Skills	Truth or Lies	Divergent discussion task
04-06	General English: Speaking Skills	Crime and Punishment: Life	Divergent discussion task
07-09	General English: Speaking Skills	Making decisions	Convergent task
10-12	General English: Speaking Skills	Making decisions	Convergent task
13	Academic English: Seminar Skills	Issue in student's discipline	Presentation and discussion
14	Academic English: Seminar Skills	Contemporary Issues: Capital Punishment	Divergent discussion
15	Academic English: Seminar Skills	Contemporary Issues: World Debt	Divergent discussion
16	Academic English: Seminar Skills	Contemporary Issues: Homelessness	Convergent discussion task

17-18	Academic English: Seminar Skills	Contemporary Issues: Censorship	Convergent discussion task