# Pedagogical Implications of the Corpus-based Investigation of Discourse Markers

Huang, Lan-fen

# Pedagogical Implications of the Corpus-based Investigation of Discourse Markers

Lan-fen HUANG

*Shih Chien University*

## Abstract

The use of discourse markers (DMs) is a prominent feature in spoken English. Most of the previous studies of DMs investigated the speech of native speakers. Little research has investigated the speech of non-native speakers. The discussion of the functions of DMs is problematic, because DMs are produced unconsciously or subconsciously: this is difficult to demonstrate unequivocally as the function of any DM may be open to interpretation, when factoring variables such as context and the unseen paralinguistic features of the discourse. This paper, based on an analysis of three corpora: The Spoken English Corpus of Chinese Learners (SECCL), MICASE and ICE-GB, reports the use of such DMs as *oh*, *well*, *like*, and *I think*. These DMs, seen in a new light, are interpreted on the basis of their co-occurring linguistic evidence and contexts. This investigation will likely have many implications for and applications to English Language Teaching, which lacks materials for the teaching of DMs. This paper suggests that some aspects can probably be taught through consciousness-raising activities and the data-driven learning (Johns, 1991) approach and that materials can be made from existing corpora, such as MICASE and LINDSEI. The pedagogical value of teaching DMs is dependent on learning context. To conclude, this paper acknowledges the limitations and weaknesses of this study and makes suggestions for future research.

## Keywords

DISCOURSE MARKER, CORPUS STUDY, SECCL, MICASE, ICE-GB

## I Introduction

With the emergence of a number of corpora dedicated to spoken English (e.g. The Michigan Corpus of Academic Spoken English (MICASE) (R. C. Simpson, S. L. Briggs, J. Ovens, & J. M. Swales, 2002), the British Academic Spoken English (BASE) Corpus

(Nesi & Thompson, 2006) and the Hong Kong Corpus of Spoken English (HKCSE) (Cheng & Warren, 1999)), a voluminous literature has grown up on the investigation of spoken English. The use of discourse markers (DMs) in speech, as opposed to, and distinct from written English, has become a popular topic for research. The use of DMs made by native speakers (NSs) has been intensively examined with fruitful results (e.g. Schourup (1985), Schiffrin (1987), Fraser (1990, 1999), Jucker (1993), Lenk (1995), Biber, Finegan, Johansson, Conrad and Leech (1999), Aijmer (2002) and Carter and McCarthy (2006)). Some of these studies are corpus-based, some based on a small set of data and others use contrived examples. My own study has been derived from three existing corpora, and one of them is in somewhat contrived contexts. In any case it seems clear that little attention has been paid to the use of DMs by learners.

The use of DMs is one of the distinct features of spoken English. At the time of writing, DM is a commonly-used term, but its terminology and definition are still open to debate. Moreover, previous studies have not empirically established the functions of DMs. Nor have they comprehensively investigated the use of DMs in Chinese learners' speech.

This paper reports a corpus-based investigation of DMs by Chinese learners and native speakers, using a process that uses collocation phenomena to empirically derive the functions of DMs rather than interpreting them intuitively.

## II Literature Review

### 2.1 Discourse Markers in Spoken English

DMs are one of the common features of spoken English[1] (Carter & McCarthy, 2006). DMs are also known by a variety of other names, such as "sentence connectives" (Halliday & Hasan, 1976), "discourse particles" (Aijmer, 2002; Schourup, 1985), "utterance particles" (Luke, 1987, 1990), "semantic conjuncts" (Quirk, Greemhaum, Leech, & Svartvik, 1985), "pragmatic expressions" (Erman, 1987), "discourse operators" (Redeker, 1991), "continuatives" (Romero-Trillo, 1997), "discourse connectives" (Blakemore, 1987, 1992) and "discourse markers" (Carter & McCarthy, 2006; Fraser, 1990, 1999). This last term is most widely used. Schourup (1999), in his later work, adopts the term discourse marker, as does Blakemore (2002). The above terms are based on different theories and varying assumptions, which are not discussed here due to lack of space. I adopt the more popular and theoretically neutral term discourse marker.

The classification of DMs and approaches to defining them are subject to debate: There is general agreement about some words/phrases, such as *well* and *you know*.

They are classified as central DMs. For these DMs, it is not difficult to give characteristics and definitions. It is less certain whether words/phrases, such as *oh*, *right*, *but* and *I think*, are in the same category. For the time being, it is almost impossible to come to a dividing line between DMs or non-DMs for the words/phrases discussed in the literature and to give defining criteria. My research (Huang, 2011) discusses the features which tend to belong to DMs rather than suggesting the criteria for deciding DMs. Based on work by Schourup (1999) and Fung and Carter (2007), I offer five characteristics for determining a DM: 1) semantic and syntactical optionality, 2) flexibility of position, 3) frequent prosodic independence, 4) connectivity and 5) multi-grammaticality.

In the literature, researchers agree that DMs are used for particular functions, with no one-to-one correspondence between form and function. In other words, a DM can have several functions and the same function can be expressed by several DMs. Furthermore, DMs are produced without conscious intention. Their uses are not easily available to introspection and the empirical evidence provided in the literature is not clear, either. For these reasons, it is difficult to demonstrate unequivocally the function of any instance of a DM. Previous studies have not empirically established the functions of DMs. Nor have they comprehensively investigated the use of DMs in Chinese learners' speech.

## 2.2 Research on Learner Corpora for English Language Teaching

Since the late 1980s, there has been a growing amount of work on learner corpora. Over the past decade, it has been recognised that learner corpora are of use in the fields of English Language Teaching (ELT) and Second Language Acquisition (SLA), mainly because they show how the language is actually used by learners (Granger, 2002, p. 5). Two research trends have emerged: first, comparative study between NS and learner language for ELT purposes and, second, error analysis and diachronic study for SLA purposes. The former is of relevance to this study and discussed in more detail below.

In learner language research, there has been a consistent focus on two types of comparison: 1) between NS and learner languages and 2) between speakers of different mother tongues. This is what Granger (1998a, pp. 12-13; 2002, pp. 12-13) terms *Contrastive Interlanguage Analysis* (CIA). Learner corpora may be compiled in conjunction with NS corpora to identify underuse and overuse of linguistic items and to gauge learners' problems with usages.

Much research using corpora has been done by comparing the International Corpus of Learner English (ICLE) with the Louvain Corpus of Native English Essays (LOCNESS), a 300,000-word corpus of essays written by NSs. The data of both

corpora are university students' argumentative writing (Granger, 1998a, p. 13). The ICLE corpus and the LOCNESS corpus have been employed to investigate aspects of lexis, discourse and the grammar of learners' English (Granger, 1998b, p. xxi) (see various studies in the volumes edited by Granger (1998c) and Gilquin, Papp and Díez-Bedmar (2009)). Other pairs of corpora have also been compared. Shirato and Stapleton (2007) compare the vocabulary in Japanese learners' conversations with that in the conversation component of BNC. Their study reveals that learners underuse certain lexical items, which are distinctive features in spoken English and overuse some auxiliary verbs and adjectives. In addition to the exploitation of the publicly available learner corpora, small specialised learner corpora and their comparable NS corpora have been compiled for particular pedagogical purposes, such as Ackerley's study (2008) on learners' report writing and Dalziel and Helm's investigation (2008) of learners' use of modal verbs in online writing.

Similar to NS corpus research, there is an increasing emphasis on phraseology and collocation in learner language research. There are many examples in the literature; for instance, Altenberg and Granger (2001), Altenberg (2002), Nesselhauf (2004) and Aerselaer (2008). Comparison between two or more groups of learners with different first languages (L1s) also attracts research attention. For example, Tankó's study (2004) of adverbial connectors and Paquot's work (2008) on the phraseological patterns.

For ELT purposes, learner corpora studies inform English learning coursebooks and reference books. For instance, the writers of the *Top Notch* series coursebooks (Saslow & Ascher, 2006) claim that their books are informed by both NSs' and learners' usages, identified in the 328-million-word Longman Corpus Network. The teacher's edition of *Top Notch* (Saslow, Ascher, & Tiberio, 2006) provides frequent learner errors, so that teachers can be alert to the vocabulary and structures which give learners difficulty. Learner corpora research also contributes to the development of reference books (e.g. dictionaries published by Longman, Cambridge, Macmillan, etc.).

Most of the comparative studies offer insights into learners' linguistic knowledge and suggest that the findings from learner corpora research have pedagogical implications and applications. Nevertheless, not many studies report empirical evidence of the actual impact of learner corpora research. Granger and Meunier (2008) have issued a call for such evidence. They stress the importance of phraseology in language teaching and learning and urge that more action should be taken in the classroom and also discuss the challenges which lie in the relevant fields of language teaching and learning.

## III   Research Design

### 3.1 Corpora under Investigation

The data for analysis originated from three corpora: The Spoken English Corpus of Chinese Learners (SECCL) (Wen, Wang, & Liang, 2005), MICASE (R. Simpson, S. L. Briggs, J. Ovens, & J. M. Swales, 2002) and ICE-GB (2006). Each of the three corpora had two subsets extracted and these six sub-corpora were processed using a standard corpus investigation software, *WordSmith 4* (Scott, 2004), to locate the relevant items and scrutinise their co-texts. Table 1 below lists the number of texts, word counts and average words per text in the six sub-corpora.

Table 1 Corpora under investigation

|  | Number of texts (texts) | Word counts (tokens) | Average words per text (tokens) |
| --- | --- | --- | --- |
| SECCL: Monologues | 1,143 | 336,303 | 294 |
| SECCL: Dialogues | 1,143 | 596,639 | 522 |
| MICASE: Highly monologic discourse mode | 13 | 134,096 | 10,315 |
| MICASE: Highly interactive discourse mode | 48 | 577,996 | 12,042 |
| ICE-GB: Unscripted monologues | 70 | 153,646 | 2,195 |
| ICE-GB: Private direct conversations | 90 | 185,000 | 2,056 |

### 3.2 Hypotheses and Research Questions

The research sets out to test hypotheses about the use which Chinese learners and NSs make of DMs. I hypothesise that the use of DMs is genre-dependent and culture-sensitive. The two general hypotheses proposed in this paper are: 1) across the monologic and dialogic genres under investigation, the more interactive the genre or type of activity is, the more DMs occur and 2) the uses of DMs in the Chinese learners' speech under investigation are different from and are possibly not as varied as those in the speech of the NSs.

I test my hypotheses within the framework of the core research questions addressed below. Question 1 asks proportion of non-discourse use of the words/phrases and discourse use in the Chinese learners' and NSs' speech and ascertains whether, or not, the Chinese learners seldom use discourse markers. Question 2 intends to find out what types of co-occurrence and contexts the DMs tend

to occur. These answers are the basis of the interpretations of functions of DMs, which are the answers to Question 3.

1.  What is the distribution of the words/phrases under investigation (*like*, *oh*, *well* and *I think*) in the Chinese learners' and NSs' speech?
2.  With what types of co-occurrence or in what contexts do the DMs tend to occur?
3.  How do the learners and NSs use the DMs under investigation?

### 3.3 Procedures for Analysing Discourse Markers

A bottom-up descriptive approach was employed. The study began with frequency data and manual classification of the instances of the words/phrases for analysis between non-discourse use (e.g. *like* as a verb and *well* as an adverb) and discourse use in the six sub-corpora in order to ascertain if the words/phrases were primarily used as a DM. The distinction between non-discourse use and discourse use of such words as *well* and *like* was clear-cut and in most cases, could be drawn without difficulty. They were straightforwardly identified by referring to their word classes and co-occurring syntactical structure. The word *oh* was always taken as a DM. The phrase *I think* was most problematic due to the fact that its syntactical role was ambiguous.

The major part of the analysis was the instances of DMs, examining their collocation phenomena. Co-occurrence of DMs was identified on the basis of linguistic evidence (e.g. hesitation markers, emphatic lexis, reported speech etc.), not an existing analytical framework or schema. In the literature, DMs have been typically described in terms of their functions, but the use of the term *function* was problematic because the researchers could read the linguistic evidence only, yet could never read the speaker's mind. The functions of DMs were easily identified by neither the speaker nor the researcher. In this study, types of co-occurrence of DMs were first examined and then taken as evidence for determining the categories for discussion, with functions being secondary interpretations. This procedure may have made the processing/production process of identification of functions clearer.

In the following example, *well* co-occurs with the quoting verb *said*. Based on this co-occurrence, it is suggested that *well* is to signal reported speech.

And the other person **said**, "**Well**, Who's that boy? There is no boy at all. That's... That's Jay's friend. That's the Bona." (SECCL: B02-01-03)

The corpus analyses helped to bring evidence of typical co-occurrence and

distribution of DMs for analysis in the six sub-corpora, making it possible to empirically establish the functions of DMs rather than interpreting them intuitively. Nevertheless, this bottom-up approach to the data revealed that genre seems to be a key factor in using DMs, but, admittedly, it was unable to give an adequate explanation for the observed under- or over-representation of the DMs across corpora and texts.

## IV   Results

### 4.1 The Proportions of the Words/Phrases as Discourse Markers

   The instances of the words and phrases have been manually classified into non-discourse use and discourse use (i.e. DMs)[2]. Table 2 below presents the percentage of the three words/phrases in the six sub-corpora. (There is no distinction between non-discourse use and discourse use in the case of *oh*.) *Like* as a DM apparently occurs more often in the highly interactive discourse mode in MICASE than in other five sub-corpora. *Well* is a typical DM, in particular in the NS dialogic genres where over 85% of the instances of *well* perform the discourse use. In contrast, *I think* as a DM appears relatively rare in both Chinese learners' and NSs' speech.

Table 2 Percentage of the words/phrases as discourse markers across corpora

| Corpus/ Discourse marker | Monologic genres | | | Dialogic genres | | |
|---|---|---|---|---|---|---|
| | SECCL: Monologues | MICASE: Highly monologic discourse mode | ICE-GB: Unscripted monologues | SECCL: Dialogues | MICASE: Highly interactive discourse mode | ICE-GB: Private direct conversations |
| *like* | 1.7 | 5.3 | 2.6 | 3.0 | **57.3** | 15.7 |
| *well* | 16.0 | **52.0** | 44.8 | 37.3 | **90.0** | **85.7** |
| *I think* | 10.0 | 22.5 | 25.3 | 7.0 | 11.7 | 19.0 |

### 4.2 The Frequencies of the Words/Phrases as Discourse Markers

   The normalised frequencies of the words/phrases as DMs are listed in Table 3 below. In most cases, there are more instances of DMs in the dialogic genres than in the monologic genres and this supports my hypothesis that the more interactive the genre or activity type is, the more DMs occur.

Table 3 Frequency comparisons of the discourse markers in the monologic and
dialogic genres under investigation (Normalised frequency per 10,000 words
(times))

| Corpus/ Discourse marker | Monologic genres | | | Dialogic genres | | |
|---|---|---|---|---|---|---|
| | SECCL: Monologues | MICASE: Highly monologic discourse mode | ICE-GB: Unscripted monologues | SECCL: Dialogues | MICASE: Highly interactive discourse mode | ICE-GB: Private direct conversations |
| *like* | **0.4** | 1.1 | **0.4** | **0.5** | **84.4** | 7.7 |
| *oh* | 11.8 | 1.8 | 2.9 | 48.3 | 48.4 | 60.7 |
| *well* | 2.4 | 7.8 | 10.3 | 8.7 | 32.9 | 70.5 |
| *I think* | 3.0 | 1.5 | 2.8 | 11.1 | 3.7 | 6.8 |

There are solid grounds for comparing the frequencies of the DMs in the monologic genre with those in the dialogic genres. In the SECCL corpus, the speakers in the monologues are the same as those in the dialogues. In the two NS corpora, MICASE and ICE-GB, the speakers across the two types of genre are not the same, but their backgrounds are similar. Arguably, the frequencies in the sub-corpora of the monologic genres can be compared with those in the dialogic genres. However, the Chinese learner corpus (SECCL) and the two NS corpora (MICASE and ICE-GB) are not designed for comparison. As the nature of these three corpora varies, differences in frequency and the use of DMs between the two groups of speakers cannot be made without controversy. In addition, taking the NS usages of DMs as the target norm for NNSs is another issue. With due consideration for these issues, in the discussion of differences between the learners and NSs, the neutral terms *over-* and *under-representation* are used, rather than using the common terms *over-* and *under-use* in most learner corpus studies. In Table 2 above, DM *like* is over-represented in the sub-corpus of the highly interactive discourse mode in MICASE, occurring 84.4 times per 10,000 words, as opposed to fewer than 7.7 times in the five other sub-corpora. *Like* in the two learner sub-corpora and the sub-corpus of the unscripted monologues in ICE-GB is under-represented, occurring fewer than 0.5 times per 10,000 words. The normalised frequency is based on fewer than 10 instances in each sub-corpus. Due to the low number of occurrences, I argue that in the Chinese NNSs' speech and in the British NSs' unscripted monologues, *like* is almost never used as a DM.

### 4.3 Co-occurrence and Suggested Functions of Discourse Markers

Table 4 below summarises the most frequent types of co-occurrence of the DMs, which present particular uses of each DM in relation to the two types of genre. Most of the instances of *like* co-occur with exemplifications, as in Example (1), in the highly monologic discourse mode in MICASE, while they often co-occur with expressions of uncertainty, as in Example (2), in the highly interactive discourse mode. In Example (1), *like* introduces exemplifications, *a play*, *a speech* and *a debate* of the general term or description, *some performances*. In Example (2), *like* co-occurs with vague language *kind of* to express uncertainty or imprecision.

(1) She usually let us give her **some performances**, in the class, **like** with sometimes give her **a play**, **a speech** or **a debate** and things like that. She stressed the group work and we found that you may ask more active in the class
(SECCL: B01-08-15)

(2) B: I just just realised that it was actually the st study of architecture I really enjoyed
B: And uh you just **kind of like** get a a few hints at what actually working in the profession's like
(ICE-GB: S1A-034)

*Oh* and *well* in the monologic genres primarily mark reported speech, as shown in Examples (3) and (4). These instances co-occurring with reporting verb *SAY* mark the boundary between the speaker's own utterance and the speech being quoted.

(3) We played very happy. And my grandmother **said**: "**oh**, this is the most happy time I have"
(SECCL: B00-11-01)

(4) And the other person **said**, "**well**, **who's that boy**? There is no boy at all. That's... That's Jay's friend. That's the Bona."
(SECCL: B02-01-03)

In the dialogic genres, *oh* and *well* are often used as a (preface to a) response, as in Examples (5) and (6). *Oh* frequently occurs in such fixed expressions as *oh yes (yeah)*, *oh no*, *oh thank you* and *oh really* as a response, as exemplified in Example (4). Similarly, *well* often prefaces a response to a question, as in Example (6). The underlined responses are not simply a *yes* or *no*. These utterances are non-answers to

the question. It is possible that *well* is used as a device for gaining more thinking time or as a mitigator to soften the impact of an insufficient answer.

(5) A: Um... it is easy for you to be fired. Do you think so?
    B: **Oh**, yes. That's a problem.
    (SECCL: C99-25-32)

(6) SU-f: <OVERLAP1> **do we have to, say exactly what** </OVERLAP1> **they are?**
        we can't say they're, something
    SU-f: **well** <u>**we can't ignore 'em**</u>, <OVERLAP1> <u>**we're doing an inventory**</u>.
        </OVERLAP1>
    SU-f: <OVERLAP1> <u>**look let's talk about it afterwards**</u>. </OVERLAP1>
    (MICASE: LAB175SU026)

*I think* seems to be used as a delaying device by the learners, because it co-occurs most often with hesitation markers, pauses and restarts, as in Example (7), while it co-occurs frequently with factual information either to express uncertainty or to appear less assertive in the NSs' speech, as in Example (8).

(7) Um... ener... I **er...** I **I think um...** the piano is my favorite, is my favorite game.
    (SECCL: 99-35-19)

(8) SU-f: <OVERLAP1> that one that just flew off </OVERLAP1> had a white tail
        band. white
    SU-f: i have seen some white.
    SU-f: i saw some white on <OVERLAP1> its tail. </OVERLAP1>
    SU-f: <OVERLAP1> **i think the head** </OVERLAP1> **is darker than the body**, **i**
        **think**.
    SU-f: (does coffee come out)
    SU-f: okay
    (MICASE: LAB175SU026)

Table 4 Summary of the most frequent types of co-occurrence of the discourse markers in the monologic and dialogic genres under investigation

| Corpus/ Discourse marker | Monologic genres | | | Dialogic genres | | |
|---|---|---|---|---|---|---|
| | SECCL: Monologues | MICASE: Highly monologic discourse mode | ICE-GB: Unscripted monologues | SECCL: Dialogues | MICASE: Highly interactive discourse mode | ICE-GB: Private direct conversations |
| *like* | n/a | Exemplifications | n/a | n/a | Expressions of uncertainty | Expressions of uncertainty/ explanations |
| *oh* | *Reported speech* | *Reported speech*/ showing emotions | Showing emotions | *As a (preface to) response* | *As a (preface to) response* | *As a (preface to) response* |
| *well* | *Reported speech* | Transitions; shifts of topic | Transitions; shifts of topic | Disagreement; negative evaluation/ *as a preface to response* | Transitions; shifts of topic/ *as a preface to response* | *As a preface to response*/ disagreement; negative evaluation/ as a continuer |
| *I think* | *Hesitation markers; pauses; restarts* | *Factual information* | *Factual information*/ Personal opinions & evaluation | *Hesitation markers; pauses; restarts* | Personal opinions & evaluation | *Factual information* |

The collocation phenomena are used as categories for discussion and they lead to the interpretations of the functions of DMs. Table 5 below lists the types of co-occurrence of the DMs and the interpretations of their functions.

It can be seen that all the four DMs for analysis are multi-functional. This characteristic of multi-functionality causes difficulty in interpreting the use of DMs, as highlighted at the beginning of this paper. In most cases the uses of DMs are not even easily available to introspection by the speaker if the chance had been given to ask the users. This paper uses collocation phenomena to categorise the uses of DMs and clarify the logic of the identification of their functions. However, occasionally more than one type of co-occurrence is found in the same instance. In cases of this kind, the classification has to be subjectively judged.

In Example (9), *oh* occurs between the quoting verb *said* and the reported speech

and it can also be interpreted as part of the fixed expression *oh dear* to show emotions. Though it is unknown whether *oh dear* is produced by the person being quoted or the person being recorded, the co-occurrence of reported speech seems to be stronger evidence and therefore this instance of *oh* is coded as in the category of reported speech.

(9) The... the absent-minded teacher **said**,... **oh dear**, I suppose you are right. I remember now, when I... eh..., when I... came out of the car<card>
  (SECCL: B01-50-05)

The functions listed in Table 5 can be divided into two broad categories. The first ten items are primarily for textual organisation. The use of DMs to perform these functions helps the process of comprehension. For example, the use of *oh* and *well* signalling a repair and *well* marking a transition give the listeners a hint about the coming change. The last nine items primarily contribute to the interpersonal aspect of interaction. For instance, *well* and *I think* can be used as mitigators to avoid sounding too assertive and soften the impact of criticism.

Table 5 Functions of discourse markers identified on the basis of co-occurrence

|   | Co-occurrence | Function | Discourse markers* |
|---|---------------|----------|---------------------|
| 1 | Hesitation markers; pauses; restarts | • To suggest a search for contents or lexis; to hold the floor <br> • To sound less direct <br> • To reformulate due to being interrupted | *like*, *oh*, *well*, *I think* |
| 2 | Exemplifications | • To introduce exemplifications | *like* |
| 3 | Explanations; clarifications | • To introduce explanations/clarifications | *like* |
| 4 | Repaired/replaced items | • To signal a repair | *oh*, **well** |
| 5 | Reported speech | • To mark the boundary between the mode of the speaker and reported speech | *like*, *oh*, *well* |
| 6 | Opening/changing of a topic | • As a topic changer | *oh*, *well* |
| 7 | Concluding remarks | • To indicate a conclusion | *I think* |
| 8 | Shifts of topic | • To mark a transition | *well* |
| 9 | Continuation of the earlier topic; | • As a continuer | *well* |

| | | | |
|---|---|---|---|
| | elaborations | | |
| 10 | Prefacing a question | • To sound less direct and imposing | *well*, *I think* |
| 11 | Prefacing responses | • As a (preface to) response to a question and new information | *oh*, *well* |
| | | • To mitigate indirect/insufficient answers | |
| 12 | Disagreement and negative evaluation | • As a mitigator | *well* |
| 13 | Personal opinions and evaluation | • To avoid being too assertive with positive evaluations | ***I think*** |
| | | • To mitigate negative evaluations | |
| 14 | Factual information | • To express uncertainty | ***I think*** |
| | | • To appear less assertive | |
| | | • To reduce commitment | |
| 15 | Numerical expressions | • To make approximations | *like* |
| | | • To focus the coming information | |
| 16 | Expressions of certainty; key points | • To focus a key point | *like*, *well* |
| | | • To draw attention | |
| 17 | Expressions of uncertainty; vague language | • To express uncertainty | *like* |
| 18 | Expressions of emotions | • To show emotions | *oh* |
| 19 | Cognition-related verbs | • To indicate a cognitive process has been done | *oh* |

*The discourse markers highlighted are relatively frequent in the learners' speech and those underlined and in bold are relatively frequent in the **NSs**' speech.

In most contexts, the speaker has more than one DM to choose from. The choice may be affected by genre and activity type. For instance, the speaker may use *well*, *I think* and other DMs before raising a question. A faculty member in the classroom setting is likely to use *now* in order to mark a shift of topic and to sound confident, in particular when s/he knows the answer. The question may not be posed to elicit responses but for other purposes, such as marking the boundary in discourse and engaging the listener(s). A faculty member in an office hour session with a student may use *well* and *I think* prefacing a question to sound less imposing and to downplay her/his academic power and status.

In Table 5 above, some DMs are more frequent in one group of speakers than the other. The highlighted DMs are more frequent in the learners' speech and those

underlined and in boldface are relatively frequent in the NSs' speech. The differences in frequency are mainly due to types of activity they occur in. The different uses of DMs in the speech of the Chinese learners and NSs are summarised in the next section.

## V   Discussions

### 5.1 Factors in the Use of Discourse Markers

The corpus methodologies demonstrate that the use of DMs correlates with genres and types of activity. In the analyses of the four DMs, it is found that the DMs occur more often in the dialogic genres than in the monologic genres. For a frequency comparison between the two types of genre, the log-likelihood test and $z$ test for two proportions are calculated to assess the significance of differences. Except the case of *like* in SECCL, the values indicate significance between the monologic genres and dialogic genres as well as under-representation in the monologic genres. This supports my hypothesis that the use of DMs is affected by genre. The more interactive the genre is, the more DMs occur.

In addition to genre, context and type of activity are also factors in the use of DMs. For example, DM *like* occurs much more frequently in the American NSs' highly interactive discourse mode than in the highly monologic discourse mode (84.4 vs. 1.1 times per 10,000 words). It is used mostly by fellow students in informal contexts, such as study group discussion, rather than by faculty members in lectures. Another example is *oh*. *Oh* is used more than four times as frequently in the learners' dialogues as in the monologues and it is more than twenty times as frequent in the NSs' dialogic genres as in the monologic genres. The use of *oh* is found to be context sensitive. About two thirds of the instances of *oh* in the NSs' unscripted monologues in ICE-GB are used to show the speaker's emotions. A further look reveals that almost all (23 out of 26) instances are from sports commentaries, in which commentators use *oh* to show their emotions.

In the case of DM *well*, there are marked differences in the distribution of the types of co-occurrence of *well* across the six sub-corpora (see Appendix). This can be attributed to the variations in the type of activity. For example, the learners' monologues are mainly accounts of personal experience, in which there are more opportunities for using *well* to mark reported speech. Their dialogues are for exchanging opinions and therefore more chances are for the use of *well* prefacing disagreement to soften the speech.

5.2 Different Uses of Discourse Markers in the Chinese Learners' and NSs' Speech

Some uses of DMs in the Chinese learners' speech are different from those in the NSs' speech. Some of them may be attributed to generic constraints. The learner speech under investigation has the nature of test language. Similar topics to the NNSs' speech under investigation are likely to be practised before the recording, which could lead to less use of some DMs, such as *I mean*, co-occurring with clarifications, explanations and elaborations. The NS monologues in MICASE are mainly lectures, in which the speakers may use DM *now* to preface a question. The speakers do not usually expect an answer or any response from the students. In contrast, in conversations, speakers could use *well* and *I think* to preface a question in order to sound less direct and imposing.

Some differences in the use of DMs between the learners and NSs resist reasonable explanation. The investigation of *like* reveals that the learners tend to employ non-discourse use of *like* in their speech and DM *like* represents only 1.7 and 3 per cent of the instances of *like* in the monologues and dialogues respectively. In contrast, DM *like* is highly represented in the NSs' dialogic genres in MICASE and ICE-GB. It can be argued that the Chinese learners perceive *like* as an inappropriate DM in the test-taking setting and therefore provide almost no instance of it. However, it is more likely that the Chinese NNSs do not know how to use *like* as a DM, while using the other DMs under investigation.

Another distinction can be made relates to *oh* as a (preface to a) response to a question and to new information. Both the learners and NSs use *oh* in turn-initial position as a (preface to a) response. Further examination reveals that the Chinese learners tend to use *oh* as a (preface to a) response to a question and the implications conveyed by *oh* are probably different from NSs' understanding of *oh*. In the NS speech, *oh* is used as a marker of change-of-state, indicating "a problem about a question's relevance, appropriateness, or presuppositions" (Heritage, 1998, pp. 294-295). However, it is found that the Chinese learners use *oh* in a neutral tone as a token of acknowledgement.

The speech of the learners and NSs shows a marked difference in the use of *I think* as a DM. In terms of the types of co-occurrence, in the NNSs' speech, most of the instances of *I think* (63.5% on average) co-occur with hesitation markers, pauses and restarts. This could suggest that the learners use *I think* as a filler in their speech. Some of the NSs also use *I think* in this way, but the percentage is much lower (18% on average). In the NSs' speech, factual information is the most frequent type of co-occurrence, representing 45% and 28.6% respectively of the instances of DM *I think* in the two sub-corpora in MICASE and 41.9% and 50.9% respectively in the two subsets in ICE-GB. However, *I think* co-occurring with factual information is seldom

used by the Chinese NNSs: only two instances in their monologues.

## VI  Pedagogical Implications

This section seeks to address the practical needs of the learners of English in a more practical way. It can be argued that it is probably not necessary for Chinese learners to sound like NSs in their use of DMs when they communicate with other NNSs of different L1s. Admittedly, it is an ideal situation that every group of English speakers should use English as the way they do and other groups will be able to adjust to it. In practice, however, some measures have to be taken in the teaching and learning of English in Chinese-speaking countries. It makes little sense to tell learners that they can simply speak English in any way they like so long as they keep their national identity. For one thing, many learners of English wish to speak in a native-like way (Timmis, 2002). In addition, there are certainly those (e.g. Svartvik (1980, p. 171), Erman (1987, p. 1) and Fung and Carter (2007)) who believe that an inappropriate use of DMs may cause misunderstanding and lead to negative effects in communication. Therefore, it is worth raising the Chinese learners' awareness of the native-like usages of DMs and enabling them to use DMs appropriately, as well as, if they wish, helping them to speak like NSs.

With limited exposure to naturally-occurring spoken English among the Chinese learners and few opportunities to use DMs in their classroom discourse, I would argue that creating space in the classroom for the teaching and learning of DMs is necessary. Teachers should evaluate, depending on learners' needs, to what extent learners have to understand NSs' use of DMs and to speak in a native-like way.

Some pedagogical aspects of the use of DMs and approaches to the instruction of the use of DMs in the classroom setting are examined. The suggested approaches and activities below aim to raise learners' awareness and to enhance their understanding in order to improve their receptive competence. Their productive competence of using DMs, as McCarthy (1998, p. 60) maintains, should be allowed room to be displayed in a more natural context in the future, rather than in immediate production in the classroom.

How the Chinese learners acquire the use of DMs and why they frequently use DMs in the exam context remain a mystery. It is probable that DMs can be acquired without consciously learning them in the classroom. Nevertheless, the corpus studies of the NSs' and learners' speech reveal some areas for learners to become aware of.

The selection of the types of discourse features to be taught is based on local contexts, such as learners' age, proficiency level and the needs and objectives of the language programmes. Not all learners need to be native-like in using DMs, but some learners would certainly benefit from understanding DMs better and using them

appropriately. In this section, the relevant research outcomes are highlighted for pedagogical use.

## 6.1 Making Learners Aware of the Use of Discourse Markers in Speech

In most contexts where English is used as a foreign language, the mode of written English has been the norm, probably firmly rooted in learners since the outset of their English learning. Learners often have a bias towards the grammar of written English. Because of this, I would anticipate a certain degree of difficulty in addressing the issue of learners' expectations and prejudices. It may be a good choice to provide authentic data as strong and positive evidence and to begin with the features of spoken English and then introduce DMs as a prominent feature in speech.

To make learners aware of the use of DMs in speech, a quick starting point can be a consciousness-raising activity of comparing an academic word list (e.g. Coxhead's *academic word list* (1998, 2000)) and frequent words in spoken English (e.g. the most frequent words in spoken English in O'Keeffe *et al.* (2007, p. 35)). It is clearly evident that words, such as *yeah*, *so*, *like*, *well* and *right*, listed in the first 50 most frequent words in the 5-million-word spoken section of the CANCODE corpus (O'Keeffe et al., 2007, pp. 33-36), are absent from an academic word list. These words are likely to be used as DMs by NSs. The Chinese learners under investigation in this paper, however, are more likely to employ the non-discourse uses of these words, in particular *like* and *well*. The learners seem to be competent in using *like* as a verb and a preposition and *well* as an adverb in their speech. They seldom use *like* and *well* as DMs. It is also probable that they do not know how to use them as DMs. When NNSs encounter the use of DMs in real life, they are likely to experience, to some extent, linguistic shocks, because these words are frequently used as DMs in the NS speech.

To make learners aware of NSs' use of DMs, consciousness-raising activities can be used in the classroom. It would be feasible to extract an excerpt from MICASE containing the instances of DMs. This excerpt would provide good learning material for some students. For Chinese learners and other groups of learners with different L1s, the Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin, Cook, & Granger, 2010) and its native counterpart, the Louvain Corpus of Native English Conversation (LOCNEC), would be good sources for preparing materials for data-driven learning (Johns, 1991).

## 6.2 Areas Which May Require Pedagogical Interventions

The corpus-based investigation of DMs across the monologic and dialogic genres shows that there are more occurrences of DMs in the dialogic genres and reveals differing usages across the two types of genre. These findings imply that

context-appropriateness in using DMs, rather than frequency, should be the focus in the pedagogical interventions, if needed.

The frequency information derived from corpus studies tell us that certain types of co-occurrence of DMs are most frequent in an NS corpus, but this does not necessarily raise the use to prominence in a beginners' class or an introductory session to DMs. Similarly, when corpus studies tell us that a certain use of DMs is not frequent in itself compared with other uses, it is not a reason for ignoring this use. It is suggested that the pedagogical interventions of the use of DMs should aim to enable learners to use DMs appropriately on the basis of their priorities, competing with fluency, the construction of relationships, creation of solidarity, etc.

One of the areas that the Chinese learners may improve is using DMs to signal a repair. It is found that the Chinese learners seldom use DMs to signal a repair but merely pause for thought and restart, whereas NSs would use a DM (e.g. *well* and *you know*) to give listeners a signal that a correction is coming.

DMs prefacing dis-preferred responses is another use that the Chinese learners can give attention to. It is common for the Chinese learners to give dis-preferred responses without any preface. This kind of direct speech tends to be interpreted by NSs as aggressiveness, over-assertiveness and a lack of consideration for people's feelings. Nevertheless, in Chinese, the learners' L1, dis-preferred responses are usually prefaced with hesitation markers and DMs just as those in English are by NSs. Instruction may begin by referring to learners' communication strategies in their L1 and then demonstrate how NSs use such DM as *well* and hesitation markers to introduce dis-preferred responses in order to make speech less direct and face-threatening.

Although it is nothing to do with right or wrong grammars if the Chinese learners frequently use *I think* in their speech, they might use it at the expense of other alternatives and this probably requires some pedagogical interventions for them. Chinese learners tend to use *I think* with personal opinions, whereas NSs use *I think* for hedging (O'Keeffe et al., 2007, pp. 174-176). The Chinese learners' awareness can be raised for the various alternatives to *I think* in the NS speech.

Similarly, the NS use of *like* is another area to introduce to learners to aid their process of comprehension. Since *like* as a DM is a fairly recent use, it is suspected that the Chinese learners do not know how and when to use *like* in this way and they lack familiarity with the NSs' usage of *like*. It is probable that the NSs' constant use of *like* distracts learners' attention from the proposition that the speaker aims to deliver. In the NS data, it is found that DM *like* could be used as an in-group marker to express solidarity in certain groups. In some contexts, learners may need to communicate in English with NSs and assert in-group membership, for example, overseas students in

programmes of English for Academic Purposes in English-speaking countries. For this group of learners, instruction can focus on raising learners' awareness of NSs' use of *like* and further encourage learners to observe the use of *like* in the group they wish to integrate with. The native-like use of *like* probably enables learners to express their solidarity.

The last area that probably requires pedagogical intervention is not directly related to the use of DMs but the language surrounding them. In the analysis of *oh*, it is found that *oh* co-occurs with indicators of misplacement in both Chinese learners' and NSs' speech. It is found that the language of misplacement in the Chinese learners' speech may be interpreted as too direct (e.g. *I have another point*), while the NSs' language is more hedging (e.g. *the other thing I wanted to just mention*). Other less direct uses of language for softening speech can be instructed together with the use of DMs.

## VII  Conclusion

### 7.1 Summary

This paper examines a Chinese NNS corpus (SECCL) and two NS corpora (MICASE and ICE-GB) in order to shed light on the use of DMs across the monologic and dialogic genres and between the Chinese learners and NSs.

The objective of this paper is using collocation phenomena and co-text analyses to empirically derive the functions of DMs. The functions of the four DMs under investigation, *like*, *oh*, *well* and *I think* have been empirically established, making a contribution to the investigation of DMs in spoken English.

The six sub-corpora extracted from SECCL, MICASE and ICE-GB are by no means comparable. However, it is less problematic to have comparability across the two types of genre in each of the three corpora and to test my hypothesis that the more interactive the genre or type of activity is, the more DMs occur. One great advantage, beyond my expectations, of using three different corpora is the result of identifying the factors in using DMs. If a small comparable NS corpus had been compiled, factors such as genre and type of activity might not have been uncovered.

The implications of this research for pedagogy should benefit the English language teaching for Chinese speakers of English, help raise their awareness, prevent misunderstanding between speakers and facilitate inter-cultural communication in English.

### 7.2 Limitations and Weaknesses of the Study

No approach is without its weaknesses and neither kind of research is free of

limitations. In this section, I would like to acknowledge five limitations and weaknesses of this study. The first two limitations relate to the corpora in use. The remaining three weaknesses are about the methodology used in this study.

The first limitation relating to the corpora under investigation is that using three corpora with mark-ups and annotations in different ways makes it difficult to conduct an investigation across corpora with regard to certain questions. The Chinese NNS corpus (SECCL) under investigation clearly identifies speaker change in the texts, but similar information in the two NS corpora (MICASE and ICE-GB) is included in the mark-up and cannot be easily located. For instance, if a group of speakers uses a particular DM (e.g. *well* co-occurring with hesitation markers, pauses and restarts) in turn-initial position more often than another group of speakers does, it is difficult to find the answer but the topic remains for future research.

The second limitation is inherent in the Chinese learner data, which are contrived and collected in a restricted test-taking setting. The speakers' use of DMs is likely to be controlled and affected by the un-naturalness of the means of getting information and role-play activity, for example, *oh* prefacing a simple question at the beginning stage of the dialogue. This limitation has been imposed on studies of Chinese learners' speech, as the publicly available corpora of this group of speakers mostly consist of elicited data.

The third limitation, which relates to methodology, has been widely acknowledged. Since the corpus-based approach is based on the data of production, it is inevitably restricted to the evidence which is present in linguistic forms. In other words, this approach misses out absent features. It is not impossible, but would be extremely difficult to investigate any item which was not present in the corpus. (For example, such feature as ellipsis in a corpus can be manually tagged before the corpus is processed by software tools.) In this paper, the corpus-based approach to the investigation of DMs looks at areas where the four DMs are used, but the areas where the DMs might have been used have been left out. It is likely that there are some areas which lack DMs and which cannot be found with corpus methodologies.

Another weakness in methodology in this study is the process involved in identifying and describing collocation phenomena surrounding DMs is sometimes unavoidably dependent on the knowledge and intuitions of the researcher. In addition, some may raise doubts whether the functions of DMs can be fully supported by the identification of co-occurrence. In the discussion of the contexts where DMs tend to occur, it has been pointed out that, in the cases where more than one type of co-occurrence is identified, intuition-based judgements have to be made about which co-occurrence is stronger. In a few instances, no linguistic evidence can be found and I have had to resort to interpreting based on intuition. The reliance on intuition and

the making of subjective judgements may result in a slightly different frequency in duplicated studies in the future. However, since such examples account for a very small proportion, the overall distribution of the co-occurrence of DMs is expected to be reliable.

The last weakness of this paper is the possibility of generalisation. Further investigation shows that the distribution of DMs in each text varies. In other words, high users and low users of DMs can be identified. Moreover, the frequencies of DMs are affected by genre, type of activity and other contextual factors. These all make it difficult to generalise the results.

## 7.3 Implications for Future Research

The research and research methodologies reported in this study point towards some promising lines of inquiry for further research. First of all, more work can be done to investigate DMs in the NS speech, pointing towards the definition, functions and subtle implications that they carry. There is no agreement on the definition of DMs, what items are DMs and what their functions are. Although some central DMs, such as *well* and *you know*, have been studied intensively and extensively, the subtle implications of DMs still need to be uncovered. To facilitate interpretation of the use of DMs, the recent availability of multi-modal corpora (e.g. the Nottingham Multi-modal Corpus (Knight, Adolphs, Tennent, & Carter, 2008)) will be of great help, as they provide audio and visual resources in addition to co-texts.

Future research can be extended to DMs in other types of activity in the Chinese learners' speech. Studies of this kind will help us better understand the uses of DMs among Chinese speakers of English. It will also be interesting to apply the analysis modelled here on the speech of other learners with different mother tongues, so that non-native varieties of English may be compared and discussed together.

The pedagogical implications discussed in this paper open up considerable scope for further work in the classroom setting. Even though I have concluded that it is probably not necessary to formally teach DMs, a certain amount of pedagogical interventions may be of use for some learners. Additionally, there has been little empirical research on the efficacy of inductive approaches, such as DDL (Johns, 1991), in the classroom.

## 7.4 Concluding Remarks

Corpus approach is very promising to the investigation of learner language, because it can throw new light on language acquisition, language teaching and learning and varieties of English as well as other neighbouring branches. A corpus-based approach to the investigation of DMs in Chinese learner and NS English, such as I have

demonstrated in this paper, is relatively under-explored, compared with investigations into features in written English, and this approach may still be unfamiliar to most researchers, language material developers, practitioners, learners and relevant parties. Hopefully, this study of the use of DMs, a prominent feature in spoken English, may make a contribution to the investigation of learner language and that the implications of corpus research will be carefully considered and examined in the future.

## Notes

[1] Other features are 1) deictic expressions, 2) situational ellipsis, 3) headers, tails and tags, and 4) polite and indirect language, vague language and approximations (Carter & McCarthy, 2006).

[2] Whether the instances of the words/phrases under investigation are of the non-discourse use and discourse use has to be manually classified. When the incidence is over 400 times, three sets of 100-line concordance samples are used for analysis. For the cases of random sampling, the normalised frequencies of discourse use of the words/phrases per 10,000 words are based on an extrapolation of the percentages of discourse use. For example, in the sub-corpus of the dialogues in SECCL, the raw frequency of *like* is 984. The manual classification is less possible for the sheer number of the raw frequency. Three sets of 100 samples were extracted for manual analysis. 9 out of the 300 instances of *like* were categorised into discourse use and therefore there were 3 per cent of the instances were used as a DM (9/300x100=3).

## Appendix

Table 6 Distribution of co-occurrence of *like* as a discourse marker

| | SECCL: Monologues | SECCL: Dialogues | MICASE: Highly monologic discourse mode | MICASE: Highly interactive discourse mode | ICE-GB: Unscripted monologues | ICE-GB: Private direct conversations |
|---|---|---|---|---|---|---|
| Occurrences Percentage Co-occurrence | 5 out of 300 random samples | 9 out of 300 random samples | 15 | 172 out of 300 random samples | 6 | 47 out of 300 random samples |
| 1. Hesitation markers; pauses; restarts | 80.0 | 55.6 | 6.7 | 15.1 | 16.7 | 19.1 |
| 2. Numerical expressions and locations | 0 | 0 | 0 | 7.6 | 0 | 0 |
| 3. Reported speech | 0 | 0 | 0 | 1.7 | 0 | 17.0 |
| 4. Expressions of uncertainty | 0 | 0 | 13.3 | 22.7 | 0 | 21.3 |
| 5. Expressions of certainty/ key point | 0 | 11.1 | 13.3 | 16.9 | 16.7 | 10.6 |
| 6. Exemplifications | 20.0 | 22.2 | 46.7 | 11.0 | 33.3 | 8.5 |
| 7. Explanations | 0 | 11.1 | 13.3 | 12.2 | 33.3 | 19.1 |
| Unclassified | 0 | 0 | 6.7 | 12.8 | 0 | 4.3 |
| | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 7 Distribution of co-occurrence of *oh* as a discourse marker

| | SECCL: Monologues | SECCL: Dialogues | MICASE: Highly monologic discourse mode | MICASE: Highly interactive discourse mode | ICE-GB: Unscripted monologues | ICE-GB: Private direct conversations |
|---|---|---|---|---|---|---|
| Occurrences Percentage Co-occurrence | 397 | 300 random samples | 24 | 278* random samples | 45 | 300 random samples |
| 1. Reported speech | **51.4** | 0.3 | **33.3** | 2.5 | 17.8 | 3.3 |
| 2. Hesitation markers; pauses; repetitive words | 10.1 | 3.7 | 4.2 | 0 | 4.4 | 0 |
| 3. Repairs; rephrasing | 7.1 | 2.0 | 0 | 1.4 | 2.2 | 1.7 |
| 4. Opening and changing a topic | 2.3 | 6.7 | 4.2 | 4.3 | 2.2 | 5.7 |
| to a question | 0 | 15.3 | 12.5 | 5.0 | 0 | 4.3 |
| to a question (fixed phrase) | 0 | 2.7 | 4.2 | 4.3 | 0 | 3.3 |
| 5. As a (preface to a) response — to new information | 0 | **44.0** | 0 | **37.8** | 0 | **27.7** |
| to new information (fixed phrase) | 0 | 19.3 | 0 | **29.1** | 0 | **33.0** |
| 6. Showing emotions | 22.7 | 4.3 | 25.0 | 6.1 | 66.7 | 11.7 |
| 7. Implying a cognitive process has been completed | 5.0 | 1.3 | 16.7 | 8.3 | 6.7 | 8.3 |
| Unclassified | 1.5 | 0.3 | 0 | 1.1 | 0 | 1.0 |
| | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

*22 instances of *oh* in the 300 random samples are the number zero.

Table 8 Distribution of co-occurrence of *well* as a discourse marker

| | SECCL: Monologues | SECCL: Dialogues | MICASE: Highly monologic discourse mode | MICASE: Highly interactive discourse mode | ICE-GB: Unscripted monologues | ICE-GB: Private direct conversations |
|---|---|---|---|---|---|---|
| Occurrences Percentage Co-occurrence | 48 out of 300 random samples | 112 out of 300 random samples | 104 | 270 out of 300 random samples | 158 | 257 out of 300 random samples |
| 1. Hesitation markers; pauses; repetitive words; restarts | 18.8 | 19.6 | 6.7 | 10.7 | 10.1 | 14.8 |
| 2. Reported speech | **33.3** | 0 | 14.4 | 4.1 | 3.8 | 4.3 |
| 3. Repairs; rephrasing | 0 | 0.9 | 7.7 | 3.3 | 4.4 | 3.1 |
| 4. Opening/closing of a topic; concluding remarks | 8.3 | 7.1 | 1.9 | 3.0 | 7.6 | 0.8 |
| 5. Questions | 0 | 3.6 | 4.8 | 7.4 | 1.9 | 5.1 |
| 6. Transitions; shifts of topic | 6.3 | 4.5 | 44.2 | 24.4 | 44.9 | 14.8 |
| 7. Disagreement; negative evaluation | 4.2 | 30.4 | 1.0 | 15.9 | 2.5 | 17.5 |
| 8. As a preface to a response | 0 | 23.2 | 1.0 | **19.3** | 0 | **18.3** |
| 9. Key information | 10.4 | 0.9 | 0 | 1.9 | 0 | 0.4 |
| 10. As a continuer | 16.7 | 7.1 | 18.3 | 8.9 | 24.7 | 17.5 |
| Unclassified | 2.1 | 2.7 | 0 | 1.1 | 0 | 3.5 |
| | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 9 Distribution of co-occurrence of *I think* as a discourse marker

| | SECCL: Monologues | SECCL: Dialogues | MICASE: Highly monologic discourse mode | MICASE: Highly interactive discourse mode | ICE-GB: Unscripted monologues | ICE-GB: Private direct conversations |
|---|---|---|---|---|---|---|
| Occurrences Percentage Co-occurrence | 30 out of 300 random samples | 21 out of 300 random samples | 20 | 35 out of 300 random samples | 43 | 57 out of 300 random samples |
| 1. Hesitation markers; pauses; restarts | **70.0** | 57.1 | 25.0 | 20.0 | 16.3 | 14.0 |
| 2. Personal opinions & evaluation | 13.3 | 33.3 | **25.0** | 42.9 | 41.9 | **31.6** |
| 3. Factual information | 6.7 | 0 | **45.0** | 28.6 | 41.9 | 50.9 |
| 4. Concluding remarks | 10.0 | 4.8 | 5.0 | 2.9 | 0 | 0 |
| 5. Questions | 0 | 4.8 | 0 | 5.7 | 0 | 3.5 |
| Unclassified | 0 | 0 | 0 | 0 | 0 | 0 |
| | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

# References

Ackerley, Katherine. (2008). Using comparable expert-writer and learner corpora for developing report-writing skills. In C. T. Torsello, K. Ackerley & E. Castello (Eds.), *Corpora for University Language Teachers* (pp. 259-273). Bern: Peter Lang.

Aertselaer, JoAnne Neff van. (2008). Contrasting English-Spanish interpersonal discourse phrases: A corpus study. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 85-99). Amsterdam: John Benjamins.

Aijmer, Karin. (2002). *English Discourse Particles: Evidence from a Corpus*. Amsterdam: John Benjamins.

Altenberg, Bengt. (2002). Using bilingual corpus evidence in learner corpus research. In S. Granger, J. Hungand & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 37-54). Amsterdam: John Benjamins.

Altenberg, Bengt, & Granger, Sylviane. (2001). The grammatical and lexical patterning of *make* in native and non-native student writing. *Applied Linguistics, 22*, 173-194.

Biber, Douglas, Finegan, Edward, Johansson, Stig, Conrad, Susan, & Leech, Geoffrey. (1999). *Longman Grammar of Spoken and Written English*. Essex: Pearson Education Limited.

Blakemore, Diane. (1987). *Semantic Constraints on Relevance*. Oxford: Blackwell.

Blakemore, Diane. (1992). *Understanding Utterances: An Introduction to Pragmatics*. Oxford: Blackwell.

Blakemore, Diane. (2002). *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

Carter, Ronald, & McCarthy, Michael. (2006). *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Cheng, Winnie, & Warren, Martin. (1999). Facilitating a description of intercultural conversations: The Hong Kong Corpus of Conversational English. *ICAME Journal, 23*, 5-20.

Coxhead, Averil. (1998). *An Academic Word List* (Vol. English Language Institute Occasional Publication). Wellington, New Zealand: Victoria University of Wellington.

Coxhead, Averil. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Cutting, Joan. (2006). Spoken grammar: Vague language and EAP. In R. Hughes (Ed.), *Spoken English, TESOL and Applied Linguistics* (pp. 159-181). London: Palgrave Macmillan.

Dalziel, Fiona, & Helm, Francesca. (2008). Exploring modality in a learner corpus on online writing. In C. T. Torsello, K. Ackerley & E. Castello (Eds.), *Corpora for University Language Teachers* (pp. 287-302). Bern: Peter Lang.

Erman, Britt. (1987). *Pragmatic Expressions in English: A Study of You Know, You See and I Mean in Face-to-face Conversation*. Stockholm: English Department, Stockholm University.

Fraser, Bruce. (1990). An approach to discourse markers. *Journal of Pragmatics, 14*, 383-395.

Fraser, Bruce. (1999). What are discourse markers? *Journal of Linguistics, 31*(7), 931-952.

Fung, L., & Carter, Ronald. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics, 28*(3), 410-439.

Gilquin, Gaëtanelle, Cook, Sylvie De, & Granger, Sylviane (Eds.). (2010). *LINDSEI Louvain International Database of Spoken English Interchange*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gilquin, Gaëtanelle, Papp, Szilvia, & Díez-Bedmar, María Belén (Eds.). (2009). *Linking up Contrastive and Learner Corpus Research*. Amsterdam: Rodopi.

Granger, Sylviane. (1998a). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp. 3-18). Harlow: Longman.

Granger, Sylviane. (1998b). Introduction. In S. Granger (Ed.), *Learner English on Computer* (pp. xxi-xxii). London: Longman.

Granger, Sylviane. (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* (pp. 3-33). Amsterdam: John Benjamins.

Granger, Sylviane (Ed.). (1998c). *Learner English on Computer*. London: Longman.

Granger, Sylviane, & Meunier, Fanny. (2008). Phraseology in language learning and teaching: Where to from here? In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 247-252). Amsterdam: John Benjamins.

Halliday, M. A. K., & Hasan, Ruqaiya. (1976). *Cohesion in English*. London: Longman.

Heritage, John. (1998). *Oh*-prefaced responses to inquiry. *Language in Society, 27*(3), 291-334.

Huang, Lan-fen. (2011). *Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers.* (PhD thesis), University of Birmingham, UK. Retrieved from http://etheses.bham.ac.uk/2969/

*The International Corpus of English The British Component*. (2006). London: University College London.

Johns, Tim. (1991). "Should you be persuaded": Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing ELR Journal 4*. Birmingham: Birmingham University Press.

Jucker, Andreas H. (1993). The discourse marker *well*: A relevance-theoretical account.

*Journal of Pragmatics, 19*, 435-452.

Knight, Dawn, Adolphs, Svenja, Tennent, Paul, & Carter, Ronald. (2008). *The Nottingham Multi-Modal Corpus: A demonstration*. Paper presented at the The 6th International Conference for Language Resources and Evaluation (LREC), Marrakech, Morocco.

Lenk, Uta. (1995). Discourse markers and conversational coherence. *Anglicana Turkuensia 14*(Special Issue Organization in Discourse), 341-352.

Luke, K. K. (1987). *The establishment of common ground in natural conversation: The case of the utterance particle LA in Cantonese*. Paper presented at the The 1987 International Pragmatics Conference, Antwerp.

Luke, K. K. (1990). *Utterance Particles in Cantonese Conversation*. Amsterdam: John Benjamins.

McCarthy, Michael. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Nesi, Hilary, & Thompson, Paul. (2006). *The British Academic Spoken English Corpus Manual*.

Nesselhauf, Nadia. (2004). *Collocation in a Learner Corpus*. Amsterdam: John Benjamins.

O'Keeffe, Anne, McCarthy, Michael, & Carter, Ronald. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Paquot, Magali. (2008). Exemplification in learner writing: A cross-linguistic perspective. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 101-119). Amsterdam: John Benjamins.

Quirk, Randolph, Greemhaum, Sidney, Leech, Geoffrey, & Svartvik, Jan. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Redeker, Gisela. (1991). Linguistic markers of discourse structure. *Linguistics, 29*(6), 1139-1172.

Romero-Trillo, Jesús. (1997). Your attention, please: Pragmatic mechanisms to obtain the addressee's attention in English and Spanish conversations. *Journal of Pragmatics, 28*(2), 205-221.

Saslow, Joan, & Ascher, Allen. (2006). *Top Notch 1*. White Plains, NY: Pearson Education.

Saslow, Joan, Ascher, Allen, & Tiberio, S. C. (2006). *Top Notch 3 Teacher's Edition and Lesson Planner*. White Plains, NY: Longman.

Schiffrin, Deborah. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.

Schourup, Lawrence. (1985). *Common Discourse Particles in English Conversation: Like, Well, Y'know*. New York: Garland.

Schourup, Lawrence. (1999). Discourse markers. *Lingua, 107*, 227-265.

Scott, Mike. (2004). WordSmith Tools (Version 4). Oxford: Oxford University Press.

Shirato, Junko, & Stapleton, Paul. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an

empirical study in Japan. *Language Teaching Research, 11*(4), 393-412.

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). The Michigan Corpus of Academic Spoken English. from Ann Arbor, MI: The Regents of the University of Michigan

Simpson, Rita, Briggs, S. L., Ovens, J., & Swales, John M. (2002). The Michigan Corpus of Academic Spoken English.   Retrieved February 18, 2009, from Ann Arbor, MI: The Regents of the University of Michigan http://quod.lib.umich.edu/m/micase/

Svartvik, Jan. (1980). *Well* in conversation. In S. Greenbaum, G. Leech & J. Svartvik (Eds.), *Studies in English Linguistics for Randolph Quirk* (pp. 167-177). London: Longman.

Timmis, Ivor. (2002). Native-speaker norms and international English: A classroom view. *ELT Journal, 56*(3), 240-249.

Wen, 文秋芳 Qiu-fang, Wang, 王立非 Li-fei, & Liang, 梁茂成 Miao-cheng. (2005). *中國學生英語口筆語語料庫 1.0 版 [Spoken and Written English Corpus of Chinese Learners, Version 1.0]*. 北京 [Beijing]: 外語教學與研究出版社 [Foreign Language Teaching and Research Press].