A Corpus-Based Study on Discourse Markers in Native and Non-native Spoken English

Huican Huo(✉)

Wuhan University, Wuhan, Hubei, China
2021201020001@whu.edu.cn

Abstract. This study investigates the difference in the usage of spoken discourse markers by native English speakers and L2 learners. Previous studies have shown that L2 learners are very different from native speakers in terms of the frequencies and functions of using discourse markers, but corpus-based research is limited, and most of them are limited to planned and manipulated speech. This gap is solved by using a mixed research method to analyze two comparable corpora composed of real languages and conversations in daily life. The results show that compared with English native speakers, L2 learners are not able to use DMs equally frequent and underused certain kinds of discourse markers significantly. In addition, L2 learners use these items mainly to perform a textual function, while native speakers use a higher proportion of interpersonal and interactive functions. And the reasons behind this phenomenon may attributed to cultural differences and the inadequacy of foreign language teaching.

Keywords: Discourse markers · L2 English speakers · Corpus-based study

1 Introduction

A discourse marker (DM) is a particle that is used to direct the flow of conversation without adding any significant paraphrasable meaning to the discourse (Carter 2007) [5]. In communication, speakers externalize their communicative intentions through different forms of discourse, so that communicative participants reason about their discourse in order to obtain the intentions and verbal information. The choice of the form of the discourse output depends on the purposefulness of the message conveyed in the communication (Buysse 2010) [2]. The speaker organizes the content of the discourse in accordance with his or her choice of output form guided by his or her intention. Based on this understanding, different kinds of DM can be utilized by speakers with a subjective choice, which has something to do with its specific functions. In most cases, DMs are syntactically independent, which means that the sentence can still remain complete in structure or even in content without the existence of DMs (Croucher 2004) [3]. In spoken languages, DMs are easily to be found, which is mainly because these words can help the speaker organize the wording and play an important and indispensable role to express the speakers' attitudes or emotions.

L. F. Ying et al. (Eds.): ICELA 2022, ASSEHR 730, pp. 167–176, 2023.
https://doi.org/10.2991/978-2-38476-004-6_23

Previous studies have shown that a divergence does exist in the use of DMs between different groups, like non-native speakers and native speakers or among different levels of second language learners (L2 learners) (Liao, S., 2009) [11]. The specific performance is that different groups have various frequencies when using specific DM, and different contexts and functions can also be found in this process. To be more specific, non-native speakers maybe incapable to use certain DM as accurately as native speakers do, or they may use it more or less frequently (Aijmer 2002) [1], which may cause misunderstandings and ambiguities in their language expressions that can lead to them not being confident enough when speaking the foreign language. An inquiry into the specific different usage of DMs by L2 learners and native speakers and exploring the reasons behind this divergence can be much valuable for improving the way that L2 learners express themselves and has pedagogical insights for foreign language teachers.

Although a plenty of research on the usage and function of DMs have been carried out previously, few can be found that are based on the real conversations happened in daily life. In other words, most of them have been studied in the context of manipulated languages like prepared interviews or questionnaires so speakers may perform deliberately to use these lexical terms (Gilquin 2016) [6]. This limitation is expected to be solved in this study by adopting the corpus-based method. To be more specific, two corpora are utilized and five DMs (i.e. *well*, *like*, *you know*, *I mean*, *but*) are focused in the present study in order to solve the following three questions:

1). Which kind of DMs are the most frequently used by native and non-native speakers respectively?
2). In what way do L2 learners differ from native speakers in terms of frequency and function of DM usage?
3). What could be the underlying reason(s) behind this divergence?

2 Data and Methodology

The present study adopts a mixed research method, including both quantitative and qualitative analysis. It should be mentioned that the specific five discourse markers (i.e. *well*, *like*, *you know*, *I mean*, *but*) investigated in this paper are determined based on both the pilot study in the corpus and the previous literature.

First of all, in order to investigate differences in terms of the frequency and function of the five discourse markers between different groups (non-native and native speakers), two selected corpora have been used. For the group of L2 learners, an online corpus (C1) named the Treebank of Learner English (TLE) is used for most of the cases in this corpora are of the medium-high level. Since DMs are not easy enough for the beginners to acquire, it is appropriate to carry out this study among relatively more proficient learners. Then, for native speakers, Santa Barbara Corpus of Spoken American English (C2) is utilized for this corpus is transcribed the real languages and conversations happened in America, which is consistent with the intention of the present study. The two corpora used in this study are comparable in size so it can be expected that reliable findings are able to be explored at the end of the research.

All cases of the five DMs token were retrieved by using the concord function of the AntConc Corpus Analysis Toolkit (3.5.8) in the two comparable corpus, and then the author manually checked and cleaned the data to exclude any use other than as DMs. In the process of quantitative analysis, the frequencies are statistically processed in an online available tool in order to find whether there are significant difference between the two groups. Then, after retrieving frequencies and manually examining instances to rule out other usages, the function of the discourse markers should also be identified. In this paper, the specific functional definition criteria used in the qualitative analysis part is the paradigm proposed by Schiffrin [17] in 1987, whose model of coherence in talk could also be considered a model of discourse. Her multifunctional model focuses on local coherence, which is "constructed through relations between adjacent units in discourse, but it can be expanded to take into account more global dimensions of coherence" (1987). Following this paradigm, DMs in the present study are classified confirming with the functional headings proposed by Schiffrin.

It is noteworthy that any linguistic item could perform more than one function (Han 2010) [7]. The four functional categories proposed by the model of coherence are divided as follows:

1. Referential: this function normally conveys a semantic meaning as well as information about how the discourse is sequenced or coordinated by marking what has been said and what is going to be said, but also indicating different relations (cause, contrast, reorientation/digression, alternative etc.) in the discourse; it also helps on the conversational organization.
2. Structural: deals with the sequence in a discourse, but on a transitional level – from one topic to another, as a turn-transition device (e.g. initiating or taking turns, providing responses) or to inform the hearer of how units of discourse are sequenced. They can also be used to continue or summarize topics.
3. Cognitive, which includes the organization of knowledge of individuals and dynamic internal processes (e.g. opinions, intentions, disagreement, comparisons) by which inferences can be made. Also, it can refer to the state in which speaker/hearer "has information about something" – i.e. meta-knowledge, regarding what speakers and hearers know about each other's knowledge; this function also indicates a thinking process (e.g. erm, I think, I guess etc.).
4. Repair/clarification: utterance activities that allow speakers to locate/replace previous information and designate the speaker's intention to reformulate a thought, an idea etc. The nature of this function enables speakers to adjust their orientation to what has been said until the next conversational unit comes up. Schiffrin [17] reminds us that "almost anything that anyone says is a candidate for repair either by the speaker him/herself or by a listener".

   Still with regard to the functions of DMs, in addition to the model proposed by Schiffrin, this study also follows the suggestion made by Müller (2005) [12] on their grouping into two main categories:

1. Interpersonal and interactional mode, which expresses attitude, evaluations, judgments, expectations, and demands of the speaker. This mode deals with the nature of the social exchange – i.e., the roles assigned to both the speaker and the hearer.

2. Textual mode, which refers to the ways that the speaker creates cohesive passages of discourse to structure meaning as text, "using language in a way that is relevant to context".

3 Results and Discussion

3.1 Quantitative Analysis

In order to explore the quantitative differences, this study adopts a data analysis tool that has been proved to be effective in the previous study to calculate each DM frequencies and makes comparison. After the statistical calculation of the collected data, results of the frequencies are shown in Table 1. As shown in the table, *well* is the DM with the highest frequency in both two corpora, which reflects the same trend in the use of DM by the two groups of people. Similarly, *like* and *you know* are the second and third DMs that have been frequently used in the two investigated corpora. Apart from that, it is worth mentioning that in C1, the use frequency of *I mean* is slightly higher than that of *but*, and it is not the case in C2, which is exactly the difference occurring.

First of all, *well* is the DM that most frequently used by both two groups, with the raw frequency of 405 and 1375 in C1 and C2 respectively. And the normalized frequencies of this DM is 6.77 in C1 against 11.60 in C2, which show a significant difference ($p < .0001$). This means that L2 learners are incapable in using as much as their English native counterparts. This finding is also in line with the findings of previous study (Baiat et al. 2013) [10].

In terms of *like,* the second most frequently used DM in the two corpora with the raw frequencies of 302 and 635 respectively in C1 and C2, which shows that native speakers use it slightly more frequently than L2 learners. In the first test, the difference is not statistically significant ($G2 = 0.91$, $p > 0.05$), indicating that the frequency of L2 learners was generally similar to that of the native speakers. However, the results of the second test show a statistically significant difference ($U = 691.051$, $p = .0001$), referring that when examining the frequency of a single text that constitutes each corpora,

Table 1. Statistical Results in the C1 and C2

	Corpus 1		Corpus 2	
	Raw	Normalized	Raw	Normalized
but	57	0.95	526	4.45
I mean	85	1.39	432	3.66
you know	248	4.16	606	5.11
like	302	4.95	635	5.25
well	405	6.77	1375	11.60
Total	1097	18.22	3574	30.07

(Table credit: Original)

Table 2. Results of the tests for the difference between C1 and C2

	well	*like*	*you know*	*I mean*	*but*
LL (log-likelihood)	94.03	0.91	7.53	60.55	102.36
p value	<.0001	>.05	<.01	<.0001	<.0001
Mann-Whitney U	201.497	691.051	203.040	356.001	475.003
p value	<.0001	<.0001	<.0001	<.0001	<.0001

(Table credit: Original)

the difference is significant. And the possible explanation may be that some L2 learners use this DM quite often, while others never or rarely use it in their daily speaking.

As for *you know*, L2 learners use it slightly less frequently than English native speakers with a significant difference (G2 = 7.53, $p < .01$, U = 203.040, $p < .0001$). And the raw frequencies of this DM are 248 and 606 respectively. According to the comparison results, it can be understood in the way that L2 learners are significantly insufficient in using this DM compared to the group of native speakers. A similar situation can also be seen in the use of *I mean*, which also shows a significant difference with native speakers use it more frequently (G2 = 60.55, $p < .01$, U = 356.001, $p < .001$).

Finally, the total number of DMs in the discourse produced by native speakers is almost five times that of second language learners. This particularly large difference is also statistically significant, showing that compared with native language speakers, L2 learners use this discourse marker much insufficiently (G2 = 102.36, $p < .0001$; U = 475.003, $p < .0001$) (Table 2).

The possible reasons for L2 learners significantly under-use of DMs can be generally demonstrated, but for specific discourse markers, some viewpoints can be adopted to explain the use of variation. Huang (2018) [9] pointed out that the extensive use of *well* in Swedish groups usually does not have a negative impact, but the lack of representation of the latter may indicate that Chinese learners sound too direct in some contexts. This can be seen as a cultural difference, which leads to the difference in the use of discourse markers.

3.2 Qualitative Analysis of the Functions

Thus, in this study, DMs are understood to be lexical items with interpersonal and textual purposes. The author agrees with Schiffrin (1987) and Muller (2005) when they assert that the use of markers is optional, and that their removal from an utterance would not greatly alter (if alter at all) its structure or its propositional content. In this view, DMs may be grammatically optional and semantically empty, but they have great pragmatic relevance, with important implications in teaching practices.

Therefore, findings about the functionality of L2 learners and native speakers reported and discussed using the five discourse markers examined in this study and discussed in two modes: interpersonal and interactional mode as well as textual mode.

Results show that a higher percentage in the domain of textual functions can be found in L2 learner group, while native speakers have a higher percentage in terms of the interpersonal and interactive function. Native speakers use interactive features more frequently than L2 learners, and second-language learners using this utterance marker in text functions more frequently than native speakers.

(1) <B> well… she is one of the best employee *so* <\B> (C2-11)

When examining differences in individual function, significant differences were found between the two groups when using *so* to represent the result (p < .01), to conclude (p < .01), and to give the lower bound (p < .001). L2 learners use *so* to express results and draw conclusions significantly more frequently than native speakers, while native speakers significantly use it to make interlocutors speak more frequently than second language learners, as shown in example 1.

(2) so can you speak Germany
a little <laughing>
how?
they taught me. they also offer Germany classes as well *like* school curriculum[…] (C1-41)

(3) Uh..like I said before, the people there are very friendly, *like* they are eager to help their neighbors when they are in trouble (C1-07)

Among the five DMs investigated in two corpora, *like* is the second most frequently used, most of the instances (about 92% in C1 and about 90% in C2) playing a role in the textual field. But when it comes to individual features, significant differences are shown between L2 learners and native speakers in using *like* as an introductory explanation and providing citation. L2 learners tend to use *like* to introduce explanations (see in example 2) and provide citations (see in example 3), which are used significantly more frequently than native speakers. On the other hand, native speakers use it to introduce examples significantly more often than L2 learners. Other similar features identified in the study were approximate numbers, tagging vocabulary focus, searching for appropriate expressions, etc.

(4) Alright, for example, like on weekends, y'know, what I liked to do was probably a little girl used to do. (C1-31)

(5) It will never too late to…start for the… *you know.* what was it?
Lifelong learning? (C1–12)

(6) *You know*, everything there is just…slow and pleasant…
People're relaxed.
Yeah. (C2-12)

(7) It looks like too outdated, so *you know,* this one didn't work well (C2-42)

(8) Yeah… but although the students are taught in many different ways and. they have been suggesting… whoever speak to it's often… *you know* they liked it but I don't agree with that.' (C2-14)

As for *you know*, there are slightly more instances can be recognized as textual function (about 54%) than instances that act in the interaction domain (about 46%) in C1. In C2, on the other hand, more instances (about 64%) act in the interaction domain than instances that act in the textual filed (about 34%). L2 learners use far more instances than native speakers in terms of *you know* to play a role in the textual realm. As for the function of interaction, native speakers use more examples of roles in the field. When examining differences in individual functioning, it was found that *you know* as a discourse marker plays the role in providing relevant information ($p < .001$), seeking confirmation ($p < .05$), tagging vocabulary/content search ($p < .001$), tagging hesitation/uncertainty ($p < .001$), and self-revising ($p < .01$). L2 learners use this discourse marker in that relevant background information (see in example 4) and tag vocabulary/content search (see in example 5) is significantly more frequent than native speakers, while native speakers use it to seek confirmation (see in example 6), flag hesitation/uncertainty (see in example 7), and self-healing (see in example 8) significantly more frequently than L2 learners.

(9) I still don't know what that was, how ridiculous is that # *I mean* for real, on W48 I was on council and had emails sent for new (C1-29).

(10) … but it would be urgent to trace back to their original source, *I mean* back beyond the start point. (C2-50).

As for *I mean*, the least frequent of five discourse markers used by native speakers, the functionality in the textual field has the highest percentage in both corpora. When examining differences in individual function, significant differences were found between the two groups when using *I mean* to express hesitation ($p < .001$) and correct listener hypothesis ($p < .001$). L2 learners use *I mean* to flag hesitation, as in example 9, significantly more frequently than native speakers, while native speakers use it to correct the listener's assumptions, as in example 10, significantly more frequently than in L2 learners. It is worth mentioning that Mei (2012) [14] also reported a similar finding, in which non-native speakers use this feature more often, while native speakers use it more often to correct the assumptions of the audience proportionally. It can be speculated that non-native speakers have a harder time spotting potentially erroneous assumptions made by their listeners, or that they don't know *I mean* by this function in discourse.

Of the use of *well*, the least frequent of five discourse markers used by L2 learners, the frequencies between the two corpora vary widely (54 times in C1 versus 531 times in C2), so it may not seem very meaningful to make a significant comparison. Overall, however, the significant difference across the two functional domains is observed in the interaction domain. L2 learners use features much less often in the interactive realm than native speakers. As for the textual field, L2 learners use slightly more examples. In terms of the individual function, findings reveal significant differences between L2 learners and native speakers when using *well* to mark continuations ($p < .05$), open discourse units ($p < .01$), and re-express previous utterances ($p < .01$).

In general, L2 learners use features that serve primarily in the textual field. In the literature, interpersonal and international functions are often reported to be used more frequently by native speakers. For example, Mei (2012) [14] reports that *I mean* that the British use hypothesis correction functions more frequently than Chinese EFL learners, while Dutch and German EFL learners use consequential and conclusive functions more frequently (Müller 2005; Buysse 2012) [2, 12].

3.3 Possible Explanations for the Difference

Through the above analysis, we know that difference does exist in the use of DMs by L2 learners and native speakers. Possible explanation for this may be attributed into two points - cultural difference (Han, 2015) [7] and insufficient teaching in class (Fernandes, 2020) [4].

Firstly of all, different cultures can contrast in various ways, some more obvious and observable than others (Nikula, 2013) [13]. For example, cultures differ in language and social greetings. If explained from the perspective of psychology, there are differences in many aspects in culture. In other words, people from different cultural backgrounds view the world differently, and the way people communicate with each other is also different. This leads to different tendencies when people express themselves. For example, native English speakers prefer to express their uniqueness in their culture (Hellermann, J., & Vergun, A., (2007) [8], so they will more frequently use some mood markers that can express their independent personality, which is also one of the characteristics of "low context" culture. In other cultures, people do not always choose to express themselves in this way. For example, in eastern cultures, people tend to integrate themselves into the collective, so they try to avoid using modal particles that make them "different", which is also one of the characteristics of "high context" culture.

There exist different thought patterns in different cultural background. The dissimilarities of thought patterns deter mine the different modes of texts (Shimada 2014) [16]. A better, clearer understanding of the differences in the rules and features of text organization would benefit language learners greatly in terms of improving their intercultural communication skills. Non-native speakers generally have a pragmatic knowledge base in their native language and will find the equivalent in a foreign language to express themselves, though the examination of frequencies and diversity of items used in class showed the pragmatic knowledge and DMs could have been, potentially, better explored.

In addition, some experimental research has been carried out to investigate the teaching pattern in foreign language class, and results show that discourse markers are used in a relatively low frequency (O'Keefe 2017) [14]. Besides, functions of discourse markers used in classroom also differentiate from those are used by native speakers in the daily life, not to mention that some adult learners may be troubled by the "fossilization" in the process of learning, which is undoubtedly make their acquisition more difficult (Romero-Trillo, J., 2002) [15]. However, the possible reason for this may be explained by the fact that more formal languages are expected to be used during the class and the teacher's pedagogic goals did not specifically involve the introduction and practice of DMs. Nevertheless, lessons should have a clear interactive approach, which are excellent opportunities to make use of a wider variety of discourse markers.

4 Conclusion

The results show that significant differences can be found between L2 learners and native speakers in the spoken discourse containing DMs, which can be reflected in the frequency and function of their use. Specifically, when the use of some DMs is concerned, there is no significant difference between L2 learners and native speakers, which also indicates that second language learners are capable of reaching the level similar to native speakers. On the basis of reviewing the previous literature, the author believes that there are two reasons behind the differences, one is the differences between different cultures, the other is the deficiencies in foreign language teaching classroom.

Considering the methods and results of this study, the author puts forward some shortcomings of this study and some prospects and suggestions for future research. First of all, the genres involved in this study only include spoken language, which may lead to the neglect of L2 learners' performance in written language. In future research, this is worth further exploring. Secondly, this study did not further classify the second language learners according to their learning stages and levels, that is to say, it did not conduct a deeper classification and comparative study of the research objects, which is worth further exploring in future research.

References

1. Aijmer, K. (2002). English discourse particles: Evidence from a corpus. Amsterdam: John Benjamins Publishing.
2. Buysse, L. (2010). Discourse markers in the English of Flemish university students. In I. Witz-cakPlisiecka (Ed.), Pragmatic perspectives on language and linguistics, Vol. 1: Speech actions in theory and applied studies (pp. 461–484). Newcastle upon Tyne: Cambridge Scholars Publishing.
3. Croucher, Stephen. (2004). I uh know what like you are saying: An analysis of discourse markers in limited preparation events. National Forensics Journal, 21, 38–57
4. Fernandes, A. (2020). DISCOURSE MARKERS IN AN ENGLISH AS A FOREIGN LANGUAGE (EFL) CLASSROOM SETTING: A REFLEXIVE STUDY ON TEACHING DISCOURSE. Radiation Protection Dosimetry, 5, 723–742.
5. Fung, L., Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. Applied Linguistics, 28(3), 410–439.
6. Gilquin, G. (2016). Discourse markers in L2 English: From classroom to naturalistic input. In O. Timo feeva, A. C. Gardner, A. Honkapohja, S. Chevalier (Eds.), New approaches to English linguistics: Building bridges (pp. 213–250). Amsterdam, Philadelphia: John Benjamins.
7. Han, S. (2010). Cultural Differences in Thinking Styles. In: Glatzeder, B., Goel, V., Müller, A. (eds) Towards a Theory of Thinking. On Thinking. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03129-8_19
8. Hellermann, J., Vergun, A. (2007). Language which is not taught: The discourse marker use of beginning adult learners of English. Journal of Pragmatics, 39(1), 157–179.
9. Huang L. A Corpus-Based Exploration of the Discourse Marker Well in Spoken Interlanguage. Language and Speech. 2019; 62(3): 570–593. doi: https://doi.org/10.1177/0023830918798863
10. G. E. Baiat, M. Coler, M. Pullen, S. Tienkouw, L. Hunyadi, 2013, "Multimodal analysis of "well" as a discourse marker in conversation: A pilot study," IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom), pp. 283–288, doi: https://doi.org/10.1109/CogInfoCom.2013.6719257.
11. Liao, S. (2009). Variation in the use of discourse markers by Chinese teaching assistants in the US. Journal of Pragmatics, 41(7), 1313–1328.
12. Müller, S. (2005). Discourse markers in native and non-native English discourse. Amsterdam: John Benjamins.
13. Nikula, T. (2013). The use of lexical certainty modifers by non-native (Finnish) and native speakers of English. In I. F. Bouton, Y. Kachru (Eds.), Pragmatics and language learning (Monograph series No. 4) (pp. 126–142). Urbana-Champaign: University of Illinois.
14. O'Keefe, A., McCarthy, M., Carter, R. (2017). From corpus to classroom. Cambridge: Cambridge University Press.
15. Romero-Trillo, J. (2002). The pragmatic fossilization of discourse markers in non-native speakers of English. Journal of Pragmatics, 34, 769–784.

16. Shimada, K. (2014). Contrastive interlanguage analysis of discourse markers used by non-native and native English speakers. JALT Journal, 36(1), 47.
17. Schiffrin, D. (1987). Discourse Markers. New York: Cambridge University Press.