

JALT Journal

JALT Journal is the research journal of the Japan Association for Language Teaching (JALT). It is published semiannually, in May and November. As a nonprofit organization dedicated to promoting excellence in language learning, teaching, and research, JALT has a rich tradition of publishing relevant material in its many publications.



Links

- JALT Publications: <http://jalt-publications.org>
- *JALT Journal*: <http://jalt-publications.org/jj>
- *The Language Teacher*: <http://jalt-publications.org/tlt>
- *Conference Proceedings*: <http://jalt-publications.org/proceedings>

- JALT National: <http://jalt.org>
- Membership: <http://jalt.org/main/membership>

Provided for non-commercial research and education.
Not for reproduction, distribution, or commercial use.

Contrastive Interlanguage Analysis of Discourse Markers Used by Nonnative and Native English Speakers¹

Kazunari Shimada

Takasaki University of Health and Welfare

In this study, the use of discourse markers (DMs) in the speech of Japanese learners of English was investigated. To explore the features of their DM use, corpora of nonnative and native English speakers' speech were analysed using the methodology called Contrastive Interlanguage Analysis. A frequency analysis of DMs revealed significant differences between Japanese learners' and native speakers' speech, supporting earlier findings. Quantitative and qualitative analyses of the learner corpus data suggest that Japanese learners may use the marker *so* more frequently than other nonnative English learners, while also using certain interpersonal or cognitive function markers such as *you know*, *I mean*, and *just* less frequently. The findings suggest the need for language instructors and materials writers to understand the characteristics of Japanese learners' interlanguage and to provide them with appropriately designed DM input.

本研究は、日本人英語学習者の話し言葉における談話標識 (discourse markers: DMs) の使用を調べたものである。日本人英語学習者のDMs使用の特徴を探るために、対照中間言語分析の手法に基づき、非英語母語話者と英語母語話者の話し言葉コーパスを分析した。まず、日本人英語学習者と英語母語話者の話し言葉におけるDMsの使用頻度を分析したところ、先行研究と同じく、大きな差が見られた。次に、非英語母語話者の話し言葉を量的・質的の両面で分析した結果、日本人英語学習者が、他の非英語母語話者に比べて*so*を多く使用し、*you know*, *I mean*, *just*などの対人関係的、認知的機能をもつDMsをあまり使用しないことが明らかになった。その結果は、教師や教材作成者が日本人英語学習者の中間言語の特徴を理解し、学習者に対して慎重にDMsをインプットしていく必要があることを示唆している。

Discourse markers (DMs) are lexical items whose pragmatic functions play a crucial role in speech communication: Speakers use them to create textual coherence in interaction, as well as to express their own feelings or stances (Carter & McCarthy, 2006). For example, *OK/okay, really, and right* are used to respond to a speaker's utterance and to suggest agreement, alignment, or active listening. *But, first, and then* serve to organise discourse structure. Words like these are tools that enable speakers to convey their meanings to their listeners. Additionally, even if spoken sentences or phrases are grammatically correct, the lack of DMs may make it difficult to attract listeners' attention in a polite way (Romero-Trillo, 2002) and may create a negative impression of being uncollaborative or awkward in conversation (Svartvik, 1980). Therefore, DMs are of special importance to nonnative speakers (NNSs), who can use them to compensate for limited English language proficiency and to improve the comprehensibility of their messages (e.g., Tyler, Jefferies, & Davies, 1988; Williams, 1992).

Considerable interest has emerged in the roles and functions of individual DMs such as *because, oh, and well* (e.g., Blakemore, 2002; Fraser, 1999, 2009; Schiffrin, 1987). The development of corpus linguistics has enabled data-driven quantitative and qualitative analyses of the use of DMs by native speakers (NSs) of English (e.g., Lenk, 1998; McCarthy & Handford, 2004). However, a relatively limited amount of research has been conducted concerning DM use in terms of second language acquisition, especially in the Japanese EFL context (see Hays, 1992; Shimada, 2011).

Positioned against this contextual background, the present study was focused on DM use in the speech of Japanese English learners. The methodology followed Granger's (1996, 2002) Contrastive Interlanguage Analysis (CIA), a corpus-based approach that employs two types of comparisons: "between native language and learner language (L1 vs L2) and between different varieties of interlanguage (L2 vs L2)" (Granger, 2009, p. 18). The CIA approach has been applied in a number of corpus studies (e.g., Ádel, 2006; Granger & Tyson, 1996), and it offers insights into the nature of interlanguage as well as aids in the identification of usage trends (e.g., overuse, underuse, and misuse) in learners' speech and writing. Thus, the aim of the present study was to investigate differences in the use of DMs (a) between Japanese L2 speakers and NSs of English, and (b) between nonnative English learners with different L1 backgrounds (Japanese, Chinese, Dutch, German, French, and Spanish).

Literature Review and Research Questions

DMs in Spoken English

DMs have been defined by researchers in a number of different ways; however, there is generally a consensus that they mainly serve syntactic and pragmatic functions in discourse. Fraser (1999, 2009) addressed their syntactic functions and considered them to be linguistic items signalling a relationship between two segments of discourse. He argued that a DM must be included as an integral syntactic part of its next discourse segment. The DMs are italicized in the following examples:

1. a. Jones died last night. *But* he had been very ill for a long time.
- b. I went to Boston first *and* later on, went to Cape Cod.
- c. The water wouldn't boil, *so* we couldn't make any tea. (Fraser, 2009, p. 294)

In other words, the purpose of each marker in examples 1a, 1b, and 1c is to make coherent links between one discourse segment and another.

In spoken English, DMs often execute pragmatic functions. Schiffrin (1987) stated that they serve as contextual coordinators for establishing or maintaining a relationship between speaker and hearer.

2. Zelda: Are you from Philadelphia?
 Sally: *Well* I grew up uh out in the suburbs. And then I lived for about seven years up in upstate New York. And then I came back here t'go to college. (Schiffrin, 1987, p. 106)

In example 2, Sally uses *well* as a signal that she cannot give a clear answer to Zelda's yes-no question—in other words, that her pragmatic contribution is at odds with her interlocutor's expectations. Thus, as Schiffrin pointed out, the marker *well* plays the role of contextual coordinator, marking a juncture between a speaker's intention and a hearer's interpretation.

Additionally, Schiffrin examined discourse particles such as *I mean*, *you know*, *oh*, and *like*. These items do not serve essential syntactic functions; rather, they are optional devices through which speakers can shape their utterances to affect hearers' knowledge.

3. a. *I mean* I may be wrong, but I'm—I *mean* that's what I'm—that's my opinion.
- b. We have some *y'know*. (Schiffrin, 1987, pp. 34-35)

Despite the fact that their predominant function is pragmatic instead of syntactic, markers such as those in examples 3a and 3b are ubiquitous in everyday spoken English. The markers in 3a and 3b play a role in indicating the speakers' intention to keep conversation going, and help the hearers focus on the upcoming words. Schiffrin's definition of DMs, then, was broader than Fraser's (1999, 2009), and her model illustrated features of the spoken mode in more detail.

Fung and Carter (2007) also examined the spoken mode, and they incorporated Schiffrin's (1987) model while proposing a functional paradigm of DMs drawn from their analysis of spoken English data produced by NSs and

Table 1. A Functional Paradigm of DMs in Speech

Category	Discourse functions and markers
Interpersonal	Marking shared knowledge, indicating attitudes, or showing responses: <i>absolutely, actually, basically, exactly, great, I see, I think, just, kind of, like, listen, obviously, oh, oh great, OK/okay, really, right/alright, see, sort of, sure, to be frank, to be honest, well, yeah, yes, you know, you see</i>
Referential	Indicating relationship between utterances: <i>and, anyway, because/'cause, but, cos, however, likewise, nevertheless, or, similarly, so, yet</i>
Structural	Organising or managing the direction of conversations: <i>and, finally, first, firstly, how about, let me conclude the discussion, let's discuss, let's start, next, now, OK/okay, right/alright, second, secondly, so, then, well, what about, yeah</i>
Cognitive	Denoting thinking process, or reformulating utterance: <i>and, I mean, I see, I think, in other words, like, sort of, that is, to put it in another way, well, what I mean is, you know</i>

Note. Adapted from "Discourse markers and spoken English: Native and learner use in pedagogic settings," by L. Fung and R. Carter, 2007, *Applied Linguistics*, 28, p. 418. Some DMs such as *and*, *I think*, and *well* have multiple functions in discourse.

NNSs. They identified 57 common English DMs and classified them into four categories: interpersonal, referential, structural, and cognitive (see Table 1). This taxonomy is an extensive one, useful for characterising a large number of DMs in spoken English.

Learner Corpus Analysis of DM Use

Despite the widely recognised importance of DMs in spoken discourse, there have been only a limited number of studies examining the use of DMs by language learners. Romero-Trillo (2002) and Müller (2004) conducted corpus-driven comparisons of DM use by NSs and NNSs, and their results suggested that the use of certain DMs was influenced by the L1 of NNSs. Romero-Trillo quantitatively analysed spoken English data from Spanish children and adults. He found that Spanish children overused the English word *listen* due to the influence of its high-frequency counterpart in their L1 speech. Similarly, Müller compared the use of *well* and *so* by German speakers of English with their use in the speech of American NSs and found that German speakers used *well* much more frequently, and *so* much less frequently, than American NSs did. Müller pointed out that both DMs were translated as the German adverb *also*, and that German speakers might have a preference for *well* in order to avoid confusing English *so* and German *so*. In addition, Aijmer (2004) and Fung and Carter (2007) conducted corpus-based analyses revealing significant differences in the distributions of certain DMs between NS and NNS speech. Aijmer found that Swedish learners of English overused *I don't know* in order to signal uncertainty or hesitation, and Fung and Carter showed that learners in Hong Kong underused many markers, such as *right*, *yeah*, *well*, and *you know*, compared to the frequencies found in British NS data.

Only a few researchers have empirically investigated DM use in the speech of Japanese English learners. Hays (1992) described the acquisition of DMs by Japanese college students of various English proficiency levels. His analysis of the spoken data revealed that although the markers *and*, *but*, and *so* were frequently used, *you know* and *well* were rarely uttered by Japanese students learning English. In other words, his results indicated that the Japanese learners had greater difficulties acquiring pragmatic markers such as *you know* and *well*. Likewise, Miura (2011) compared the frequency of DMs used by Japanese learners of English to those of English NSs and found that certain markers such as *well*, *I mean*, *kind of*, and *like* were underused by novice and lower level learners. Additionally, Shimada (2011) conducted a corpus-based analysis of English DM use by Japanese learners and NS chil-

dren and adults. The results revealed that as speakers' proficiency improved, they used many items more frequently, regardless of their L1. However, the quantitative analysis confirmed significant differences in the distributions of DMs between Japanese learners and NSs. One of the notable findings was that Japanese learners overused relatively simple types of DMs such as *OK/okay, so, and yes*.²

Most studies on learners' use of spoken DMs have revealed that learners use certain items much more or less frequently than NSs do. However, the differences in DM frequency between NS and NNS speech are not enough to fully explain the features of DM use in interlanguage—that is, researchers have not yet determined whether the differences are due to the specific influences of individual L1 backgrounds or whether they are common to language learners in general. In order to address the issue, as Granger (2002) argued, it is necessary to construct a comparison of learner languages that incorporate speakers of different L1 backgrounds.

In addition, many comparative studies are based on disparate databases. For example, Shimada (2011) compared three spoken corpora, but there were considerable differences in the ways the data were collected. In that study, the Japanese learner corpus comprised a collection of interviews from a speaking test, but the speech data of NS children and adults were extracted from naturally occurring conversations in daily situations. These different situations may affect how speakers use DMs to facilitate communication, and different types of data collection may generate different results.

Research Questions

In the present study, features of DM use in the speech of Japanese learners of English were explored. The following research questions were addressed using the methods of CIA:

- RQ1: How do levels of use of spoken English DMs by Japanese learners differ from those of NSs of English?
- RQ2: How do levels of use of spoken English DMs by Japanese learners differ from those of other English language learners with different L1 backgrounds?

RQ1 is intended to replicate previous studies but using homogeneous databases. RQ2, on the other hand, is designed to explore the features of Japanese learners' DM use by comparing interlanguages of different L1 backgrounds.

Method

Databases

In order to make comparisons based on the CIA approach, the present study used two corpus databases. Data for EFL learners were from the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin, De Cock, & Granger, 2010), and data for native English speakers were from the NICT JLE Corpus (Izumi, Uchimoto, & Isahara, 2004).

The former database, LINDSEI, is a spoken corpus consisting of interviews produced by university undergraduates with different L1 backgrounds. All are higher intermediate and advanced learners of English. The spoken corpus consists of 11 subcorpora, classified according to learners' L1, and the data collection was performed using the same procedure for all subcorpora. Each interview lasts about 15 minutes and contains three tasks: (a) warm-up questions on a set topic (e.g., the most impressive country they have visited, their favourite film or play), (b) free and informal discussion with the interviewer, and (c) a picture description. The present study drew on six of the subcorpora, which are characterised in Table 2 below.

Table 2. Number of Interviews and Words per Subcorpus

L1 subcorpus	Language family	n of interviews	n of words
Japanese (JP)	Asian	51	37,126
Chinese (CH)	Asian	53	63,542
Dutch (DU)	Germanic	50	79,652
German (GE)	Germanic	50	85,950
French (FR)	Romance	50	91,402
Spanish (SP)	Romance	50	64,804
Totals		304	422,476

Note. Adapted from *LINDSEI: Louvain international database of spoken English interlanguage* by G. Gilquin, S. De Cock, and S. Granger (Eds.), 2010, p. 25. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Each subcorpus is made up of about 50 interviews, but the number of words in the Japanese subcorpus is much lower than that in the other subcorpora.³

NS data from the NICT JLE Corpus consisted of 20 interviews (94,845 words) produced by American speakers aged 20-24. Each interview lasts about 15 minutes. The interview tasks are also similar to those of LINDSEI, comprising warm-up questions, a single picture description task, and a role-play with the interviewer. The aim of the present study, therefore, is to address gaps in earlier work, ensuring the homogeneity of databases in order to permit an effective comparison of NS and NNS speech.

Procedure

The present study was focused on the 57 DMs listed in Fung and Carter's (2007) functional paradigm, which embraces the features of DMs in spoken English. In the first procedure, the corpus analysis software WordSmith Tools 5.0 (Scott, 2008) was used to obtain frequencies for each of the 57 items. Concordance lines were also viewed to differentiate words used as DMs from those playing other grammatical roles. Some examples are as follows:

Words used as DMs:

They are advertising by the week, so I found it. (The NICT JLE Corpus, N_file00006.stt)

... *well* first of all it's her expression she's got this really sour expression. (LINDSEI-GE050)

Words not used as DMs:

... I ... wouldn't be able to come back *so* early. (LINDSEI-FR006)

... but now I cannot speak English very *well*. (LINDSEI-JP051)

The categorization was carried out by the author. In order to test the reliability of the coding, a post-hoc intra-coder reliability check was conducted based on Müller (2004) at an interval of about 2 years. Despite the long interval, the simple agreement rate of the coding of *like*, *so*, and *well* was 94%, 99%, and 98%, respectively. Thus, the reliability of the coding process is considered high.

Statistical analyses of the frequencies of DMs were conducted to answer RQ1 and RQ2. The raw frequency of each item was standardized as a frequency per 10,000 words, and then used to calculate the log-likelihood ratio⁴ and chi-square value for comparison between corpora of different sizes. In corpus studies, although chi-square tests have often been performed to

compare word frequencies across corpora, log-likelihood tests are considered to have higher reliability than other statistical methods when comparing different-sized datasets (Rayson & Garside, 2000). When researchers compare two datasets with a single degree of freedom, significance is statistically tested by the log-likelihood ratios. If the log-likelihood ratio is ± 3.84 or more, a significant difference exists between the two datasets at a 5% significance level (Rayson, Berridge, & Francis, 2004). Additionally, Mann-Whitney tests were employed to compare the frequency of DMs by each functional category, following Fung and Carter (2007).

In addition to these quantitative analyses, the study included qualitative observations about the context, situation, and discourse function of spoken DMs. These observations serve to complement the quantitative analyses, providing vital details on the functions of DM use in actual learner speech.

Results and Discussion

Comparisons of DM Use Between Japanese EFL Learners and NSs of English

In order to answer RQ1, a comparative analysis was conducted using the frequency of DMs in two subsets of speech data: the Japanese subcorpus of LINDSEI (i.e., LINDSEI-JP) and the NS subcorpus of the NICT JLE Corpus (i.e., NICT-NS). Table 3 provides the standardized frequency of each marker, the log-likelihood ratios, and chi-squared values. If the occurrence rate of DMs was 0.01% or below in either database, the items were not included in the analysis.

Chi-square tests revealed that significant differences existed between the two databases in the frequencies of 21 out of 27 DMs with an occurrence rate of more than 0.01%. Additionally, log-likelihood ratios were added to the results obtained with the chi-square tests. If the ratio applied to the two databases was $+3.84$ or more, the item was considered to be used more frequently in LINDSEI-JP than in NICT-NS. On the other hand, when the ratio was -3.84 or less, the item was considered to be used less frequently in the Japanese learner data. The tests revealed that Japanese learners more frequently used relatively simple markers such as *yes*, *so*, and *I think*, while they used some interpersonal or cognitive markers such as *like*, *really*, *you know*, *kind of*, and *I mean* less frequently than NSs of English. Moreover, Mann-Whitney tests showed that significant differences existed between the two databases in the frequency of DMs in the interpersonal category ($U = 110$, $p = .040$). Therefore, the results support those of previous studies (e.g., Hays,

1992; Miura, 2011; Shimada, 2011), in finding that there was a significant discrepancy between Japanese learners and NSs of English in the frequency of DMs.

Table 3. Comparisons of DM Use Between Japanese EFL Learners (LINDSEI-JP) and NSs of English (NICT-NS)

Frequency per 10,000 words					
DM	Category	LINDSEI-JP	NICT-NS	LLR	Chi-square value
yes	IP	71.92	14.55	248.791	287.012**
so	Ref/Str	206.86	133.38	88.213	95.000**
I think	IP/Cog	88.35	51.66	54.020	58.292**
but	Ref	145.72	101.22	44.215	46.994**
now	Str	13.47	3.58	35.907	40.969**
first	Str	2.96	0.11	21.678	23.961**
finally	Str	2.96	0.74	8.470	9.684**
yeah	IP/Str	86.46	72.54	6.599	6.817**
and	Ref/Str/Cog	420.46	398.02	3.297	3.464
because/'cause	Ref	47.68	46.29	0.109	0.111
I see	IP/Cog	1.08	1.48	-0.326	0.311
or	Ref	50.10	54.09	-0.811	0.806
exactly	IP	2.15	3.48	-1.622	1.507
anyway	Ref	1.08	2.32	-2.356	2.090
basically	IP	0.27	4.32	-20.173	13.780**
oh	IP	7.54	21.30	-34.107	29.021**
then	Str	15.35	38.91	-53.065	46.000**
right/alright	IP/Str	0.27	11.07	-60.590	38.787**
OK/okay	IP/Str	22.90	59.25	-83.548	72.304**
actually	IP	4.85	27.94	-86.724	66.491**
I mean	Cog	2.15	25.73	-110.554	77.784**
well	IP/Str/Cog	5.39	37.32	-128.558	96.303**
kind of	IP	5.39	41.12	-148.569	110.000**
just	IP	10.77	77.39	-271.486	203.074**
you know	IP/Cog	4.31	64.32	-294.673	203.503**
really	IP	8.62	78.13	-304.263	221.379**
like	IP/Cog	28.82	140.65	-390.444	308.967**

Note. The occurrence rate of the markers *cos*, *great*, *next*, *obviously*, *sort of*, *sure*, and *what about* was 0% in either corpus. They were excluded from this analysis due to the impossibility of computing the log-likelihood ratio (LLR). Further research should be done to investigate why a certain DM occurs in one dataset but not in the other.

IP = interpersonal; Ref = referential; Str = structural; Cog = cognitive.

** $p < .01$.

Comparisons of DM Use Between Japanese EFL Learners and Other English Learners

This section addresses RC2, which was about comparing DM frequencies in NNS speech from the Japanese subcorpus with the five other subcorpora of LINDSEI (i.e., LINDSEI-OTHERS). Table 4 shows comparisons of the frequency of DMs. As in the analysis of the previous section, if the occurrence rate of a given DM was 0.01% or below in either database, the item was not included in the analysis.

The results of chi-square tests revealed that although Japanese learners often used some items such as *so* and *but*, they also used 14 out of 27 DMs less frequently than other nonnative English learners did. These findings were supported by tests of log-likelihood ratios.⁵ Although Mann-Whitney tests did not show significant differences in the frequencies of DMs according to functional category, interpersonal or cognitive function markers such as *well*, *really*, *you know*, *I mean*, and *just* were used less frequently by Japanese learners than by other English learners. Thus, the significant differences in the frequencies of DMs may represent the features of Japanese learners' DM use.

On the other hand, the results given in Table 4 reveal no significant differences between the two databases in the frequency of seven items: *exactly*, *kind of*, *or*, *OK/okay*, *anyway*, *cos*, and *basically*. There were only small differences between learners' respective frequencies of three markers—*and*, *yes*, and *right/alright*—although the differences were significant at a 5% significance level. In short, it was notable that Japanese learners used some items just as frequently as other nonnative English learners. Among these items, the use of *kind of*, *OK/okay*, *basically*, *yes*, and *right/alright* may be regarded as features of DM use in NNSs' interlanguage because the frequency of the five items differed significantly between Japanese learners and NSs of English (see Table 3).

Table 4. Comparisons of DM Use Between Japanese EFL Learners (LINDSEI-JP) and Other Nonnative English Learners (LINDSEI-OTHERS)

DM	Category	Frequency per 10,000 words			Chi-square value
		LINDSEI -JP	LINDSEI -OTHERS	LLR	
so	Ref/Str	206.86	96.04	315.280	397.358**
but	Ref	145.72	119.45	18.157	19.430**
now	Str	13.47	8.15	9.638	11.130**
finally	Str	2.96	1.09	7.093	9.470**
first	Str	2.96	1.17	6.292	8.234**
and	Ref/Str/Cog	420.46	394.14	5.815	6.164*
OK/okay	IP/Str	22.90	19.05	2.456	2.591
kind of	IP	5.39	4.88	0.173	0.178
exactly	IP	2.15	2.13	0.001	0.001
or	Ref	50.10	55.35	-1.749	1.711
anyway	Ref	1.08	2.36	-3.025	2.484
cos	Ref	4.31	6.90	-3.859	3.414
basically	IP	0.27	1.32	-4.363	3.057
yes	IP	71.92	84.57	-6.790	6.553*
right/alright	IP/Str	0.27	2.15	-9.283	6.050*
I think	IP/Cog	88.35	109.15	-14.451	13.799**
yeah	IP/Str	86.46	111.48	-20.767	19.613**
like	IP/Cog	28.82	44.56	-21.778	19.507**
actually	IP	4.85	14.07	-28.085	21.731**
oh	IP	7.54	18.42	-28.653	23.000**
because/'cause	Ref	47.68	73.26	-34.943	31.434**
then	Str	15.35	33.61	-42.937	35.367**
just	IP	10.77	47.72	-145.738	104.410**
I mean	Cog	2.15	31.30	-164.463	100.366**
you know	IP/Cog	4.31	39.91	-182.483	117.121**
really	IP	8.62	57.53	-227.775	153.006**
well	IP/Str/Cog	5.39	70.01	-357.270	221.268**

Note. The occurrence rate of the markers *sort of* and *that is* was 0% in either corpus. They were excluded from this analysis due to the impossibility of computing the log-likelihood ratio (LLR). Further research should be done to investigate why a certain DM occurs in one dataset but not in the other.

IP = interpersonal; Ref = referential; Str = structural; Cog = cognitive.

*p < .05. **p < .01.

However, these data do not address differences in DM use within the category LINDSEI-OTHERS, and distributions within individual subcorpora could boost or lower the overall frequency. To provide a clear picture, the frequencies of 12 DMs mentioned in this section were also compared across the six subcorpora of NNS speech. The further comparison was made to

confirm whether the use of *so*, *but*, *well*, *really*, *you know*, *I mean*, and *just* exhibited the features of Japanese learners' speech, and whether the use of *yes*, *kind of*, *right/alright*, *basically*, and *OK/okay* reflected the features of DM use in NNSs' interlanguage.

Figure 1 shows the frequency of *so* and *but* in each subcorpus. Although *so* was used in the Japanese subcorpus substantially more frequently than in any other nonnative subcorpus, only small differences existed among subcorpora in the frequency of *but*. Thus, the results confirm that the marker *so* is used more frequently by Japanese learners, and that the lower usage levels of *but* in the Chinese and German subcorpora lower the overall frequency of LINDSEI-OTHERS.

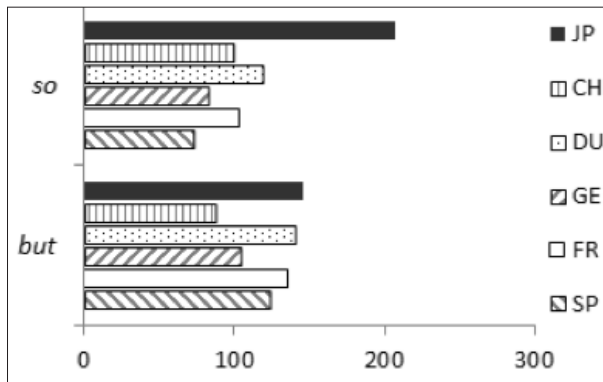


Figure 1. Frequency of *so* and *but* per 10,000 words in each subcorpus of LINDSEI.

Figure 2 shows a comparison of the frequency of *well* and *really* in each subcorpus. The analysis revealed that both Japanese and Chinese learners of English used the two markers notably less frequently than other nonnative English learners. In other words, the results suggest that English learners whose L1 belongs to an East Asian language family may be more likely to use the markers *well* and *really* much less frequently.

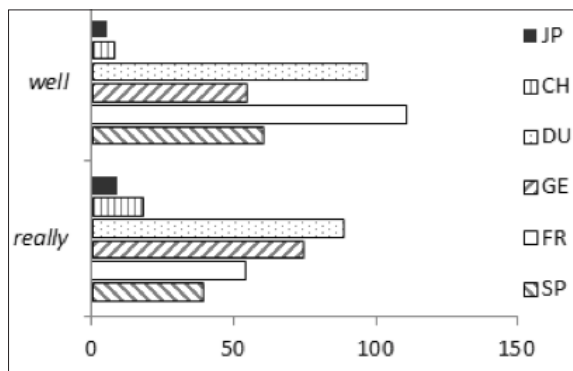


Figure 2. Frequency of *well* and *really* per 10,000 words in each subcorpus of LINDSEI.

Figure 3 shows the frequency of *you know*, *I mean*, and *just* in each subcorpus. The analysis revealed that Japanese learners used the three markers less frequently than other nonnative English learners. In other words, the results display a marked tendency for Japanese learners to use the interpersonal or cognitive function markers less often. These distinguishing features can be found only among Japanese learners of English; that is, they are not shared by nonnative English learners with different L1 backgrounds.

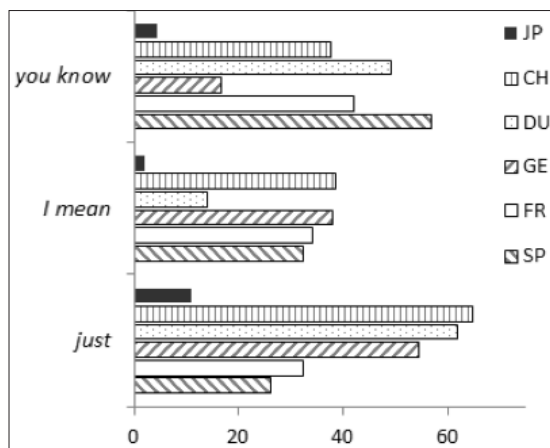


Figure 3. Frequency of *you know*, *I mean*, and *just* per 10,000 words in each subcorpus of LINDSEI.

Figure 4 shows the frequencies of *yes*, *kind of*, *right/alright*, *basically*, and *OK/okay* in each subcorpus. The marker *yes* generally displays small differences among the subcorpora except for in the French subcorpus, where it was quite frequent indeed. On the other hand, the three markers *kind of*, *right/alright*, and *basically* were infrequently used in all six subcorpora. The general frequent use of *yes* and the low frequencies of *kind of*, *right/alright*, and *basically* may be common to learners of English. With regard to the frequencies of *OK/okay*, Figure 4 shows that there is a considerable variability among the subcorpora.

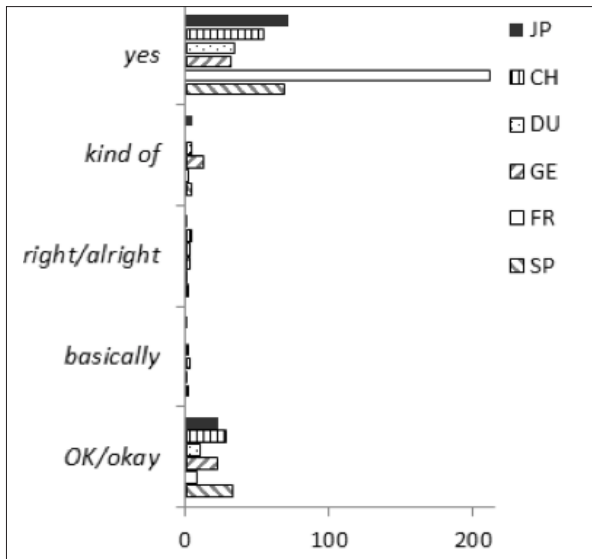


Figure 4. Frequency of *yes*, *kind of*, *right/alright*, *basically*, and *OK/okay* per 10,000 words in each subcorpus of LINDSEI.

In short, although simple items such as *yes* may be preferred by NNSs, items such as *kind of*, *right/alright*, and *basically* may be more difficult for them to acquire.

Why Do Japanese EFL Learners Overuse the Marker So?

Previous studies such as Hays (1992), Miura (2011), and Shimada (2011) have suggested that Japanese learners may infrequently use certain pragmatic markers such as *well*, *I mean*, and *you know*, but they may frequently

use simple types of markers such as *so* and *yes*. The present study yielded similar findings and distinguished features particular to Japanese learners from those seen in the speech of other NNSs. To investigate the acquisition of DMs in Japanese learners' speech, however, it is important to explore why some items are more or less frequently used. To that end, this section is focused on the marker *so*, which is frequently used by Japanese learners.

According to Fung and Carter's (2007) framework, the marker *so* has two discourse functions, referential and structural. Although the referential marker *so* serves a syntactic function to signal a relationship between one discourse segment and another, the structural marker *so* has some pragmatic functions, such as as a signal of summarising opinions and topic shifts. In the present study, as in my earlier study (Shimada, 2012), tokens of *so* were classified by functional category: referential, structural, or other. The following are illustrative examples of *so* extracted from the speech data of LINDSEI:

4. Referential: I don't think I pronounce it very well, *so* I am a bit embarrassed ... (LINDSEI-SP015)
5. Structural: ... I think that's Julia Roberts. *So* that's all. (LINDSEI-CH019)
6. Structural: *So* what do you think of the city Guangzhou? (LINDSEI-CH045)
7. Other: ... I always use bus *so* untto⁶ ... my nearest station is Ujiie Station. (LINDSEI-JP005)

In example 4, the speaker uses the referential marker *so* in order to establish a cause-and-effect link between the first clause and the second one. In example 5, the speaker tries to mark the conclusion of the topic by using the structural marker *so*. The speaker in example 6 changes the topic to the listener's impression of the city Guangzhou by using the structural marker *so*. In example 7, however, the marker *so* is neither referential nor functional; instead, it seems to be used as a filler, which can provide time for the speaker to think about what to say next.

Figure 5 shows the percentages for the three types of *so* (referential, structural, other) in the randomly sampled speech data, which comprise 10 interviews from each subcorpus. The coding of the functional categories was carried out by the author. As in the categorization of DMs described above, a post-hoc intra-coded check was conducted for the three subcorpora, LINDSEI-JP, -CH, and -DU (i.e., 30 interviews) at an interval of about

2 years. The overall agreement rate was 93%. Thus, the reliability of this analysis is considered high.

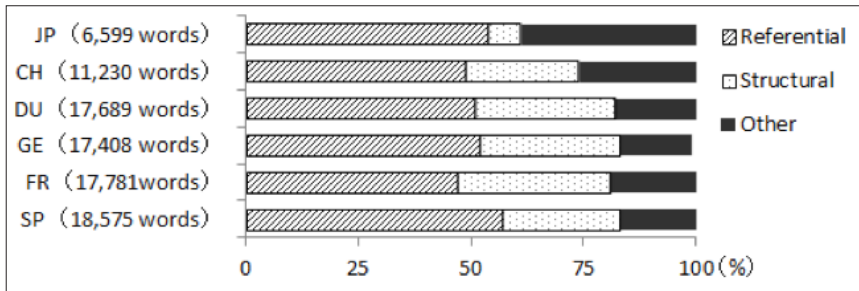


Figure 5. Percentages for the three types of *so* in randomly sampled speech data (10 interviews from each subcorpus of LINDSEI).

The results given in Figure 5 reveal that the proportion of the structural marker *so* was very low in the Japanese subcorpus. The third class of *so*, which is neither referential nor structural in function (i.e., other) was used more frequently by Japanese English speakers than by any other subcorpus group. The use of *so* as a filler may boost the frequency of the marker in Japanese English learners' speech.⁷

Conclusion

CIA was employed in this study to investigate the use of DMs in the speech data of Japanese learners of English. The results illuminate some features of these speakers' DM use.

This study's first research question was about frequencies of DMs in the speech of Japanese learners in comparison with those of NSs of English. Frequency analysis revealed significant differences between Japanese learners and NSs of English in the frequency of many DMs. Japanese learners frequently used some simple markers such as *yes*, *so*, and *I think*, yet they infrequently used certain interpersonal or cognitive function markers such as *like*, *really*, *you know*, *kind of*, and *I mean*. These findings corroborate those of previous studies, and they indicate that Japanese learners may have more difficulty acquiring particular pragmatic markers. These findings have important implications for language instructors, who may improve their students' interactional L2 skills as well as their linguistic ones through instructional focus on DMs.

The second research question was about levels of English DM use by Japanese learners in comparison with those of English learners with different L1 backgrounds. Frequency analyses revealed both similarities and differences between Japanese learners and other nonnative English learners in their use of DMs. Although Japanese learners used *so* much more frequently than other nonnative learners, they also used certain interpersonal or cognitive function markers such as *you know*, *I mean*, and *just* much less frequently. In other words, certain features of their DM use are distinguishable from those of nonnative English learners generally. This suggests the need for language instructors and materials writers to carefully provide Japanese learners with language input according to the characteristics of their interlanguage. For example, language instructors and materials writers should provide infrequent and difficult items, such as interpersonal or cognitive markers, at an intermediate or advanced proficiency level. Additionally, they should furnish Japanese learners with opportunities to use as many kinds of easy-to-use items as possible at a lower level.

This study has two basic limitations. Qualitative observations indicated that Japanese learners might use *so* as a filler, but this analysis has been far from exhaustive; more work on qualitative patterning is thus needed. As Romero-Trillo (2002) and Müller (2004) have suggested, Japanese learners' more or less frequent use of DMs may be a result of the influence of their L1. Second, some tasks to elicit speech may have an effect on learners' DM use. For example, a picture description task may not lend itself to the use of interpersonal markers such as *really* and *just*. Further research is needed to analyse learners' speech from a qualitative perspective and to investigate why Japanese learners may display different tendencies in English DM use from other nonnative English learners.

Notes

1. An earlier version of this paper was presented at the 127th Kanto Chapter Conference of the Japan Association for Language Education and Technology, Tokyo, Japan, 12 November 2011.
2. According to the online English Vocabulary Profile (<http://www.englishprofile.org/>), the markers *OK/okay*, *so*, and *yes* are classified into the Common European Framework (CEFR) level A1 or A2. Therefore, these markers can be regarded as easy items for English learners.
3. As Pritchard (1995) points out, Japanese learners of English may prefer slow, careful speech and take a long pause before answering a ques-

tion. If so, the interaction style may have a negative effect on fluency in speech production. However, LINSDEI does not contain audio data and does not provide the information necessary to find out why the Japanese students produced a much smaller number of words than any of the other nonnative English learners.

4. The tests of the log-likelihood ratios are also called G-tests.
5. The author combined the five subcorpora into one group and ran log-likelihood tests to compare the frequency of DMs between LINDSEI-JP and LINDSEI-OTHERS.
6. The Japanese word *untto* is approximately equivalent to the English marker *well*.
7. In the Japanese subcorpus, *so* as a filler was ubiquitous, although the frequency was not fully examined. Shimada (2012) also pointed out that the filler usage may contribute to Japanese learners' overuse of the marker. The present study confirms those earlier findings.

Acknowledgments

I would like to thank Professor Akira Kubota at the University of Tsukuba for his advice throughout this research.

Kazunari Shimada is an assistant professor at Takasaki University of Health and Welfare. His research interests include second language acquisition, materials development, and corpus linguistics.

References

- Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins.
- Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies*, 3(1), 173-190.
- Blakemore, D. (2002). *Relevance and linguistic meaning: The semantics and pragmatics of discourse markers*. Cambridge: Cambridge University Press.
- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide to spoken and written English grammar and usage*. Cambridge: Cambridge University Press.
- Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, 31, 931-952. [http://dx.doi.org/10.1016/S0378-2166\(98\)00101-5](http://dx.doi.org/10.1016/S0378-2166(98)00101-5)

- Fraser, B. (2009). An account of discourse markers. *International Review of Pragmatics*, 1, 293-320. <http://dx.doi.org/10.1163/187730909X12538045489818>
- Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, 28, 410-439. <http://dx.doi.org/10.1093/applin/amm030>
- Gilquin, G., De Cock, S., & Granger, S. (Eds.). (2010). *LINDSEI: Louvain international database of spoken English interlanguage*. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.
- Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast* (pp. 37-51). Lund, Sweden: Lund University Press.
- Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition, and foreign language teaching* (pp. 3-33). Amsterdam: John Benjamins.
- Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Ajimer (Ed.), *Corpora and language teaching* (pp. 13-32). Amsterdam: John Benjamins.
- Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and nonnative EFL speakers of English. *World Englishes*, 15, 17-27. <http://dx.doi.org/10.1111/j.1467-971X.1996.tb00089.x>
- Hays, P. R. (1992). Discourse markers and L2 acquisition. *Papers in Applied Linguistics-Michigan*, 7, 24-34.
- Izumi, E., Uchimoto, K., & Isahara, H. (Eds.). (2004). *Nihonjin 1200 nin no eigo su-piikingu koopasu* [A spoken corpus of 1200 Japanese learners of English]. Tokyo: ALC Press.
- Lenk, U. (1998). Discourse markers and global coherence in conversation. *Journal of Pragmatics*, 30, 245-257. [http://dx.doi.org/10.1016/S0378-2166\(98\)00027-7](http://dx.doi.org/10.1016/S0378-2166(98)00027-7)
- McCarthy, M., & Handford, M. (2004). "Invisible to us": A preliminary corpus-based study of spoken business English. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 167-201). Amsterdam: John Benjamins.
- Miura, A. (2011, September). *Discourse markers in spoken corpora of Japanese EFL learners*. Paper presented at Learner Corpus Research 2011, Louvain-la-Neuve, Belgium.
- Müller, S. (2004). 'Well you know that type of person': Functions of *well* in the speech of American and German students. *Journal of Pragmatics*, 36, 1157-1182. <http://dx.doi.org/10.1016/j.pragma.2004.01.008>

- Pritchard, R. M. O. (1995). *Amae* and the Japanese learner of English: An action research study. *Language, Culture and Curriculum*, 8, 249-264.
<http://dx.doi.org/10.1080/07908319509525207>
- Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In G. Purnelle, C. Fairon, & A. Dister (Eds.), *Le poids des mots: Proceeding of the 7th International Conference on Statistical Analysis of Textual Data* (pp. 926-936). Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.
- Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In A. Kilgarriff & T. B. Sardinha (Eds.), *Proceedings of the Workshop on Comparing Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics* (pp. 1-6). Hong Kong: Hong Kong University of Science and Technology.
- Romero-Trillo, J. (2002). The pragmatic fossilization of discourse markers in nonnative speakers of English. *Journal of Pragmatics*, 34, 769-784.
[http://dx.doi.org/10.1016/S0378-2166\(02\)00022-X](http://dx.doi.org/10.1016/S0378-2166(02)00022-X)
- Schiffrin, D. (1987). *Discourse markers*. Cambridge: Cambridge University Press.
- Scott, M. (2008). WordSmith Tools (Version 5) [Computer software]. Liverpool, UK: Lexical Analysis Software.
- Shimada, K. (2011). Discourse marker use in learners and native speakers: A corpus-based analysis of spoken English. *Annual Review of English Language Education in Japan (ARELE)*, 22, 377-392.
- Shimada, K. (2012). Discourse markers in EFL textbooks and spoken corpora: Materials design and authenticity. *Language Education & Technology*, 49, 215-244.
- Svartvik, J. (1980). 'Well' in conversation. In S. Greenbaum, G. Leech, & J. Svartvik (Eds.), *Studies in English linguistics for Randolph Quirk* (pp. 167-177). London: Longman.
- Tyler, A. E., Jefferies, A. A., & Davies, C. E. (1988). The effect of discourse structuring devices on listener perceptions of coherence in non-native university teacher's spoken discourse. *World Englishes*, 7, 101-110.
<http://dx.doi.org/10.1111/j.1467-971X.1988.tb00223.x>
- Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly*, 26, 693-711.
<http://dx.doi.org/10.2307/3586869>

