Iwasaki Linguistic Circle

c/o Kenkyusha Co., Ltd.
11-3 Fujimi 2-Chome
Chiyoda-ku, Tokyo 102–8152
Japan

Shigeru Takebayashi, Chairman

2009年7月20日印刷　2009年7月29日発行
編集兼発行者　岩崎研究会　代表者　竹林　滋
振替口座 00140–2–8737
〒102–8152 東京都千代田区富士見2–11–3　株式会社研究社内（電話 03–3288–7711）
印刷所　研究社印刷株式会社

Profiling Metadiscourse Markers in Native and Non-Native English

YUICHIRO KOBAYASHI

1. Introduction

The aim of this paper is to profile the use of metadiscourse markers in native and non-native English essays. Granger and Rayson (1998) apply Crytal's (1991) notion of "profiling," i.e. the identification of the most salient features in a particular person or register, to the field of interlanguage studies. In this paper, I will refer to the term "profile" or "profiling" in the same meaning. Starting from the assumption that every interlanguage is characterized by a "unique matrix of frequencies of various linguistic forms" (Krzeszowski 1990: 212), this study will employ three kinds of electric corpora: the Japanese EFL Learner (JEFLL) Corpus, the Japanese component of the International Corpus of Learner English (ICLE-JP), and the Louvain Corpus of Native English Essays (LOCNESS).

Corpora can provide a wide variety of linguistic information which could be useful in many different fields of language studies. Learner corpus is "a corpus, or computer textual database, of the language produced by foreign language learners" (Leech 1998: xiv). Since the 1990s, when many learner corpora were constructed, a number of language researchers carried out learner corpus-based Second Language Acquisition (SLA) studies, but most of their studies focused on learners' vocabulary and some grammatical features. Discourse analyses focus on language characteristics that extend across clause boundaries, and, as a result, discourse characteristics are more difficult to identify and analyze than lower-level lexical or grammatical features (e.g. Biber *et al.*

1998; Conrad 2002). Therefore this study will employ Hyland list of metadiscourse markers (e.g. Hyland 2005), one of computable discoursal variables, and examine the frequency-patterns of them.

2. Purpose

As already mentioned above, the purpose of this paper is to profile the frequencies and distribution in native and non-native English essays. Research questions are:

(1) How are metadiscourse markers distributed according to academic years?

(2) What is the difference of metadiscourse between native and non-native speakers of English?

To borrow Granger's (1998) terms, the former is the comparison between different stages of interlanguage (IL-IL comparison), and the latter is the comparison between native language and interlanguage (NL-IL comparison).

3. Data and methodology

3.1. Data

Three kinds of corpus data are compared in this study; the Japanese EFL Learner (JEFLL) Corpus, the Japanese component of the International Corpus of Learner English (ICLE-JP), and the Louvain Corpus of Native English Essays (LOCNESS).

JEFLL is a collection of free compositions written by learners of six different academic years at several junior and senior high schools in Japan. The corpus size is approximately 600,000 words. Now, the corpus is freely available for research via the web query system developed by Shogakukan Corpus Network (SCN).

ICLE-JP is a corpus of argumentative essays on different topics written by Japanese university students of English mainly in their third or fourth year. The corpus size is nearly 170,000 words. Since the Japanese component is not included in the ICLE CD-ROM, I obtained academic license to use it from the ICLE-JP team at Showa Women's University.

LOCNESS is a corpus of native English essays made up of British pupil's A level essays, British university students' essays, and American university students' essays. It can serve as a reference corpus for NL-IL comparison. The corpus size is about 320,000 words. We can purchase a copy of the corpus from Sylviane Granger or Sylvie De Cock.

Table 1 shows four sub-corpora analyzed in this study. JH, SH, UNI, and NS in the table stand for junior high school students, senior high school students, university students, and native speakers of English respectively.

Table 1: Corpora employed in this study

	JH	SH	UNI	NS
Corpus	JEFLL		ICLE-JP	LOCNESS
Tokens	328665	310251	168800	323985

3.2. Methodology

The definition of metadiscourse is still controversial, and it is often characterized as "discourse about discourse" (Vande Kopple 1985: 83). However, according to Crismore *et al.*'s influential definition, metadiscourse is "[l]inguistic material in texts, written or spoken, which does not add anything to the propositional content but this is intended to help the listener or reader organize, interpret and evaluate the information given" (1993: 40). Today, the most popular framework of metadiscourse may be Hyland list. While it has employed to analyze many kinds of texts, such as company annual reports (Hyland 1998), introductory academic coursebooks (Hyland 1999), undergraduate textbooks (Hyland 2000), or postgraduate dissertations (Hyland 2004), it is important in writing instruction and writing assessment (Ädel 2006; Hyland and Tse 2004; Kobayashi and Yamada 2008). Furthermore, it can be applied to corpus-based analysis of metadiscourse.

Hyland list consists of nearly 500 items, and they are categorized into ten types listed in Table 2.

Table 2: Metadiscourse categories

Category	Function
Transitions (TRA)	Express semantic relation between main clauses
Frame markers (FRM)	Refer to discourse acts, sequences, or text stages
Endophoric markers (END)	Refer to information in other parts of the text
Evidentials (EVI)	Refer to source of information from other texts
Code glosses (COD)	Help readers grasp functions of ideational material
Hedges (HED)	Without writer's full commitment to proposition
Boosters (BOO)	Emphasize force or writer's certainty in proposition
Attitude markers (ATM)	Express writer's attitude to proposition
Engagement markers (ENG)	Explicitly refer to or build relationship with reader
Self-mentions (SEM)	Explicit reference to author(s)

(Hyland and Tse 2004: 169)

In this study, data processing includes the following five steps:

Firstly, the raw and normalized frequencies of ten metadiscourse categories in four sub-corpora will be extracted.

Secondly, using the Pearson's product-moment correlation coefficient, the correlation matrix of the frequencies will be formed in order to establish the strength of relationships between sub-corpora.

Thirdly, correspondence analysis will be conducted in order to reduce the dimensionality of data matrix, and to visualize underlying complex relationships between categories, those between sub-corpora, and those between categories and sub-corpora.

Fourthly, statistical significance for each category will be tested with Pearson's chi-square test. *Post hoc* pairwise comparisons will be performed with Bonferroni's procedure when the calculated value of chi-square is significant.

Finally, each category will be qualitatively compared by means of careful examination of concordance lines.

4. Results and discussion

4.1. The distribution of metadiscourse categories

Table 3 lists the raw frequencies (RF) and normalized frequencies (NF) (per 100,000 words) of ten metadiscourse categories in four sub-corpora. The column "*p*" indicates *p*-values calculated by chi-square test. The normalized frequencies are visualized in Figure 1.

Table 3: The raw and normalized frequencies of ten metadiscourse categories

	JH		SH		UNI		NS		*p*
	RF	NF	RF	NF	RF	NF	RF	NF	
TRA	21784	6628.03	19164	6176.93	8334	4937.20	15045	4643.73	***
FRM	5235	1592.81	4357	1404.35	2527	1497.04	2319	715.77	***
END	62	18.86	41	13.22	22	13.03	14	4.32	n.s.
EVI	2	0.61	11	3.55	44	26.07	386	119.14	***
COD	950	289.05	1640	528.60	1453	860.78	2219	684.91	***
HED	3056	929.82	4219	1359.87	2787	1651.07	6356	1961.82	***
BOO	3304	1005.28	4205	1355.35	3259	1930.69	3940	1216.11	***
ATM	1904	579.31	2116	682.03	898	531.99	1109	342.30	***
ENG	7598	2311.78	6334	2041.57	3129	1853.67	4165	1285.55	***
SEM	39996	12169.23	32622	10514.71	8677	5140.40	3103	957.76	***

(***=p<.001, **=p<.01, *=p<.05, n.s.=p≧.05)

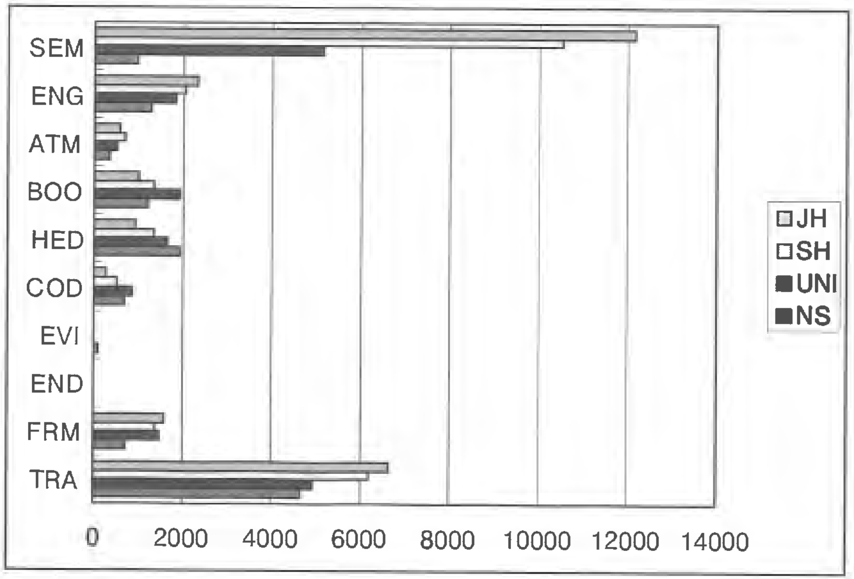

Figure 1: The normalized frequencies of ten categories

As can be seen in Figure 1, SEM is the most frequent category, and TRA is the second frequent category. These two categories, as well as FRM and ENG, are more frequent in non-native essays than in native ones. On the other hand, EVI and HED are less frequent in non-native than in native. A glance at Figure 1 may show a correlation between the frequency-patterns and academic years in non-native speakers, and a discrepancy between native and non-native speakers in the frequency-patterns. When there is a correlation between four sub-corpora, using Pearson's product-moment formula, the result falls into a convincing pattern. Correlation-coefficients range in value from +1.000, a perfect positive correlation, to –1.000, a perfect negative correlation. As Table 4 shows, the correlation-coefficients between three non-native sub-corpora run very high, ranging from 0.916 to 0.998, and those between natives and non-natives run relatively low, ranging from 0.413 to 0.708. What is meant in the figure is that there is a consistently high discoursal similarity within non-native sub-corpora and that there is a difference between natives and non-natives.

Table 4: Intercorrelations among four sub-corpora

	JH	SH	UNI	NS
JH	1.000			
SH	0.998	1.000		
UNI	0.916	0.939	1.000	
NS	0.413	0.456	0.708	1.000

In order to explore more complex relationships between categories, those between sub-corpora, and those between categories and sub-corpora, which are difficult to do by observing the contingency table, this study employs a multivariate analysis called correspondence analysis. The analysis is a technique for data-reduction, which visualizes the complex interrelationships between row variables, those between column variables, and those between row and column variables graphically in a multi-dimensional space. It computes the row and column scores in a way that permutes the original data matrix so that the correlation between the row and column variables can be maximized (e.g. Tabata 2002). Figure 2 shows the row and column scores of most powerful dimensions, which account for 99.41% of total variation in the data matrix, on a scatter diagram. The coordinates in the diagram reflect the relationships between variables, and similar variables are plotted close to each other.

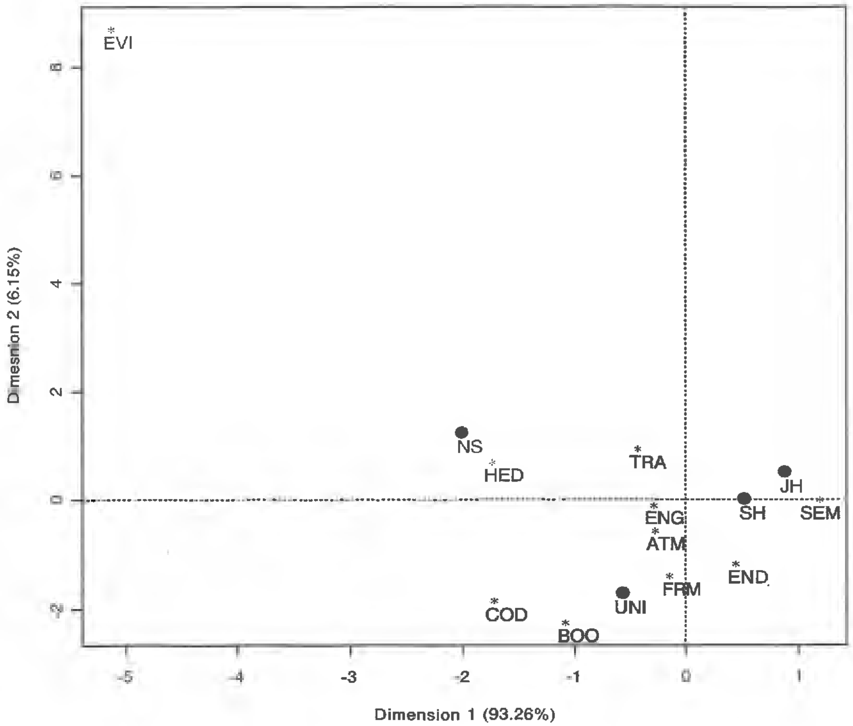

Figure 2: Correspondence analysis (Dimension 1 and 2)

The most prominent feature of this figure is that the proficiency level of English is reflected in Dimension 1. As far as the horizontal axis is concerned, NS is plotted apart from the other three sub-corpora, JH, SH, and UNI. While UNI, the most advanced group of non-natives, is plotted relatively close to NS, JH, the most novice group, is plotted furthest from NS. In other words, the frequency-patterns of metadiscourse categories serve as a developmental index, a "developmental yardstick against which global (i.e. not skill or item specific) second language proficiency could be gauged" (Larsen-Freeman 1983: 287). Moreover, we see from Figure 2 that SEM and END are characteristics of JH and SH, that BOO and COD are those of UNI, and that EVI and HED are those of NS. Although the profiling using the technique for data-reduction can be understood intuitively, it often blurs some of subtle differences within the data matrix. To solve the problem, more detailed analysis will be the subject of the following section.

4.2. Profiling metadiscourse categories

In order to discover significant overuse / underuse of metadiscourse categories, it is necessary to investigate the frequency-patterns not only quantitatively but also qualitatively. Furthermore, the frequencies of each item should be scrutinized if necessary. Although it is reasonable that all categories will be examined in detail, as space is limited, this paper will concentrate on TRA, FRM, and SEM in the following sections. This is also because novice non-natives prefer to use these three categories in the sentence-initial position in order to connect one sentence with another.

4.2.1. Transitions

Transitions (TRA) are mainly conjunctions and adverbial phrases which help readers interpret pragmatic connections between steps in an argument (Hyland 2005: 50). As shows in Table 3, the frequency of TRA decreases significantly as the proficiency level of English raises (chi-square = 2673.50, df = 3, $p < .001$). Subsequently, the results of *post hoc* pairwise comparisons with Bonferroni's procedure show that statistical differences are found between JH and UNI ($p < .05$), JH and NS ($p < .001$), SH and UNI ($p < .001$), SH and NS ($p < .001$), and UNI and NS ($p < .001$). In other words, non-natives overuse significantly TRA in comparison to natives.

Although TRA consists of 48 items, the sum total of frequencies of the top three items ("and," "but," and "because") makes up nearly 90% of word-tokens of all items in TRA. The distribution of the top three items appears in Figure 3.

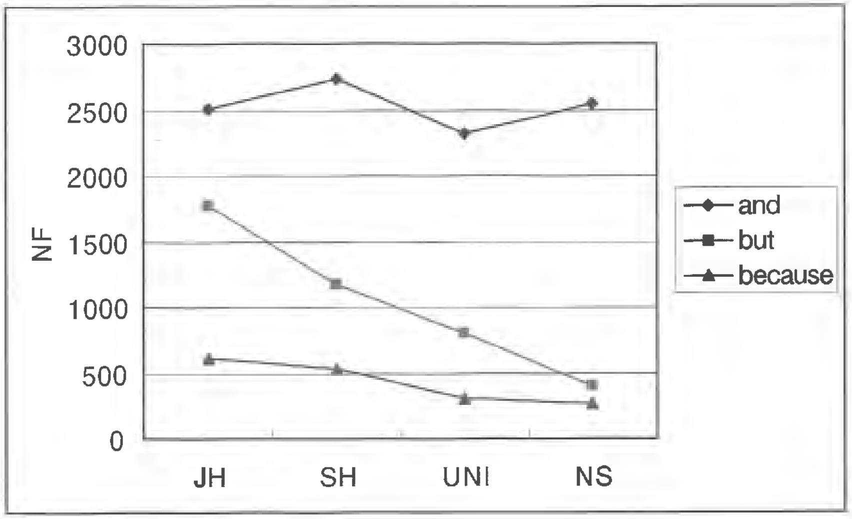

Figure 3: The distribution of top three TRA ("and," "but," and "because")

This figure tells us that the NF of "but" is inversely proportional to the proficiency level of English. Take an essay written by a JH student for example. In JH and SH components of the JEFLL Corpus, in order to ensure fluency, the subjects are allowed to use Japanese words whenever they cannot hit upon the right words in the writing task (Tono 2002: 160).

> I usually have bread.
> **But** I like rice.
> I eat sometimes.
> **But** it's [JP: metta ni nai].
> And I don't like milk.
> **But** I like [JP: misoshiru].
> **But** milk and [JP: misoshiru] drink [JP: metta ni nai]. (JH)

This clearly shows that novice non-natives tend to overuse "but" in the sentence-initial position. Figure 4 summarizes the proportion of top three TRA ("and," "but," and "because") in the sentence-initial position.

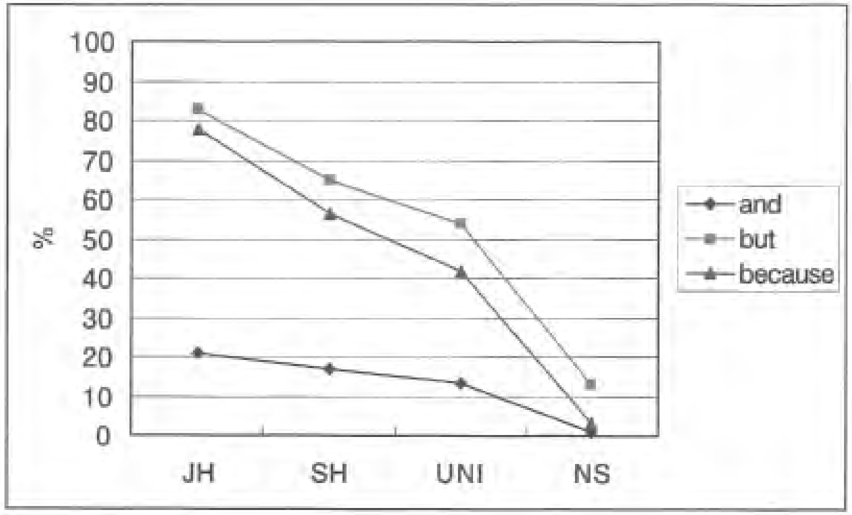

Figure 4: The distribution of top three TRA in the sentence-initial position

As this figure indicates, non-natives are apt to overuse conjunctions, especially "but" and "because," in the sentence-initial position. Biber *et al.* demonstrate quantitatively that "but" in the beginning of sentences is characteristic of conversation (1999: 83–84). In addition, it seems that some occurrences of "but" don't bear the meaning of contrast (e.g. "**But** I like rice."). As Crewe (1990) and Altenberg and Tapper (1998) remark, relations that can be inferred from the text do not have to be marked explicitly, which means that a high frequency of connectors in a text does not necessarily improve its cohesive quality. Overuse and misuse of connectives are likely to reduce the comprehensibility of the text.

4.2.2. Frame markers

Frame markers (FRM) signal text boundaries or elements of schematic text structure (Hyland 2005: 51). As shows in Table 3, significant differences are found among sub-corpora for the frequency of FRM (chi-square = 207.52, df = 3, $p < .001$). The results of *post hoc* pairwise comparisons show that statistical differences are found between JH and SH ($p < .01$), JH and UNI ($p < .001$), SH and UNI ($p < .001$), and UNI and NS ($p < .001$). To put it briefly, non-natives use more FRM than natives.

All items in FRM are classified into four subcategories: sequencing, label stages, announce goals, and shift topic. The distribution of the subcategories appears in Figure 5.

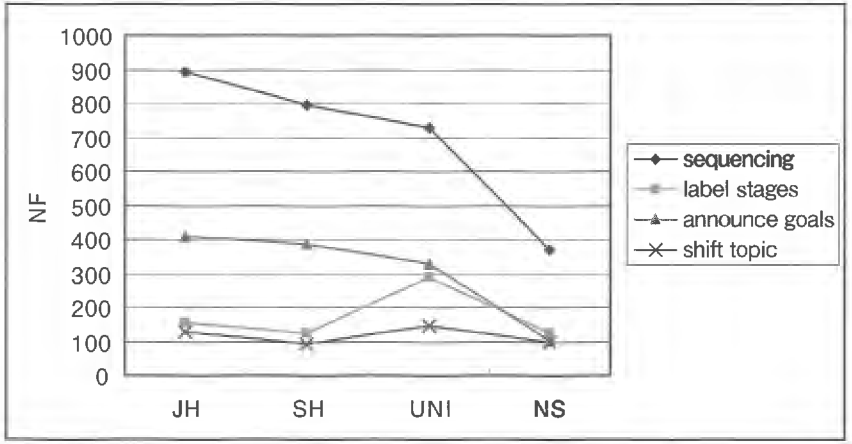

Figure 5: The distribution of four subcategories in FRM

As this figure indicates, sequencing is the most common subcategory, and its frequency accounts for approximately 50% of word-tokens of all items in FRM. Moreover, novice non-natives, especially JH students, tend to overuse sequencing markers, as can be seen in the following essay. In the JEFLL Corpus, misspellings produced by non-natives are left for error analysis.

I'll bring our [JP: yokin tucho] **first**.
I'll bring our [JP: arubamu] **second**.
I'll bring our [JP: kane] **thrid**.
I'll bring our [JP: sohuto] **seventh**.
I'll bring our [JP: huku] a bittle **fourth**.
I'll bring our computer **fifth**.
I'll bring our [JP: tyotto no] food **sexth**.
These are in some of my fag. (JH)

In this essay, all of the sentences except for the final include sequencing markers although it is not clear why there is the "seventh" in the fourth sentence. Perhaps "thrid" and "sexth" are misspellings of "third" and "sixth" respectively. The writer of this essay may try to compose logically, but his or her attempt proves fruitless. Such an overuse is also reported by Tankó (2004) who investigates the use of adverbial connectors in Hungarian university students' essays. It often results from superficial attention to logical forms (Intaraprawat and Steffensen 1995: 271), and results in "artificial, mechanical prose" (Zamel 1983: 27).

4.2.3. Self-mentions

Self-mentions (SEM) refer to the degree of explicit author presence in the text measured by the frequency of first-person pronouns and possessive adjectives (Hyland 2005: 53). As shows in Table 3, the frequency of SEM decreases significantly as the proficiency level of English raises (chi-square = 20569.21, df = 3, $p < .001$). Subsequently, the results of *post hoc* pairwise comparisons show that statistical differences are found between JH and SH ($p < .001$), JH and UNI ($p < .001$), JH and NS ($p < .001$), SH and UNI ($p < .001$), SH and NS ($p < .001$), and UNI and NS ($p < .001$). That is to say, non-natives overuse significantly SEM in comparison to natives.

The corpora employed in this study include seven items in SEM: "I," "my," "me," "mine," "we," "our," and "us." The distribution of seven items appears in Figure 6.

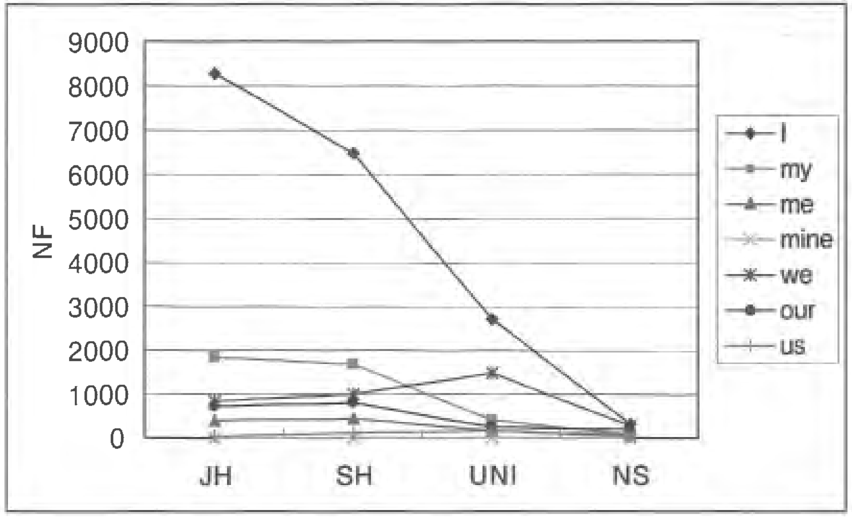

Figure 6: The distribution of seven items in SEM

This figure illustrates that the frequency of "I" decreases steeply as the proficiency level of English raises. Consider an essay written by a JH student for example.

I often eat rice in the morning.
I drink milk everyday.
I like milk very much.
I sometimes eat bread.
I like bread a lettle.
I eat breakfast everyday. (JH)

What this example makes clear is that novice non-natives tend to overuse "I" in the sentence-initial position. Such an overuse is also reported by Kobayashi and Yamada (2008) who examine the use of metadiscourse markers in the spoken corpus of Japanese EFL learners. Furthermore, it is not particular to Japanese speakers of English, and it is fairly common among non-natives (Petch-Tyson 1998). As Biber *et al.* illustrate statistically that first person pronouns are very frequent in spoken language, and less frequent in academic writings (1999: 333–334), non-natives write English as if they were speaking.

5. Final remarks

This study has shown that, as for the use of metadiscourse markers, there is a difference between natives and non-natives. Non-natives, especially novice ones, can use very few varieties of metadiscourse markers so that they overuse or underuse significantly the most of metadiscourse categories in comparison to natives. However, this study has been exploratory in character and its limitations are obvious. In this study, my approach has been mainly quantitative, and categories examined qualitatively are only three: TRA, FRM, and SEM. While automated quantitative analysis is "a very accurate quick 'way in' for any researchers confronted with large quantities of data with which they are unfamiliar" (Thomas and Wilson 1996: 106), it calls for qualitative analysis which can offer a rich and detailed perspective on the data.

Finally, to investigate the characteristics of non-native English is informative for compiling ELT (English Language Teaching) dictionaries. In 1987, Longman started collecting samples of learners' writing to build a corpus of learners' English. This corpus (the Longman Learners' Corpus, LLC) was intended to help in compiling ELT dictionaries, such as the third edition of *Longman Dictionary of Contemporary English* published in 1995 (e.g. Gillard and Gadsby 1998). Error information given in most ELT dictionaries is limited to word-level errors, such as collocational errors, spelling errors, or errors over countability. A further direction of ELT dictionaries will be to give some information on pragmatic / discoursal errors shown partly in this study.

6. References

Ädel, A. (2006) *Metadiscourse in L1 and L2 Writing*. Amsterdam: John Benjamins.

Altenberg, B., and M. Tapper (1998) "The Use of Adverbial Connectors in Advanced Swedish Learners' Written English." In Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. 80–93.

Biber, D., S. Conrad, and R. Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999) *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Conrad, S. (2004) "Corpus Linguistic Approaches for Discourse Analysis." *Annual Review of Applied Linguistics* 22: 75–95.

Crewe, W. J. (1990) "The Illogic of Logical Connectors." *ELT Journal* 44: 316–325.

Crismore, A., R. Markkanen, and M. Steffensen (1993) "Metadiscourse in Persuasive Writing: A Study of Texts Written by American and Finnish University Students." *Written Communication* 10: 39–71.

Crystal, D. (1991) "Stylistic Profiling." In Aijmer, K., and B. Altenberg (eds.), *English Corpus Linguistics*. London: Longman, pp. 221–238.

Gillard, P., and A. Gadsby (1998) "Using a Learners' Corpus in Compiling ELT Dictionaries." In Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. 159–171.

Granger, S. (1998) "The Computer Learner Corpus: A Versatile New Source of Data for SLA Research." In Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. 3–18.

Granger, S., and P. Rayson (1998) "Automatic Profiling of Learner Texts." In Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. 119–131.

Hyland, K. (1998) "Exploring Corporate Rhetoric: Metadiscourse in the CEO's Letter." *The Journal of Business Communication* 35: 224–246.

Hyland, K. (1999) "Talking to Students: Metadisourse in Introductory Coursebooks." *English for Specific Purposes* 18: 3–26.

Hyland, K. (2000) *Disciplinary Discourses: Social Interactions in Academic Writing*. London: Longman.

Hyland, K. (2004) "Disciplinary Interactions: Metadiscourse in L2 Postgraduate Writing." *Journal of Second Language Writing* 13: 133–151.

Hyland, K. (2005) *Metadiscourse: Exploring Interaction in Writing*. New York: Continuum.

Hyland, K., and P. Tse (2004) "Metadisourse in Academic Writing: A Reappraisal." *Applied Linguistics* 25: 156–177.

Intaraprawat, P., and M. Steffensen (1995) "The Use of Metadiscourse in Good and Poor ESL Essays." *Journal of Second Language Writing* 4: 253–272.

Kobayashi, Y., and H. Yamada (2008) "The Use of Metadiscourse Markers in Japanese EFL Learners' English." *English Corpus Studies* 15: 161–173 (in Japanese).

Krzeszowski, T. (1990) *Contrasting Language: The Scope of Contrastive Linguistics*. Berlin: Mouton de Gruyter.

Larsen-Freeman, D. (1983) "The Importance of Input in Second Language Acquisition." In Andersen, R. (ed.), *Pidginization and Creolization as Language Acquisition*. Rowley: Newbury House, pp. 87–93.

Leech, G. (1998) "Preface." In Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. xiv–xx.

Petch-Tyson, S. (1998) "Writer / Reader Visibility in EFL Written Discourse." In Granger, S. (ed.), *Learner English on Computer*. London: Longman, pp. 107–118.

Tabata, T. (2002) "Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution." In Saito, T., J. Nakamura, and S. Yamazaki (eds.), *English Corpus Linguistics in Japan*. Amsterdam: Rodopi, pp. 165–182.

Tankó, G. (2004) "The Use of Adverbial Connectors in Hungarian University Students' Argumentative Essays." In Sinclair, J. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Thomas, J., and A. Wilson (1996) "Methodologies for Studying a Corpus of Doctor-Patient Interaction." In Thomas, J., and M. Short

(eds.), *Using Corpora for Language Research*. London: Longman, pp. 92–109.

Tono, Y. (2002) *The Role of Learner Corpora in SLA Research and Foreign Language Teaching: The Multiple Comparison Approach*. Unpublished Ph.D Dissertation. Lancaster: Lancaster University.

Vande Kopple, W. J. (1985) "Some Explanatory Discourse on Metadiscourse." *College Composition and Communication* 36: 82–93.

Zamel, V. (1983) "Teaching Those Missing Links in Writing." *ELT Journal* 37: 22–29.