A Discriminant Analysis of Non-native Speakers and Native Speakers of English[1]

Masatoshi Sugiura,[2] Masumi Narita,[3] Tomomi Ishida[2]
Tatsuya Sakaue,[2] Remi Murao[4] and Kyoko Muraki[2]

1. Introduction

Since the late 1980s learner corpus analyses have greatly contributed to our understanding of the characteristics of L2 learners' interlanguage. A variety of large-scale learner corpora have been increasingly compiled and processed using powerful computer technology, thus accelerating second language acquisition (SLA) research. Examining the potential of corpus-based SLA research, Granger (2002) proposed two methodological approaches: Contrastive Interlanguage Analysis and Computer-aided Error Analysis. Both approaches can highlight quantitative and qualitative comparisons between non-native and native languages, or among L2 learners with different L1 backgrounds. The results of these comparisons seem to be convincing and generalizable because they are based on a large amount of systematically analyzed L2 learners' data.

1.1 Well-designed learner corpus: Learner corpus 1.0

Our 20-year history of learner corpus research, however, suggests that corpus design criteria need to be reconsidered because they are likely to vary depending on not only the corpus builder's research purposes, but also the possible constraints on data collection. In her pioneering book in 1998, Granger pointed out that "... in other words the quality of the investigation is directly related to the quality of the data. It is especially important to have clear design criteria in the case of learner language, which is a very heterogeneous variety: there are many different types of learners and learning situations" (1998: 7). Nesselhauf also mentioned that "whether the full

[1] NICE: Learner Corpus 2.0 to come
[2] Nagoya University
e-mail: sugiura@nagoya-u.jp, ishida_tomomi@nagoya-u.jp, sakaue@nagoya-u.jp, muraki@nagoya-u.jp
[3] Tokyo International University
e-mail: mnarita@tiu.ac.jp
[4] Waseda University
e-mail: murao@aoni.waseda.jp

potential of learner corpora can be used, critically depends on the availability of well-designed corpora. But despite the fact that a number of such corpora already exist, there is still great scope for further corpora and for improvement of the existing ones" (2004: 132) It is evident that even in the early days of learner corpus-based research, both of these researchers emphasized the significance of well-defined corpus design criteria for extracting meaningful results from corpus data.

Based on the suggestions of Atkins *et al.* (1992), Granger launched the International Corpus of Learner English (ICLE) project, which has collected essay data written by English learners in a number of countries. The ICLE has played an important role in the development of learner corpus research. Its design criteria are shown in Table 1, and it includes two kinds of features: six attributes common to all the ICLE project's subcorpora, seven other attributes that can vary in each subcorpus.

Shared features	Variable features
Age	Sex
Learning context	Mother tongue
Level	Region
Medium	Other foreign languages
Genre	Practical experience
Technicality	Topic
	Task setting

Table 1: ICLE design criteria (Granger, 1998: 9)

1.2 Problematic attributes in the ICLE's data

When you try to use the ICLE data in practice for SLA research, however, you may become entangled in its problematic attributes, such as "level," "topic" or "task setting." These attributes represent, respectively, the subjects' level of English proficiency, the topic they are expected to write about and the conditions under which the data are collected. The ICLE project allegedly targeted subjects with an advanced level of proficiency, "a notion which is defined on the following external ground: they are university undergraduates in English Language and Literature in their third or fourth year" (Granger, 1998: 10). The problem is that it is difficult to ensure that those who meet this criterion are always advanced English learners.

"Topic" is a linguistic attribute pertaining to the subjects' lexical choices. The ICLE data cover a wide variety of 922 topics. Because the types of content words in "nuclear power" essays and "learning English" essays, for example, are quite different from each other; not only content words, but also function words, such as modal

auxiliaries, are sometimes affected by different topics (Aijmer, 2002: 60). Because the topics are so various, it is difficult to draw a conclusion about their effect on essay writing.[5]

"Task setting" is also problematic because the settings can vary in a way that requires some subjects to write timed essays and others to write non-timed essays (e.g., take-home essays), or that allows some subjects, but not others, to use reference materials. Variations in the task setting greatly affect L2 writing performance. The attributes of the data produced by the same L1 English learners can differ if some of the task settings differ. In order to illustrate this point, we have searched the ICLE database. Table 2 shows the differences in the average number of words (tokens) in argumentative essays written by the same L1 English learners, from the ICLE data. Whether the essay writing is timed or not does affect the results. By the same token, the use of reference tools also affects the result greatly, as shown in Table 3.

L1	Timed (N)	Not Timed (N)
Dutch	1461.5 (11)	815.7 (215)
German	387.0 (182)	617.0 (216)
Swedish	581.5 (251)	598.6 (155)

Table 2: Average number of words written under different conditions (timed and not timed)

L1	With references (N)	Without references (N)
Dutch	863.8 (226)	722.2 (11)
German	623.8 (215)	415.6 (194)
Swedish	605.2 (115)	579.8 (255)

Table 3: Average number of words written under different conditions (with and without references)

If learners are allowed to use as much time as they want to write an essay, the essay becomes longer than timed essays. In the case of Dutch students in Table 2, on the contrary, their timed essays are much longer than their untimed essays, which seems rather ridiculous. These data should have included some special factors. Reference tools can also boost the number of words. Thus, task settings are regarded as important factors which affects the corpus data's attributes. Therefore, grouping all

[5] The ICLE data may not distinguish "topics" from "titles of the essays." If not, the notion of "topic" in the ICLE seems rather vague as a criterion.

the data together because they came from the same L1, regardless of task setting differences, is too rough to allow analysis of the corpus as SLA data. In SLA research, data control has always been the key to the analysis.

1.3 Well-controlled learner corpus: Learner corpus 2.0

As we have reviewed above, the ICLE can be regarded as well-designed, but its data are not as well-controlled as SLA data. In order to pursue SLA research using learner corpora, we need well-controlled corpus data, which can be called "Learner corpus 2.0."

Based on the ICLE project's advancements, we have decided to compile a new learner corpus, the Nagoya Interlanguage Corpus of English (NICE). This is a collection of argumentative essays produced by Japanese learners of English in which writing topics and task settings are well-controlled. The subjects' English proficiency level is to be precisely recorded by means of standard English test scores such as TOEIC (Test of English for International Communication) scores, as long as the students have a score. The next section gives a detailed description of the NICE learner corpus and its comparable corpus of native English speakers' argumentative essays. If the data are well-controlled, it is much easier to compare factors which may affect the students' performance on the writing activities.

In order to demonstrate the characteristics of our NICE learner corpus, we would like to examine the following research questions:

1. How effective is the NICE learner corpus at identifying the distinctive features between learners' and native speakers' language use?
2. Does controlling English proficiency levels clarify these distinctions?
3. Does controlling Topic differences clarify these distinctions?

The paper is structured as follows: the second section presents how the NICE and its control corpus of native English speakers were designed and constructed. The third section describes our discriminant analysis of essay data from non-native speakers and native speakers of English, and presents our statistical findings. The fourth section discusses our research findings, and the concluding section summarizes the present study.

2. Nagoya Interlanguage Corpus of English (NICE)

Based on the critical review above of the ICLE's corpus design , we have compiled a new English learner corpus, NICE-NNS, and a comparable corpus of native English speakers, NICE-NS. Each corpus consists of 200 essays. The task settings are: 1) timed, 60 minutes and 2) no reference tools; thus, all of the essays are collected under the same conditions. The summary of the corpus' size is shown in Table 4.

	NICE-NNS	NICE-NS
Total number of words	69,858	117,571
Total number of essays	207	200
Average number of words/essay	337	588

Table 4: Summary of the NICE

2.1 Corpus design

As the previous section pointed out, the ICLE data have three major problems in their attributes; 1) proficiency level, 2) topic and 3) task setting. First, although the ICLE user manual states that the learners are at an advanced proficiency level, none of the proficiency tests objectively evaluates them. Simply assuming that all third and fourth year university undergraduates majoring English Language and Literature have advanced English proficiency may be an overgeneralization. In order to estimate the learners' English proficiency level, NICE-NNS data include a description of all learners' TOEIC, TOEFL or STEP score if they have taken these tests. Also recorded is detailed information concerning students' a) length of English study, b) length of stay abroad, c) daily use of English reading, writing, speaking and listening, d) experience of English essay writing, and e) self evaluation of their ability to write English essays. Table 5 shows the learners' and native speakers' attributes.

NICE-NNS	NICE-NS
1) English study history 2) Language other than English 3) Length of studying other language 4) Qualifications: TOEIC, TOEFL, STEP 5) Experience going abroad 6) Daily amount of English reading, writing, listening, speaking 7) Essay writing (in Japanese or English) proficiency self-estimation Japanese essay	1) Mother tongue: British, American, Canadian, Australian, etc. 2) Parents' mother tongue 3) Academic background 4) Foreign language learning experience 5) Essay writing proficiency self-estimation

Table 5: Attributes of NICE-NNS and NICE-NS

Second, because ICLE data include as many as 922 different topics, it is difficult to compare the language used between the topics. By limiting the number of topics to eleven, NICE data, in this respect, use a considerable number of essays written on the same topic, which enables us to directly compare language usage between learners and native speakers, as well as between topics. The eleven topics and the number of files are shown in Table 6. These topics have been decided based on the comparative consideration among the topics found in LOCNESS, ICLE and TOEFL TWE, avoiding the influence of cultural background.

Topic	Files	Ratio (%)
sports	61	29.5
school education	51	24.6
money	19	9.2
violence on TV	15	7.2
recycling	13	6.3
death penalty	13	6.3
suicide	10	4.8
divorce	8	3.9
crime	7	3.4
teenagers	5	2.4
water pollution	5	2.4
TOTAL	**207**	**100**

Table 6: The number of files and ratio of eleven topics

Third, ICLE data were collected under different task settings; that is, some essays were written as in-class examinations, and others were written as homework. Furthermore, some essays were written under time pressure, others were not, and the use of reference materials or dictionaries was not fully controlled. These essay data differ in the kind of language performance they reflect, which again makes it difficult to directly compare between any suggested groups. Aiming to collect data which reflect subjects' language output from their conceptual knowledge, the use of dictionaries and other resources is restricted in the NICE data. By not letting students use dictionaries, we can reveal their language production ability. The time for essay writing is also controlled at 60 minutes, which enables us to compare and analyze the amount of words produced when all students are under the same time pressure.

In summary, for any particular SLA research, because these variables easily affect learners' outputs, the NICE controls them.

2.2 Data collection

The NICE's output data were collected in an experimental context. The subjects were 207 Japanese university undergraduate and graduate students from all majors. As for the native speaker data, 200 files were gathered from universities and agencies. After signing the contract, subjects were asked to choose one of the eleven topics that seemed easy for them to write about. The writing time was 60 minutes. *Microsoft Word* was adopted for the writing task; the only tool students were allowed to use was the spell checker, and the use of any other referential tools, including dictionaries, was strictly prohibited. When writing unknown English words in Japanese, subjects were told to write them in the Roman alphabet. One serious problem with the ICLE data is that because the ICLE is a collection of essays written by learners from many different language backgrounds, the character code is not consistent throughout the data. The mixture of two or more character codes makes it extremely difficult to process the language data. For this reason, NICE data were collected with special care for the character code, excluding the two- byte characters used for Japanese. After completing their essays, the subjects answered a questionnaire concerning their language background, which later was encoded in each datum of the subjects' attributes, as described in Table 5. The experiment took approximately 90 minutes.

2.3 Text formatting

All the text in Microsoft Word files was revised as follows:

1. Spelling by the spell check tool
2. Converting two-byte characters, if any, into single-byte characters

As a second step in text formatting, *Microsoft Word* files were converted into plain text files for further data analysis.

Text files were formatted to be one sentence per line (CHAT-format). In the CHAT-format, each line begins with either @, * or %.[6]

- @ line: a header which gives information about the participants and the task setting
- * line: text that the participants produced
- % line: additional information (e.g., %par stands for paragraph delimiter)

[6] This format is basically the same as CHILDES' CHAT format. See MacWhinney (2000).

```
@Begin
@Participants:  JPN201
@Age:   23
@Sex:   F
@YearInSchool:  M1
@Major: other
@StudyHistory:  10
@OtherLanguage: French=4.0;Chinese=2.0
@Qualification: TOEIC=915(2005);TOEFL=570(2004);none=
@Abroad:       UK=1.0;none=
@Reading:      5
@Writing:      3
@Listening:    4
@Speaking:     1
@JapaneseEssay: 3
@EnglishEssay:  3
@Difficulty:   3
@Topic: death penalty
@Comments:
@Coder: 2006-10-11 DataInputBy SAKAUE Tatsuya;

*JPN201: Is Death Penalty Really Necessary?

%par:

*JPN201:  Some people say that death penalty is necessary, others not.

*JPN201:  It is quite controversial to discuss death penalty for it is a matter
of ethics or values that one nation or one culture has.

*JPN201: Japan is the country that still carries out death penalty.

…

@End
```

Figure 1: Sample data from the NICE-NNS

3. Analysis

Because the two corpora, NICE-NNS and NICE-NS, are regarded as two sets of data produced by two groups whose membership is known beforehand, i.e., non-native speakers and native speakers, it is possible to build a predictive model based on the data's characteristics by means of discriminant analysis, which is a kind of multivariate analysis.

In order to examine our research question, we have conducted the following three kinds of discriminant analyses:

1. NICE-NNS to NICE-NS as a whole
2. NNS subgroups divided by English Proficiency test
   (TOEIC level A, B and C) to NS
3. Subgroup of NNS to NS with limited number of topics;
   namely, "school education" and "sports"

For the analysis we have selected the following attributes of the two corpora as variables for the analysis:

1. Type
2. Token
3. Type/token ratio (TTR)
4. Number of sentences (Ss)
5. Average Word Length (AWL)
6. Average Sentence Length (ASL)

All these are mechanical text features. The reason for this selection is twofold. First, the calculation is easy. Second, we would like to try to find some simple features, if possible, which can identify differences in text between non-native speakers and native speakers.

Basically, all the procedures of the discriminant analyses we have conducted this time are the same.[7] A forward stepwise procedure was used to select the variables with a *P* value less than 0.01.[8]

3.1 Result: NICE-NNS to NICE-NS as a whole

First, we conducted a discriminant analysis for all the data to figure out the general tendencies of the corpus data. Table 7 shows the averages of the variables.

	NNS(207)	NS(200)
Type	131.84	247.06
Token	337.48	587.86
TTR	0.41	0.43
Ss	26.14	31.36
AWL	4.45	4.56
ASL	13.50	20.05

Table 7: Averages of variables for all the data

Based on the result of the forward stepwise procedure of the discriminant

[7] We used an Excel macro program created by Dr. Aoki at Gunma University, Japan (http://aoki2.si.gunma-u.ac.jp).

[8] Because some of the data do not follow a normal distribution, non-linear discriminant analysis is theoretically preferable to our method. We did try non-linear discriminant analysis, and found its results to be virtually the same as ours, with a *P* value less than 0.01. Thus, for convenience, we used linear discriminant analysis here.

analysis, two variables, Type and Token, turned out to be statistically significant to distinguish non-native speakers from native speakers, as shown in Table 8.

	Partial *F*-value	*p*-value	discriminant coefficient
Type	402.46	< 0.001	-0.14
Token	47.19	< 0.001	0.02
constant			17.91

Table 8: Variables selected by discriminant analysis for all the data

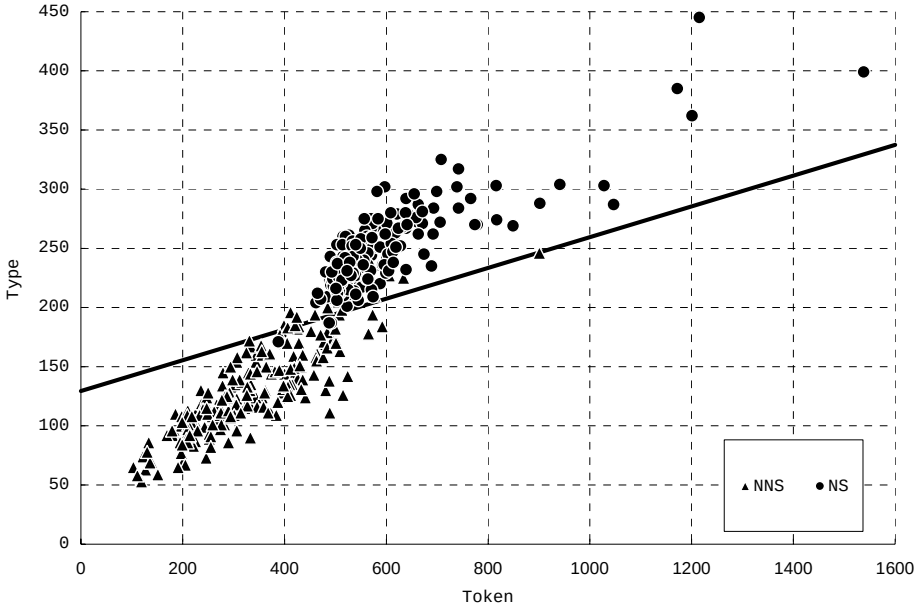

Figure 2: Scatter diagram of all data and the discriminant function

With this result the discriminant function can be obtained below, and this function can accurately classify 94.8 percent of the cases.

Z = -0.14 * Type + 0.02 * Token + 17.91

3.2 Result: English proficiency levels

Out of the 207 NICE-NNS data, 121 NNS data have TOEIC scores. We classified the 121 data into three subgroups based on the TOEIC score classification: 990 - 860 for level A, 855 - 730 for level B, and 725 - 470 for level C.[9]

[9] Actually, ten subjects have TOEIC scores below 470, and are excluded from this analysis.

	NNS TOEIC score holders				NS(200)
	All (121)	Level			
		A (33)	B (19)	C (59)	
Type	144.05	168.12	172.63	129.12	247.06
Token	359.34	412.09	429.58	326.93	587.86
TTR	0.42	0.42	0.41	0.42	0.43
Ss	25.83	22.33	27.95	27.14	31.36
AWL	4.50	4.70	4.51	4.43	4.56
ASL	14.68	19.07	16.36	12.55	20.05

Table 9: Average of variables for the three levels of NNS data

We conducted the following four discriminant analyses between each and all of the three NNS subgroups and NS:

Case 1: All NNS TOEIC score holders vs. NS
Case 2: NNS Level A vs. NS
Case 3: NNS Level B vs. NS
Case 4: NNS Level C vs. NS

3.2.1 Case 1: All NNS TOEIC score holders

In Case 1, Type and Token were selected by means of a forward stepwise procedure with a *P* value of 0.01. The discriminant function for Case 1 is as follows:

Z = -0.13 * Type + 0.02 * Token + 17.28

Its classification accuracy rate is 95.3 percent. This result is quite similar to the result of the discriminant analysis between all of the 207 NNS data and the NS data. Type and Token are the two variables which discriminate non-native speakers from native speakers.

3.2.2 Case 2: Level A

As well as Type and Token, TTR was selected in the second case. The discriminant function is as follows:

Z = -0.17 * Type + 0.04 * Token + 41.60 * TTR -0.10

This function's classification accuracy rate is 94.0 percent.

3.2.3 Case 3: Level B

TTR was also selected in Case B. The discriminant function is as follows:

Z = -0.22 * Type + 0.05 * Token + 66.65 * TTR -8.62

The classification accuracy rate is 95.9 percent.

3.2.4 Case 4: Level C

Instead of TTR, Case 4 selected the average number of sentences (Ss) and the average sentence length (ASL), in addition to Type and Token. The discriminant function is as follows:

Z = -0.17 * Type + 0.03 * Token -0.50 *ASL -0.23 * Ss + 31.92

The classification accuracy rate is 97.7 percent.

3.3 Result: Topics

Because "School Education" and "Sports" were most popular among the eleven topics for learners of English, we selected the corpus data written about these two topics and conducted discriminant analyses. If the topics are all the same for both NNS and NS, the factor "different topics," which may affect writing production, can be ignored, and we can focus on the remaining factors.[10]

[10] In order to compare the differences between the two topics strictly, we used the most strict sets of data, where the same seventeen writers in each group, NNS and NS, produced both essays on "School Education" and "Sports."

3.3.1 “School Education”

	NNS(17)	NS(17)
Type	150.82	249.06
Token	387.94	581.82
TTR	0.40	0.44
Ss	22.94	31.53
AWL	4.81	4.74
ASL	17.27	19.85

Table 10: The average of variables for “School Education” data

The result of the discriminant analysis for “School Education” data seems interesting, because only Type was selected. The discriminant function is as follows:

Z = -0.09 * Type + 17.65

The classification accuracy rate is 94.1 percent.

3.3.2 “Sports”

The data concerning “Sports” lead to similar results.

	NNS(17)	NS(17)
Type	153.18	250.29
Token	368.76	587.71
TTR	0.43	0.43
Ss	23.35	31.12
AWL	4.54	4.50
ASL	16.89	20.63

Table 11: The average of variables for “Sports” data

The discriminant function is as follows:

Z = -0.11 * Type + 22.49

The classification accuracy rate is 94.1 percent.

4. Discussion

Through the series of discriminant analyses above, it is obvious that Type and Token are the key variables among the six examined with the NICE-NNS data and the NICE-NS data. In the detailed analysis from the point of view of learners' English proficiency levels in the three subgroups of NNS data, TTR was additionally selected for levels A and B. In the case of level C, however, two variables concerning sentences were selected, instead of TTR.

4.1 TTR

The reason that TTR was selected for the advanced and intermediate learners is not clear. Actually, the average TTR scores for levels A, B and C, and NS, are 0.42, 0.41, 0.42 and 0.43, respectively. These seem almost the same. Although TTR has been used as an index to indicate the lexical richness of text, many researchers have also criticized the reliability of this index. The so-called MTTR, Mean TTR, is sometimes used instead of TTR. For example, Meunier (1998: 32) pointed out the problem with TTR by saying, "The shorter the text, the higher the type/token ratio," and discussed the possibility of using MTTR. In the present study, however, MTTR is not appropriate because the texts or essays are not long enough to be cut into smaller chunks to calculate their means.[11] Vermeer (2000) recommended the Guiraud index as an adequate measure of lexical richness.[12] The comparison of these indices, however, is one of the future research themes.

4.2 Precise English proficiency levels

One interesting point about the results of the discriminant analyses in terms of the three levels of English proficiency is that, in the case of level C, the two variables concerning sentences are selected. This means that the beginning levels of NNS and NS can be distinguished by taking a look at the features concerning sentences. This is naturally true, because beginners usually cannot produce long sentences. For all levels of learners, two lexical variables, Type and Token, are distinctive features; but for low level learners, sentence production itself seems difficult so the products of writing, that is, essays, must be shorter than the more advanced groups. Actually, the average

[11] Meunier suggests taking every 1,000 running words as a unit.
[12] The Guiraud index can be calculated by dividing the number of types by the square root of the number of tokens.

numbers of sentences produced among the tree subgroups of learners are almost the same, but the average sentence lengths are quite different, as shown in Table 9.

In order to scrutinize the differences among the three subgroups of learners, let us conduct another discriminant analysis here. This time, without the NS data, a multiple discriminant analysis with the combinations of Levels A and B, Levels B and C, and Levels C and A should be used. The result of this analysis can be summarized in the following three discriminant functions:

1. Between Level A and B: Z = -0.20 * ASL + 0.01 * Type + 2.34
2. Between Level B and C: Z = -0.23 * ASL + 0.04 * Type + 9.18
3. Between Level C and A: Z = -0.43 * ASL + 0.03 * Type + 11.52

The classification accuracy rate is 64.0 percent. Thus, all of the three NNS subgroups divided by English proficiency level can be distinguished by two variables: ASL and Type. In other words, the higher the proficiency level, the longer sentences the learners can write, and the more different words they can use.

The classification accuracy rates among the three NNS subgroups are interesting, too. The lower the level, the higher the classification accuracy rate, which means the lower the level, the easier it is to distinguish the learners' data from native speakers'.

4.3 Controlled topics

The discriminant analyses using the topic-controlled data show a rather simple and consistent result. The only variable selected by the analysis was Type. In addition, although the two topics are different from each other, the results of the analysis are almost the same. This indicates that, regardless of the topic, the important variable to distinguish learners from native speakers is their variety of words. If the topics are different, the words used may vary, so it is difficult to narrow down the possibilities. However, if we can use data obtained under the same conditions (in this case, the same topic), we can exclude some unnecessary possibilities.

5. Conclusion

This paper has demonstrated several discriminant analyses based on data from the well-controlled learner corpus, NICE. After critically reviewing the former learner corpus studies, the outline of the newly designed learner corpus, NICE, was described.

The third section reported on the three kinds of discriminant analyses we conducted. First, the general tendency was analyzed using all the data from NICE, when Type and Token were selected as the distinctive features. The second analysis was conducted with precise English proficiency level data. On one hand we found that, for all the proficiency levels, Type and Token, again, are the two variables that consistently distinguish general learners from native speakers. On the other hand, only the beginners showed distinctive features concerning sentences. The third analysis was concerned with the controlled topics, and it found that only Type could significantly distinguish NNS from NS if the topics were the same.

Section 4, based on the results, first discussed the problematic characteristics of TTR and introduced some alternatives, such as MTTR and the Guiraud index; but the analysis using these measures is left to future studies. After that, in order to scrutinize differences in the three NNS subgroups' proficiency levels, a multiple discriminant analysis was conducted, which found that two variables, ASL and Type, could distinguish differences in the learners' proficiency levels. Last, the discussion section emphasized the importance of controlling the variables, and gave controlled topics as examples.

Throughout this paper we have been insisting on just one point: the importance of controlling the variables in SLA research using learner corpora. We need to pay close attention to the attributes recorded in the corpus data, because SLA is a field of research that analyses not just recorded words, but also the process by which human beings acquire a second language.

After twenty years of learner corpus research, we are expecting Learner Corpus 2.0, the second generation, to come. We hope that this paper can demonstrate the possibilities of SLA research using a Learner Corpus 2.0, NICE.[13]

References

Aijmer, K. (2002) Modality in Advanced Swedish Learners' Written Interlanguage, in Granger, S. *et al.* (eds) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 56–76. Amsterdam: John Benjamins Publishing.

Aston, G., S. Bernardini and D. Stewart (eds) (2004) Corpora and Language Learners. Amsterdam: John Benjamins.

Atkins, S., J. Clear, and N. Ostler (1992) 'Corpus design criteria'. *Literary and*

[13] The Nagoya Interlanguage Corpus of English, NICE, will be available on the web in 2008 at http://bill.gsid.nagoya-u.ac.jp.

Linguistic Computing vol. 7 (1), 1–16.

Douglas, F. (2003) 'The Scottish corpus of texts and speech: Problems of corpus design'. *Literary and Linguistic Computing vol. 18 (1)*, 23–37.

Garside, R., G. Leech and T. McEnery (eds) (1997) Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Addison Wesley Longman.

Granger, S. (2002) A bird's-eye view of learner corpus research, in Granger, S. *et al.* (eds) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, pp. 3–33. Amsterdam: John Benjamins Publishing.

Granger, S. (ed) (1998) Learner English on computer. London: Addison Wesley Longman.

Granger, S., J. Hung and S. Petch-Tyson (eds) (2002) Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam: John Benjamins.

Granger, S., E. Dagneaux, and F. Meunier (eds) (2002) International Corpus of Learner English. (CD-ROM and Handbook). Presses universitaires de Lovain.

MacWhinney, B. (2000) The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum Associates.

Meunier, F. (1998) Computer Tools for the Analysis of Learner Corpora, in S. Granger (ed.) Learner English on Computer. London, pp. 19–37. London: Addison Wesley Longman.

Nesselhauf, N. (2004) Learner Corpora and their Potential for Language Teaching, in J. M. Sinclair (ed.) How to Use Corpora in Language Teaching, pp. 125–52. Amsterdam: John Benjamins.

Renouf, A. (ed.) (1998) Explorations in Corpus Linguistics. Amsterdam: Rodopi.

Sinclair, J. M. (ed.) (2004) How to Use Corpora in Language Teaching. Amsterdam: John Benjamins.

Vermeer, A. (2000) 'Coming to grips with lexical richness in spontaneous speech data'. *Language Testing 17 (1)*, 65–83.