Barbara Lewandowska-Tomaszczyk
Patrick James Melia
(eds.)

# PALC'99: Practical Applications in Language Corpora

Papers from the International Conference
at the University of Łódź, 15-18 April 1999

Offprint

2000

PETER LANG
Europäischer Verlag der Wissenschaften

# A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora

YUKIO TONO

## 1. Introduction

In the past several years, on-line corpora and corpus analysis tools have become increasingly accessible, and corpus-based research has become increasingly common not only in the field of linguistics but also in the field of second language acquisition and English language teaching. General corpora have been used for learners' dictionary making (cf. COBUILD project), evaluating ELT materials (cf. Kennedy 1987; Mindt 1992; Ljung 1990), grammar books (cf. COBUILD grammar series; Biber *et al.* 1999) and classroom teaching as well (cf. Bernadini 1998).

There has been an increasing awareness among second language acquisition researchers that the computational processing of large quantity of language performance data will shed new light on previously intractable research questions in this area. Thus more and more attention has been recently paid to compiling a corpus of learner language (cf. Granger 1998). The first international symposium of computerized learner corpora was held in Hong Kong in December, 1998, which indicates clearly that corpus-based research in second language acquisition will contribute significantly to our understanding of second language acquisition or learning processes.

In this study, I would like to present the results of my corpus-based analysis of the interlanguage development by English as a foreign language (EFL) learners. I will focus especially on the analysis of part-of-speech tag sequences found in developmental learner data. Automatic grammatical tagging and automatic acquisition of tag n-gram statistics has made it possible to describe the developmental patterns of learner language in the manner that was not thought of until a decade ago. This study is basically a replication of the pioneering study conducted by Jan Aarts and Sylviane Granger (1998). While their primary interest is on the difference between advanced learners' writing and native speakers' writing, my main focus in this paper is on the differences between learners at different developmental stages.

Aarts and Granger (1998) explained the rationale for this study as follows: "If tag sequences can help us discover writers' fingerprints (here they were talking about stylometry), one can assume that they can help uncover EFL learners' fingerprints." (ibid.:132) They decided to compare tag trigrams (i.e. sequences of three tags in succession) in learner and native corpora with a view to discovering distinctive interlanguage patterns. They compared four similar-sized corpora of c. 150,000 words. Three of those corpora were from the International Corpus of Learner English (ICLE) database and contained argumentative essay writing by Dutch, Finnish and French-speaking advanced learners of English. The fourth corpus was an extract from a native speaker corpus, called the LOCNESS corpus. This covers similar types of writing by American students.

The four corpora were tagged with the TOSCA tagger, using the TOSCA-ICE tagset, which contains 270 different tags. For their study, however, the information contained in the TOSCA-ICE tagset was drastically reduced by replacing all the tags with a much simpler tagset. Table 1 shows the list of tags they used:

| | | | |
|---|---|---|---|
| ADJ | adjective | IT | cleft/ant/prop it |
| ADV | adverb | N | noun |
| ART | article | NUM | numeral |
| AUX | auxiliary | PREP | preposition |
| CONJUNC | conjunction | PROFM | proform |
| CONNECT | connective | PRON | pronoun |
| DISC | discourse item | PRTCL | particle |
| EXTHERE | existential there | PUNC | punctuation |
| GENM | genitive marker | V | verb |

**Table 1. List of tags used in Aarts and Granger (1998)**

Standard UNIX tools were used to extract the tag trigrams from the tagged corpora. In the output files, each trigram is presented with the following information: absolute frequency, expected frequency, relative frequency (per 100,000 words), and chi-square value. The threshold for statistical significance was set at 6.63 ($p < 0.01$). The native speaker (NS) corpus served as normative for the non-native speaker (NNS) corpora and was taken as standard for expected frequencies and chi-square values.

Aarts and Granger found possible distinguishing features of learner writing in terms of the following:

(1) top-ranking trigrams (top 20)
(2) distinctive trigrams: trigrams which have a significantly higher or lower frequency in NNS writing than in NS writing

Further details of their findings will be given in the data analysis section. It will suffice to say here that while Aarts and Granger's focus was mainly on the difference between NS and NNS writings, the primary purpose of my study is to compare advanced learners' writing with less advanced learners' in terms of the top 20 tag sequences.

## 2. Research design

### 2.1. Research questions

Following the study by Aarts and Granger (1998), I aimed to answer the following questions:

(1) Are there any differences in the frequency of part-of-speech tag trigrams in the writing by EFL learners at different proficiency levels? If so, where does this difference come from?
(2) What are distinctive interlanguage patterns like? What marks off those that are common to all learners from those that are specific to a group of learners at a particular developmental stage?

Aarts and Granger (1998) have already provided some partial answers to these questions, but the corpora used in this study were based on learners at beginning and intermediate stages. Therefore, it would be very interesting to replicate their study in order to describe the overall developmental patterns of part-of-speech tag trigrams throughout different proficiency groups for a better understanding of the entire L2 learning process.

### 2.2. Corpora

I used a subsection of the JEFLL (Japanese EFL Learner) Corpus[1]. The JEFLL Corpus consists of the three sections: a written corpus of Japanese learners of English at different proficiency levels (c. 200,000 running words), a spoken counterpart (c. 50,000 running words), and a corpus of English textbooks officially used at schools in Japan (c. 170,000 running words). The

[1] For further detail of the JEFLL Corpus, see my web page (http://www.lancs.ac.uk/postgrad/tono/).

fourth component, a Japanese corpus of essays with the same titles as those used for the written corpus, is not yet available but I would like to add such comparable L1 corpus to my data. It would be potentially useful for analysing L1 transfer effect in detail.

The written corpus section is about 200,000 words at the moment but since we have seven different subcorpora (JH1 through UNIV), each subcorpus size is much smaller. The JEFLL Corpus is constantly growing, however, in collaboration with secondary school teachers in Japan and it is one of the few developmental interlanguage corpora in the world. I should mention that the writing task given to the students is a little different from what people usually understand by the word 'essay.' It is actually a simple free composition task. Students are asked to write their opinions or ideas about a certain topic in only 20 minutes without the help of dictionaries. The task was given in class so that they could show their own writing ability, without depending on dictionaries or other persons' help.

I selected a sample of 1,709 timed essays totalling about 100,000 words. It should be noted that the writing data for SH1 and SH3 was missing this time because those data for SH1 and SH3 were taken from a private school whose academic level was not as high as the national schools, and thus might affect the overall data. The size of each corpus is rather small, but it was hoped that the number of observations for the top 20 trigrams would be large enough to yield a statistically significant result.

| Group | Age | Size | School | Task |
|---|---|---|---|---|
| JH1 | 12–13 | 8,548 | National | in-class essay w/o dictionary/ 20 min. |
| JH2 | 13–14 | 22,598 | National | in-class essay w/o dictionary/ 20 min. |
| JH3 | 14–15 | 27,596 | National | in-class essay w/o dictionary/ 20 min. |
| SH2 | 16–17 | 24,758 | National | in-class essay w/o dictionary/ 20 min. |
| UNI | 18–19 | 18,038 | Private/National | in-class essay w/o dictionary/ 20 min. |
| Total | | 101,538 | | |

**Table 2. EFLL subcorpora used for the study**

## 2.3. Data Analysis

The seven different corpora were tagged with the CLAWS tagger, using the C7 tagset[2] (Garside, Leech and McEnery 1997). The CLAWS tagger is a stochastic

[2] Granger suggested that it was dangerous to compare the tagged data using different tagsets. (personal communication) I will try to compare the data using the TOSCA-ICE tagset in the future.

tagger developed at Lancaster University. It operates with a lexicon which contains about 10,000 items and the idiom lists of word sequences. The C7 tagset contains c. 130 different tags. For this study, however, I reduced the number of tags drastically to make it comparable with the study done by Aarts and Granger (See Table 1 for the list).

The program was written in C to extract the tag trigrams from the tagged corpora[3]. In my analysis, I did not process any native corpus data for comparison. Instead an additional statistical analysis on trigram data was made. Besides the frequency analysis, the trigram statistics were further processed, using correspondence analysis (Meulman 1997). One of the goals of correspondence analysis is to describe the relationships between two nominal variables in a correspondence table in a low-dimensional space, while simultaneously describing the relationships between the categories for each variable. An analysis of contingency tables often includes examining row and column profiles and testing for independence via the chi-square statistic. However, the number of profiles can be quite large, and the chi-square test does not reveal the dependence structure. The Crosstabs procedure offers several measures of association and tests of association, but cannot graphically represent any relationships between the variables. Factor analysis is a standard technique for describing relationships between variables in a low-dimensional space. However, factor analysis requires interval data, and the number of observations should be five times the number of variables.

Correspondence analysis, on the other hand, assumes nominal variables and can describe the relationships between the categories of each variable, as well as the relationships between the variables. In addition, correspondence analysis can be used to analyze any table of positive correspondence measures. The module has been provided on SPSS 7.5 on later versions.

## 3. Results

### 3.1. Overall frequency

Table 3 shows the top 20 trigram types in the four subcorpora of the JEFLL Corpus and Table 4 shows the ICLE subcorpora along with LOCNESS Corpus (NS corpus).

[3] I owe special thanks to Izumi Tanaka for his support on tag trigram extraction.

| Rank | JH1 | f | JH2 | f | JH3 | f | SH2 | f | UNI | f |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PUNC # # | 1233 | PUNC # # | 2419 | PUNC # # | 30 | PUNC # # | 2407 | PUNC # # | 1459 |
| 2 | # # PRON | 741 | N PUNC # | 1475 | N PUNC # | 18 | N PUNC # | 1313 | N PUNC # | 808 |
| 3 | N PUNC # | 740 | # # PRON | 1104 | # # PRON | 14 | # # PRON | 1095 | ADJ N PUNC | 444 |
| 4 | # PRON V | 383 | # PRON V | 778 | # PRON V | 9 | # PRON V | 758 | CONJUNC PRON V | 437 |
| 5 | # PRON N | 324 | ART N PUNC | 383 | PRON V ADV | 5 | CONJUNC PRON V | 524 | # # PRON | 375 |
| 6 | N N PUNC | 303 | PRON V ADV | 383 | ART N PUNC | 5 | PRON V ADV | 448 | PRON V ADV | 341 |
| 7 | N N N | 301 | # # CONNECT | 373 | # # CONNECT | 4 | ADJ N PUNC | 429 | V PREP V | 311 |
| 8 | N N V | 238 | ADJ N PUNC | 369 | CONJUNC PRON V | 4 | ART N PUNC | 334 | PRON AUX V | 308 |
| 9 | ADJ PUNC # | 216 | CONJUNC PRON V | 366 | V ART N | 4 | V ART N | 332 | V ADJ N | 294 |
| 10 | PRON N N | 211 | # # ADV | 315 | ADJ N PUNC | 4 | # # CONNECT | 315 | # PRON V | 275 |
| 11 | PRON V N | 196 | V ART N | 314 | PREP ART N | 3 | PRON AUX V | 315 | PRON V PREP | 246 |
| 12 | V ART N | 190 | PRON V N | 306 | ADJ PUNC # | 3 | PREP ART N | 308 | V ADV V | 237 |
| 13 | ADV ADJ PUNC | 184 | V N PUNC | 305 | # # ADV | 3 | # # ADV | 291 | # # ADV | 232 |
| 14 | V ADV ADJ | 165 | ADJ PUNC # | 290 | N N PUNC | 3 | V ADV ADJ | 289 | # # CONNECT | 228 |
| 15 | N V N | 162 | N N PUNC | 284 | PRON V PREP | 3 | PRON V PREP | 281 | V ART N | 219 |
| 16 | # # N | 149 | PREP ART N | 279 | # CONNECT PRON | 3 | PUNC PRON V | 261 | PUNC PRON V | 216 |
| 17 | PRON N V | 137 | PRON V PREP | 267 | V ADV ADJ | 3 | ADJ PUNC # | 259 | PRON V ADJ | 212 |
| 18 | ART N N | 130 | V ADJ N | 244 | V N PUNC | 3 | N PREP N | 256 | V ADV ADJ | 196 |
| 19 | N V ART | 124 | V ADV ADJ | 235 | PRON V N | 3 | PRON V ADJ | 254 | # # N | 194 |
| 20 | V N N | 118 | V ADV V | 232 | PRON AUX V | 2 | PREP N PUNC | 253 | ADJ PUNC # | 193 |

**Table 3. Top 20 trigrams of JEFLL Corpus**

| Rank | Dutch | Finnish | French | NS |
|---|---|---|---|---|
| 1 | PUNC ## | PUNC ## | PUNC ## | PUNC ## |
| 2 | N PUNC # | N PUNC # | N PUNC # | PREP ART N |
| 3 | PREP ART N | PREP ART N | PREP ART N | N PUNC # |
| 4 | ART N PREP | N PREP N | ART N PREP | ART N PREP |
| 5 | N PREP N | ART N PREP | ART ADJ N | N PREP N |
| 6 | ART ADJ N | ART ADJ N | ADJ N PUNC | N PREP ART |
| 7 | N PREP ART | # # PRON | N PREP ART | ART ADJ N |
| 8 | ADJ N PUNC | N PREP ART | N PREP N | V ART N |
| 9 | V ART N | ADJ N PUNC | # # PRON | ADJ N PUNC |
| 10 | ART N PUNC | V ART N | PREP N PUNC | PREP N PUNC |
| 11 | # # PRON | PRON AUX V | V ART N | ADJ N PREP |
| 12 | PRON AUX V | ADJ N PREP | PRON AUX V | # # PRON |
| 13 | PREP N PUNC | PREP N PUNC | N AUX V | ART N PUNC |
| 14 | ADJ N PREP | ART N PUNC | ADJ N PREP | PREP PRON N |
| 15 | N AUX V | N AUX V | ART N PUNC | # # N |
| 16 | PREP PRON N | PREP PRON N | PREP ART ADJ | N PREP PRON |
| 17 | AUX V PREP | N PREP PRON | PREP PRON N | N AUX V |
| 18 | V PREP ART | N CONJUNC N | N PREP PRON | PREP ART ADJ |
| 19 | N CONJUNC N | V PRON N | PRON N PUNC | V PREP ART |
| 20 | N PREP PRON | AUX V PREP | V PREP ART | PRON AUX V |

**Table 4. Top 20 trigrams of the ICLE and NS corpora (Aarts and Granger 1998: 141)**

As I compared the ICLE subcorpora with the NS corpus, the trigram types in the top 20 were roughly the same. The French corpus shared 19 of the 20 trigrams with the NS corpus, while the Finnish and Dutch corpora shared 17 and 18. In the case of the JEFLL subcorpora, JH1 shared only 5, and JH2 and JH3 shared 7 and 8 respectively. SH2 shared 10 and UNI shared 7[4]. The fact that there are more sentence boundary markers (#) in this top 20 trigrams than in the list by Aarts and Granger (1998) indicates that the sentence length on average is much shorter and therefore the sentence boundary markers are more likely to appear in the trigrams. This lack of similarity between my data and the ICLE data clearly shows that the proficiency level of the learners in my corpus was much lower than those in the ICLE data. Thus, even university students' data in my corpus should be considered to be in the intermediate stage of development with respect to the more mature stage shown in the ICLE data.

Looking at the lists of the top 20 trigrams, there are very striking differences in the way learners at different proficiency levels produce sentences. Let me first

[4] Only university students were given essays on different topics from the other four groups, which might affect the relative frequencies of top 20 trigrams.

focus on the trigrams which are in common among all the learners. The top 3 trigrams in the list (PUNC # #; N PUNC #; # # PRON) do not give much information. They are obviously more frequent if more sentences appear in a corpus. The trigram V ART N appeared throughout the lists of all the JEFLL subcorpora, which indicates that this pattern is introduced from the very beginning and is used quite intensively. The same thing can be said about other shared trigrams such as PREP ART N and ART N PUNC. These patterns will appear even though the sentences produced are very short and have a very simple syntactic structure.

### 3.2. Features of unshared trigram patterns

A comparison of the frequencies of the top trigrams showed that only a minority were similar in the JEFLL and the ICLE subcorpora. Since learners in the ICLE data seemed to be much more advanced than average Japanese university students in my corpora, the analysis of the JEFLL corpus data should provide a very interesting picture of the interlanguage grammar. Before going into the details of the distinctive trigrams, it should be noted that the patterns appearing in the corpus of JH1 are quite distinct from those appearing in the rest of the groups. Among those 15 trigram patterns which are different from the NS list, 9 patterns appeared only in the JH1 corpus. Many of them contain the sequence N N. For example, N N N (f = 301), N N V (f = 238), PRON N N (f = 211), ART N N (f = 130) and V N N (f = 118). Interestingly these patterns never appeared elsewhere in the upper level corpora tag frequency lists (except for N N PUNC). As you can probably imagine, this is due to their frequent use of the JH1 students' mother tongue words in their writing. In order to ensure fluency, the subjects were allowed to use Japanese words whenever they could not hit upon the right word in the writing task. Every Japanese word is tagged with a <JP> tag, but unfortunately CLAWS recognized this as a kind of proper noun and automatically assigned noun tags. This is why there are so many N N patterns in the top list. For future analyses, I will replace this by another tag in order to properly describe the sequences containing Japanese words.

I will focus on the three features of trigrams which are not shared with either the NS or the ICLE corpora. First, the underuse of prepositional patterns will be reported (3.2.1.). Second, I will describe the verb-related trigrams as a feature of early interlanguage grammar (3.2.2.). Finally, the patterns involving articles and auxiliaries will be described (3.3.3.).

### 3.2.1. Underuse of prepositional patterns

The most striking feature common to all the JEFLL subcorpora is the underuse of patterns involving prepositions. 7 out of the 9 trigrams missing in all of my learner corpora are somehow related to patterns with prepositions: ART N PREP, N PREP ART, ADJ N PREP, PREP PRON N, N PREP PRON, PREP ART ADJ and V PREP ART. Aarts and Granger (1998) also reported that for all three groups of learners, Dutch, Finnish and French, there was a consistent, significant underuse of PREP (ibid: 138). The results of my study even more clearly show that many PREP-related trigrams were constantly underused so that they did not appear in the top 20 list.

The underlying structures for these seriously underused trigram patterns are NP or PP. Some of these trigrams show that learners at lower proficiency levels have difficulty using NP+PP constructions (for instance, the underuse of ART N PREP, N PREP ART, ADJ N PREP, N PREP PRON all testify to this phenomenon). Learners know how to construct a prepositional phrase itself, which is shown in the high frequency of the trigram PREP ART N, but they find it difficult to combine the prepositional phrase with other elements such as NP or verbs as a head of VP. This tendency was observed throughout all five categories of learners. As will be seen in the correspondence analysis of section 3.3, a group of university students showed improvement in terms of PREP related trigrams although, as far as the top 20 trigrams are concerned, their writing was still not as good as that of the ICLE groups.

### 3.2.2. Verb-related trigrams

Another constant among learners in the JEFLL subcorpora is the high frequencies of V-related trigrams. In the top 20 trigrams in the NS data, for example, the percentage of V-related trigrams is only 20% (n=4) and the Dutch, Finnish, and the French learners of English in the ICLE have 20% (n=4), 25% (n=5) and 20% (n=4) respectively. Thus the rate of V-related trigrams in the advanced NNS and the NS corpora were roughly the same, whereas in the JEFLL subcorpora, the percentages are JH1: 45% (n=9), JH2: 50% (n=10); JH3: 40% (n=8); SH2: 45% (n=9); UNI: 60% (n=12). There was a marked difference in the frequency of V-related trigrams between the NS/ advanced NNS and the less advanced NNS corpora.

As was seen in the previous section, learners at the beginning and intermediate stages begin to acquire basic syntactic rules, but still have difficulty in constructing long NPs or PPs. If NPs are predominantly short, then such patterns as ART N PREP, N PREP ART, ART ADJ N, and ADJ N PREP do not tend to appear very frequently in the data. Thus, the high frequency of V-related trigrams has a close relationship with the low frequency of PREP and NP-related trigrams.

### 3.2.3. The underuse of AUX and ART

The patterns including AUX and ART show another discriminatory feature between lower proficiency groups and higher proficiency groups. The JH3 data was the first level in terms of proficiency at which patterns which include AUX appeared in the list. The only pattern that appeared throughout the JEFLL subcorpora is PRON AUX V (the rank was 20 (JH3); 11 (SH2); 8 (UNIV)), whereas the ICLE data has three different tags (PRON AUX V; N AUX V; AUX V PREP) in the top 20 lists. Therefore, the proper use of AUX (especially modal auxiliary) proves to be a good indicator of advanced learners' writing.

The underuse of ART was also very striking throughout the JEFLL subcorpora. Since the article system is found to be one of the most difficult grammatical morphemes for Japanese learners (See, for instance, Tono 1998), constant avoidance of the use of definite or indefinite articles was also observed in the JEFLL corpus. It is worth mentioning that the underuse of articles was especially noticeable in PREP-related trigrams. While ART appeared quite frequently in such trigrams as V ART N, PREP ART N, it did not appear frequently in PREP-related trigrams such as ART N PREP, N PREP ART, and PREP ART ADJ. In most cases in those two trigrams (V ART N, PREP ART N), the article was part of a short NP or PP, whereas in other cases like ART N PREP and N PREP ART, it was part of a much longer NP or PP, which made it difficult for learners to use those articles. Thus, the low frequency of ART is closely related to the construction of NP and PP.

### 3.3. Correspondence analysis

How can one capture the relationship between tag trigrams and learner proficiency levels? Is there any better way to explore the relationship than simply comparing the frequency data using Chi-square or log-linear analysis? In this study, correspondence analysis was used in order to explore the relationships between the two nominal variables: different language proficiency groups and particular trigram patterns. This statistical technique was developed by the Data Theory Scaling System Group (DTSS), Faculty of Social and Behavioral Sciences, Leiden University and was first incorporated into SPSS 7.5 onwards (cf. Meulman 1997). It is a data reduction procedure like factor analysis, and basically describes the relationships between two nominal variables while simultaneously describing the relationships between the categories for each variable. It is suitable for nominal variables like the trigram frequencies used in this study, does not require the conditions for parametric statistics and provides a good graphical representation of the relationship among variables.

In order to make a correspondence table for the analysis, I selected 3 parts of speech categories which seemed to have a crucial role in each developmental stage of acquisition: i.e. N, V and PREP. First, the trigram frequencies were normalised to a rate per 100,000 words, then the frequencies of all the trigrams which included either N, V, or PREP respectively were summed. I will call each of these broader categories of trigram: N-related trigrams, V-related trigrams and PREP-related trigrams. Some trigrams contain more than one of the three categories. For instance, V N PREP contains all the three categories in one trigram, in which case it was counted one for each of the three categories. I reduced the trigram categories in this way for the top 100 trigrams[5]. Table 5 shows the correspondence table, which shows the cross tabulation of the input variables with row and column marginal totals.

| School Year | Trigram types | | | |
|---|---|---|---|---|
| | N-related | V-related | Prep-related | Active Margin |
| JH1 | 17102 | 7861 | 1301 | 26264 |
| JH2 | 11255 | 8874 | 3330 | 23459 |
| JH3 | 10778 | 8853 | 3266 | 22897 |
| SH2 | 9654 | 8595 | 3543 | 21792 |
| UNI | 8797 | 8714 | 4361 | 21872 |
| Active Margin | 57586 | 42897 | 15801 | 116284 |

**Table 5. Correspondence table**

Explaining the details of the meaning of each score is a little beyond the scope of this paper, but correspondence analysis basically tries to capture the relationship between the two nominal categories by specifying 2 dimensions. After obtaining the measure of distance based on Chi-square scores, the technique produces row and column points for the categories in each variable and shows the relationship between each two variables as well as the categories for each variable by producing a matrix of joint plots of row and column points. See Appendix 1 for further details of the statistics.

[5] I am grateful to Michael Oakes for his suggestion on the nature of the input data to the correspondence analysis.

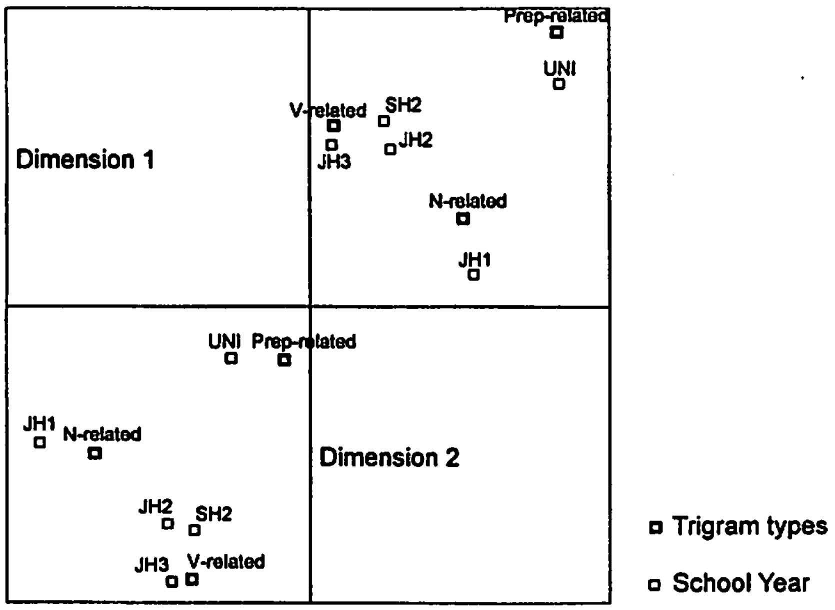

**Figure 1. Row and column points symmetrical normalization**

Let me move on to the biplots shown in Figure 1. For each variable, the distances between category points in a plot reflect the relationships between the categories with similar categories plotted close to each other. Projecting points for one variable on the vector from the origin to a category point for the other variable describes the relationship between the variables.

The biplots of row and column points in Figure 1 illustrate the relationship between the school year and tag trigram patterns. As far as the row variable (i.e. the school year) is concerned, JH1 and UNI were both plotted apart from the other three groups, JH2, JH3 and SH2. UNI and JH1 were also quite distant from each other. Since Dimension 1, which is the upper left box in Figure 1, explains almost 99% of the total inertia of the dimension, it can be said that UNI is close to these three intermediate groups, JH2, JH3 and SH2, and JH1 is rather far from the rest of the groups. In the case of the trigram categories, the PREP-related trigram and the N-related trigram were further apart and the V-related trigram was located in between.

What can then be observed in the relationship between the two variables? The plots show that the group of JH1 is close to the category "N-related" trigrams and that the group of university students has a close relationship with the PREP-related trigrams. In the top 20 trigrams, the UNI group did not have as many trigrams including PREP, but the correspondence analysis of the top 100 trigrams showed that the UNI group, which is considered to be the most advanced in all the groups in my corpus, indeed had the tendency to use more PREP-related constructions in their interlanguage. The other three groups JH2, JH3, and SH2, all gathered together near the V-related category. This seems to be reasonable when so many V-related trigrams were found in the top 20 trigrams as compared with the ICLE subcorpora.

## 4. Discussion

First of all, the results of correspondence analysis indicate that the analysis of tag trigrams of interlanguage corpora proves to have implications for describing L2 acquisition stages. Figure 2 illustrates the interlanguage development stages in terms of part-of-speech tag sequences. At the very beginning stage, L2 learners seem to have difficulty expressing themselves solely in the target language. In the JH1 corpus, whenever they found it difficult to put their ideas in English, they resorted to using Japanese words in English sentences. It is quite natural that they could not write in English very well at this stage of learning. This tendency to produce N N constructions becomes less as they move on to the next grade. The intermediate groups, the JH2, JH3, and SH2 groups, showed a marked tendency to use Verb-centred trigrams. The students had mastered the basic sentence constructions and could use a fixed pattern like V ART N as a simple verb phrase or PREP ART N for a simple prepositional phrase. More complex structures, however, such as NP containing NP + PP were constantly avoided. The essay quality of the UNIV group was still lower than the Dutch, Finnish, and French learners of English in the ICLE subcorpora. The markedly low frequency of PREP-related trigrams is the evidence of this gap. And yet, the UNIV group used PREP-related trigrams most of all the groups in the JEFLL subcorpora. It can be said, therefore, that the use of complex prepositional phrases is one of the most salient characteristics of fully developed interlanguage. Even the learners in the ICLE subcorpora, however, show the constant underuse of prepositional phrases. Aarts and Granger (1998) quoted Biber *et al.* (1994), in which they found that "prepositional postnominal modifiers received the least attention in grammars although they proved to be much more common than relative or participal clauses in all three registers (editorials, fiction and letters) (Aarts and Granger 1998: 138). The treatment of postnominal prepositional phrases in English grammar books in Japan is also very rudimentary. It is therefore desirable to reconsider the treatment of some grammar items based upon these findings.

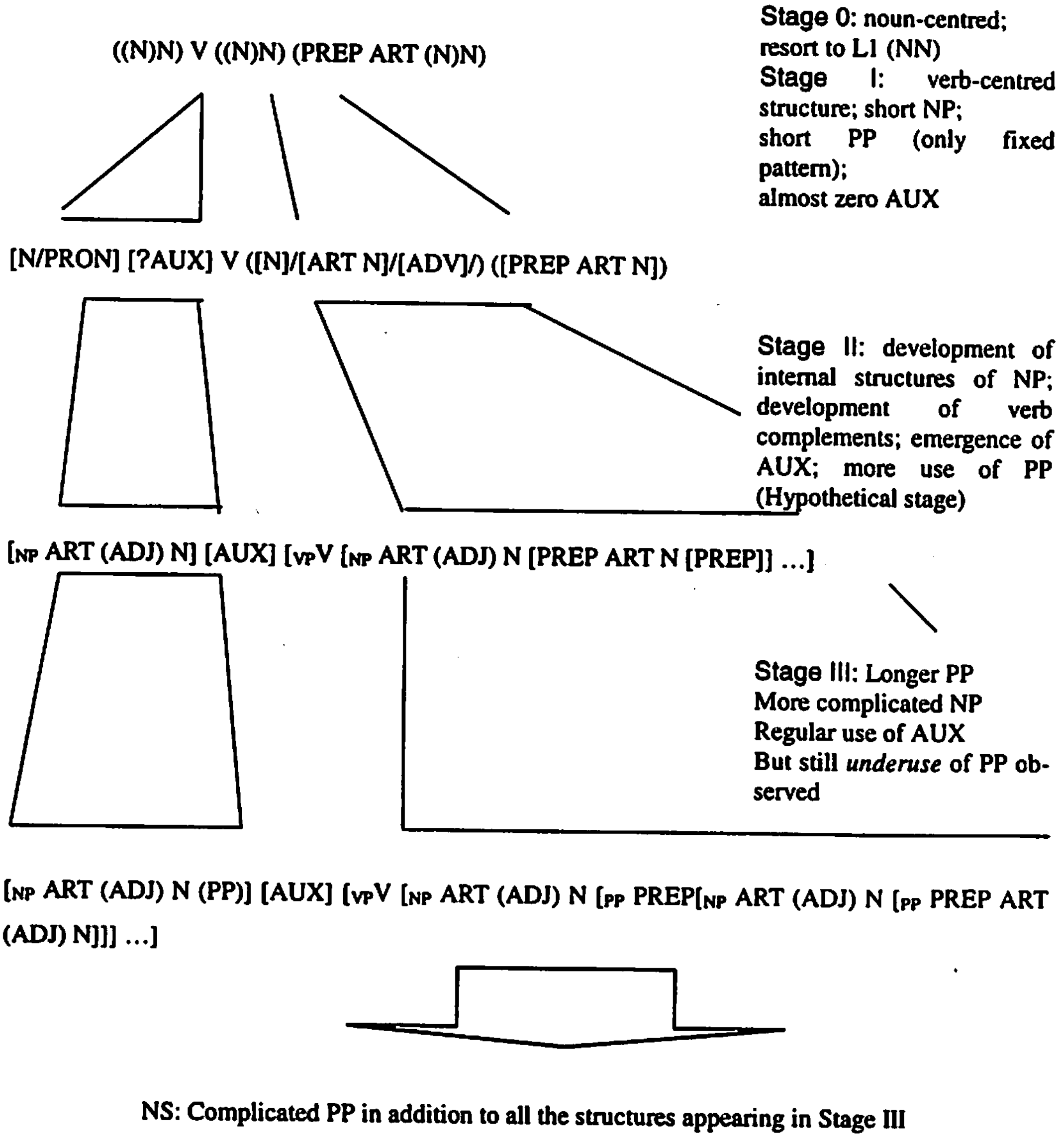
((N)N) V ((N)N) (PREP ART (N)N)
Stage 0: noun-centred; resort to L1 (NN)
Stage I: verb-centred structure; short NP; short PP (only fixed pattern); almost zero AUX
[N/PRON] [?AUX] V ([N]/[ART N]/[ADV]/) ([PREP ART N])
Stage II: development of internal structures of NP; development of verb complements; emergence of AUX; more use of PP (Hypothetical stage)
[$_{NP}$ ART (ADJ) N] [AUX] [$_{VP}$V [$_{NP}$ ART (ADJ) N [PREP ART N [PREP]] ...]
Stage III: Longer PP
More complicated NP
Regular use of AUX
But still *underuse* of PP observed
[$_{NP}$ ART (ADJ) N (PP)] [AUX] [$_{VP}$V [$_{NP}$ ART (ADJ) N [$_{PP}$ PREP[$_{NP}$ ART (ADJ) N [$_{PP}$ PREP ART (ADJ) N]]] ...]
NS: Complicated PP in addition to all the structures appearing in Stage III

**Figure 2. Hypothetical developmental patterns of POS tag sequences**

The fact that most missing trigrams in the top 20 lists either contained PREP or ART indicates that the article system is really another problematical area for Japanese learners. The use of articles in complex noun phrases or prepositional phrases is particularly difficult. The same thing can be said about the use of modal auxiliaries. The severe underuse of AUX shows that the less advanced learners tend to make their sentences economical as one of their communication strategies. It is often possible to communicate ideas without modals even though it does not sound natural or the intention may not necessarily be fully understood. For lower-or intermediate level learners, there are more things to learn than auxiliaries and the priority is rather low. This tendency is very clear in grammar books, too. Modal auxiliaries are first introduced with "*can, may, must.*" But after the past tense was introduced, it became too complicated to learn the difference between *can* and *could*, *may* and *might*, and so on. Many students give up trying to understand the differences in modal auxiliaries. The matter will get worse if subjunctive moods are introduced and they have to learn such a pattern as "*should have been.*" In this study, the serious underuse of AUX and ART was empirically verified, which I believe has very important pedagogical implications.

What implication does this study have for SLA (second language acquisition) research? First, this study described the overall patterns of syntactic development in interlanguage. Much research has been devoted to each syntactic item such as grammatical morphemes, passives, negatives, relative clauses, and so on, but very few studies have described the SLA process in its entirety. The ZISA project (Clahsen 1984) and Pienemann's processability hypothesis (Pienemann 1997) would be among the few studies to attempt this range of analysis. A corpus-based analysis of interlanguage is strong in the sense that it is firmly evidence-based. On the other hand, this kind of research is weak because it does not generate any theory of acquisition. It can provide a description of developmental patterns, but it can not explain how those patterns emerge. For my findings to be explained, I need to explore available SLA theories of acquisition and examine whether those SLA theories truly fit into my data. Therefore, a corpus exploitation should go hand in hand with theory-driven research. Since this kind of tag trigram analysis could be understood with the learning of probability of word combinations, the possible application of the Connectionist approach or the probabilistic model of grammar (cf. Halliday 1991) might be promising. Further research is needed to answer this question.

Finally, a few methodological considerations are in order. Firstly, analysing part-of-speech tags is genuinely interesting in its own right, but part-of-speech tags are just like a skeleton of the language. Further research on each of the tag sequences and what is actually happening at real lexical level will shed more light on the nature of these tag trigrams.

Secondly, my learner data is still modest in size. There is always the possible pitfall of representativeness. I must answer seriously the following question: "Am I sure that the patterns emerging from my learner corpora are really due to the difference in language proficiency?" "Are there any essay task influences?" "How can I distinguish L1-related patterns of development from universal ones?"

and so on. In this study, however, I am quite confident that the number of observations of tag sequences was sufficient enough to make a meaningful statistical comparison possible among different groups. The size of corpora depends upon the use to which they are intended to be put. I am also aware that replication of the study using larger corpora is also needed in order to ensure the reliability of the findings.

Thirdly, the JEFLL Corpus as well as other major learner corpora in the world are mainly collections of written essays whereas in most SLA and L1 acquisition studies the primary focus has been on spontaneous or controlled speech production data. The acquisition process could appear quite different if the written data was investigated primarily. It also depends upon data elicitation techniques. My corpus data is based upon timed free writing tasks without dictionaries, but if the data were taken from homework essays and the use of dictionaries was permitted, then the quality of essay data could be very different. This should not discourage learner corpus researchers from continuing to gather data in a well-defined objective way. The standardization of learner profiles and refinement of corpus design criteria is definitely necessary.

Last but not least, the approach taken in this paper, namely correspondence analysis, to my knowledge, has rarely been applied in corpus linguistics. As this technique is applied to the tag trigrams of the various parts of speech, it would be possible to describe the development of other constructions such as adjectival or adverbial phrases, conjunctions, negatives among others. In the future, I plan to parse the learner data and see how higher syntactic structures develop throughout the stages of learner development, using correspondence analysis. There is a great potential to apply this statistical technique to higher-level structures such as verb complementation or subordination. I hope that this type of study will provide better understanding of interlanguage grammar and improvement of pedagogy for the future.

## References

Aart, J. and S. Granger (1998). "Tag sequences in learner corpora: a key to interlanguage grammar and discourse". In Granger, S. (ed.) *Learner English on Computer*. Addison Wesley Longman: 132–141.

Bernadini, S. (1998). "Systematising serendipity: proposals for large-corpora concordancing with language learners". *TALC98 Proceedings*: 12–16.

Biber, D., Conrad, S. and R. Reppen (1994). "Corpus-based approaches to issues in applied linguistics". *Applied Linguistics* 15 (2): 115–36.

Biber, D., Johansson, S., Leech, G., Conrad, S. and E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Addison Wesley Longman.

Clahsen, H. (1984). "The acquisition of German word order: a test case for cognitive approaches to L2 development". In: R. W. Andersen (ed.). *Second Languages: a Cross-linguistic Perspectives*. Rowley, MA: Newbury House: 219–42.

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

Granger, S. (ed.) (1998). *Learner English on Computer*. Addison Wesley Longman.

Halliday, M. A. K. (1991). "Corpus studies and probabilistic grammar." In: K. Aijmer and B. Altenberg (eds). *English Corpus Linguistics*. Longman: 30–43.

Kennedy, G. (1987). "Quantification and the use of English: a case study of one aspect of the learner's task". *Applied Linguistics* 8:264–86.

Ljung, M. (1990). *A Study of TEFL Vocabulary. Stockholm.* Almqvist and Wiksell.

Meulman, J. J. (1997). "Optimal scaling methods for multivariate categorical data analysis". The white paper written for SPSS web page (http://www.spss.com/cool/papers/ swpopt.htm).

Mindt, D. (1992). *Zeitbezug im Englischen: eine didaktische Grammatik des englischen Futurs.* Tübingen: Gunter Narr.

Pienemann, M. (1997). "A unified framework for the study of dynamics in language development – applied to L1, L2, 2L1 and SLI". Manuscript.

Tono, Y. (1998). "A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes". In *TALC'98 Proceedings*: 183–187.

# Appendix 1

## Correspondence Analysis: Statistics

Overview Row Points[a]

| School Year | Mass | Score in Dimension | | Inertia | Contribution | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Of Point to Inertia of Dimension | | Of Dimension to Inertia of Point | | |
| | | 1 | 2 | | 1 | 2 | 1 | 2 | Total |
| JH1 | .226 | −.772 | .074 | .026 | .687 | .073 | .999 | .001 | 1.000 |
| JH2 | .202 | .069 | −.062 | .000 | .005 | .045 | .935 | .065 | 1.000 |
| JH3 | .197 | .100 | −.161 | .000 | .010 | .098 | .817 | .183 | 1.000 |
| SH2 | .187 | .252 | −.073 | .002 | .260 | .059 | .993 | .007 | 1.000 |
| UNI | .188 | .498 | .218 | .009 | .238 | .525 | .984 | .016 | 1.000 |
| Active Total | 1.000 | | | .039 | 1.000 | 1.000 | | | |

[a] Symmetrical normalization

Overview Column Points[a]

| Trigram types | Mass | Score in Dimension | | Inertia | Contribution | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Of Point to Inertia of Dimension | | Of Dimension to Inertia of Point | | |
| | | 1 | 2 | | 1 | 2 | 1 | 2 | Total |
| N-related | .495 | −.402 | .058 | .016 | .408 | .096 | .998 | .002 | 1.000 |
| V-related | .369 | .229 | −.157 | .004 | .099 | .532 | .961 | .039 | 1.000 |
| Prep-related | .136 | .844 | −.216 | .019 | .493 | .371 | .994 | .006 | 1.000 |
| Active Total | 1.000 | | | .039 | 1.000 | 1.000 | | | |

[a] Symmtrical normalization

Summary[a]

| Dimension | Singular Value | Inertia | Chi Square | Sig. | Proportion of Inertia | | Confidence Singular Value | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Accounted for | Cumulative | Standard Deviation | Correlation |
| | | | | | | | | 2 |
| 1 | .196 | .038 | | | .993 | .993 | .003 | .141 |
| 2 | .017 | .000 | | | 1.007 | 1.000 | .003 | |
| Total | | .039 | 4509.416 | .000[a] | 1.000 | 1.000 | | |

[a] 8 degrees of freedom.

# LEXICAL SEMANTICS AND LEXICOGRAPHY