Automatic extraction of L2 criterial lexico-grammatical features across pseudo-longitudinal learner corpora: using edit distance and variability-based neighbour clustering

Yukio Tono
Tokyo University of Foreign Studies

The aim of this study is to offer a generic technique of extracting lexico-grammatical features that serve as criteria for distinguishing one CEFR level from the others in pseudo-longitudinal learner corpora. Semi-automatic error tagging for surface error taxonomy was performed on a written corpus of 10,038 Japanese EFL learners by comparing the original essays against the proofread ones, by using edit distance and automatic POS tagging. The output was further processed using multivariate statistics called correspondence analysis and variability-based neighbour clustering to examine whether those automatically assigned errors could serve as criterial features. The results show that this new approach of error annotation and clustering is useful to identify criterial features for lower levels that are not provided by the English Profile Programme and suggest an alternative classification of features for all CEFR levels.

1. Introduction

In SLA, it is becoming increasingly popular to use techniques and resources developed in the field of corpus linguistics and natural language processing. The use of learner corpora, systematically sampled collections of learner speech or writing in a machine-readable format, is rapidly gaining ground among ELT materials developers, practitioners and SLA researchers (Granger, 1998; Granger, Hung, & Petch-Tyson, 2002). Behind all of this, there is a growing awareness that frequency of items to be acquired in input plays an important role in L1 and L2 acquisition processes (Gries & Divjak, 2012). According to Goldberg (1995, 2006), the Saussurian concept of a symbolic unit, that is a form-meaning pair, is assumed to cover not only the level of words, but also applies to constructions at all levels of semantic linguistic representation from morphemes and words to increasingly complex syntactic configurations. This symbolic unit is acquired through the exposure to the target language in context. I would argue that with the advent of corpus linguistics and natural language processing, SLA researchers should once again

Table 1. Possible criterial feature types

Type of feature	Descriptions	Examples (based on Hawkins & Buttery 2010)
Positive linguistic properties of the L2 levels	Correct properties of English that are required at a certain L2 level and that generally persist at all higher levels. E.g. property P acquired at B2 may differentiate [B2, C1 and C2] from [A1, A2 and B1] and will be criterial for the former.	The ditransitive NP-V-NP-NP structure (*she asked him his name*) appears at B1, and is thus criterial for [B1, B2, C1, C2]. The object control structure, NP-V-NP-AdjP (*he painted the car red*) is criterial for [B2, C1, C2].
Negative grammatical properties of the L2 levels	Incorrect properties or errors that occur at a certain level or levels, and with a characteristic frequency. Both the presence versus absence of the errors, and the characteristic frequency of error can be criterial for the given level or levels. E.g. error property P with a characteristic frequency F may be criterial for [B1 and B2].	Errors involving incorrect morphology for determiners, as in Derivation of Determiners (abbreviated DD) *She name was Anna* (instead of *Her name* ...), show significant differences in error frequencies that decline from B1 > B2 > C1 > C2.
Positive usage distributions for correct L2 properties	Positive usage distributions for a correct property of L2 that match the distribution of native speaking (i.e. L1) users of the L2. The positive usage distribution may be acquired at a certain level and will generally persist at all higher levels and be criterial for the relevant levels.	The distribution of relative clauses formed on indirect object/oblique positions (e.g. *the professor that I gave the book to*) to relativizations on other clausal positions (subjects and direct objects) appears to approximate that of native speakers at the C levels, but not at earlier levels. Hence this is a positive usage distribution that is criterial for [C1, C2]
Negative usage distributions for correct L2 properties	Negative usage distributions for a correct property of L2 that do not match the distribution of native speaking (i.e. L1) users of the L2. The negative usage distribution may occur at a certain level or levels with a characteristic frequency F and be criterial for the relevant level(s).	The distribution of relative clauses formed on indirect object/oblique positions is the negative usage distribution, criterial for B2 and below.

focus on descriptive aspects of IL processes, in addition to already available introspective and experimental methods. By identifying the use/misuse of language features and their relative frequencies at different developmental stages in more detail, one can take into account frequency effects in language acquisition and learning.

To this end, a very unique project called the English Profile Programme (EPP) has started. It is sponsored by the Council of Europe and is maintained by the research team including Cambridge ESOL Examinations, Cambridge RCEAL, and University of Bedfordshire. The aim of the EPP is to create a 'profile' or set of Reference Level Descriptions (RLDs) for English linked to the Common European Framework of Reference (CEFR). The EPP website (http://www.englishprofile.org/) states that reference level descriptions

> will provide detailed information about the language that learners can be expected to demonstrate at each CEFR level (A1 & A2: basic user; B1 & B2: independent user; C1 & C2: proficient user), offering a clear benchmark for progress that will inform curricula development as well as the development of courses and test material to support learners, teachers and other professionals in the teaching of English as a foreign language.

What is unique in the EPP is its corpus-based method of finding 'criterial features' from learner corpora sampled from the subjects at different CEFR levels. Salamoura and Saville (2009) defined a 'criterial feature' as follows (Salamoura & Saville, 2009, p. 34).

> A 'criterial feature' is one whose use varies according to the level achieved and thus can serve as a basis for the estimation of a language learner's proficiency level. So far the various EP research strands have identified the following kinds of linguistic feature whose use or non-use, accuracy of use or frequency of use may be criterial: lexical/semantic, morpho-syntactic/syntactic, functional, notional, discourse, and pragmatic.

Hawkins and Buttery (2010), for example, have identified four types of feature that may be criterial for distinguishing one CEFR level from the others. Table 1 shows the classifications.

The English Profile (EP) researchers have done preliminary studies with regard to the criterial features, using the Cambridge Learner Corpus (CLC) (Williams, 2007; Parodi, 2008; Hendriks, 2008; Filipovic, 2009; Hawkins & Buttery, 2010). The CLC currently comprises approximately 50 million words of written learner data, roughly half of which is coded for errors. It has also been parsed using the Robust Accurate Statistical Parser (RASP) (Briscoe, Carroll & Watson, 2006). Salamoura and Saville (2009) state that the CLC mainly covers A2 level and above, which is the reason why the EP researchers started to build a new corpus called the Cambridge English Profile Corpus (CEPC), mainly focusing on lower-proficiency level students' writing and speech.

Considering the sheer size of the CLC with error annotations and the CEFR as a framework, this EP programme seems to create a new research paradigm in learner corpus research. Those who are interested in using learner corpora in SLA research can relate their findings to the EP researchers' findings in terms of criterial features. Those who are involved in syllabus/materials design will find the RLDs for English very informative once those items are actually identified. Test developers will make full use of the results of the EP research for improving their test design and contents.

Some may argue that this whole approach is affected by the 'comparative fallacy' (Bley-Vroman, 1983). Bley-Vroman warned that L2 speakers' interlanguage systems should be seen as independent of their L1s and target languages and should thus be studied in their own right. This implies discarding the notion of 'target-like' performance. Most learner-corpus-based IL studies rely on the comparison between L2 learners and their mother tongues or target-like performance by native speakers of the target languages. In my opinion, this again depends on research purposes. If one wishes to describe interim states of IL systems, independent of both L1s and target languages, Bley-Vroman's position makes perfect sense. However, as Kasper (1997) said, SLA researchers have legitimate and important interests in assessing learners' IL knowledge and actions not just as achievements in their own right but also measured against some kind of standard (ibid: 310). From pedagogical and assessment viewpoints, there is nothing wrong with setting native speakers' well-formed sentences as a goal, because that is the language taught in the classroom. Therefore, L2 profiling research is worth the effort, as long as we properly understand its aims.

One of the issues of identifying criterial features is deciding how to extract errors from learner data and judge whether they serve as criterial features or not. The CLC is manually tagged for errors, but it would be quite difficult to extract learner errors from generic learner data without error annotations. There are two main purposes of this paper; to propose a new approach of annotating errors semi-automatically by comparing the original learner data against the proofread data, by using edit distance and automatic POS tagging, and to judge whether or not those errors can serve as criterial features by employing multivariate statistics called correspondence analysis and variability-based neighbour clustering. This is especially useful because it provides a set of criterial features for lower levels that are not provided by CLC, in order to identify a set of features for Japanese learners of English in specific L2 contexts, to suggest an alternative classification of features for all CEFR levels, and to offer a generic technique of extracting criterial features from any learner corpora.

2. Method

2.1. The JEFLL Corpus and its parallel version

The JEFLL Corpus is a corpus of 10,038 Japanese students' written compositions in English, totalling 669,281 running words (available online at http://scn02.corpora.jp/~jefll04dev/). The subjects were sampled across six school years (from Year 7 to 12 in terms of the U.S. school system). In Japan, English is generally introduced in Year 7 for the first time, so JEFLL consists of samples from beginning to lower-intermediate levels. The students were asked to write a short in-class essay in English in 20 minutes without the help of a dictionary. Essay topics were also controlled; there were six different topics in total (3 argumentative and 3 narrative/descriptive). The corpus can be queried on the basis of learner profile information such as school year, school type, and school level, as well as task variables (e.g. topics).

Using the JEFLL Corpus, my research team conducted a series of studies for identifying features characterising different stages of acquisition. Table 2 summarises the results.

Table 2. Previous studies using the JEFLL Corpus

Language features	References	Main findings
Morpheme orders	Tono (1998)	• Article errors are persistent and the development of accurate article use is much slower than reported in previous research. • Possessive *-s* is easier than the universal order proposed in previous research.
N-gram[1] analysis	Tono (2000, 2009)	• The early stages are characterized by trigrams associated with V.
Verb subcategorization	Tono (2004)	• Subcategorization errors are influenced by inherent verb semantics and are not affected so much by input from the textbooks. • Overuse/underuse phenomena are related to textbook input.
Verb & noun errors	Abe (2003, 2004, 2005) Abe & Tono (2005)	• Verb errors are more frequent at lower proficiency levels. • Noun errors occur more frequently at higher levels.
NP complexity	Kaneko (2004, 2006); Miura (2008)	• Internal structures of NP are closely related to developmental stages. • Clause modifiers (relative clauses and *that*-clauses) are associated with the most advanced level.

[1] N-gram is a contiguous sequence of n items from a given sequence of text. In corpus linguistics, items in question can be words, parts-of-speech, or combinations of those. An n-gram of size 3 is called 'trigram.'

One of the methodological problems is the difficulty in error annotations. Some studies (Tono, 1998; 2004; Abe, 2003; 2004) examined errors in the JEFLL Corpus, but only smaller sets of texts, approximately 10,000 words for each subset, were used for manual error tagging. It is very time-consuming to tag the entire corpus for all types of errors, so we focused on certain grammatical errors only and performed so-called 'problem-oriented' tagging for errors. Currently, there are not very many fully error-tagged corpora available. The Cambridge Learner Corpus may be the only exception but again the corpus sampling tends to be skewed toward intermediate to advanced learners of English and unfortunately it is for in-house use only.

Instead of manually annotating every error in the files, a proofread version of the JEFLL Corpus was prepared. For this, one educated adult native speaker, who worked as an English instructor at a university in Tokyo, was hired to read through and correct errors in all the essays in the JEFLL Corpus. A single person did the job, because previous experiences show that annotation by a single person was more consistent than several people working together, although sufficient training was needed. A one-month training session was conducted, in which the proofreader was asked to correct several essays at different levels. The proofreader then discussed with the researcher the way errors were identified and corrected. Only local sentence-level lexico-grammatical errors were corrected. No corrections were made beyond sentence levels, such as coherence, connectivity, or the use of discourse markers across sentence or paragraph levels, for these error corrections usually involve a change in converting sentence orders or putting two sentences into one or vice versa. The sentence alignments in the essays were maintained strictly. One of the difficulties of proofreading the data in the JEFLL Corpus is that the compositions contain Japanese words or phrases. In the composition tasks, the use of Japanese was allowed especially for learners at the very beginning-level. Therefore, a proofreader competent in Japanese was chosen in order to produce corrected versions of the corpus.

2.2. Edit distance

A metric called an edit distance was employed. The edit distance between two strings of characters is the number of operations required to transform one of them into the other. There are several different ways to define an edit distance (for instance, Hamming distance, longest common subsequence, Levenshtein distance). Usually, an edit distance produces the actual number (e.g. the dis-

[2] Differences between the two words are positions No. 2, 3, 5 and 7 in the letter sequence of "sitting". Thus the distance is 4.

tance is 4, between “seaten” and “sitting”[2]), showing the amount of difference between the two sequences, but in the present study, I used this heuristic for identifying the same and different parts in the aligned sentences. My colleague, Hajime Mochizuki, helped to implement the program into the programming language Ruby, and the algorithm he used was basically the same as the so-called Levenshtein distance (Levenshtein, 1966). A commonly-used bottom-up dynamic programming algorithm for computing the Levenshtein distance involves the use of an (n + 1)×(m+1) matrix, where n and m are the lengths of the two strings. Figure 1 illustrates the matrix. The two sequences can be aligned in three possible ways, as (1) shows.

(1) a. Two elements are identified as the same and aligned to each other (“\” path in the matrix)
b. X is aligned to a gap (“|” path)
c. Y is aligned to a gap (“–” path)

Suppose X has a sequence “ABCE” and Y has “ACDE,” the thick black line in Figure 1 indicates the optimal path for alignments. There is possibly more than one path from the starting point (0,0) to the end point (4,4). A Dynamic Programming (DP) algorithm checks all available paths from the start to the end and calculates each cost to identify the optimal path.

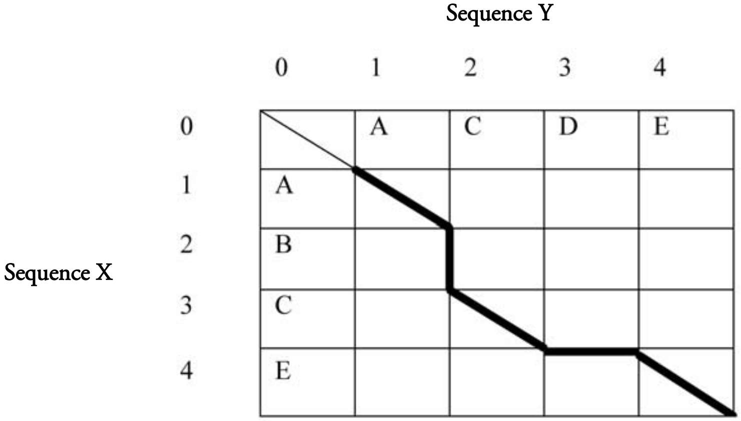
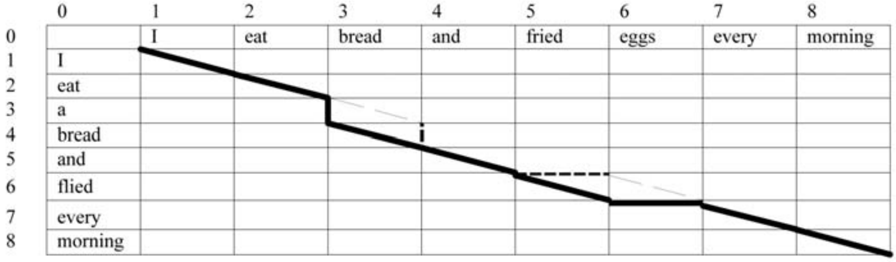


Figure 1. Dynamic Programming matrix

In our case, two aligned sequences correspond to two sentences, and the parts in the sequences (A to E in Figure 1) are actual words in the sentences. Figure 2 shows in matrix form how this algorithm checks the two aligned sentences, an original sentence (vertical) and its corrected counterpart (horizontal).

Figure 2. DP matrix for sentence examples



In Figure 2, two possible cases of alignment are illustrated. The alignments are described in (2) and (3) below:

(2) a . I eat * bread and fried eggs every morning.
b. I eat a bread and flied * every morning.
(3) a. I eat bread * and fried eggs every morning.
b. I eat a bread and * flied every morning.

The alignment result in (2) is better than that in (3) in the sense that missing items in the sentence pairs (a) and (b) are correctly matched in (2), compared to the results in (3). Each of the paths in Figure 2 shows these alignment results, with thick black lines showing the case in (2) and dotted lines, showing the case in (3). Each edit distance in (2) and (3) is calculated and the optimal path (in this case, (2)) produces the highest score. Look at (2) once again. There are three allowable edit operations in the Levenshtein distance, which is described in (4):

(4) a. I eat * bread and fried eggs every morning.
b. I eat a bread and flied * every morning.
↑ ↑ ↑
Operations: [insertion] [substitution] [deletion]

In error analysis, these three edit operations correspond to the types of errors identified in the so-called Surface Strategy Taxonomy (Dulay, Burt, & Krashen, 1982, p. 150; see also the "surface modification" typology proposed by James, 1998), as shown in (5):

(5) a. substitution → misformation errors
b. insertion → addition errors
c. deletion → omission errors

Therefore, using the Levenshtein distance, similarity scores were calculated between each word in two aligned sentences. The program gave as output the best tagged alignment results with the highest total of individual scores as an optimal alignment. The three error types are identified automatically based on the alignment results, and then tagged for each error type: <msf> for misformation, <add> for addition, and <oms> for omission. Correction candidates are specified in the case of misformation tags, as in <msf crr= "correct answer">. The output of the program is shown in (6):

(6) I eat <add>a</add> bread and <msf crr=fried>flied</msf> <oms>eggs</oms> every morning.

If the alignments are accurate, chances are that surface strategy taxonomy errors can be extracted fairly accurately and automatically.

2.3. Procedure

Using the heuristics described in 2.2., the parallel (i.e. original and proofread) version of the entire JEFLL Corpus was processed for the Levenshtein distance and then automatically tagged for three types of surface strategy taxonomy error: omission, addition and misformation. The output of the program was checked manually, and problematical cases of word order errors were identified and corrected. In order to capture an overall tendency of extracted errors, all the tagged surface strategy taxonomy errors were processed for part-of-speech (POS) information, using an automatic POS tagger. This made it possible to analyse extracted errors in terms of their parts of speech. At this level, the error annotation in the corpora is only related to the surface strategy taxonomy errors and their POS information. I am fully aware of the limitations of dealing with errors using the surface taxonomy and POS only. It needs further analysis in terms of linguistic classification, e.g. agreement errors, tense errors, verb subcategorization errors, among others. Furthermore, a POS tagger developed for analysing native speakers' data may not be entirely suitable for interlanguage data. But I have the following justifications for my approach. First, the main purpose of this chapter is to propose a method of annotating errors semi-automatically in learner language and not to propose comprehensive criterial features from learner data. Using the approach described in this paper, researchers can work on their learner data and make further analysis of each error type they are interested in. Second, the overview of POS-related errors based on the surface strategy taxonomy still provides a very interesting summary regarding the state of ILs at each stage and helps to generate new hypotheses related to different aspects of acquisition. For instance, omission errors of determiners are quite frequent across all the stages of acquisition in the JEFLL Corpus, while the repertoire of nouns in lexicon will also increase as the level increases. This means that the use of articles improves for particular noun groups, but the knowledge of the article system is not fully acquired as more lexical items are introduced in the lexicon. This kind of microscopic analysis can be done for each error type, but this should be dealt with elsewhere. Third, automatic annotation described in this paper can be used to annotate large samples of learner corpora, which is cost-effective, and helps to conduct profiling research such as EPP to provide a bird's eye view of how learner performance will change from one stage to another.

The frequency distributions of the above error types in terms of POSs were obtained across the school years. Multivariate statistics were used in order to capture complex relationships between school years and different error types. Correspondence analysis was used first to obtain biplots between major error types and school years, which was supplemented by clustering techniques called "variability-based neighbour clustering (VNC)" (Gries & Stoll, 2008). Both are techniques of data reduction and summarisation. Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information which is similar in nature to that produced by Factor Analysis techniques, and they allow one to explore the structure of categorical variables included in the table. Graphical representations of two variables mapped onto the two extracted dimensions are especially useful in order to see relative proximity of the items in each variable. VNC differs from standard approaches because it only clusters neighbouring data points, thus preserving the data points' temporal sequence. This is important because the order of school years needs to be taken into account as we cluster linguistic features characterising each level.

3. Results

3.1. The performance of edit distance

The results of the Levenshtein distance show that this technique seems to work well. The precision and recall[3] rates for omission errors were 98.25% and 100% respectively (F measure is 0.9911 at α= 0.5). For the addition errors, the precision rate was 96.83% and the recall was 100% (F=0.9839). Only misformation errors were found to be less accurate. The number of incorrectly analysed items

[3] Precision is defined as a measure of the proportion of selected items that the system got right: precision = (true positive)/((true positive)+(false positive)). Recall is defined as the proportion of the target items that the system selected: recall = (true positive)/((true positive)+(false negative)) (Manning & Schutze 1999: 268).

was 179 out of 641 (precision = 72.07%), which shows that alignment of misformation was very difficult in comparison to the other two error types. Consequently, F measure was also low (F= 0.8373).The sample output is shown in (7), where no error was found in the analysed sentence:

```
(7) <result>
<sentence id= “ns”>
Today I ate bread and milk
</sentence>
<sentence id= “st”>
Today I ate bread and milk
</sentence>
<trial no= “01a”>
Today I ate bread and milk
</trial>
</result>
```

The first sentence labelled “ns” is the one proofread by a native speaker. The second sentence labelled “st” is the student’s original sentence and the third one is the output of comparing the pair (“ns” and “st”). If there is no error in the sentence, the output is the same as the two sentences above.

The sentences in (8) show the case in which the sentence pair (“ns” and “st”) has several differences. In the first output labelled “trial No. 01a”, differences between the pair were identified in terms of omission, addition and misformation (tagged <oms>, <add>, and <msf> respectively) along with suggested corrections shown in the attribute “crr=”. The edit distance program works in such a way that the first trial was retained as long as there was no overlapping word found in the identified error items. If there was any overlapping word, for example, “breakfast” in the output “01a”, additional analysis was made to reclassify the two overlapped words into a single case of transposition from one position to another in a sentence. Thus, in the output “02”, the word “breakfast” is tagged as <trs_add> for the first one and <trs_oms> for the second one, showing that these two words both belong to the same misordering error.

```
(8) <result>
<sentence id= “ns”>
I like breakfast but I don’t eat rice and miso soup for breakfast
</sentence>
<sentence id= “st”>
I like breakfast but I don’t eat in breakfast rise and misosoup
</sentence>
```

```
<trial no= “01a”>
I like breakfast but I don’t eat <add>in</add> <add>breakfast</add> <msf crr=
“rice”>rise</msf> and <oms>miso</oms> <msf crr= “soup”>misosoup</msf>
<oms>for</oms> <oms>breakfast</oms>
</trial>
<trial no= “02”>
I like breakfast but I don’t eat <add>in</add> <trs_add crr=
“breakfast”>breakfast</trs_add> <msf crr= “rice”>rise</msf> and <oms>miso</oms>
<msf crr= “soup”>misosoup</msf> <oms>for</oms> <trs_oms crr=
“breakfast”>breakfast</trs_oms>
</trial>
```

This technique of dealing with transpositions is quite similar to Damerau-Levenshtein distance, but the algorithm used here is a partial implementation of the formula, developed by Hajime Mochizuki (Tono & Mochizuki, 2009).

3.2. Distributions of surface strategy taxonomy errors

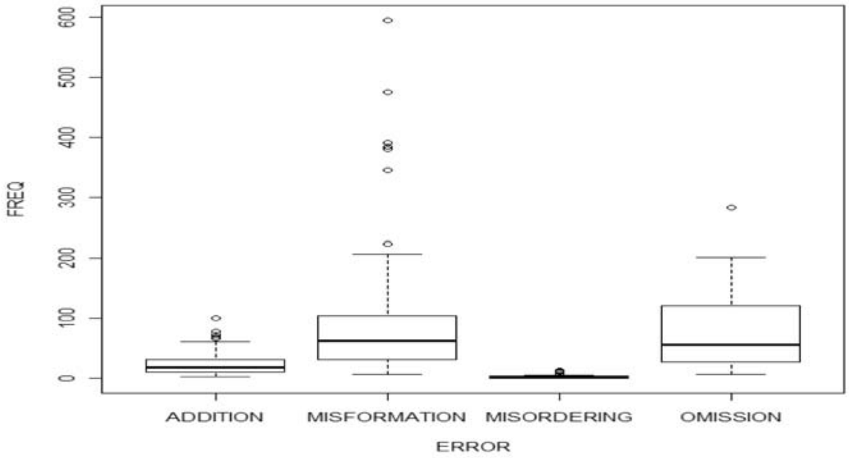
Figure 3 shows overall distributions of four types of surface strategy taxonomy errors (addition, omission, misformation and misordering). In terms of the number of error tags, misformation errors were found to be most frequent (n = 67,176), followed by omission errors (n = 49,077)[4], addition errors (n= 16,156) and misordering errors[5] (n= 2,082). Table 3 shows the breakdown of four types of errors across school years and parts of speech. This time, the frequencies are normalised per 10,000 words for comparison across different subcorpora. Overall, noun and verb errors are very frequent, followed by determiner errors. This has to be interpreted with caution because the total number of occurrences of nouns and verbs is usually greater than the other parts of speech. In this study, normalization was done for corpus size, but not for POS categories, so it is difficult to say exactly the error frequencies for nouns and verbs are greater than those of the other parts of speech. A relative measure will be needed in the future study to tease these possibilities out. Interestingly, the number of noun misformation errors (n=594.8) in Year 7 decreased dramatically through Year 7 to 9, and stayed the same across Year 10-12. One of the reasons is that Year 7 students overused Japanese words in the essays, which happened to be tagged as nouns since a POS tagger did not recognise Japanese words. There are also

[4] Please note, however, that this figure is based on the automatic extraction, whose precision is roughly 72%.

[5] The number of misordering errors has to be interpreted carefully because this feature was added after the first evaluation was done for the other three types of errors and the accuracy rate was not checked against manually corrected data.

many misformation and omission errors on verbs. However, verbs behave differently from nouns in several respects. First, the number of verb misformation errors stays almost the same throughout the school years while noun misformation errors decrease in the first three years. This may be again related to the use of Japanese words in the compositions. Second, verb omissions are very high in year 7, they decrease considerably in Year 8 and after another slight decrease in Year 9 they tend to remain constant; noun omission errors seem to follow a U-shaped curve, with a high initial proportion gradually shrinking in Years 8 and 9, to then grow again in later years. Verbs are also different from nouns in the way addition errors occur. While the number of noun addition errors decreases constantly from Year 7 to 10, verb addition errors increase from Year 7 to 10. This is mainly due to the increasing overuse of "have" as an auxiliary besides its use as a lexical verb, as learners experiment with more complex grammatical constructions.

Figure 3. Distributions of surface strategy taxonomy errors



Determiner errors are especially frequent in the case of omissions. The frequencies of omission errors are five to six times higher than addition errors, which shows that Japanese-speaking learners of English tend to omit determiners rather than oversupply them. Error rates remain almost the same throughout the school years, which shows that determiner omission errors are quite persistent in nature. Prepositions are also problematical and they are frequently omitted. Interestingly, preposition omission errors have a typically U-shaped error curve, where the errors decrease for the first three years and then increase again in a later stage. Although the number is relatively smaller, addition errors of prepositions also increase steadily as the school year increases. Preposition errors

Table 3. Normalised frequencies of 4 types of errors across school years and POSs (per 10,000 words)

Addition

YEAR	DET	NOUN	PRN	ADV	ADJ	BE	VERB	PRP	MODAL	TO	CONJ	TOTAL
7	28.8	100.8	12.0	13.7	10.0	26.4	18.6	10.2	5.5	6.4	3.5	242.8
8	25.6	67.0	14.4	15.1	9.7	22.6	23.5	19.3	3.4	11.5	3.4	223.5
9	23.7	60.8	12.4	16.3	7.1	20.9	29.0	16.3	5.6	8.6	5.0	214.7
10	32.3	38.6	19.1	35.8	6.8	29.3	78.8	30.4	16.7	11.8	6.0	315.4
11	36.7	41.2	25.4	32.9	11.7	26.6	73.5	33.5	20.3	12.3	7.3	332.3
12	33.6	42.0	25.6	35.8	13.0	28.0	69.5	32.0	18.4	11.7	7.5	329.2
												1658.0

Omission

YEAR	DET	NOUN	PRN	ADV	ADJ	BE	VERB	PRP	MODAL	TO	CONJ	TOTAL
7	176.7	283.7	138.2	56.2	79.7	80.4	200.8	126.4	24.8	32.3	23.5	1229.7
8	165.6	188.8	81.8	39.7	47.9	51.0	126.3	97.8	10.2	22.8	12.8	852.7
9	119.8	103.7	53.0	33.6	27.7	40.2	98.6	69.2	9.8	16.7	7.2	588.5
10	193.7	154.2	61.4	51.6	44.0	56.1	102.6	131.2	14.0	32.3	16.1	867.4
11	149.8	145.6	62.3	58.4	42.2	52.3	85.8	125.1	15.4	22.2	14.1	784.2
12	157.9	191.9	67.7	56.2	53.5	47.7	109.6	120.7	14.0	27.0	12.2	870.5
												5193.0

Misformation

YEAR	DET	NOUN	PRN	ADV	ADJ	BE	VERB	PRP	MODAL	TO	CONJ	TOTAL
7	46.9	594.8	104.5	62.2	63.6	134.2	223.9	38.3	11.3	7.1	16.2	1309.9
8	45.9	475.0	77.3	75.3	73.5	86.0	207.1	62.5	13.4	14.4	15.0	1153.4
9	44.1	380.4	63.2	69.6	53.2	61.7	200.0	57.2	14.8	10.5	21.6	985.3
10	60.4	391.2	61.1	151.6	79.5	67.5	202.1	95.8	24.0	15.3	34.7	1193.2
11	61.9	345.9	60.9	132.7	66.6	61.6	193.4	79.0	20.2	18.0	31.7	1082.7
12	54.9	383.7	64.7	124.2	76.7	57.9	199.8	78.8	26.0	15.7	26.7	1121.0
												6845.6

Misordering

YEAR	DET	NOUN	PRN	ADV	ADJ	BE	VERB	PRP	MODAL	TO	CONJ	TOTAL
7	1.1	14.0	2.9	2.4	4.2	0.4	5.1	1.3	0.4	0.4	0.9	40.2
8	2.6	11.7	2.8	3.4	2.9	1.0	3.6	1.0	0.2	0.8	1.2	39.2
9	1.0	8.5	2.7	2.8	2.3	1.2	2.8	1.0	0.4	0.4	1.1	33.3
10	3.7	12.1	5.1	4.4	2.5	1.6	3.5	4.7	0.5	1.1	2.8	51.9
11	4.2	11.3	3.2	5.0	3.3	1.9	4.9	2.8	0.8	1.0	1.7	51.1
12	3.9	8.8	3.4	4.4	3.5	2.3	4.8	3.0	0.4	0.8	1.7	49.0
												264.6

will become more frequent as learners learn more prepositions and try to use them to express more complex ideas in English.

It is noteworthy that errors observed with a frequency analysis based on the surface strategy taxonomy have some general characteristics, which may point to some general interlanguage developmental trends. First, omission errors are more common than additions. Naturally, L2 learners start with simplified structures, which lack required elements such as determiners, prepositions, verbs, and nouns to form well-formed sentences. As their proficiency levels go up, however, the ratio of addition errors to omission errors will become higher. This indicates that the more proficient L2 learners become, the more varieties of language they will use and they will thus take increasingly more risks in expressing themselves, which will lead to more errors. This is clearly shown in the increasing frequencies of errors related to verbs, adverbs, adjectives, prepositions, conjunctions and modals (see Table 3). This tendency is closely related to lexical choice errors with major content words and is known to have an inverted U-shaped curve (Hawkins & Buttery, 2010), which indicates that errors of this type will continue to increase as learners become proficient from the beginning to the intermediate levels and as the repertoire of language becomes wider and errors will decrease or disappear when they reach near-native proficiency levels. In JEFLL, because of the lower proficiency levels, most addition errors continue to grow in number or stay the same throughout the six years.

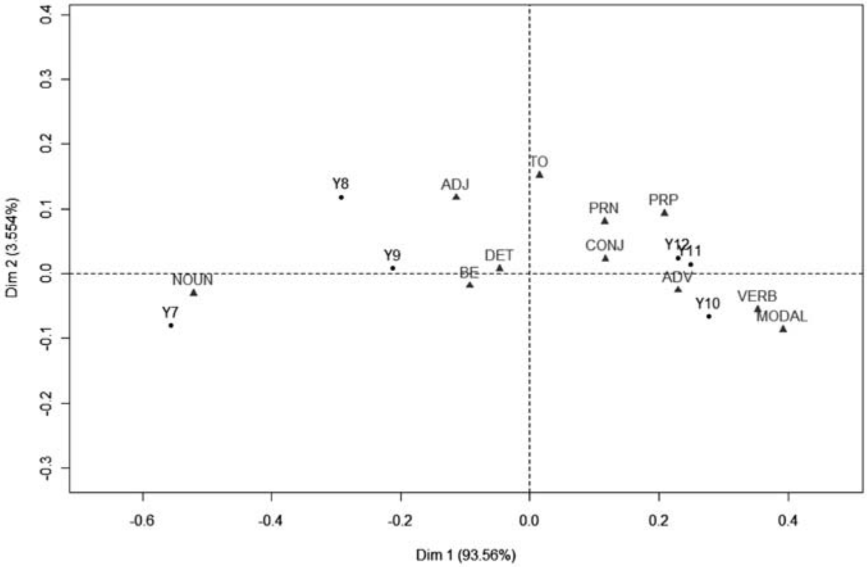
The statistics, however, have to be interpreted carefully in the case of misformation errors, given that the identification of misformation errors by edit distance has lower precision/recall scores in comparison to the other error types. There is also an influence of the use of Japanese words in the essays, which boosted the frequencies of noun errors, especially in Year 7.

3.3 Correspondence Analysis

There are many ways to approach multifactorial data. The primary purpose of this study is to identify criterial features that distinguish one proficiency level from another. What is meant by criterial features here is a set of surface strategy errors classified according to parts of speech. Therefore, what needs to be done is to extract error categories that are salient enough to serve as criteria for distinguishing learners' proficiency levels. Hawkins and Buttery (2010) examined error frequencies across different CEFR levels by setting thresholds of error ratio to determine the significance of errors as criteria. Since the JEFLL Corpus was not categorised for CEFR levels, a different approach had to be taken. The simplest way to analyze contingency tables like Table 3 is the Chi-square test, but unfortunately, the Chi-square test does not provide a solution to the problem of identifying detailed relationships among column and row variables. Though it tests whether two variables are independent of each other, it does not allow us to characterize the school years in terms of the distribution of POS errors. Answers to the question are provided by correspondence analysis. Correspondence analysis is a statistical visualization method for picturing the associations between the levels of a two-way contingency table. In this case, the two variables were school years (row variables) and POS errors (column variables). This technique plots together in a bi-dimensional space groups of texts (Years 7-12) and features, thus representing graphically which features are more significant in identifying each group. Dimension scores were first calculated independently for the two variables, thus the distance between column or row variables is meaningful in independent row or column plots, which are not listed here. On the biplots like Figures 4 onwards, only the dimensions between row and column points are meaningful, because the elements for the two variables were plotted at the same time on the bi-dimensional space using a technique called symmetrical normalization. The simplest way to interpret the biplots is to draw a line on the plot through the origin (0,0) and the point corresponding to the POS error in question (NOUN, for instance). Perpendiculars to this line are dropped from each school year's position on the plot. Look at how close each POS error is on this line to the point, NOUN. One can see Y7 is the closest, Y8 and Y9 follow, and the other three (Y10, 11, and 12) are furthest. The relative positions between the school years and the POS errors show that NOUN is the most closely associated with Year 7 and VERB, MODAL, PRP, ADV tend to be related to more advanced levels (Years 10-12). DET, on the other hand, is positioned almost in the center (0,0), which means that DET is relatively the same in frequency across school years. An analysis was made independently for each of the four error types, due to the complexity of multiple correspondence analysis. Figure 4 shows the results of correspondence analysis for addition errors.

The horizontal axis (Dimension 1) explains 93.56% of the overall Chi-square value (or inertia), which means that we can interpret the results almost exclusively with regard to their positions on the first axis. Regarding the positions of the school year, Year 7 was placed on the leftmost edge, Year 8 and Year 9 were close together on the left side, much closer to the origin for the first axis, while Year 10, Year 11, and Year 12 appeared very closely together on the right side of the origin for the first axis. Therefore, it is fair to conclude that the first axis separates essays written by junior high school students from those by senior high school students, which means the first axis basically shows the differences in proficiency levels. Interestingly, all three groups in senior high school (Year 10-12) were very close in position, which indicates that as far as addition errors are concerned, the three groups were very similar. The same thing can be

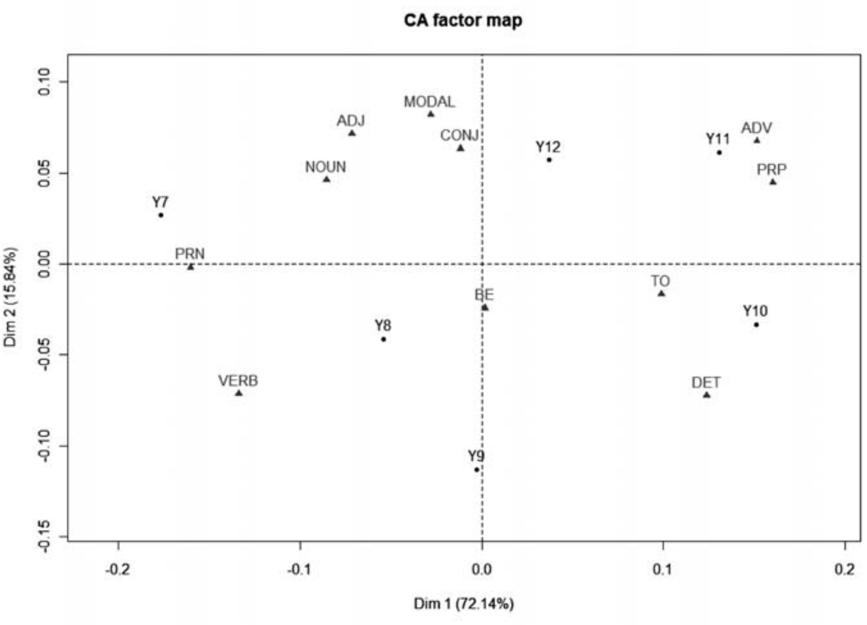
Figure 4. Correspondence analysis (addition errors)



said about Year 8 and Year 9. Year 7 was apart from the other groups, showing that the group behaved very differently. The positions of POS errors in relation to the school years revealed interesting patterns. Noun errors (NOUN), for example, were close together with Year 7, far from the other error groups. As can be seen from Table 3, noun errors were very high in frequency for Year 7, mainly due to the fact that Year 7 students used Japanese words very often in the compositions, which were analysed as nouns by a POS tagger. Thus, high frequencies of noun errors involve the use of Japanese words in the passages. Another reason why noun errors were located far from the other groups is that their frequencies kept going down significantly from Year 7 to 9 until they became stable for higher levels. On the other hand, verb errors (VERB) and modal auxiliary errors (MODAL) showed opposite tendencies, with their frequencies continuing to increase toward Year 12. Figure 5 shows the results of correspondence analysis for omission errors.

The overall picture here is different from addition errors. The relationship between the two variables (POS omission errors X school year) summarised in the biplots in Figure 5 can be interpreted by looking at Table 3 again. The students' groups were not plotted in the order of the school years. Rather, Year 12 was placed toward the centre, and Year 10 and Year 11 were on the rightmost end. This is partly due to the fact that error frequencies reported in Table 3 suddenly increased in Year 10 after a gradual decrease from Year 7 to 9. It seems that omission errors did not simply decrease as the school year went up. In

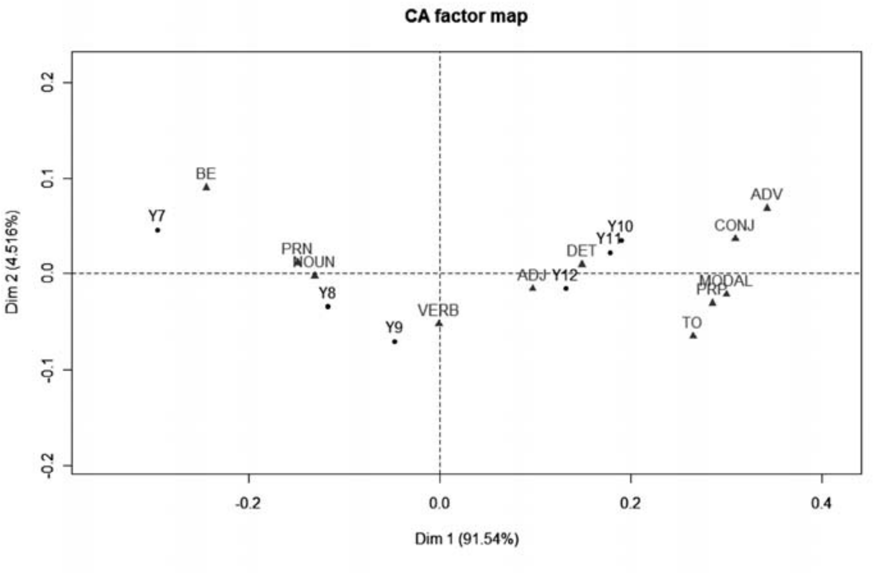
Figure 5. Correspondence analysis (omission errors)



many cases, omission errors decreased in frequency from Year 7 to 9, rose again in Year 10 and either stayed the same toward Year 12 or fluctuated through the three years in senior high, which explains why the points for these years do not follow a straight line from left to right in the biplot. Also there were two different groups of POS errors, divided by the origin of the axis. Those placed on the left side of the origin for the first axis (PRN, NOUN, VERB, and ADJ) all shared the same tendency that their frequencies in Year 7 were much higher, compared to the other errors (ADV, PRP, DET, and TO), whose frequencies were not very high in Year 7 and gradually became higher in Year 10 - 12. The former group consists of parts of speech that are primary components of constructions and open class in nature (except for PRN) whereas the latter group belongs to closed class and their primary functions are connecting components in a sentence. This shows that learners at the beginning stage of acquisition fail to supply major elements such as verbs or nouns, but these omission errors tend to decrease as they progress. On the other hand, they will have more errors on function words such as prepositions, determiners, infinitives, and adverbs, which help to modify principal elements in a sentence to make it more complex.

Figure 6 illustrates the way misformation errors occurred and their relationship with school years.

Figure 6. Correspondence analysis (misformation errors)



For misformation errors, Dimension 1 explains 91.5% of the inertia, thus this horizontal axis tells us most of the relationship between error types by POS and the school years. As is shown in Figure 6, the school years were basically plotted in the order of the progression of the grades, but again the senior high school groups (Year 10 to 12) appeared close together in almost the same area, which shows that error patterns in the upper-grade groups were quite similar. A striking difference was found in two groups of POS errors. By examining frequencies in Table 3 to interpret the plot, the group plotted on the left side of the origin for the first axis (BE, PRN, NOUN) all had the tendency to be very high in frequencies in Year 7, gradually decrease to Year 9, and then stay at the lower level throughout Year 10 to 12. On the other hand, the group plotted on the right side of the origin for the first axis (ADV, CONJ, MODAL, PRP, TO) all showed the similar tendency that the error frequencies increased constantly toward Year 12. The other POS errors (VERB, ADJ, DET) showed almost the same error frequencies throughout the six years. Misformation errors showed a tendency similar to addition errors in the sense that the growth of learners' vocabulary and their repertoire, as they move from the beginning to the lower-intermediate stages of learning, will lead to taking more risks to use newly learned items, thus resulting in more errors. This also has something to do with the syntactic elaboration of sentences, which is shown in the errors of closed system such as CONJ, MODAL, PRP and TO.

3.4. Refining the analysis by using neighbour clustering

Even though correspondence analysis shows a graphical image of the relationship between the variables in terms of distances, it does not give us any information about how items in the variables can be clustered meaningfully. Cluster analysis is usually a common technique for classification tasks, but it has a serious problem in the sense that standard cluster analysis cannot take into account the 'time factor'. The present data is pseudo-longitudinal in nature, and it is desirable to find meaningful clusters based on error frequencies, but at the same time sensitive to the order of data points along the time sequence.

Gries & Stoll (2009) dealt with these 'variability problems' of children's mean MLUs over time as 'developmental problems'. He rightly commented that "one cannot simply lump together all utterances with a particular MLU value because this procedure would be completely blind to the order of elements and the developmental implications this may have" (ibid: 222). This problem is similar to mine, and his solution was to employ 'variability-based neighbour clustering (VNC)'. VNC is a hierarchical cluster-analytic approach, which takes into account the temporal ordering of the data (Hilpert & Gries, 2009, p. 390). What VNC basically does is to access the first and the second time period (Year 7 and Year 8, for instance) and compute the similarity measures of their respective two values (using e.g. variation coefficients or summed standard deviations, depending on the nature of the data), then proceed to do the same for all successive pairs of values, the second and the third, the third and the fourth, etc. always storing the similarity measures. After that, VNC identifies the largest similarity score, which indicates the values that are most similar to each other and thus merit being merged into one group. After the first iteration, there are only five data points, the first two groups having been merged. This process will be repeated until only one data point is left.

Figure 7 shows the result of VNC for noun addition errors. The left panel of Fig. 7 plots the distance in summed SD as an analogue to scree plots in principal component analysis, where they are used as a guideline to determine how many factors should be included in a model. The plot indicates how many different stages could be identified within a developmental progression, as in our case, the series of school years. The plot shows substantial distances between the first three largest clusters, i.e. steep slopes between the first three points. After the third cluster, the curve levels off to the right and becomes nearly horizontal. This suggests a division into three separate developmental stages, each represented by a cluster. The dendrogram (right panel) illustrates what these clusters are.

Dendrograms are best read from the bottom, since they join together groups starting from those having the lowest distance. The distance is represented not in the horizontal but in the vertical axis, which means that a short vertical line represents closely associated points while a long one represents a greater distance between them. Cluster 1 distinguishes Year 7 from the rest. Cluster 2 ranges from Year 8 and Year 9, and cluster 3 ranges from Year 10 to Year 12.

Figure 7. VNC for noun addition errors (LEFT: scree plots; RIGHT: dendrogram)

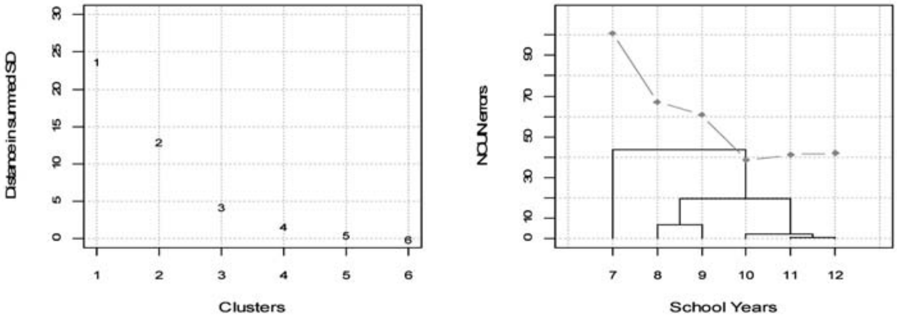
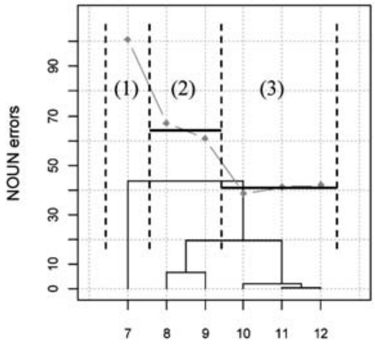


Figure 8 shows the three clusters by dividing them by vertical dotted lines. Horizontal lines under the numbers (2) and (3) indicate the mean frequencies that are observed in the data for the three clusters.

Figure 8. Three clusters in the dendrogram of noun addition errors



Dendrograms of VNC for addition and omission errors sub-classified by POS are reported in a separate file which can be accessed online at the URL http://eurosla.org/monographs/EM02/tono_fig9-10.pdf. Misformation and misordering errors were not examined because of lower precision/recall scores.

The analysis revealed that some POS errors could not produce meaningful clusters. When the scree plots did not show any steep slope between the points, the results were not very useful even though the dendrograms in Figures 9 and 10 made two clusters anyway, just for the sake of giving an idea of where the division could be made. Regarding the addition errors in Figure 7, only nouns, adverbs, verbs, modals and prepositions made two meaningful clusters. Except for noun addition errors, which produced three clusters due to the effects of the intensive use of Japanese in Year 7, the first cluster ranges from Year 7 to Year 9, and the second ranges from Year 10 to Year 12, thus clearly dividing the junior high group and the senior high group in terms of the error occurrence patterns. This confirms the findings observed in correspondence analysis in Figure 4, and without VNC it was difficult to state which POS errors actually contributed to the divisions.

The omission errors show slightly more complicated pictures. As was shown in Figure 5, there is a tendency for omission errors to decrease throughout Year 7 and Year 9, and increase again in Year 10 toward Year 12, which is due to the fact that learners took more risks to extend their repertoire of English at later stages, yielding more errors. Learners tended to master the use of basic lexis and grammar that they had learned at the early stage, but as they moved onto more advanced stages, they produced different types of omission errors. In terms of accuracy rates, this is a well-known inverted U-shaped developmental curve. Among the omission errors, only nouns, pronouns, and verbs seemed to show meaningful clusters. Interestingly, the two clusters are Year 7 and the rest in most cases. It is worth pointing out again in this connection the results of correspondence analysis. Those errors placed on the left side of the origin for the first axis (PRN, NOUN, VERB, and ADJ) in Figure 5 nearly correspond to the ones showing meaningful clusters in Figure 8, namely nouns, verbs, and pronouns. One should bear in mind that their frequencies in Year 7 were much higher, compared to the other errors (ADV, PRP, DET, and TO), whose frequencies were not very high in Year 7 and gradually became higher in Year 10 - 12. Therefore, the results of VNC suggest that three omission errors above all (noun, verb and pronoun) are useful in distinguishing Year 7 from the rest of the groups, while for the other POS errors the results are not conclusive.

4. Discussion

So far, I have proposed a new way of extracting errors from learner corpora and judging the status of those extracted errors as criterial features. Edit distance is a common metric to spot differences between two strings of characters. It is used intensively in other areas such as the analysis of DNA sequences. By extending its use to a comparison of learner production and target-like performance, it is possible to identify surface strategy errors semi-automatically over a large amount of learner data. The present study also shows that data reduction techniques such as correspondence analysis are useful in summarising the data. However, correspondence analysis plots do not show exactly what meaningful clusters are. In order to solve this problem, a special clustering technique called variability-based neighbour clustering was introduced. The results of the combination of these two techniques revealed the contribution of addition/omission errors for particular POSs as criterial features of the developmental stages.

Table 4 summarises the results in terms of extracted criterial features to characterise Japanese EFL learners' acquisition stages.

Table 4. Extracted criterial features for the learning stages of Japanese EFL learners

Types	POS	Criterial for:	mean error freq. of errors
Addition	nouns	[Year 7] > [Year 8 - 9] > [Year 10 -12]	58.4
	adverbs	[Year 10 - 12] > [Year 7 - 9]	24.93
	verbs	[Year 10 - 12] > [Year 7 - 9]	48.81
	prepositions	[Year 10 - 12] > [Year 7 - 9]	23.62
	modals	[Year 10 - 12] > [Year 7 - 9]	11.65
Omission	nouns	[Year 7] > [Year 8] = [Year 10 -12] > [Year 9]	177.98
	verbs	[Year 7] > [Year 8 - 12]	120.62
	pronouns	[Year 7] > [Year 8 - 12]	111.73

Note: '>' means "occur more frequently than ...";

As shown in the column of mean error frequencies, the relative frequencies of omission errors are much higher than those of addition errors. However, a closer look into the categories of omission errors by POS reveals that omission errors are only useful for distinguishing the very beginning stage of learning from the rest, as shown in the third columns in Table 4. Overall, omission errors tend to decrease toward Year 9 and then jump up again in upper grades. Since the primary purpose of this paper is to present a heuristic to identify criterial features, I will not develop this point any further. More research into omission errors at a lexical level will be needed in order to describe in more detail what is happening in this U-shaped phenomenon.

Addition errors are more sensitive to level differences and thus work as criterial features distinguishing the lower level from the upper. It is noteworthy that in all cases but noun errors, addition errors are more frequent in the upper levels (Year 10-12). Adverbs, prepositions or modals are the elements that modify main constituents of a sentence. For instance, adverbs modify either verbs, adjectives or other adverbial phrases. Prepositions usually modify nouns or verbs. Modals modify verbs to add epistemic or deontic meanings. As proficiency levels increase, learners have a wider repertoire of these lexical items and feel more confident in using basic lexis and grammar, which leads to a greater chance that they take risks to use new items to convey subtler meanings. Sometimes they fail to make the right word choices, and thus have more lexical choice errors, but in other cases they overuse and add unnecessary words to sentences, yielding non-target-like outcomes.

There are a few methodological issues related to this approach. One is the issue of "normalisation". In this study, a parallel set of the original students' essays and their proofread versions were used for edit distance. In order to produce parallel corpora, one native speaker instructor, who was trained for error corrections, worked on all of the 10,000 essays. It is a well-known fact (cf. Milton & Chowdhury, 1994) that a certain error in a sentence can be corrected (i.e. normalised) in more than one way. I am aware of such multiple interpretations of L2 learner errors and that there is also a system of multi-layered annotations, such as MMAX2 (Müller & Strube, 2006), so that one can annotate possible choices of normalisation in more than one way. In this study, however, I did not take that approach for two main reasons. First, native speakers' correction possibilities could be almost infinite if we allow for multiple possibilities of normalization. If the native speaker wanted to extend their correction to stylistic or discourse elements, a number of different ways of correcting and refinement could be possible, and it would thus be almost impossible to incorporate those into the analysis, although the variation in native speakers' judgments could be a valuable research object in its own right. The second reason is that even though there were some minor inconsistencies in normalisation patterns, corrections in more than 10,000 essays should cause some patterns of use/misuse to emerge, which help to explain the patterns of development over different school years. There is no error annotation system that can be said to be superior to others in and of itself. Error annotation adequacy is always relative to the research goals.

It would be pedagogically very significant to identify criterial features from learner corpora. If those performance features can work as 'classifiers' in the sense of text mining, it is possible to produce an automatic performance analysis system, in which the input by an L2 learner will undergo text analysis and his or her proficiency level will be determined by checking the existence of criterial features. In language testing, with such criterial features available, the assessment procedure of speech or writing can be facilitated by first automatically assessing the text based upon known criterial features and then by human intervention only on those aspects that need human judgements. What we need is a formal procedure for extracting and identifying criterial features. This paper proposes a formal, methodological procedure for identifying criterial features in IL development. Using edit distance, possible error candidates are automatically extracted. Subcategorising those errors by POS can be done by automatic POS tagging. Variability-based neighbour clustering will make it possible to aggregate similar groups and cluster variables into meaningful stages of learning. This procedure can be applied to any kinds of learner corpora if they have parallel versions of the data set ready for edit distance. A word of caution is in order here. The approach presented in this paper is only applied to extracting surface strategy taxonomy errors. It will not deal with semantic errors such as tense/aspect morphology, for this kind of information is not revealed on the surface. Also this method is only applicable to "errors" as criterial features. It will not be used to extract well-formed language features as criteria. This should not be the limitation of this study, however, because well-formed linguistic features are usually much easier to extract, using ordinary corpus analysis tools such as concordancing or n-gram analysis over different sets of learner data. I hasten to add that VNC can also be used for analysing both errors and non-errors as long as frequency information is available regarding given linguistic features across different stages.

Some final notes are in order with respect to methodological issues. The detection of misformation errors could be improved. At the moment, the accuracy of misformation errors is sufficiently high with respect to one-to-one lexical mapping relation. If the mapping is between one to multiple words or vice versa, the accuracy rate suddenly drops. In order to solve this problem, ontological knowledge such as POS-labelled wordlists or something of the kind will be needed, which is more complex than simple surface character-level similarities. The results of multivariate analysis should also be further interpreted from both macroscopic and microscopic viewpoints. In macro views, my findings should be related to a much larger framework of criterial features and CEFR levels. If several dozen criterial features were identified, it would be necessary to re-classify those criterial features in terms of their relative importance. Also there are some cases in which a bundle of criterial features will work better than a single feature, thus some methods have to be proposed in order to figure out how to deal with such possibilities. I should admit that identifying criterial features is one thing, but constructing the overall framework is quite another. This whole process of identifying criterial features using learner corpora and constructing the overall theoretical framework based on those criterial features seems to me a very promising research strand, which definitely links learner corpus research to SLA and English language teaching and assessment in a meaningful way.

References

Abe, M. (2003). A corpus-based contrastive analysis of spoken and written learner corpora: the case of Japanese-speaking learners of English. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003)* (pp. 1-9). Lancaster University: University Centre for Computer Corpus Research on Language.

Abe, M. (2004). A corpus-based analysis of interlanguage: errors and English proficiency level of Japanese learners of English. In Y. Tono (Ed.), *Handbook of An International Symposium on Learner Corpora in Asia* (pp. 28-32). Tokyo: Showa Women's University.

Abe, M. (2005). A comparison of spoken and written learner corpora: analyzing developmental patterns of grammatical features in Japanese Learners of English. *The Proceedings of the NICT JLE Corpus Symposium* (pp. 72-75). Kyoto: National Institute of Communications Technology.

Abe, M. & Tono, Y. (2005). Variations in L2 spoken and written English: investigating patterns of grammatical errors across proficiency levels. *Proceedings from the Corpus Linguistics Conference Series* (Vol. 1, no.1) Retrieved from http://www.corpus.bham.ac.uk/pclc/ index.shtml

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning,* 33, 1-17.

Briscoe, E., Carroll, J., & Watson, R. (2006). *The second release of the RASP System.* Retrieved January 15, 2012, from http://acl.ldc.upenn.edu/P/P06/P06–4020.pdf

Dulay, H., Burt, M., & Krashen, S. (1982). *Language Two.* Oxford: Oxford University Press.

Filipovic, L. (2009). *English Profile – Interim report.* Internal Cambridge ESOL report, April 2009.

Goldberg, A. E. (1995). *Construction: A Construction Grammar Approach to Argument Structure.* Chicago: University of Chicago Press.

Goldberg, A.E. (2006). *Constructions at Work: the nature of generalization in language.* Oxford: Oxford University Press.

Granger, S. (Ed.). (1998). *Learner English on Computer.* London/New York: Addison Wesley Longman.

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

Gries, S. Th. & Divjak, D. (2012). *Frequency Effects in Language Learning and Processing.* Berlin: Mouton de Gruyter.

Gries, S. Th. & Stoll, S. (2009). Finding developmental groups in acquisition data: variability-based neighbor clustering. *Journal of Quantitative Linguistics 16*(3), 217-242.

Hawkins, J. A. & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal, 1*(1), 1-23.

Hendriks, H. (2008). Presenting the English Profile Programme: In search of criterial features. *Research Notes, 33*(3), 7-10.

James, C. (1998). *Errors in Language Learning and Use: Exploring Error Analysis.* London: Longman.

Kaneko, E. (2004). Development of noun phrases in the interlanguage of Japanese EFL learners. Poster session presented at the 6th Conference of the Japanese Society for Language Sciences (JSLS 2004), Nagoya.

Kaneko, E. (2006). Corpus-based research on the development of nominal modifiers in L2. Paper presented at the American Association of Applied Corpus Linguistics (AAACL), Flagstaff, Arizona.

Kasper, G. (1997). "A" stands for acquisition: A response to Firth and Wagner. *Modern Language Journal, 81*(3), 307-312..

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707-710.

Manning, C. & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge MA: MIT Press.

Milton, J.C.P. & Chowdhury, N. (1994). Tagging the interlanguage of Chinese learners of English. *Proceedings of the joint seminar on corpus linguistics and lexicology* (pp. 127-143). Hong Kong: Language Centre, HKUST.

Miura, A. (2008). Kaiwa (NICT JLE) vs. Sakubun (JEFLL) Corpus no hikaku to bunseki [A comparison of spoken and written corpora]. *English Corpus Studies*, 15, 135-148.

Müller, C. & Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In S. Braun, K.Kohn, & J. Mukherjee (Eds.), *Corpus Technologgy and Language Pedagogy: New Resources, New Tools, New Methods* (pp. 197-214). Frankfurt: Peter Lang.

Parodi, T. (2008). L2 morpho-syntax and learner strategies. Paper presented at the Cambridge Institute for Language Research Seminar, Cambridge, UK.

Salamoura, A. & Saville, N. (2009). Criterial features of English across the CEFR levels: evidence from the English Profile Programme. *Research Notes*, 37, 34-40.

Tono, Y. (1998). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In *TALC (Teaching and Language Corpora) '98 Proceedings* (pp. 183-187). Oxford: Seacourt Press.

Tono, Y. (2000). A corpus-based analysis of interlanguage development: Analysing part-of-speech tag sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk, B., & J.P. Melia (Eds.), *PALC'99: Practical Applications in Language Corpora* (pp. 323-340). Frankfurt: Peter Lang.

Tono, Y. (2004). Multiple comparisons of IL, L1 and TL corpora: the case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and Language Learners* (pp. 45-66). Amsterdam: Benjamins.

Tono, Y. (2009). Variability and invariability in learner language: A corpus-based approach. In Y. Kawaguchi, M. Minegishi, & J. Durand (Eds.), *Corpus Analysis and Variation in Linguistics* (pp. 67-82). Amsterdam: Benjamins.

Tono, Y. & Mochizuki, H. (2009). Toward automatic error identification in learner corpora: A DP matching approach. Paper presented at Corpus Linguistics 2009, Liverpool, UK.

UCLES-RCEAL Funded Research Projects. Retrieved January 15, 2012, from http://www.englishprofile.org/images/pdf/ucles_rceal_projects.pdf.

Williams, C. (2007). *A preliminary study into the verbal subcategorisation frame: Usage in the CLC*. Unpublished manuscript.