Lou Burnard/Tony McEnery (eds.)

Rethinking Language Pedagogy from a Corpus Perspective

Papers from the third international conference on Teaching and Language Corpora

Offprint
2000

PETER LANG
Europäischer Verlag der Wissenschaften

A computer learner corpus based analysis of the acquisition order of English grammatical morphemes

Yukio Tono

Lancaster University

1. Introduction

The examination of learner language has always been of importance to second language acquisition (SLA) researchers. Learner language provides such researchers with insights into the process of second language acquisition, with the aim of applying any findings to a variety of practical goals in language teaching: improved syllabus design, material design, task design, testing, and so on.

A number of different approaches have been taken to the description of learner language. Ellis (1994:44) identified four major approaches:

- the study of learners' errors
- the study of developmental patterns
- the study of variability
- the study of pragmatic features

The study of learners' errors was conducted quite intensively in the late 1960s and 70s after Pit Corder (1967) made a significant claim that L2 learners, like L1 learners, have a 'built-in-syllabus,' which guides their progress in language . Selinker (1969) coined the term 'interlanguage' to refer to the special mental grammars which learners constructed during the course of their development. Interlanguage theory treats learner behaviour, including their errors, as rule-governed.

Error analysis went out of fashion in the 1980s as a number of methodological and theoretical problems with it were identified. In short, error analysis did not provide a complete picture of how learners acquire an L2 because it described learner language solely as a collection of errors (ibid: 73). As a consequence, more and more attention was paid to the entirety of learner language. Central to this enterprise is the description of developmental patterns of interlanguage (point 2).

Dulay and Burt were among the first to conduct empirical research on the acquisition order of certain grammatical features of English. They carried out their research on grammatical morpheme acquisition order, which was first investigated by Roger Brown in L1 acquisition (Brown 1973). Throughout their papers, it was their claim that L2 acquisition proceeds quite systematically and that the acquisition order is not rigidly invariant but is remarkably similar irrespective of the learners'

L1 backgrounds, their age, and whether the medium of data-collection is speech or writing (Dulay and Burt 1975). Since then, more than fifty L2 morpheme studies have been reported, many using a variety of data collection and analysis procedures (see Larsen-Freeman and Long 1991; Ellis 1994 for review).

The methodology utilised in the early morpheme studies was criticised (see Long and Sato 1984). However, as Larsen-Freeman and Long (1991) stated, in spite of their limitations, the morpheme studies provide strong evidence that interlanguages exhibit common accuracy/acquisition orders. Contrary to what some critics have argued, there are many studies conducted with sufficient methodological rigour and showing sufficiently consistent general findings for this finding to be investigated and taken seriously (ibid: 92). It is in the light of such suggestions that I am seeking to verify their findings by analysis computer learner corpora.

2. Learner corpora: new horizons

Although a corpus-based approach to SLA research and foreign language teaching is still in its infancy, there has been a growing interest in this new field. There are several reasons for this. Firstly, the practical prerequisites for corpus-based teaching and learning have improved dramatically. Smaller, cheaper computers with more data storage space have become widely available to the public. The amount of data available in machine-readable form becomes greater and greater every year. Secondly, a growing awareness of the usefulness of quantitative data provided a major impetus to the re-adoption of the corpus-based language study as a methodology in linguistics (McEnery and Wilson 1996: 18). Many SLA researchers have found it very difficult to take what theoretical linguists or psycholinguists say about an L1 acquisition model and check if the same abstract linguistic principle is still applicable in L2 learning. More and more researchers now prefer to look at real language performance data instead of relying too much on intuitive or introspective data. Thirdly, the role of teachers and learners has been changing. Students are increasingly encouraged to take charge of their own learning. The data-driven learning paradigm has gained acceptance in language teaching classrooms (Johns 1993)

The benefit of large corpora is being fully appreciated amongst the EFL community as witnessed by the publication of 'corpus-based' pedagogical dictionaries such as the *Collins COBUILD English Dictionary* (1987, 2nd edition in 1995) or the *Longman Dictionary of Contemporary English* (1978, 3rd edition in 1995). Native speaker corpora showed us how words are used together in naturally occurring texts and helped us establish criteria for learners to define what it means to be 'target-like.'

Recently there has been a growing awareness that it is useful to investigate learner language by collecting a large amount of learner performance data on computer. The term 'learner corpus' was first used for data gathered for the Longman learners' dictionaries, in which the information on EFL learners' common mistakes was provided based upon the Longman learners' corpus. A project called the International Corpus of Learner English (ICLE) was launched as a part of ICE (International

Corpus of English) project (Granger 1998) in 1990. Now more than a dozen projects constructing learner corpora have been underway around the world.[1]

In this paper, I will revisit the once popular topic of SLA research — English grammatical morpheme acquisition studies — and try to see how computer learner corpora could possibly shed some new light on this research question. There are two reasons why I chose morpheme studies as my primary topic for investigation. First, as Ellis (1990) said, morpheme acquisition studies were a kind of performance analysis in the sense that it aimed to provide a description of the L2 learner's language development and looked not just at deviant but also at well-formed utterances (Ellis 1990:46). Performance analysis provides a basis for investigating the following important questions:

- Is there any difference between the order of instruction and the order of acquisition?
- Is it possible to alter the 'natural' order of acquisition by means of instruction?
- Do instructed learners follow the same order of acquisition as untutored learners or a different order? (Ellis 1990: 139)

Computer learner corpora, if used properly with a suitable research design, prove to be an effective means of answering these interrelated questions by providing the evidence of learner language in a more systematic and exhaustive way. Secondly, although there is a criticism of morpheme studies to the effect that its 'accumulated entities' view of L2 acquisition (Rutherford 1987) is misplaced (Ellis 1990:141), morpheme order studies are still a good starting point to see how effective learner corpora could be in describing interlanguage.

3. Morpheme studies: short review

In the early 1970s, it was discovered that English children learn grammatical morphemes (i.e. morphemes such as *-ing* and *the*, which play a greater part in structure than content words such as *dog*) in a definite sequence (Brown 1973). Dulay and Burt (1975) decided to replicate the study with L2 learners. They made Spanish-speaking children learning English describe pictures and checked how often the children supplied eight grammatical morphemes in the appropriate places in the sentence. The results showed that L2 learners have a common order of difficulty for grammatical morphemes as shown below:

Table 1. An accuracy order of grammatical morphemes (Dulay and Burt 1973)

1	plural *-s*	book*s*
2	progressive *-ing*	John go*ing*
3	copula *be*	John *is* here

[1] See my WWW page (http://www.lancs.ac.uk/postgrad/tono/) for a list of learner corpora project around the world.

4	auxiliary *be*	John *is* going
5	articles	*The* books
6	irregular past tense	John *went*
7	third person -*s*	John like*s* books
8	possessive -*s*	John*'s* book

One of the problems for rank orders that Dulay and Burt first introduced is that they disguise the difference in accuracy between various morphemes. For instance, a morpheme that is just 1 percent lower than another morpheme is given a different ranking in just the same way as a morpheme that is 25 percent lower. To overcome this problem, Krashen (1977) proposed a grouping of morphemes (see Figure 1). He claimed that it was "a natural order supported by the longitudinal and cross-sectional, individual and grouped SL findings. Items in the boxes higher in the order were regularly found (80 - 90%) accurately supplied in obligatory contexts before those in boxes lower in the order." (Krashen 1977: 151)

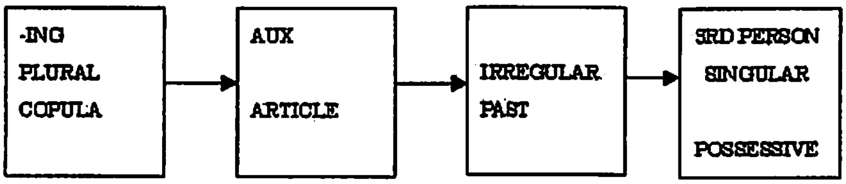

Figure 1. Proposed natural order for L2 acquisition (Krashen 1977)

The results were used to claim that there was a more or less invariant order of acquisition which was independent of L1 background and age. Although this order was slightly different from that found for the same morphemes in L1 acquisition research, it provided evidence in favour of the existence of universal cognitive mechanisms which enabled learners to discover the structure of a particular language (see Ellis 1994 for the details of review).

Although there were some strong criticisms of the morpheme studies (for example, Hatch 1978; Long and Sato 1984), the interest in morpheme acquisition grew. Several different approaches to morpheme studies developed focusing on the target-like use analysis of morphemes, as opposed to obligatory context analysis only (Pica 1982; Lightbown 1983), morpheme acquisition in different L2 contexts (Fathman 1978; Makino 1979; Sajavaara 1981), morpheme acquisition by learners with different L1 backgrounds (Mace-Matluck 1977; Fuller 1978), and morpheme acquisition in L2s other than English (Bye 1980; van Naerssen 1986). Reviewing the previous literature, Larsen-Freeman and Long (1991) concluded that these studies provided strong evidence of a stable developmental order irrespective of L1 spoken or target L2 (Larsen-Freeman and Long 1991: 92).

4. Research method

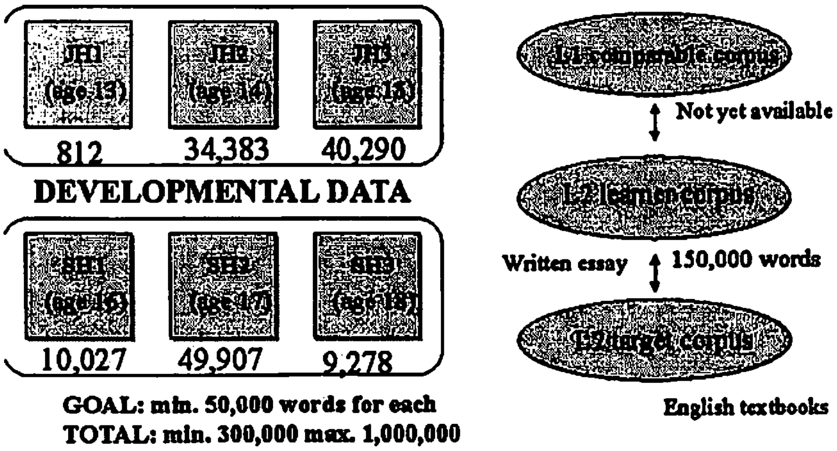

Figure 2. Components of the JEFLL Corpus, March 1999

4.1. Purpose

The purpose of this study is to investigate the accuracy order of grammatical morphemes by using a corpus of Japanese EFL learners. With this done, I will compare the results with the order proposed by Dulay, Burt and others.

4.2. Hypotheses

In carrying out my study I will explore the following hypotheses:

- The overall picture of accuracy order produced by Dulay and Burt will be confirmed.
- There will be a difference in accuracy order of some items, specific to Japanese learners of English.

By my exploration of these hypotheses I will show that a large collection of learner language data is an effective tool to re-examine the results of previous L2 acquisition order studies and establish a solid basis for more systematic and exhaustive analysis of interlanguage.

4.3. Learner corpus data

The learner corpus data used for this study was a subsection of the Japanese EFL Learner (JEFLL) Corpus (c. 300,000 words). The framework of the JEFLL Corpus as of March, 1999 is shown in Figure 2.

It primarily consists of a collection of 3000 Japanese EFL learners' written essays. The data is developmental in nature in the sense that I obtained written essays on the same topics from learners across different school years (aged from 13 to 18 years;

JH means junior high school and SH means senior high school). The subjects were asked to write in-class essays on argumentative or descriptive topics such as "Which do you prefer for breakfast, rice or bread?" or "What would you take out with you in case of a major earthquake?" The subjects were asked to write a free composition for about 20 minutes without dictionaries. The length of composition varies depending on the students, but on average it is around 50 words per essay for JH2-3 and around 100 words for SH 2-3.

4.4. Tagging schemes

I prepared the following tagset for the annotation of grammatical morphemes in the corpus:

Table 2. Annotation codes used

Correct forms	<COP>, <ART>, <PL>, <AUXBE>, <PROG>, <IRPST>, <3PS>, <POS>
Incorrect forms	<ER_COP>, <ER_ART>, <ER_PL>, <ER_AUXBE>, <ER_PROG>, <ER_IRPST>, <ER_3PS>, <ER_POS>

Each text was tagged according to the criterion set for the Bilingual Syntax Measure proposed by Dulay and Burt. In other words, I only looked at the 'obligatory context', i.e. contexts that require the obligatory use of grammatical morphemes in samples of learner language. It was almost impossible to make error judgements for each item automatically, but it was quite time-consuming to look at every occurrence of those items above manually. Therefore, I tried to make the process as automatic as possible.

First, the learner text was POS tagged with CLAWS, a part of speech tagger developed at Lancaster University. Then irrelevant POS tags were filtered out and relevant tags were converted according to the following basic rules:

- Tag <COP> or <AUXBE> based on the following context:
    - Look for every occurrence of *be* verb (_VB*)
    - If those *be* verbs are followed by either verbs with a progressive marker (VVG) or a past participle marker (VVN), then assign the tag <AUXBE>
    - If those *be* verbs are followed by adjectives (J*) or nouns (N*), then assign the tag <COP>.
- Assign the tag <PL> to all the nouns with the tag _NN2
- Assign the tag <POS> to all the parts tagged _GE
- Assign the tag <PROG> to all the words tagged _VVG
- Assign the tag <3PS> to the words tagged _VVZ
- Assign the tag <ART> to *the_AT, a_AT1, an_AT1*

- <IRPST> should be tagged by looking at each verb labelled VVD. Both regular and irregular verbs were labelled VVD, so there is no way of distinguishing them.

After these tags were automatically assigned, errors were manually tagged by checking the concordance lines retrieved with the above tags. Since there is the possibility of omission errors, which would be missed by this technique, all of the files were manually checked to see if the correct form tags were properly assigned and if there were any omission or misformation errors. After the tagging was done, I calculated the accuracy with which the morphemes were actually used in context. I followed the measurement method adopted by Dulay and Burt, namely looking only at the obligatory context, so information on overuse of the forms was not used even though the corpus itself contained that information.

The following is a sample of tagged text:

```
I have hardly had <ART> a </ART> bad dream.
I <AUXBE> am </AUXBE> often followed by someone in <ART> a </ART> dream.
But I can always flee from him as if I <COP> were </COP> <ART> a </ART> main
charactor in <ART> a </ART> movie.
Do I see <ER_ART> the </ER_ART> <PL> movies </PL> too much?  And also I
never die.
If I <IRPST> fell </IRPST> down from <ER_ART> the </ER_ART> cliff, I would
never <AUXBE> be </AUXBE> injured and I should keep smiling.
I don't have <ART> an </ART> enemy in <ART> a </ART> dream.
```

4.5. Data analysis

As was described above in 4.3, the corpus data available for JH1 and SH1 was rather small in size. Therefore the actual analysis was made on subcorpora for JH2, JH3, SH2, and SH3 only. The tagged corpus data was processed and the frequency data and concordance lines were obtained for each morpheme type by using WordSmith and TXTANA[3]. I obtained the accuracy rate by dividing the frequencies of correct forms by the sum of the frequencies of correct and incorrect forms. I also defined the state of acquisition as '90% correct' in the same way as defined in the Bilingual Syntax Measure.

5. Results and discussion

The following table shows the frequency of each grammatical morpheme type and the distribution of correct and incorrect use. Frequencies are given as rates per 10,000 words. <ART> occurred most often while <POS> and <3PS> occurred relatively infrequently.

[3] TXTANA is a commercial Windows-based concordancer developed by Shiro Akasegawa, URL: http://www.biwa.or.jp/~aka-san/index.html (This site is in Japanese).

Table 3. Overall Frequencies of 8 Grammatical Morphemes

	JH-2	*JH-3*	*SH-2*	*SH-3*
<COP>	311	410	477	379
<PL>	207	240	198	187
<3PS>	34	46	30	22
<POS>	20	25	31	32
<ART>	360	410	342	256
<IRPST>	207	167	166	137
<AUXBE>	86	191	86	117
<PROG>	39	43	44	27
<ER_COP>	19	15	23	21
<ER_PL>	52	56	45	24
<ER_3PS>	14	20	9	3
<ER_POS>	6	8	2	2
<ER_ART>	212	174	226	66
<ER_IRPST>	45	43	44	27
<ER_AUXBE>	11	8	13	10
<ER_PROG>	15	9	5	2

The next table shows the results of the accuracy order of eight grammatical morphemes. The morphemes are ordered according to accuracy rate. The students had least difficulty with copula *be*, and the most difficult items were definite and indefinite articles, *the* and *a*. Among the eight morphemes, copula *be*, auxiliary *be*, possessive *-s* and progressive *-ing* reached the 90% accuracy rate and were regarded as 'acquired' items, but the other four morphemes could not be produced with this accuracy rate even at the second year of senior high school.

Table 4. Results of Morpheme Acquisition in JEFLL Corpus

	JH-2	*JH-3*	*SH-2*	*SH-3*
Copula -be	94.17%	96.26%	95.5%	94.74%
Aux be	89.0%	96.1%	86.7%	92.45%
Possessive -s	76.67%	76.19%	94.8%	95.24%
Progressive -ing	72.1%	82.3%	89.8%	94.3%
Plural -s	80.0%	81.04%	81.4%	88.51%
3rd person -s	70.83%	69.57%	76.7%	89.36%
Irregular past	82.28%	79.62%	78.9%	83.69%
Article	63.02%	70.24%	60.2%	79.62%

Figure 3 shows diagrammatically the comparison of our results with the order obtained by Dulay and Burt (1975). The noteworthy difference is that the articles *the/a* are the most difficult items for Japanese learners and showed the lowest accuracy rate of all of the morphemes. Since the Japanese language does not attach articles to nouns, the proper use of articles should be very difficult for Japanese learners to acquire. Genitive *-s*, in contrast, was the item which was relatively easy for Japanese learners and is ranked higher than the order given by Dulay and Burt. Therefore, hypothesis one is not fully supported.

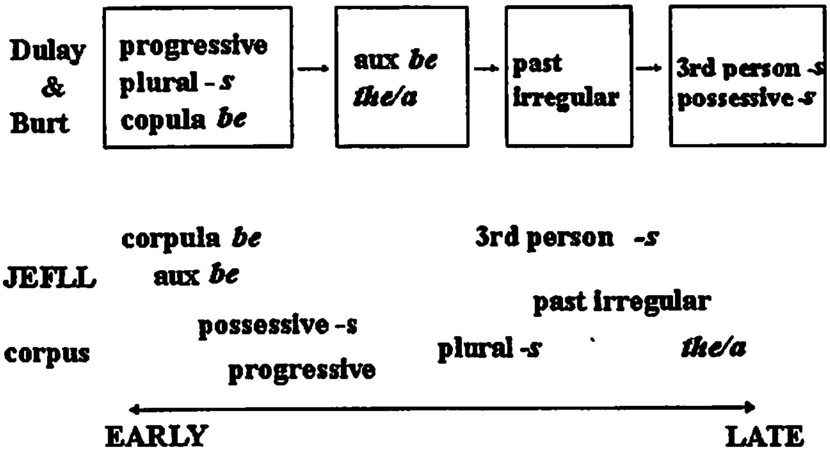

Figure 3. Grammatical morpheme order for the two studies

Spearman rank order correlation coefficients were used to see how deviant the morpheme order found in my study is from the original list by Dulay and Burt, as shown below. I tested how the morpheme order in the present study approximates to the original order by removing deviant items in turn. By this procedure, I was able to see which item was the primary factor in disturbing the original order.

Table 5. Spearman rank order correlation coefficients

D&B X Tono	no adjustment	.53 (p = 0.18)
D&B X Tono	[-ART]	.54 (p = 0.22)
D&B X Tono	[-ART/PL]	.71 (p = 0.11)
D&B X Tono	[-ART/PL/POS]	.80 (p = 0.10)

This table indicates that removing only the article (ART) from the order does not significantly improve the correlation coefficient. Removing both the articles and the plural *-s* (PL) helps improve the correlation. The additional removal of possessive *-s* (POS) improves the coefficient, although the overall correlation between the two lists is not significant (p > .05). This result shows that when we look at order only, the article system seems to be the primary factor causing distortion, but plural *-s* and possessive *-s* may also be sources of deviation.

To recapitulate, the hypothesis that there is an overall agreement in the acquisition order of grammatical morphemes between the study by Dulay and Burt and mine is not fully supported by this study. In a large collection of written essay data in L2

English, some grammatical morphemes such as possessive *-s* are more salient for learners. In spoken data, the primary source for Dulay and Burt, however, possessive *-s* might have been omitted more often because of difficulty in pronunciation. There are some L1-related difficult items, too. For instance, plural *-s* is supposed to be acquired at a quite early stage, but Japanese learners do not have morphological markers for plurals and thus produce more errors with this morpheme in English. The same thing can be said about the article system. Japanese learners always find it difficult to use determiners properly for reasons outlined.

6. Conclusion

This study shows the possibility of verifying available SLA findings with computer learner corpora. The advantage of a corpus-based approach to SLA is that by sharing the data with other researchers, we increase the chance of replication/falsification of hypotheses. While I look only at Japanese learners' written composition data in this paper, comparative research on learner corpus data for other L1s may shed more light on the nature of morpheme acquisition order in the future.

The use of learner corpora opens up the possibility of filling the gap between small-scale, tightly controlled experimental research and large-scale, but impressionistic, survey-questionnaire type research. The description of learner language with a large collection of learner corpus data has great potential for making a great contribution to re-examining previous research findings in SLA from a more comprehensive empirical perspective. Learner corpus research is still in its infancy and the refinement of sampling frames, elicitation tasks, and tagging schemes is still necessary. We hope the development of learner corpora will truly be a 'revolution in applied linguistics.' (Granger 1994)