A Computer Learner Corpus-based Analysis of the Acquisition Order of English Grammatical Morphemes

Yukio Tono, Ph.D. student, Lancaster University, UK, Formerly Tokyo Gakugei University, JAPAN

Introduction

The description of learner language has always been of primary concern to second language acquisition (SLA) researchers. Learner language provides the researcher with insights into the process of acquisition. If we have a better understanding of second language (L2) acquisition process, then we can apply the findings to a variety of practical aspects of language teaching: 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:

1. the study of learners' errors
2. the study of developmental patterns
3. the study of variability
4. the study of pragmatic features

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

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

Dulay and Burt were among the first to conduct empirical research on the acquisition order of certain grammatical features of English. They investigated the 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 proceed quite systematically and that the acquisition order is not rigidly invariant but is remarkably similar irrespective of the learners' L1 backgrounds, of their age, and of whether the medium of data-collection is speech or writing (Dulay and Burt 1975).

Although there were some stringent criticisms about morpheme studies (for example, Hatch 1978; Long and Sato 1984), Larsen-Freeman and Long (1991) conclude that they provide strong evidence of a developmental order (Larsen-Freeman and Long 1991: 92).

Computer learner corpora: new horizons

In 1990s, there has been remarkable progress in computer technology, especially PC architecture and throughput capacity, which lead to a growing interest in processing natural language on computer. Computational linguistics is a buzz word for 90s and the word "corpus" or "corpora" has become increasingly familiar to language teachers and researchers as well as linguists.

The benefit of large corpora was, for the first time, fully appreciated among the EFL sector, by the publication of "corpus-based" pedagogical dictionaries such as Collins COBUILD English Dictionary (1987, 1995) or Longman Dictionary of Contemporary English (1978, 1995). Native speakers' 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." As Granger (1994) pointed out, however, we should not exaggerate the impact of native corpora on foreign language teaching. "Having access to comprehensive frequency lists may well help course designers compile better lexical syllabuses, but it will not give them access to learners' actual lexical problems." (ibid: 25)

Recently there has been a growing awareness that it is necessary to investigate learner language by collecting a large amount of learner performance data on computer. The term "learner's corpus" was first used for Longman's learners' dictionaries, in which the information on EFL learners' common mistakes was provided based upon Longman's learners' corpus. The project called International Corpus of Learner English (ICLE) was launched as a part of ICLE project (Granger 1993). Researchers around the world started to collect learner data to compile computerized learner corpora (e.g. Milton, Asao et al., Tono 1996).

In this paper we will revisit the once popular topic of SLA research, English morpheme acquisition studies, and try to see how computerized learner corpora could possibly shed some new light on this old problem. There are a couple of reasons why we chose morpheme studies as our 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:

1. Is there any difference between the order of instruction and the order of acquisition?

2. Is it possible to alter the 'natural' order of acquisition by means of instruction?

3. Do instructed learners follow the same order of acquisition as untutored learners or a different order? (Ellis 1990: 139)

Computerized learner corpora, if used properly with a suitable research design, will prove to be an effective tool for us to answer these interrelated questions by providing the evidence of learner language in a more systematic way.

Secondly, although there is a criticism against 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.

Morpheme Studies: Short Review

In the early 1970s it was discovered that English children learn grammatical morphemes (morphemes such as "-ing" and "the" that play a greater part in structure than content words such as "dog") in a definite sequence (Brown 1973). Dulay and Burt (1973) 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 such a common order of difficulty for grammatical morphemes as follows:

1	plural "-s"	"books"
2	progressive "-ing"	"John going"
3	copula "be"	"John is here"
4	auxiliary "be"	"John is going"
5	articles	"The books"

6	irregular past tense	"John went"
7	third person "-s"	"John likes books"
8	possessive "-s"	"John's book"

Later researchers such as Dulay, Burt, and Krashen (1982) put the morphemes at the same accuracy order level together and see them as groups rather than individual items.

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).

Research Method

Purpose

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

Hypotheses

We will test the following hypotheses:

(1) An overall picture of accuracy order will coincide with the one obtained by Dulay and Burt.

(2) There will be a difference in accuracy order, due to either L1 background or elicitation tasks.

We hope to show that a large collection of learner language will be an effective tool to verify the results of acquisition order studies, the data of which was collected rather fragmentally.

Learner Corpus Data

The learner corpus data used for this study was based on TGU Learner Corpus (600,000 tokens in size), which was compiled from Japanese EFL learners' written essays. The data is developmental in nature in the sense that we obtained the written essays of the same topics from learners in different school years (Junior High 1st (age 13) to Senior High 3rd (age 18)). 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 big earthquake?" In this paper, we will report the results of our pilot research, which should be supplemented by follow-up research in my poster session.

Tagging Schemes

In the pilot study, we prepared the following tagset for grammatical morphemes:

Correct form: <ART>, <POS>, <3PS>, <IRPST>, <AUXBE>, <PL>, <COP>

Incorrect form: <ER_ART>, <ER_POS>, <ER_3PS>, <ER_IRPST>, <ER_AUXBE>, <ER_PL>

Each text was tagged according to the criterion set for the Bilingual Syntax Measure by Dulay and Burt. In other words, we only looked at the "obligatory context," i.e. contexts that require the obligatory use of grammatical morphemes in samples of learner language. And then we calculated the accuracy with which the morphemes were actually supplied in these contexts. The following is a sample tagged text:

> I love the money. So I saved it. I don't release it. Time <COP> is</COP> the money, but the money <COP> is </COP> not time. It <COP> is</COP> real? I don't think so. Money <COP>is </COP> <ART>the </ART> best thing in <ART>the </ART> world. <ART>A </ART> rich man doesn't have to work, but <ART>a </ART> poor man has to work. So, <ART>a </ART> rich man who has <ART>a </ART> lot of money has <ART>a </ART> lot of time. I love the money. I' <AUXBE>m </AUXBE> going to <ER_COP> <! = be> </ER_COP> <ER_ART> the <! = a> </ER_ART> rich man. And I play all of <ART> the </ART> life.<p>

Data Analysis

We processed the tagged corpus data and obtained the frequency and concordance lines for each morpheme occurrence by WordSmith and TXTANA. We obtained the accuracy rate by dividing the frequencies of correct forms by the sum of frequencies of correct and incorrect forms. We also defined the state of acquisition as "90% correct" as defined in the Bilingual Syntax Measure.

Results and Discussion

Table 1 shows the results of the pilot study (the results of the actual study will be presented at posters). They have least difficulty with copula "be," most difficulty with definite and indefinite articles "the" and "a." Among the seven morphemes, copula "be," auxiliary "be," and possessive "-s" reached 90% accuracy rate and were regarded as "acquired" items, but the other four morphemes could not reach sufficient accuracy rate even at the second year of senior high school level.

	JH-2	JH-3	SH-2
Copula be	94.17%	96.26%	94.74%
Aux be	89%	95.10%	92.45%
Possessive -s	76.67%	76.19%	95.24%
Plural -s	80%	81.04%	88.51%
3rd person -s	70.83%	69.57%	89.36%
Irregular past	82.28%	79.62%	83.69%
Article	63.02%	70.24%	` 79.62%

Table 1 The results of morpheme accuracy order in the pilot study

Table 2 shows the comparison of our results with the order obtained by Dulay and Burt (1975). The noteworthy difference is that articles "the/a" are the most difficult items for Japanese learners and showed the lowest accuracy rate in all the morphemes. Since Japanese language does not have the notion of articles attached to nouns, the proper use of articles should be very difficult for us to acquire. Possessive "-s", in contrast, was the item which was relatively easier for Japanese learners and ranked higher as compared with the order by Dulay and Burt. Therefore, hypothesis 1 is not fully supported.

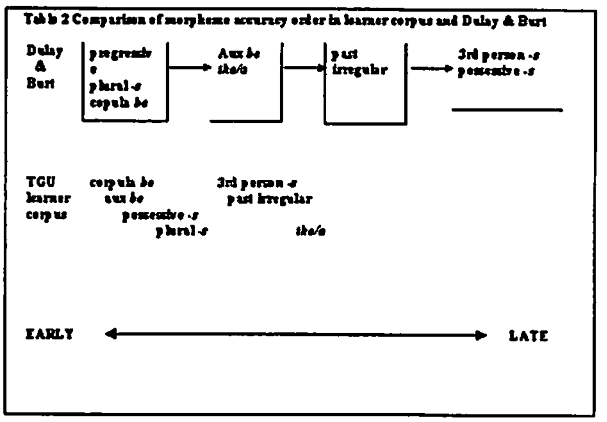

Table 2 Comparison of morpheme accuracy order in learner corpus and Dulay & Burt

Conclusion

We will provide more detailed statistical analysis of the results at poster sessions. In the meantime, it would suffice to say that the use of learner corpora opened up a new possibility of filling the gap between small-scale, tightly controlled experimental researches and large-scale impressionistic questionnaire type researches. The description of learner language with a large collection of learner corpus data has great potential for making a great contribution to verifying the previous results in SLA research from a more large-scale, data-driven perspective. Learner corpus research is still in its infancy and the refinement of sampling frames, elicitation tasks, and tagging schemes would be indispensable. We hope the development of learner corpora will be truly a "revolution in applied linguistics" (Granger 1994).

References

Brown, R. (1973) *A First Language: the Early Stages.* Cambridge, Mass: Harvard University Press.

Corder, S.P. (1967) "The significance of learners' errors." *International Review of Applied Linguistics 5:* 161-9.

Dulay, H. and M. Burt (1973) "Should we teach children syntax?" *Language Learning* 23: 37-53.

Dulay, H. and M. Burt (1975) "Creative construction in second language learning and teaching." In M. Burt and H. Dulay (eds.) *On TESOL 75: New Directions in Second Language Learning, Teaching and Bilingual Education.* Washington, D.C.: TESOL.

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

Ellis, R. (1990) *Instructed Second Language Acquisition.* Oxford: Basil Blackwell.

Ellis, R. (1994) *The Study of Second Language Acquisition.* Oxford University Press.

Granger, S. (1993) "The international corpus of learner English" In J. Aarts, P. De Haan & N. Oostdijk (eds.) *English Language Corpora - Design, Analysis and Exploitation.* Amsterdam & Atlanta: Rodopi.

Granger, S. (1994) "The learner corpus: a revolution in applied linguistics" *English Today* 39 (3): 25-29.

Hatch, E. (1978) "Acquisition of syntax in a second language." In J. Richards (ed.) *Understanding Second and Foreign Language Learning: Issues and Approaches.* Rowley, Mass.: Newbury House.

Larsen-Freeman, D. and M. Long (1991) *An Introduction to Second Language Acquisition Research.* London: Longman.

Long, M. and C. Sato (1984) "Methodological issues in interlanguage studies: An interactionist perspective." In A. Davies, C. Criper and A. Howatt (eds.) *Interlanguage.* Edinburgh: Edinburgh University Press.

Rutherford, W. (1987) *Second Language Grammar: Learning and Teaching.* London: Longman.

Selinker, L. (1969) "Language transfer." *General Linguistics* 9: 67-92.

Tono, Y. (1996), *Using Learner Corpora for L2 Lexicography.* LEXICOS 6 (AFRILEX SERIES 6) Stellenbosch: Universiteit van Stellenbosch.