Compiling Learner Corpus Data of Linguistic Output and Language Processing in Speaking, Listening, Writing, and Reading

Katsunori Kotani
Kansai Gaidai University / Osaka, Japan
kkotani@kansaigaidai.ac.jp

Takehiko Yoshimi
Ryukoku University / Shiga, Japan
yoshimi@rins.ryukoku.ac.jp

Hiroaki Nanjo
Ryukoku University / Shiga, Japan
nanjo@rins.ryukoku.ac.jp

Hitoshi Isahara
Toyohashi University of Technology / Aichi, Japan
isahara@tut.jp

Abstract

A learner's language data of speaking, writing, listening, and reading have been compiled for a learner corpus in this study. The language data consist of linguistic output and language processing. Linguistic output refers to data of pronunciation, sentences, listening comprehension rate, and reading comprehension rate. Language processing refers to processing time and learners' self-judgment of their difficulty of processing in speaking, listening, and reading and the fluency of their writing. This learner corpus will contribute to making the language learning process more clearly visible.

1 Introduction

Learner corpora have contributed to second language acquisition (SLA) research. For instance, SLA research using learner corpora examines learners' proficiency on the basis of what vocabularies/grammars learners actually use (Tono 2009, among others). Thus, most learner corpora are compiled of linguistic outputs that learners produce in speaking and/or writing.

In order to enhance SLA research, a learner corpus should be compiled of a learner's language data of the four modalities (speaking; listening; writing; reading). Language data of each modality are further classified into two types: linguistic output and language processing data. Language output data consist of linguistic objects that learners produce. Language processing data indicate how they produce linguistic outputs. Thus, we have eight types of language data that are useful for the SLA research on the development of learners' proficiency (Hinkel 2010, Segalowitz 2003). Among these eight types of language data, the previous studies (Granger et al. 2009, Izumi et al. 2004, Meurers et al. 2010, Wen et al. 2008) compiled the language output data of speaking, writing, and reading for constructing a learner corpus. See Section 2 for further detail. On the other hand, the other previous studies (Zechner & Bejar 2006, Arthur 1979, Hirai 1999, Kotani et al. 2010, Chang 2010) compiled the language processing data not for constructing a learner corpus but for examining learners' performance. See Section 3 for further detail. Thus, there is a shortage of language output data of listening, and furthermore we have to construct a learner corpus that integrates all these eight types of data. Hereafter, we refer to this corpus as I(ntegrated)-Learner Corpus. In order to construct I-Learner Corpus, we have compiled data of linguistic output and language processing of the four modalities when learners actually use the target language, which in this study is English.

2 Background

Written data are compiled in the International Corpus of Learner English (ICLE) (Granger et al. 2009). The written data are taken from essays written by learners of English as a foreign language (EFL). ICLE consists of learners' essays (approximately 500 words long) and learner information, but has no error tags.

Spoken data are compiled in the National Institute of Information and Communication Technology Japanese Learner English (NICT JLE) Corpus (Izumi et al. 2004). The spoken data are obtained by transcribing one-to-one interviews of EFL learners whose native language is Japanese. The NICT JLE Corpus includes error tags and reference data spoken by native speakers, but has no sound data for phonetic/phonological analyses.

Both spoken and written data are compiled in the Spoken and Written Corpus of Chinese

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1418–1422,
Chiang Mai, Thailand, November 8 – 13, 2011. ©2011 AFNLP

Learners (SWECCL) (Wen et al. 2008). The spoken data of SWECCL consist of both sound data and transcription data of retelling, monologues, and role plays. This corpus also includes three years' worth of longitudinal data.

Read data are compiled in the Task-based Corpus (Meurers et al. 2010). This corpus consists of written answers for text comprehension questions. Since written answers for text comprehension questions can demonstrate both learners' reading and writing proficiency, the Task-based Corpus is taken as a learner corpus that integrates written and read data.

3 Corpus design

Table 1 summarizes the design criteria of the I-Learner Corpus on attributes of learners and those of language. The criteria follow the attributes of a learner corpus (Granger 2007).

		Attributes
Age		18 years old and older
Sex		male, female
Mother tongue		Japanese
Level		beginner, intermediate, advanced
Learning context		EFL
Experience		36 months or more
Modality		spoken (S), written (W), listened (L), read (R)
Medium	S, L, R	news broadcast
	W	question answering, description
Genre		expository language use
Topic	S, L, R	general news topic/topic related to university life
	W	learning profiles, daily events
Technicality		general
Task setting		paid task, no dictionary

Table 1. Design criteria.

The target learners are EFL learners of university students whose native language is Japanese. In Japan, students study English at least three years in junior high school.

The modality of language data covers spoken, written, listened, and read data. Spoken, listened, and read data are taken from news broadcast. Written data are taken from question answering and picture description. The genre is expository contexts in daily-life language use. Though there are other contexts such as academic/professional contexts, these contexts contain more non-linguistic aspects. Thus, we chose daily-life contexts in order to minimize non-linguistic aspects such as background knowledge. Hence, the topics in news broad cast are general news topic and the ones related to university life. The topic of writing covers learners' learning profiles and daily events.

We basically use the compiling procedure stated in the previous studies reviewed in Section 2. Following the Task-based Corpus (Meurers et al. 2010), we compile read and listened data from answers to comprehension questions.

Although the comprehension-question-based procedure is suitable for compiling comprehension rate of a whole text or that of some part(s) of a text, it unfits for compiling comprehension rate at a sentence level. Of course, it is possible to compile comprehension rate at the sentence level by preparing comprehension questions for each sentence, but this is just not realistic. However, we have to compile read and listened data at the sentence level just like for spoken and written data.

Our solution of this problem is to compile language processing. It is reported that language performance can be evaluated on the basis of language processing: speaking performance (Zechner & Bejar 2006), writing performance (Arthur 1979), listening performance (Hirai 1999), and reading performance (Kotani et al. 2010, Chang 2010). An advantage of language processing is the possibility to measure at the sentence level. In addition, language processing (speaking speed) is compiled in native speaker corpora (Braun 2006, Gut 2009). Hence, we compile data of both language processing and linguistic output across the four modalities.

Language processing has two parts. One is the processing time how long a learner takes for linguistic output. The other is the judgment how difficult a learner judges processing in speaking, listening, and reading to be and how fluent a learner judges his or her writing to be. Table 2 lists the data to be stored in the I-Learner Corpus.

	Linguistic Output	Language Processing
Speaking	sound output	time, difficulty
Writing	sentence output	time, fluency
Listening	comprehension rate	difficulty
Reading	comprehension rate	time, difficulty

Table 2. Data specification.

Spoken data of the I-Learner Corpus consist of recordings of oral reading (linguistic output), and oral reading time and a learner's judgment of pronunciation difficulty on a five-point scale (1: Very Easy, 2: Easy, 3: Moderate, 4: Difficult, 5:

Very Difficult) (language processing). In addition to learners' data, we prepare reference sound data read by native speakers.

Written data of the I-Learner Corpus consist of sentences of question answering, sentences of picture description (linguistic output), writing time, and a learner's judgment of his or her fluency on a five-point scale (language processing).

Listened data of the I-Learner Corpus consist of comprehension rate (linguistic output) and a learner's judgment of listening difficulty (language processing).

Read data of the I-Learner Corpus consist of the comprehension rate (linguistic output), reading time, and a learner's judgment of reading difficulty on a five-point scale (language processing).

4 Data compiling

4.1 Procedures

Data compiling proceeded in the following order: listening task, reading task, speaking task, and writing task.

In the listening task, learners listened to four news articles that were read by native speakers of English sentence-by-sentence using a data collecting tool described in Section 4.4. Learners judged listening difficulty of a sentence after listening to it. When learners finished listening to an article, they answered five comprehension questions on the data collecting tool. Learners could listen to a sentence only once.

Listened data are often gathered in a situation where learners listen to sentences in an article from start to finish without a stop. However, learners in this study listened to a news article sentence-by-sentence in order to report their judgments for listening difficulty.

In the reading task, learners silently read four news articles sentence-by-sentence using the data collecting tool. Learners judged reading difficulty of a sentence after reading it. When learners finished reading an article, they answered five comprehension questions. Learners could not read a sentence again nor use a dictionary.

Read data are often taken in a situation where learners see a news article as a whole. However, learners in this study read a news article sentence-by-sentence so that processing time and their judgments of reading difficulty could be kept track of.

In the speaking task, learners read aloud four news articles sentence-by-sentence using the data collecting tool. Learners judged pronunciation difficulty of a sentence after reading it aloud. Learners had no comprehension questions in the speaking task, because the speaking task and the reading task used the same articles in order for learners to grasp the contents of the articles before reading aloud.

Spoken data are often taken from utterances in actual discourse (Izumi et al. 2004). However, we chose an oral reading task in which learners read the same sentences, because we can directly compare phonetic/phonological properties between learners.

The writing task had two sub-tasks. In the first, learners wrote answers for twenty questions on their profiles. In the second, learners wrote sentences describing four pictures of a series of events on the data collecting tool. Learners were instructed to write at least five sentences for a picture. In both tasks, learners judged the fluency of a sentence after writing it. Learners could not rewrite a sentence after moving on to another sentence nor use a dictionary.

Although written data are often taken from essays (Granger et al. 2009), we chose question answering and picture description in order to minimize the non-linguistic aspects. While essay writing depends on logical, analytical, and critical thinking, learners can answer profile questions and describe pictures without depending too much on non-linguistic skills as long as questions and pictures are simple enough.

4.2 Participants

Ninety EFL learners took part in the data compiling (48 Male, 42 Female: mean age 21.6 years old, ranging from 19–40 years old). They were university students in Tokyo, Japan. Their practical experience ranged 53 months to 216 months. The learners were paid for their participation.

The learners submitted scores of the Test of English for International Communication (TOEIC) taken within a year before the data started to be compiled. On the basis of the TOEIC scores, they were classified into three proficiency levels: beginner (N=30, TOEIC score range, 280-495), intermediate (N=30, TOEIC score range, 500-725), and advanced (N=30, TOEIC score range, 730-985) levels.

4.3 Materials

The following are questions for learner profiles in the writing task: "Which languages do you speak and read, and how well?" "What language did you learn?" (Ehrman 1996, Eignor et al.

1998). Pictures (Figure 3) described in the writing task were cited from Hughes (2003).

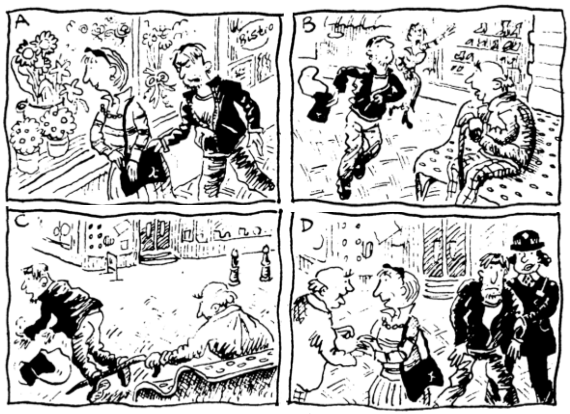

Figure 3. Pictures for description.

News articles used in the speaking, listening, and reading tasks were taken from Voice of America (VOA) (http://www.voanews.com). The length of articles was approximately 350 words (within plus/minus 5%). Each article had five comprehension questions that were made by the authors following question formats (Nation & Malarcher 2007).

These articles had two difficulty levels. Low-difficulty articles were taken from Special English program developed for learners of English, e.g., "Grading Grades." These articles are written in short, simple sentences that contain only one idea, and the sentences consist of a core vocabulary of 1500 words without idiomatic expressions. Low-difficulty articles were limited to the topic "studying in the U.S." High-difficulty articles were taken from VOA editorials that present differing points of view on a wide variety of issues, e.g., "Educating Marginalized Children."

4.4 Data compiling devices

The data collecting tool that learners used presents a sentence on a computer screen in the speaking and reading tasks. This tool keeps track of processing time for a sentence in the speaking, writing, and reading tasks. This tool provides comprehension questions and saves answers in the listening and reading tasks. In the writing task, this tool presents a question/a picture, and provides a blank space in which to write a sentence. The tool further keeps scores (1-5) of sentence difficulty/fluency judged by a learner in all the tasks.

In addition to this data collecting tool, the following devices were used. In the listening task, learners listened to audio files of news articles with headphones. In the speaking task, each learner reads aloud news articles in a recording booth. The recording booth is a sound-attenuated chamber (1700mm, 1900mm, 2100mm (approximately WDH)). A learner sat on a chair at a desk. The oral reading was recorded using a unidirectional electric-condenser microphone on a solid-state stereo. The sampling rate used was 44.1KHz, and quantization was set to 16 bits.

5 Application of the corpus

One application of the I-Learner Corpus is to use the corpus data as a language resource for examining learners' performances across multiple modalities, because the I-Learner Corpus includes linguistic output and language processing of the four modalities. This examination will reveal whether a learner's proficiencies in these modalities have developed equally. It will also enable us to examine how learners' proficiency develops from beginner to advanced levels, because the I-Learner corpus includes data of learners at these levels. These linguistic analyses constitute an important part of the I-Learner Corpus.

Another application is to use the corpus data as training data for a machine learning algorithm to construct an automatic evaluation method for learners' performances of the four modalities. When this automatic evaluation method and the data compiling devices are implemented in a computer-assisted language learning (CALL) system, the CALL system becomes able to compile learners' data. The CALL system can add new corpus, especially, longitudinal data if learners use the system for a certain period.

6 Conclusion

This paper described data compiling for constructing an I-Learner Corpus, a learner corpus that is compiled of data of linguistic output and of language processing of the four modalities. The I-Learner Corpus enables us to examine learners' performances in more detail and serves as a language resource for a learner model that predicts learners' performance.

A future work is to provide annotation data such as error information of pronunciation and written sentences. Another work is to enlarge corpus data by adding data of learners with different native languages.

Acknowledgments

This work was supported in part by Grant-in-Aid for Scientific Research (B) (22300299).

References

Arthur, Bradford R. 1979 Short-term changes in EFL composition skills. In Carlos A. Yorio, Kyle Perkins, and Jacquelyn Schachter, editors, *On TESOL '79: The Learner in Focus*. TESOL, Washington D.C., pages 330-342.

Braun, Sabine. 2006. ELISA: A pedagogically enriched corpus for language learning purposes. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Peter Lang, Frankfurt, pages 25-47.

Chang, Anna C.-S. 2010. The effect of a timed reading activity on EFL learners: Speed, comprehension, and perceptions. *Reading in a Foreign Language*, 22(2): 284-303.

Ehrman, Madeline E. 1996. *Understanding Second Language Learning Difficulties*. SAGE Publications, London.

Eignor, Daniel, Carol Taylor, Irwin Kirsch, and Joan Jamieson. 1998. Development of a scale for assessing the level of computer familiarity of TOEFL examinees. Research Reports RR98-7, Educational Testing Service, Princeton, New Jersey.

Granger, Sylviane. 2007. The computer learner corpus: A versatile new source of data for SLA research. In Wolfgang Teubert and Ramesh Krishnamurthy, editors, *Corpus Linguistics: Critical Concepts in Linguistics*, volume 2. Routledge, London, pages 166-182.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot. 2009. *International Corpus of Learner English*, version 2. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Gut, Ulrike. 2009. *Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German*. Peter Lang, Frankfurt.

Hinkel, Eli. 2010. Integrating the four skills: Current and historical perspectives. In Robert B. Kaplan, editor, *Oxford Handbook in Applied Linguistics*, 2nd edition. Oxford University Press, New York, pages 110-126.

Hirai, Akiyo. 1999. The relationship between listening and reading rates of Japanese EFL learners. *The Modern Language Journal*, 83(3): 367-384.

Hughes, Arthur. 2003. *Testing for Language Teachers*, 2nd edition. Cambridge University Press, Cambridge, UK.

Izumi, Emi, Kiyotaka Uchimoto, and Hitoshi Isahara, editors. 2004. *Nihonjin 1200 nin no Eigo Spiking Koopasu* [A Speaking Corpus of 1200 Japanese Learners of English]. ALC Press, Tokyo, Japan.

Kotani, Katsunori, Takehiko Yoshimi, and Hitoshi Isahara. 2010. A prediction model of foreign language reading proficiency based on reading time and text complexity. *US-China Education Review*, 7(10): 1-9.

Meurers, Detmar, Niels Ott, and Ramon Ziai. 2010. Compiling a task-based corpus for the analysis of learner language in context. In Sam Featherston and Britta Stolterfoht, editors, *Proceedings of Linguistic Evidence 2010*, pages 214—217.

Nation, Paul and Casey Malarcher. 2007. *Reading for Speed and Fluency*. Compass Publishing, Seoul, Korea.

Segalowitz, Norman. 2003. Automaticity and second languages. In Catherine J. Doughty and Michael H. Long, editors, *The Handbook of Second Language Acquisition*. Blackwell, Oxford, pages 382-408.

Tono, Yukio. 2009. Integrating learner corpus analysis into a probabilistic model of second language acquisition. In Paul Baker, editor, *Contemporary Corpus Linguistics*. Continuum International Publishing Group, London, pages 184-203.

Wen, Qiufang, Maocheng Liang, and Xiaoqin Yan. 2008. *Spoken and Written Corpus of Chinese Learners (SWECCL)* 2.0. Foreign Language Teaching and Research Press, Beijing, China.

Zechner, Klaus and Isaac I. Bejar. 2006. Towards automatic scoring of non-native spontaneous speech. In Robert C. Moore, Jeff A. Bilmes, Jennifer Chu-Carroll, Mark Sanderson, editors, *Proceedings of the 2006 Human Language Technology Conference and the North Linguistics Annual Meeting (HLT/NAACL 2006)*, pages 216-223.