Application of a Corpus to Identify Gaps between English Learners and Native Speakers

Katsunori Kotani
16-1 Nakamiyahigashino-cho, Hirakata,
Osaka, Japan 573-1001
kkotani@kansaigaidai.ac.jp

Takehiko Yoshimi
1-5 Yokotani, Seta, Otsu,
Shiga, Japan 520-7729
yoshimi@rins.ryukoku.ac.jp

Abstract

In order to develop effective computer-assisted language teaching systems for learners of English as a foreign language, it is first necessary to identify gaps between learners and native speakers in the four basic linguistic skills (reading, writing, pronunciation, and listening). To identify these gaps, the accuracy and fluency in language use between learners and native speakers should be compared using a learner corpus. However, previous corpora have not included all necessary types of linguistic data. Therefore, in this study, we aimed to design and build a new corpus comprising all types of linguistic data necessary for comparing accuracy and fluency in basic linguistic skills between learners and native speakers.

1 Introduction

Learners of English as a foreign language (EFL) frequently demonstrate a level of language ability that differs from that of native speakers (gap between learner and native speaker). Compared with native speakers, learners more frequently generate grammatically incorrect sentences and speak at a slower rate (Brand and Götz 2011, Chang 2012, Thewissen 2013). To develop tools and methods for effective learning of EFL, gaps in the four basic linguistic skills (reading, writing, pronunciation, and listening) need to be clearly identified and bridged.

A comparative learner corpus is a promising linguistic resource for identifying the gaps. To identify the gaps, a learner corpus should cover the basic linguistic skills (Treiman et al. 2003), because these skills are prerequisite for developing a level of ability adequate for effective communication with English speakers in a global society (Ono 2005).

A learner corpus should address gaps in both accuracy and fluency. Gaps in accuracy result from a lack of linguistic knowledge and manifest as misunderstandings when reading, grammatically incorrect usage when writing, mispronunciations when speaking, and misunderstandings when listening. Gaps in fluency result from a limited ability to perform cognitive-linguistic operations (Juffs and Rodríguez 2015), and manifest as slower rates of reading, writing, and pronunciation. In addition, fluency gaps also tend to result in a lack of confidence among learners.

Some learner corpora have been developed for the purpose of comparative analysis with native speakers (Sugiura 2007, Friginal et al. 2013, Barron and Black 2014); however, these corpora have only focused on writing and speaking, not reading or listening. A learner corpus compiled by Kotani et al. (2011) was composed of reading, writing, pronunciation, and listening data, but did not include data from native speakers.

In this study, we aimed to construct a comparative learner corpus to analyze gaps in the accuracy and fluency of the four basic linguistic skills. Specifically, this study collected corpus data from native speakers and merged these data with those from learners in the corpus compiled by Kotani et al. (2011). For this study, English speakers were categorized into four proficiency levels as follows: Learners in Kotani et al. (2011) were classified into three groups based on their level of proficiency, and native speakers were designated as the fourth and most advanced-level.

With the goal of supporting EFL teachers who use "authentic" materials such as web-pages that are used in English speakers' daily life, we also constructed a statistical model for calculating the

Proceedings of the Eighth Workshop on Building and Using Comparable Corpora, pages 38–42,
Beijing, China, July 30, 2015.

difficulty of each sentence in authentic materials, which demonstrates the effectiveness of our corpus. Because the difficulty level of authentic materials is not always clear, teachers must personally inspect all such materials in order to verify that they are appropriate for the proficiency level of learners. Therefore, we developed a statistical model to automatically measure sentence difficulty and thereby reduce the effort required by teachers for this preparatory task.

2 Corpus between learners and native speakers

2.1 Corpus data

This study collected corpus data from native speakers following the method of Kotani et al. (2011). The corpus data of Kotani et al. (2011) consisted of data collected from learners for analyzing the accuracy and fluency of reading, writing, pronunciation, and listening, and the data are summarized in Table 1.

Language use	Perspective	
	Accuracy	Fluency
Reading	Comprehension rate	Silent-reading rate
		Difficulty judgment score
Writing	Written sentence (Correct rate)	Writing rate
		Difficulty judgment score
Pronunciation	Speech sound (Correct rate)	Reading-aloud rate
		Difficulty judgment score
Listening	Comprehension rate	Difficulty judgment score

Table 1: Summary of corpus data

The accuracy of reading (comprehension rate) was assessed by calculating the percentage of correct answers to comprehension questions based on written text. The accuracy of writing and pronunciation (the correct rate) was assessed by calculating the percentage of correctly written words or pronounced speech sounds from the total number of words in written sentences or spoken words, respectively. The accuracy of listening was assessed in terms of comprehension rate for spoken text, similarly to that of reading.

Fluency in terms of reading, writing, and pronunciation was assessed based on silent-reading, writing, and reading-aloud rates, respectively.

Fluency was also assessed based on a difficulty judgment score. Difficulty judgment scores for reading were assessed in terms of learners' judgment on the difficulty of reading comprehension, which they indicated on a five-point Likert scale (1: easy; 2: somewhat easy; 3: average; 4: somewhat difficult; or 5: difficult). Scores for writing were assessed in terms of learners' confidence in accuracy on a five-point Likert scale (1: confident; 2: somewhat confident; 3: average; 4: not very confident; or 5: not confident). Those for pronunciation and listening were assessed on the five-point Likert scale in terms of learners' judgment on the difficulty of pronunciation and listening comprehension, respectively.

2.2 Data collection method

Corpus data were collected through a series of reading, writing, pronunciation, and listening tasks. In the reading task, learners silently read 80 sentences in four news articles sentence-by-sentence, selected a difficulty score for each sentence, and answered five multiple-choice comprehension questions for each article. In the writing task, learners wrote sentences to describe four pictures and answered 20 questions about their background and computer skills, and then selected a difficulty score for each sentence. The pronunciation task proceeded similarly to the reading task: learners read aloud 80 sentences in four news articles, and selected a difficulty score for each sentence. Their voices were recorded in a sound-attenuated recording booth. In the listening task, similar to the reading task, learners listened to 80 sentences from four audio news clips sentence-by-sentence, and then selected a difficulty score for each sentence. After a clip was finished, the learner answered five multiple-choice comprehension questions for each clip.

The learner corpus of Kotani et al. (2011) compiled corpus data from three different proficiency groups of learners (beginner-level, intermediate-level, and advanced-level) based on TOEIC (Test of English for International Communication) scores; each group comprised 30 learners. Hence, for this study, we chose to collect corpus data from 30 native speakers (16 male, 14 female; mean age ± standard deviation [SD], 22.5 ± 2.0 years; age range, 20–27 years) to represent a level higher than that of advanced-level learners. The native speakers were recruited from among university students living in areas in and around Tokyo. All native speakers were compensated for their participation.

2.3 Descriptive statistics

All distributions shown in Tables 2, 3, and 4 followed our expectation that the difficulty of a task would decrease from the beginner to the native speaker level. This outcome suggests the validity of our corpus data.

Mean comprehension rates (± SD) of 120 instances collected from each group (*n* = 30) of learners and native speakers in four articles and clips involving the reading and listening tasks, are summarized in Table 2.

Group	Task	
	Reading	Listening
Beginner	47.3(23.6)	43.3(22.2)
Intermediate	52.0(22.7)	49.8(20.9)
Advanced	65.8(24.0)	67.0(20.0)
Native-speaker	76.5(21.8)	75.7(17.4)

Table 2: Comprehension rates of the four groups (%); mean (SD)

Group	Task			
	Reading	Writing	Pronunciation	Listening
Beginner	3.26 (0.84)	3.07 (0.94)	3.61 (0.89)	3.63 (0.76)
Intermediate	2.72 (0.81)	3.02 (0.60)	3.29 (0.80)	3.18 (0.72)
Advanced	2.18 (0.92)	2.36 (1.00)	2.73 (1.05)	2.28 (0.96)
Native-speaker	1.92 (0.84)	1.56 (0.73)	2.15 (0.88)	1.87 (0.82)

Table 3: Difficulty judgment scores of the four groups; mean (SD)

Mean difficulty judgment scores (± SD) of 2400 instances collected from each group (*n* = 30) in 80 sentences involving reading task, are summarized in Table 3. Mean difficulty judgment scores (± SD) of 30**m* instances collected from each group (*n* = 30) in *m* sentences involving the writing task, in which the number of written sentences (*m*) differed for each individual, are also summarized in Table 3. Mean difficulty judgment scores (± SD) of 2400 instances collected from each group (*n* = 30) in 80 sentences involving pronunciation and listening tasks, are also shown.

Mean processing rates (± SD) of 2400 instances collected from each group in 80 sentences involving reading task, are summarized in Table 4. Mean writing rates (± SD) of 30**l* instances collected from each group (*n* = 30) in *l* sentences involving the writing task, in which the number of written sentences (*l*) differed for each individual, are also summarized in Table 4. Mean processing rates (± SD) of 2400 instances collected from each group in 80 sentences involving pronunciation task, are also shown. Processing rates were calculated as the number of words read/written/pronounced in one minute (WPM: words per minute).

Group	Task		
	Reading	Writing	Pronunciation
Beginner	86.91 (42.19)	9.21 (3.50)	66.28 (13.10)
Intermediate	97.17 (32.11)	10.21 (3.96)	76.97 (15.27)
Advanced	128.32 (44.99)	13.35 (5.42)	91.68 (12.87)
Native-speaker	206.21 (61.15)	17.34 (5.78)	119.91 (14.73)

Table 4: Processing rates of the four groups (WPM); mean (SD)

3 Measurement of sentence difficulty

3.1 Goal

In order to select online materials that are appropriate for the proficiency level of learners, a teacher must personally assess the difficulty of the materials, which is often unclear. A method that would enable the automatic measuring of sentence difficulty of online materials would thereby be expected to reduce the burden of this preparatory task.

To achieve this, we constructed a statistical model based on our corpus data. Our statistical model calculates sentence difficulty in terms of gaps in language use between learners and native speakers on the basis of linguistic features of sentences.

3.2 Methods

We carried out a multiple regression analysis of our corpus data using sentence length (number of words), and mean length of words in a sentence (mean number of syllables), as independent variables.

For the dependent variable, we used the gaps in the silent-reading rate, which were derived for each sentence (*n* = 80) by subtracting the mean silent-reading rate of advanced-level learners (*n* = 30) from that of native speakers (*n* = 30). The distribution of these gaps is summarized in Figure 1. The gaps ranged from < 25 to > 125 WPM, and the distribution of silent-reading rates followed a normal distribution according to the Kolmogorov-Smirnov test (K = 0.49, p < 0.01).

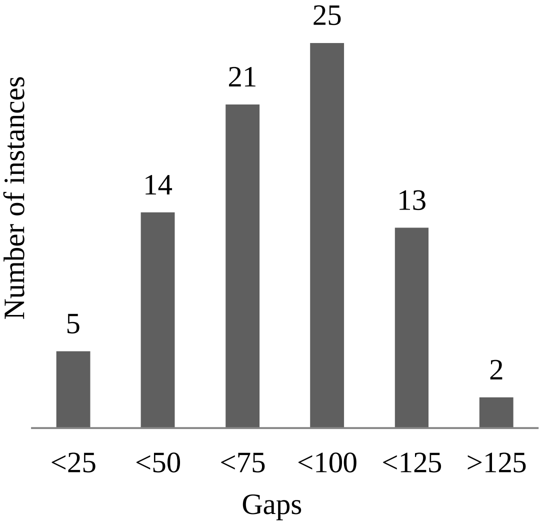

Figure 1: Gaps in the silent-reading rate between learners and native speakers (WPM)

3.3 Results

A significant relationship was observed between the linear combination of linguistic features and gaps in the silent-reading rate (F (2, 77) = 17.42, p < 0.01). The sample multiple correlation coefficient adjusted for degrees of freedom was 0.54, indicating that approximately 31.1% of the variance in the gaps in the sample could be accounted for by the linear combination of linguistic features.

We then assessed our method using a leave-one-out cross-validation test. In this test, our method was examined *n* times (*n* = 80) by using one instance as test data and *n*−1 instances as training data. Spearman’s correlation coefficient was used to compare the gaps predicted using our method with those that were actually measured. The correlation coefficient (r = 0.48) was statistically significantly different from zero (p < 0.01).

Errors in the cross-validation test results are summarized in Figure 2. Errors were calculated as absolute values of the differences between gaps predicted using our method and the actual gaps. Our method was associated with a lower error rate (0 to 25 WPM).

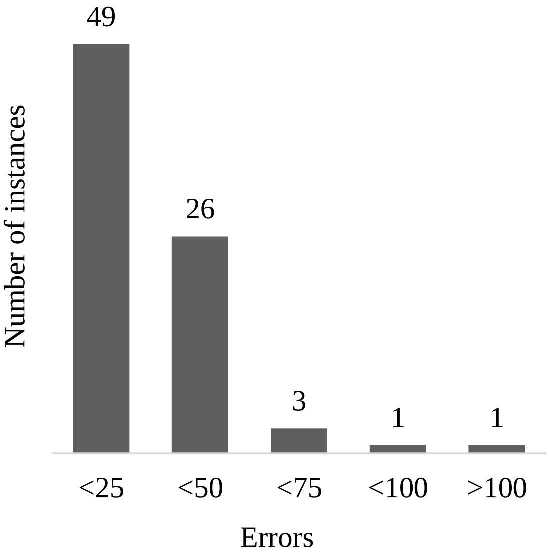

Figure 2: Errors in the cross-validation test results (WPM)

4 Conclusion

This paper described the construction of a corpus comprising all types of linguistic data necessary for comparing accuracy and fluency in basic linguistic skills between learners and native speakers. We expect that this corpus will enable teachers to more accurately assess the performance of learners in greater detail through a comparison with native speakers. We also expect our statistical model to serve as an effective method for measuring the difficulty of online materials, thereby reducing the burden of this preparatory task and allowing teachers to more easily select online materials that are appropriate for the proficiency level of learners.

Acknowledgments

This work was supported in part by Grant-in-Aid for Scientific Research (B) (22300299) and (15H02940).

Reference

Christiane Brand and Sandra Götz. 2011. Fluency versus accuracy in advanced spoken learner language. *Errors and Disfluencies in Spoken Corpora. Special Issue of International Journal of Corpus Linguistics*, 16(2): 255–275.

Anne Barron and Emily Black. 2014. Constructing small talk in learner-native speaker voice-based telecollaboration: A focus on topic management and backchanneling. *System*, 48: 112-128.

Anna C.-S. Chang. 2012. Improving reading rate activities for EFL students: Timed reading and repeated oral reading. *Reading in a Foreign Language*, 24(1): 56-83.

Eric Friginal, Man Li, and Sara C. Weigle. 2013. Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing*, .23: 1-16.

Alan Juffs and Guillermo A. Rodríguez. 2015. *Second Language Sentence Processing*. New York: Routledge.

Katsunori Kotani, Takehiko Yoshimi, Hiroaki Nanjo, and Hitoshi Isahara. 2011. Compiling learner corpus data of linguistic output and language processing in speaking, listening, writing, and reading. *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011):* 1418-1422.

Hiroshi Ono. 2005. A development of placement test and e-learning system for Japanese university students: Research on support improving academic ability based on IT. *Research Report, National Institute of Multimedia Education 2005-6.*

Jennifer Thewissen. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal*, 97(1): 77–101.

Rebecca Treiman, Charles Clifton Jr., Antje S. Meyer, and Lee H. Wurm. 2003. Language comprehension and production. *Comprehensive Handbook of Psychology 4: Experimental Psychology*. New York: John Wiley & Sons, Inc.: 527-548.

Masatoshi Sugiura, Masumi Narita, Tomomi Ishida, Tatsuya Sakaue, Remi Murao, and Kyoko Muraki. 2007. A discriminant analysis of non-native speakers and native speakers of English. *Proceedings of the 2007 Corpus Linguistics Conference*.