Lexical analyses of native and non-native English language instructor speech based on a six-month co-taught classroom video corpus

Noriaki Katagiri¹, Goh Kawai²

¹Hokkaido Sapporo Kaisei Senior High School, Hokkaido University, Graduate School of Media and Communication
²kotonomichy@nifty.com, goh@kawai.com

Abstract

We developed and analyzed a videotaped corpus of Japanese and English instructors who taught a high school English language class together over a 6-month period. Distributions of lexical items spoken by native and non-native speakers show they (1) both concentrated within the 2000 most common words found in written language, (2) both skewed towards classroom context, but (3) unequal with regards to interrogative forms (such as what and why) and verbs (such as imperatives used in directing student behavior). These findings respectively suggest that (1) non-native instructors desiring to teach wholly in the target language should focus on common words, (2) students expect classroom context, and (3) native instructors favor content-oriented instruction that demands higher proficiency of students, while non-native instructors prefer step-wise instruction suited for learners across various levels. The last finding in particular may benefit the training of both kinds of instructors.

Index terms: classroom video corpus, oral tokens, native vs non-native instructor comparison

1. Introduction

English as a foreign language is an intensely-sought but elusive goal in Japan's secondary education. Although learners understand oral English commands that are used continuously and repeatedly [1], a nationwide survey shows only approximately 30% of middle school instructors and 55% of high school instructors use some English in their oral communication classes [2]. To augment target language exposure, Japanese public schools employ native English speakers in approximately 15% of all English classes [2] and [3]. If immersion is the objective, the shortage of 85% must be taught by Japanese teachers trained to teach wholly in English. (For brevity, we will refer to Japanese teachers of English as “JTs,” and English teachers of English as “ETs.”)

Apart from the vaguely defined notion of native speaker, the differences in language usage between JTs and ETs are largely unknown. Explicit understanding of ET and JT language use can help ETs and JTs mimic each other’s behavior such that JTs become more proficient and ETs become more attuned to their students’ capabilities and needs. Training ETs is necessary because the qualifications of ETs are somewhat relaxed. For instance, a major recruiter of ETs requires "excellent pronunciation, rhythm, intonation and voice projection skills" in English, but no formal training in pedagogy [4]. One step towards such bilateral teacher training is analyzing the qualitative and quantitative lexical differences between ETs and JTs [5]. To this end, we analyzed ETs and a JT teaching a class together. This paper focuses on analyses of word-level transcriptions, which are hoped to yield scaffolding vocabulary suited for JTs to teach almost wholly in the target language. The remainder of this paper describes data collection (section 2), results and analyses (section 3), and implications for teacher training (section 4).

2. Materials and methods

2.1. The video corpus

We videotaped 3 ETs and 1 JT co-teaching at a Japanese high school English language class. The videotape captured instructor utterances, and student reactions (raising hands, nodding, taking notes, practicing conversations with classmates, etc) [1]. The ETs and JT almost exclusively spoke English during 49 classroom sessions over a 6-month period between 2007-04 and 2007-09. The JT was present throughout all 49 classroom sessions. ETs appeared one at a time (although there were 3 individuals) in a total of 9 class periods. When an ET was present, they were present throughout the classroom session.

The videotapes were digitized and transcribed at the word level. We used XML tags to distinguish speech between ET, JT, and CD (utterances prerecorded on CDs accompanying textbooks). Non-English words such as student names and other proper nouns were marked so they could be excluded from later processing. Transcription errors were minimized through proofreading. The total duration of the corpus is over 2000 minutes, and contains over 120,000 words.

2.2. Spoken lexical tokens

We decided to analyze both lexical tokens and types because lemmatizing (i.e., grouping different lexical tokens into lexical types) using computer software can be unreliable [5]. Due to the lack of ET data, ETs were grouped together for analysis.

In addition to comparing frequency distributions of ET and JT lexical tokens, we compared our data with the JACET 8000 most-common wordlist [6]. The JACET wordlist is based on the British National Corpus, English language textbooks used in Japan, newspaper articles, movie scripts, and TOEFL exams.
3. Results and analyses

3.1. Word frequencies

Excluding non-English words, there were 126,210 tokens total (JT 96,460, ET 18,572, CD 11,178). Figures 1 and 2 show lexical token frequencies for ET, JT and CD for each class period. Contrary to initial assumptions, ET and JT token distributions were similar at a gross level, except for rarely occurring words. The JT may have produced less rare words because the JT could switch to L1 (Japanese) when students did not understand. The ETs, having minimal command of L1, may have been forced to repeat difficult words in L2 (English).

Figure 1: Histogram of lexical tokens per classroom session.
The JT was present in all 49 classroom sessions. 3 ETs, one at a time, were present in 9 sessions. The number of word tokens dropped when the JT used L1 to explain abstract concepts such as syntax or a midterm review. Each time a new ET taught a class, the JT demonstrated teaching, resulting in roughly equal spoken word token output between the ET and JT. In the following class sessions, ETs dominated speech output. CDs were not used extensively.

Figure 2: Descriptive statistics of number of lexical tokens over all lexical types. The median, mean, and maximum numbers of tokens per type were roughly equal for ETs and JT (shown respectively as the Median, Mean, and Max columns above). This means ETs and JT produced roughly the same number of word tokens for most of their word types (but the word types may differ). The minimum number of tokens differs substantially because ETs (being monolingual) were forced to repeat words not understood by students, whereas the JT could switch to L1. This is reflected in JT’s slightly greater standard deviation.

There were 764 word types common between the ETs and JT. The relative frequencies of these common words were similar between ETs and JT, suggesting that these words are regarded as equally useful between ETs and JT. Table 1 shows correlation coefficients of frequencies of word tokens produced by both ET and JT. Figure 3 shows a scatter plot of the common words.

Table 1: Correlation coefficients of frequencies of word tokens produced by both ET and JT ranked by JACET wordlist

<table>
<thead>
<tr>
<th>Rank range</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>59</td>
<td>511</td>
<td>105</td>
<td>43</td>
<td>28</td>
<td>13</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.97</td>
<td>0.91</td>
<td>0.81</td>
<td>0.08</td>
<td>0.72</td>
<td>0.10</td>
<td>-0.72</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>Determination correlation coefficient</td>
<td>0.97</td>
<td>0.84</td>
<td>0.85</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Figure 3: Scatter plot of ET and JT relative frequencies of common word tokens. ETs and JT produced 764 common words, mostly roughly less than 20 times each. Given that only half of word types are common between ETs and JT, the relative low frequency of common word types suggests that ETs and JT differ in their choice of vocabulary, even though the words they have in common correlate highly in their likelihood of occurrence.

3.2. Comparison with JACET wordlist

To study how our data differs from language use at large, we compared ET and JT tokens with the JACET wordlist. Among all word tokens, a total of 103,218 words (82%) were found in the JACET wordlist (breakdown is JT 77,765 words (81%), ET 15,765 (85%), CD 9688 (87%)). Figure 4 shows the cumulative percentages of ET and JT tokens for each 1000-word band of the JACET wordlist. Approximately 85% of both ET and JT tokens are found among the 2000 most-common words in the JACET list. This is consistent with language use in general [5]. Furthermore, ETs and JT differ predominantly in the 2-2000 range (Figure 5). This suggests that, as far as dictionary citation forms are concerned, JTs desiring to teach wholly in L2 should focus on the top 2000 words, with special attention to the top 1000.
ET and JT cumulative percentages are roughly similar between the 1st and 3000th places in the JACET wordlist. Below 3000, ETs and JT gradually separate, with the JT using relatively more tokens below 4000. This may be due to the JT’s role of introducing vocabulary items in class (most new words are dependent on courseware content, and often fall outside the JACET wordlist). The ETs’ role is motivating students to speak with classmates and instructors.

With regards to teacher training, JTs aspiring to teach in L2 should be prepared to (1) talk about themselves, and compare their views with their students, and (2) ask what, who, why questions, and accept and/or correct incomplete responses. Students will benefit from classroom management language spoken by ETs; to do this, ETs should familiarize themselves with courseware planning. These kinds of training are probably feasible, and undoubtedly will give both ETs and JTs a sense of accomplishment and professional growth.

We compared word tokens that appeared more often in either ET or JT speech. Normalized frequencies are shown for each of the JACET word ranks. Rank 1.001-1.137 includes essential words that JACET deems equally worthy of instruction regardless of frequency in language (e.g., cardinal numbers, days of the week). The JT uses such words more often than ETs (e.g., when referring to textbook page numbers). Conversely, ETs strongly prefer words in the 2-1001 rank. The ET normalized frequency is almost double that of JT. Words in the 1002-2001 rank appear at roughly identical frequencies. Hence, JTs should produce more words in the 2-1001 rank to become ET-like.

We identified words with high keyness (i.e., occurring with unusual frequency compared to a benchmark corpus) [5]. Figures 6 through 8 show distributions of such words. The length and direction of the vertical bars indicate keyness. Tables 2 and 3 list the top 30 words that appear unusually frequently in JT and ET speech.

The pattern that emerges from keyness observations is that the JT is comfortable teaching from the textbook and chooses to do so almost exclusively, whereas the ETs are not trained in the courseware and engage in verbal exchange with students.
Table 2: Top 30 words used more frequently by JT than ETs. Words that were used more frequently by the JT are shown in order of relative frequency difference (rightmost column). Left to right, the columns show (1) JACET wordlist rank, (2) word type, (3) classroom action verb flag, (4) ET frequency, (5) JT frequency, (6) normalized JT frequency, (7) ET-JT, and (8) rank order of ET-JT frequency.

Table 3: Top 30 words used more frequently by ETs than JT. The format of this table is identical to Table 2. Compared to the JT, who seems to encourage students to work silently (action verbs look, listen, check), ETs seem to encourage open-ended dialog between them and their students (action verbs think, say). Further indication of speaking motivation is seen in pronouns (you, it, I), and interrogative words (what, who, why).

4. Discussion
Our findings suggest that JTs desiring to teach wholly in L2 might consider focusing on the 2000 most-common words, with emphasis on familiarity with producing the verb counterpart of common nouns, and vice versa (e.g., drive or fork used as verbs, and take or call used as nouns). Avoiding less descriptive verbs such as be and do increases ET-ness.

Our data also seem to show that native instructors favor open-ended questions that demand higher proficiency of students, whereas non-native instructors prefer to introduce vocabulary and syntax one at a time to accommodate learners across various levels. JTs desiring to teach wholly in L2 might consider including ET-like behavior in their practice. ETs might become aware of their students' abilities.

Given the gross similarity between ET and JT language use, we believe that students expect classroom-oriented language, and that classroom L2 can be as natural and authentic as any other context.

We plan to analyze our video corpus with regards to (1) word chunks (e.g., n-grams), (2) utterance complexity (e.g., number of words per sentence, syntactic structure), and (3) disfluencies (particularly false starts and repairs). Both (2) and (3) are expected to show larger means and variance among ETs than JTs. Knowledge gained will be used to train JTs and students to more ET-like behavior, and to train ETs to plan their choice of vocabulary and syntax.

5. Acknowledgments
The authors are grateful to our students and colleagues who agreed to participate in this study.

6. References