Sophia Junior College Faculty Journal
Vol. 29, 2009, 25-45

Assessing Lexical Production in NNS-NNS Casual Conversations: A Mini-Corpus Approach

Timothy Gould

This paper is a general introduction to and rationale for the construction of a linguistic corpus based exclusively on casual L2 English conversations between female L1 Japanese junior college students. As an English teacher to this narrow population of learners, my motivation is to try to gain a deeper understanding of how our students use their verbal English skills when they are not speaking in a classroom environment or guided by learning oriented tasks. In other words, I want to begin to address the question, "Of all the English our students have learned, what are the words and constructions they use when they are on their own?" Although I refer to the data gathered and prepared thus far as a corpus, it might more accurately be called an interim, or "mini" corpus. As such, this is a work in progress and the data presented below is an exploratory precursor to analysis using the larger and more representative corpus pointed to here. One of my main goals is to illustrate to other teachers who work with these students the nature of a corpus and to attempt to show how they might find this to be a valuable resource helpful in their own teaching and research efforts. Additionally, simply browsing the corpus may lead to a better sense of our students' knowledge and provide insights into how better to approach teaching them. To this end, the paper proceeds as follows.

In the first part I lay out the basic design and methodology being used to capture and transcribe the data that makes up the corpus. In the second section I detail the preparation of the transcripts that provide the raw data of the corpus and I discuss some of the issues and theoretical decisions that have been made in this effort. In the third section I present some basic statistics extracted from the corpus and I also give a brief overview of some concordance capabilities available to assist in analyzing the corpus. Let me make two comments about the nature of this paper. Since I intend to make this corpus available for use by other teachers, I think its construction should be documented in a way that allows anyone to clearly see how it is being put together. I hope that this detailed view of the process will generate some constructive comments and criticism that will form the starting point for further discussions to help guide its development as the corpus grows. Secondly, in the third section I have included quite a bit of raw, unanalyzed data. This data is offered as an initial glimpse into the kinds of words our students employ and may help trigger questions that teachers might want to use the corpus to help answer.

Data Collection and Methodology

In gathering the data, I have tried to simulate, as closely as possible, the conditions of a ‘natural’ conversational environment in order to capture the type of free form conversation the participants might be called upon to join in the wider, non-pedagogical ‘real’ world. The general notion behind this methodology is essentially that if we can witness our students using their L2s when they are not guided or influenced by us (their EFL teachers), we can target our own pedagogical interventions much more specifically and to greater effect. The main problem we encounter, however, when we try to ‘witness’ our students’ use of language, is that our presence during a conversation removes precisely the spontaneity of interaction and naturalness we are interested in capturing. This issue is known as the ‘observer’s paradox,’ and to escape its influence, the conversations that comprise this corpus were videotaped without the presence of a teacher. Specifically, students in my required English classes were allowed to self-select from among their classmates into groups of three participants each. Since students in these required English classes engage in other activities and classes together, the students knew each other and were easily able to divide themselves into small groups. They were then given ‘free conversation’ time in class to help get them used to talking to each other in English before the actual videotaping.

When it came time to videotape the individual groups, I started the video camera recording and immediately left the room for the ten-minute duration of the taping. Although the presence of a video camera may have had a slight effect on the naturalness of the conversational environment, I attempted to put the students at ease and alleviate any nervousness they may have felt. They were informed that their performance during the videotaped conversation would not be part of their course grade and that I was not going to use the tape to evaluate them in any way. Viewing the results, the students, if they appeared anxious at all, were more likely to express concern about their English skills and choosing conversation topics than about the presence of the video camera, and they soon ceased to take any notice of the camera at all and seemed to be comfortable and at ease.

As stated earlier, I refer to the work here as a mini-corpus approach. This is meant to indicate that the corpus I have constructed here is in its nascent stages and has been purposefully limited in size and scope in order to be just large enough to test some of its potential uses, determine directions for future research, fine tune the nature of the corpus itself, and expose and remove as many weaknesses and limitations as possible before committing the necessary time and effort to building a larger and more robust corpus. The mini-corpus utilized in this paper is based on the transcripts of six of the above-mentioned videotaped sessions. The transcripts themselves were produced and linked to the videotapes following the conventions of CHAT (Codes for the Human Analysis of Transcripts) (MacWhinney, 2000), and subsequently modified for analysis. Details of the preparation and modification of the transcripts will be taken up in the next section.

While a corpus of this nature has many potential uses for investigating language across a wide spectrum of disciplines, in this paper I focus specifically on the construction of the corpus and its usefulness in helping to determine the nature and frequency of our students' vocabulary production. At this early stage of corpus building, I will offer some rudimentary statistics, but future research will subject the data to more sophisticated analysis. Now, I am particularly keen to establish a perspective whereby we can generate a basic profile and initial analysis of each student's lexical diversity and place their performance along a continuum ranging from the specific details of their individual contributions to the conversation, out to a global view of their performance in relation to the corpus as a whole.

In the next section, then, I will lay out in detail the process of constructing this small corpus, provide some provisional results, and set the stage for future work which will attempt to provide a resource for research and pedagogical questions to mine this ever-growing corpus for additional insights into how our students use their English skills while conducting free conversations.

Corpus Construction

The original meaning of "corpus" is, of course, "body." Generally, then, we understand a linguistic corpus to be a "body," or collection, of words. Putting aside the non-trivial issue of precisely how to define a word, for our purposes here we will consider a word as simply a string of letters separated from other strings by spaces. Even this oversimplified definition, however, belies a host of complexities for second language corpus construction, some of which I will now describe.

Preparation of a linguistic corpus from transcripts involves the making of decisions that can have both theoretical and statistical implications on the findings gained from them. In this section I will detail some of the relevant choices that were made and provide, to the extent possible, the rationale for them. The first issue that was dealt with was that many of the words spoken during the videotaped conversations were uttered in Japanese. This presented problems on a number of fronts. As detailed in Gould (2008), a common occurrence by the participants was to handle administrative issues related to the conversation in Japanese. For example, students would often converse in English about, say, what they did last weekend, but then switch to Japanese to determine what topic they would broach next or to work out conversational troubles.

Since the transcripts being made from these conversations are also used in other areas of research, an accurate account of all utterances, including the Japanese, must be maintained. It is not a possible option to simply leave out the Japanese lexical items and discourse markers during the transcription process, so they must be included in the transcript, yet excluded from the frequency analysis of the English vocabulary. To accomplish this, an “exclude” file containing all of the Japanese words found throughout all the transcripts was compiled and entered into the frequency analysis software. In this way, when the software program analyzes the transcripts, it ignores the Japanese words. On its face this seems like an unremarkable and straightforward process, but there are a surprising number of Japanese words and discourse markers which, when transcribed in Roman script, have the same form as English words. In order to find instances of these, a trial run of the frequency software was conducted on the original transcripts. The output of this process takes the form of an alphabetical list noting the frequency of each item and indicating where it appears in the transcript.

Each possibly ambiguous item between Japanese and English, then, must then be manually verified to make sure it is a legitimate English word. A problem here for native English speakers checking these files, and hence an area where much time is required to prevent mistakes, is that when reading an item, it is very difficult to look at L1 (English) words and read them as L2 (Japanese) lexical or discourse items. For clarity I will provide some illustrative examples.

The orthographic form which constitutes the English word “made” can also appear as a Japanese word, as in “itsu made.” Once the frequency list has been generated, however, each item appears on an individual line and is thus stripped of its context, so if a speaker used the English word "made" in a sentence, and the Japanese "itsu made" elsewhere in the conversation, the output list produced by the software will contain an entry that looks like this:

* made: 2

Indicating that two occurrences of the form "made" were found. When orthographic ambiguity of this type is discovered, we cannot merely enter the ambiguous string into the exclude file, because doing so would also exclude the legitimate English word–an unacceptable outcome. The solution in this case has been to scour the transcripts for these "double agents," and temporarily mark the Japanese words and then to add the new, altered word, "jmade," for example, to the exclude file.

Some of the other orthographically ambiguous strings which were discovered include:

* "men" as found in "ramen" looks like the English word "men."
* The Japanese possessive "no" has the same orthographic form as the English negative "no."
* The Japanese deictic marker "sore" is the same as the English pain indicator "sore."

Even forms that are not lexical items in Japanese can present problems when they appear in the form of English words. The first run of the trial frequency program produced a number of instances of the form "a." This was initially assumed to be the English determining article, but further investigation revealed that "a" had been used as a type of pause filling device during the conversation. Here are two examples from different speakers in different conversations:

* ST2: #1_1 a #2_0 (laughs) u:n I was junior high school student.
* ST1: #3_2 a #1_6 what will you #2_0 do ¿ #1_1 christmas #1_5 day?

In the example from student 2, the phonetic form "a" qua article would have been correctly placed before "junior high school student," but as it was uttered, it is obviously not a lexical item but, as stated above, a pause-filler. In the entire corpus, a total of forty-five instances of "a" were found, of which less than half were uttered as the English article. Given the amount of time and ink spent by second language educators trying to teach the correct usage of determiners, this points to an interesting area that can be easily researched and analyzed using a corpus approach.

Another modification of the original transcripts for corpus use concerns the manner in which non-standard English pronunciation is preserved. Some Japanese students maintain in their English speech patterns an L1 rule from Japanese which requires a vowel ending for each syllable. The manifestation of this Japanese rule (which is of necessity perpetuated by Japanese-English dictionaries published in Japan) in English speech results in "katakana" English, where a vowel is uttered at the end of every word, but it is especially prominent after full stops. Examples of this behavior from the current transcripts include "watched-u," "good-o," and "watch-i" among many others. If these forms are left in their original state, frequency counts applied to the corpus consider "watched" and "watched-u" as two different word types, which they are not. We are interested here in collecting evidence of vocabulary items that the speakers have used correctly, and "watched-u," although not adhering to prescriptive English pronunciation rules, is an unambiguously correct and meaningful use of the word "watched," so we must count it as such. As it relates to the preparation of the corpus, then, a decision was made in this case, to alter the original speakers' pronunciation to match the accepted orthography of each word in question. So while some accuracy, in terms of the transcripts' portrayal of real world speech events, has been lost, the trade off, which improves our ability to analyze correct lexical usage, has been determined to be acceptable. This solution, however, is considered ad hoc and a more elegant solution, while not available now, will be incorporated into future versions of the corpus. The solution lies in tagging certain words and families, which will be automatically altered before being submitted to the corpus for analysis.

In addition to changing non-standard orthography in order to capture correct usage, as detailed above, there are some cases in the transcript where actual English words are spoken by participants, but they are phonetic repetitions of a previous speaker's utterance and do not seem to carry the semantic load that would allow us to consider them instances of the word whose form they resemble. Consider the following exchange, which occurred during the conversation conducted by Group 5:

* ST1: Chikuabu is fish.
* ST3: raw fish?
* ST1: [rawfish]?

Student 1 is explaining about a certain type of fish, "chikuabu," and Student 3, understanding that it is a type of fish, or a way of preparing fish, asks a clarifying question to determine if “chikuabu” is raw. Student 1’s response, however, presents us with a problem; how do we treat utterances which appear to be English words, but, based on the context, do not seem to contain semantic content? In the excerpt above, I have transcribed her answer within square brackets to indicate that her utterance appears to be a phonetic approximation, repeating what she heard, not an additional clarification question. I do, however, want to maintain the connection with the previous utterance, so it has not been transcribed as [roffish], which is actually how this utterance sounds. In this case, the term ‘rawfish’ was added to the exclude file so the frequency software would ignore it.

This type of situation opens another avenue for possible research using a corpus-based approach. Namely, is the strategy by an interlocutor of repeating the phonetic shape of an item that has not been understood an effective one? And how common is this strategy? Despite teachers exhortations to persuade them to use set-piece phrases such as “Could you repeat that please?” or “I’m sorry could you say that again,” do students really use these phrases when left to their own devices? Again, preparation of the transcripts has exposed an issue which may spur further corpus-based research, and one which will be taken up in future work.

A similar situation but with a different resolution arises when a participant fails to understand part of a previous speaker’s utterance, and in their effort to clarify the trouble, they produce a word of English, but not the one that had just been spoken. Here is an example of this phenomenon, which occurred during Group 6’s conversation:

* ST3: do you like this school?
* ST2: disk?

Without diverging too far afield into a discussion about pronunciation, suffice it to say that Student 3’s question appears to have been initially interpreted by Student 2 as “do you like disk-u?” Since this question comprises a topic initiating turn by Student 3, in which she is closing the previous topic about the cuteness of someone’s daughter, there is no previous context against which Student 2 can gauge the relevancy of her interpretation, hence her clarifying question, ‘disk?’ So to return to the issue of preparation of the transcript for corpus use, the question arises as to whether or not this word should be counted as an occurrence of the English word "disk." In this case, unlike the situation above concerning "rawfish," I believe that the Student 3's question activated a real lexical item which is part of Student 2's vocabulary, so it was decided to count this particular instance of "disk" as legitimate word use by Student 2. This decision notwithstanding, I also recognize that "disuku" is a loan word from English to Japanese, so there are plausible arguments against my choice. Before closing this discussion, however, let me introduce another example with a different outcome. In the following excerpt, Student 3, in answer to a question about what traditional Japanese food she likes, introduces the word "radish," which seems to be an unknown word for the other participants.

1. * ST3: e: radish.
2. * ST1: tanish?
3. * ST2: rashu?
4. * ST3: radish.
5. * ST1: radish?
6. * ST3: Daikon.
7. * ST1: a:::.
8. * ST2: a:::.

I have included the entire exchange because I think that it reveals the nature of the participants strategy for dealing with lexical troubles, but the focal point for this discussion is Student 2's use of the form "rashu" in line 3. Student 1's "tanish" cannot be considered in any light an English word, so "tanish" was simply added to the exclude file. "Rashu," on the other hand, as uttered by Student 2, does realize the phonetic form of an English word when stripped of it final "u." In this case, I took advantage of my access to the audio and video context surrounding Student 2's lexical output during the conversation to make a determination about how to treat this item. While I cannot entirely discount the possiblility that Student 2 knows the word "rash" in English, it seems more likely to me that she is simply repeating the basic phonetic shape of Student 3's utterance of "radish." Part of my decision rests on Student 3's initial pronunciation of "radish," which is produced with a very lightly flapped "d." This light flapping makes Student 2's hearing it as the phonetic "rash" very plausible. Unfortunately, at this time the video is not accessible to the reader. To remedy this, however, as the current corpus grows, I would like to make it available online so that others can watch the interactions themselves and come to their own conclusions about my decisions. This access can possibly even provide a forum for discussion, which would add to the general knowledge base about how our students deploy their English skill as they interact in free conversation.

I will now briefly discuss the difficult issue, hinted at earlier, about how best to handle loan words from English to Japanese which appear during the conversations. Despite earlier claims that loanwords were a hindrance to L1 Japanese learners acquiring English (Simon-Maeda, 1995), research by Daulton (1999, 2007) found that, "English loanwords in Japanese greatly enhance the acquisition of the English basewords on which they originate." In light of this, I am allowing most loan words to remain in the frequency count if they were used during an English utterance. Many of these words, however, are substantially modified from their original English forms, as in utterances such as, "I watched terebi last night," or "on terebi." Instead of changing the transcription to inaccurately portray the clearly distinct phonetic form "terebi" as "television," then, I have allowed both "terebi" and "television" as distinct lexical items. This move is also considered an ad hoc solution, but at this time I am not sure how best to consistently handle this issue. On the one hand, "terebi" will appear on the frequency list, which is not accurate, but I will also be able to search out the use of loan words. Although there are good arguments to disallow Japanese derived forms of English words, I am not closing the theoretical door against future change, but for now most loan words in English contexts, even with Japanese pronunciation, will be counted. This caveat about context is intended to disallow English loanwords spoken in Japanese contexts, such as the following utterance from Student 2, Group 1:

* ST2: kino mita terebi toka.

Because "terebi" and other loanwords are allowed or disallowed based on context, each one must checked individually. This process will become unwieldy as the corpus grows, so in the future a coding system will have to be employed during the transcription process. As an aside, it goes without saying that loanwords such as "baito," which originate in languages other than English, have not been included in the corpus used here.

Although a smaller issue, the converse of the English to Japanese problem also obtains when participants use Japanese words that are loanwords to English. Again, a determination about whether to include the word must be made on a case by case basis, but here the theoretical line defining which word belongs to which language becomes even fuzzier and is beyond the scope of this paper. Nevertheless, let me lay out two illustrative cases. Here is a brief exchange concerning a participant's part-time job.

* ST2:	oh baito.
* ST3:	part time job?
* ST1:	part time job.
* ST1:	yes.
* ST3:	mmhmm.
* ST2:	sushi?
* ST1:	sushi.

We have here two instances of the word Japanese word "sushi." The context in which this word appears also contains a German loanword which, when uttered, prompts an English translation by the other participants. The example here of double confirmation by Students 3 and 1 of the English translation of "baito" – "part time job," is an interesting phenomenon in its own right and an issue for later study, but here we must determine how to handle "sushi." The context of this conversation, including the parts before the excerpt included here show that this first instance of "sushi," uttered by Student 2, really carries the illocutionary force of the question, "you work in a sushi restaurant, don't you?" To which the second utterance by Student 2, means "yes, I do work in a sushi restaurant." In this case, however, the semantic implication of these two sentences could also have been expressed in Japanese. So while the meaning is clear, the language expressing that meaning remains ambiguous, so I have chosen to disallow these utterances of "sushi." Elsewhere in the transcripts, the word "sushi" appears in the sentence, "I like sushi." In this case I have accepted it into the corpus as an English word.

In the interest of space, I will briefly introduce some other issues and their resolutions without a full explanation. Some speakers display a tendency to repeat a single word a number of times as part of the same utterance, as in the excerpt below:

* ST1:	a: do you have Christmas plan?
* ST3:	yeah yeah yeah yeah yeah.
* ST3:	no no no no no no no boyfriend no.

This hyper-repetition has the effect of inflating the frequency of the words in question. To resolve this issue, I used a convention available in CLAN which lists the number of repetitions of an item in square brackets. So Student 3's utterance of "no" seven times would appear as "no [x7]" in the transcripts. This maintains the correct lexical usage and allows us to see the affected repetition, yet ignores the repetition when calculating word frequencies.

Other areas of difficulty include Japanese band names, which often use low frequency words, and if counted as part of a participants lexicon, artificially inflate her lexical diversity. For example, Japanese band names that were excluded from the transcripts for this reason include, "Bump of Chicken," "Exile," "Boa," and "Mr. Children." Song lyrics present another area of difficulty because the lyric can be quoted by a participant without necessarily understanding the lyrics. For example, a well known phrase such as "I love you" would be accepted, but "whispering sweet nothings," without clarifying context, would not. In general, place names, brand names, and other referential items not used contextually have been excluded.

In this section, I have detailed some of the issues that became apparent during the preparation of only six transcripts for use in our mini-corpus. Many of the issues have been solved in an ad hoc fashion and await further investigation to find clearer and more efficient ways to handle them. It seems that no matter how the transcripts are prepared, however, the process is labor intensive and requires a great deal of planning and coordination if it is to be done on a larger scale. Part of my goal in articulating the task of preparing the transcripts has been to elicit comments from other potential users who might have ideas on how to streamline the entire process. In the next section I will introduce and explain some of the preliminary statistics available from the corpus.

Corpus Based Data

This section is devoted to presenting some of the basic output and information available from our mini-corpus. I begin by talking about frequency counts and then move to one of the most often used, yet still controversial, analytic devices for evaluating lexical diversity–the type-token ratio. First, frequency measures, which simply list and count the number of words in a text, provide a useful way of initially taking stock of a corpus. Frequency counts take as input a text, along with instructions about which strings should be ignored (see previous section), and output a list of words sorted alphabetically or by the frequency with which each word appears in the text.

Starting with West (1953), frequency lists based on large corpora have been used by teachers (and others) in an attempt to determine the core second language vocabulary necessary for language learners. Additionally, frequency counts provide the input for determining some basic measures of lexical diversity, which are meant to indicate the 'richness' or variety of a speaker's or group's vocabulary. Below I explain the calculation of the type-token ratio followed by the results for each student participant in the corpus under discussion here. In presenting this data, I have listed the students by the conversation group they participated in and I also give the cumulative type-token for the ratio entire group. This exhaustive listing exposes some of the weaknesses of the type-token ratio that I will also discuss.

Let me use one student as an example to explain the type-token ratio. Student 1 from Group 1 spoke a total of 92 words during the ten-minute conversation. From this performance, we can calculate the lexical variation in her speech by dividing the total number of words uttered by the number of word types used. For example, Student 1's frequency profile shows that she said "do" three times. If these were the only words she spoke during the conversation, we would figure her type-token ratio by dividing 1, the number of types, by 3, the number of instances of that type, to arrive at a type-token ratio of .33. The type-token ratio can range from infinitely small (x repetitions of one word) to one (no words repeated by a speaker) and is best used as a comparative tool to analyze samples of relatively similar sizes. While all of the conversations making up our corpus are ten minutes long, the nature of each conversants' input varies according to the distinct dynamics of that particular conversation. Some groups in general are more talkative than others, and while some distribute the conversations equally among themselves, there are cases when a dominant speaker emerges and contributes the bulk of lexical items. This distribution of speaking can be analyzed using the corpus and is an area for future research.

As stated above, Student 1 uttered a total of 92 words. Within these 92 words there are 51 types, which gives us the following:

Student 1, Group 1 type-token ratio: 51 types / 92 tokens = .554

Now let us look at the type-token ratios for all of the groups and students comprising the corpus. Each underlined heading below details the input to the type-token calculation, the type-token ratio for each participant, and the cumulative type-token ratio for the entire group's conversation. After these, we have the cumulative type-token profile for the corpus as a whole.

Type-token ratios for Group 1; 3 participants and cumulative

Student 1:	51 types / 92 tokens = .554
Student 2:	49 types / 93 tokens = .527
Student 3:	55 types / 108 tokens = .509
Group:	98 types / 293 tokens = .334

Type-token ratios for Group 2; 3 participants and cumulative

Student 1:	27 types / 39 tokens = .692
Student 2:	110 types / 296 tokens = .372
Student 3:	48 types / 76 tokens = .632
Group:	124 types / 411 tokens = .302

Type-token ratios for Group 3; 3 participants and cumulative

Student 1:	53 types / 85 tokens = .624
Student 2:	40 types / 67 tokens = .597
Student 3:	73 types / 125 tokens = .584
Group:	106 types / 277 tokens = .383

Type-token ratios for Group 4; 3 participants and cumulative

Student 1:	88 types / 180 tokens = .489
Student 2:	81 types / 161 tokens = .503
Student 3:	63 types / 103 tokens = .612
Group:	138 types / 444 tokens = .311

Type-token ratios for Group 5; 3 participants and cumulative

Student 1:	52 types / 105 tokens = .495
Student 2:	28 types / 55 tokens = .509
Student 3:	79 types / 144 tokens = .549
Group:	100 types / 304 tokens = .329

Type-token ratios for Group 6; 3 participants and cumulative

Student 1:	101 types / 269 tokens = .375
Student 2:	94 types / 174 tokens = .540
Student 3:	83 types / 187 tokens = .444
Group:	154 types / 630 tokens = .244

Type-token ratio for all participants across all 6 Groups
395 types / 2359 tokens = .167

I mentioned earlier that the type-token ratio is best used to compare lexical diversity between similar sample sizes. This dependence on sample size can be seen when we compare the results for the different levels of analysis. In Group 1, for example, the three participants each obtain similar type-token ratios between .509 and .554. From this we can see that they seem to have divided the conversational ‘labor’ between themselves relatively equally. When we look at Group 2, however, we see that Student 1 and Student 3 both obtain comparatively high type-token ratios, but surprisingly, Student 1, with the lowest total number of spoken word types, has the highest type-token ratio of all eighteen students included in the corpus. With 27 types of words used across 39 tokens, we can safely conclude that hers is not the most lexically rich conversation, so we see the caveat about sample size is well deserved. Additionally, we can see that as the sample size increases, the type-token ratio decreases, so that when we calculate the type-token ratio for the entire corpus, we obtain a .167.

Although the type-token ratio has its weaknesses, it can help us find and develop research questions in areas we might not otherwise be inclined to look. Take Group 1 again, for example. As noted, the three students have very similar type-token ratios, showing an apparently equal division of the conversation, but when looking at the video and transcripts it is quite clear that Student 2 is the weakest English speaker in the group. So how does she obtain a type-token ratio similar to the other students? While I believe that her performance is based on a strategy of repeating the other students' utterances, the important point is that again we see how corpus analysis can reveal, and then help us to explore, questions about our students' linguistic performance. It should also be noted that other measures of lexical diversity which attempt to overcome the weaknesses inherent in the standard type-token ratio have also been developed, but since here I am only introducing some basic features of corpus study, those will have to be detailed elsewhere. In the remainder of this paper I will introduce another tool which makes use of corpus based word lists, the lexical frequency profile.

Lexical Frequency Profiles

The Lexical Frequency Profile (Laufer & Nation, 1995), is a measure which displays “the percentage of words a learner uses at different vocabulary frequency levels” (p. 311). This means that a text, in our case the spoken production of L2 English speakers, is compared against frequency lists compiled from large (over 1 million words) corpora and the results give an indication of that speaker’s general vocabulary. In the tables below I have included the lexical frequency profiles for each student in Group 1 and also the cumulative profile for all six groups, essentially the entire corpus as it now stands. Although I do not intend the reader to slog through all of the data I have included here, I do think that it would be worthwhile for teachers who work with this population to peruse the output to get an idea of the types of words our students are using in unguided conversations.

The profiles below list “K1” words, which are those words spoken by a student that are also contained in the corpus-generated first 1000 most frequent words. The types and tokens are separated according to the list they appear on, so looking at Student 1’s profile, we see that she uttered 92 tokens of 47 words that appear on the list of 1000 most frequent words, which represents 94.85% of her total production. The K1 words are further broken down into function and content words with the number of each placed in parentheses in the “token” column. Next, we see the words that the student used from the “K2,” or second 1000 most frequent words. In the case of Student 1, she used three 2K words, which represent 3.09% of her total lexical production. The next row gives a total percentage of words used from both the 1K and the 2K lists – 97.94% for Student 1. AWL in the next row stands for the “Academic Word List,” developed by Coxhead (1998). The inclusion of the AWL is not absolutely necessary for participants in casual conversations, but I have included it here to give an idea of the types of words our students use that are found on that particular list. We see that Student 1 used one word from the AWL, which represents 1.03% of her total production. The final row shows the statistics for words that do not appear on any of the word lists, hence “off-list.”

For ease of reference, I have placed beneath each lexical profile the complete type or token list, which lists the words referred to in the corresponding lexical profile table. So beneath Student 1’s profile, for example, we find the token list of her total production, which is further broken down by the most frequent 500 function and content words. For Students 2 and 3 and the cumulative profile, I have included the type list, which presents each type along with its frequency in square brackets. In the cumulative chart, which represents the entire corpus, we see that 92.22% of all the words spoken were from the 1K and 2K most frequent lists.

Again, I want to reiterate that any comments and suggestions are welcome about how this data might be used to better understand and accommodate our students and improve the way we approach teaching them English. This paper has only scratched the surface in mining the data that is already available here, and future investigations, along with a larger, more robust corpus, are sure to yield additional insights.

Appendix

Lexical Frequency Profile Group 1 Student 1

	Types	Tokens	Percent
K1 Words (1-1000):	47	92	94.85%
Function:	...	(52)	(53.61%)
Content:	...	(40)	(41.24%)
K2 Words (1001-2000):	3	3	3.09%
1k+2k	...	...	(97.94%)
AWL Words (academic):	1	1	1.03%
Off-List Words:	1	1	1.03%
	52	97	100%

0-1000 about always am am am another ate bad because been country day did do do do eat ever go good good have house how how how i i i i i i i i i i i is is it last long long me me me me me morning morning on on part real september stay that thirty this time to to to too too too too too too want want watch watched what what where where why will will yes yes yes yes yes yes yes yesterday you you you you

First 500 function: about always am am am because been did do do do have how how how i i i i i i i i i i i is is it me me me me me on on that this to to to what what where where why will will you you you you

First 500 content: another bad country day ever go good good house last long long morning morning part real time too too too too too too want want yes yes yes yes yes yes yes

Second 500 content: ate eat september stay thirty watch watched yesterday

1001-2000: christmas dinner grandfather

AWL: job

OFF LIST: pasta

Lexical Frequency Profile Group 1 Student 2

	Types	Tokens	Percent
K1 Words (1-1000):	42	84	89.36%
Function:	...	(52)	(55.32%)
Content:	...	(32)	(34.04%)
K2 Words (1001-2000):	4	6	6.38%
1k+2k	...	...	(95.74%)
AWL Words (academic):			0.00%
Off-List Words:	3	4	4.26%
	49	94	100%

1k types: about_[1] and_[2] bad_[1] been_[1] come_[1] did_[3] eat_[2] event_[1] go_[3] good_[2] have_[3] high_[1] how_[1] i_[10] interesting_[1] is_[2] it_[2] last_[1] laughs_[1] me_[4] morning_[2] on_[2] school_[1] student_[1] that_[1] this_[1] to_[5] too_[4] two_[1] very_[1] want_[1] was_[2] watched_[1] weeks_[1] went_[1] what_[1] when_[1] where_[1] will_[4] yes_[4] you_[4] your_[1]

2k types: birthday_[1] christmas_[2] dinner_[2] tomorrow_[1]

AWL types: 0

OFF types: homestay_[1] illumination_[2] junior_[1]

Lexical Frequency Profile Group 1 Student 3

	Types	Tokens	Percent
K1 Words (1-1000):	47	100	91.74%

Function:	...	(62)	(56.88%)
Content:	...	(38)	(34.86%)
K2 Words (1001-2000):	2	2	1.83%
1k+2k	...	...	(93.57%)
AWL Words (academic):	1	1	0.92%
Off-List Words:	5	6	5.50%
	55	109	100%

1k types: about_[2] always_[1] and_[2] august_[1] bed_[1] been_[1] but_[1] country_[2] days_[1] did_[1] do_[5] early_[1] english_[1] foreign_[2] four_[1] friend_[1] go_[5] good_[1] have_[5] how_[4] i_[14] long_[1] morning_[1] my_[1] part_[1] plan_[1] seven_[1] so_[2] speak_[2] study_[2] there_[1] time_[1] to_[7] too_[1] twenty_[1] want_[4] was_[1] watch_[1] watched_[1] went_[1] what_[1] when_[1] will_[3] winter_[1] with_[1] yesterday_[1] you_[8]

2k types: during_[1] tomorrow_[1]

AWL types: job_[1]

OFF types: british_[1] french_[2] headache_[1] nhk_[1] vacation_[1]

Lexical Frequency Profile Groups 1 - 6

	Types	Tokens	Percent
K1 Words (1-1000):	271	2104	87.56%
Function:	...	(1156)	(48.11%)
Content:	...	(948)	(39.45%)
K2 Words (1001-2000):	41	112	4.66%
1k+2k	...	...	(92.22%)
AWL Words (academic):	6	28	1.17%
Off-List Words:	51	159	6.62%
	369	2403	100%

1k types: about_[33] actor_[3] after_[1] ago_[1] all_[1] alone_[2] always_[2] am_[30] and_[25] another_[4] answer_[4] anyway_[1] are_[11] ask_[3] at_[4] ate_[1] august_[1] back_[3] bad_[2] beautiful_[3] because_[5] bed_[3] been_[7] best_[8] big_[4] brother_[7] but_[17] by_[4] can_[5] change_[1] children_[1] choose_[1] class_[4] classes_[4] cloudy_

[1] cold_[4] college_[7] come_[3] comes_[1] could_[1] country_[5] course_[1] daughter_[1] day_[3] days_[2] december_[1] did_[15] difficult_[3] do_[82] dream_[3] drinking_[1] early_[4] eat_[5] egg_[3] eighteen_[1] english_[5] event_[1] ever_[4] everyday_[1] everything_[2] example_[3] famous_[2] far_[1] favorite_[4] february_[1] fifteen_[1] fifth_[1] figure_[1] fine_[6] finish_[2] fish_[6] five_[7] food_[4] foods_[1] for_[5] foreign_[3] forty_[2] four_[14] fourteen_[2] friend_[7] friends_[4] from_[9] future_[1] game_[2] gentleman_[1] get_[9] go_[30] going_[2] good_[9] has_[3] have_[48] he_[4] head_[1] heavy_[2] help_[1] her_[2] here_[5] high_[8] him_[7] his_[2] home_[8] hot_[3] hour_[2] hours_[4] house_[1] how_[42] hundred_[1] I_[183] if_[2] in_[11] interesting_[1] is_[55] it_[17] january_[4] kind_[6] know_[23] land_[7] last_[10] late_[1] laughing_[5] laughs_[1] let_[2] life_[1] lights_[1] like_[44] listen_[2] little_[2] live_[3] ll_[1] long_[4] look_[1] love_[22] many_[6] march_[2] married_[2] maybe_[7] me_[39] money_[2] more_[2] morning_[5] mountain_[1] much_[4] music_[2] must_[2] my_[19] name_[5] near_[1] new_[4] next_[3] night_[7] no_[21] not_[21] nothing_[2] now_[5] of_[9] old_[5] on_[6] one_[9] only_[3] or_[5] part_[9] party_[1] people_[3] plan_[12] player_[3] please_[1] pretty_[1] question_[2] real_[1] really_[10] recently_[1] red_[4] remember_[1] report_[2] reports_[2] rest_[1] right_[1] same_[8] school_[11] sea_[2] see_[5] september_[1] seven_[1] she_[8] show_[1] singer_[6] singing_[3] sister_[2] six_[3] sleep_[2] sleeping_[2] sleepy_[5] small_[5] smile_[1] so_[18] something_[1] sometimes_[3] song_[1] soon_[1] speak_[2] spend_[1] spring_[4] stay_[5] story_[2] student_[4] study_[3] summer_[2] system_[2] takes_[2] ten_[2] test_[2] tests_[1] than_[2] that_[8] the_[10] there_[3] they_[2] think_[9] third_[4] thirty_[1] this_[11] thousand_[1] three_[10] time_[11] times_[3] to_[54] today_[12] too_[48] train_[3] twelve_[2] twenty_[10] two_[12] university_[2] use_[4] very_[20] voice_[2] walking_[1] want_[16] was_[4] watch_[4] watched_[3] watching_[2] we_[9] wednesday_[1] week_[3] weeks_[1] well_[2] went_[6] what_[38] when_[9] where_[11] which_[3] who_[5] why_[9] wife_[1] will_[23] win_[1] winter_[13] with_[11] woman_[1] won_[3] work_[1] working_[2] year_[6] years_[2] yes_[85] yesterday_[2] you_[136] young_[2] your_[7] yours_[2]

<u>2k types</u>: abroad_[1] band_[1] bicycle_[4] birthday_[4] bitter_[4] boiled_[1] busy_[1] chain_[3] christmas_[10] coffee_[2] cooking_[2] cool_[5] dinner_[3] during_[2] engineer_[2] exciting_[1] funny_[1] grandfather_[1] hello_[7] hi_[2] holiday_[2] holidays_[1] hungry_[3] lunch_[6] match_[3] nice_[10] plane_[2] quickly_[1] raw_[3] recommend_[1] rice_[3] shop_[5] shopping_[2] shops_[1] sorry_[2] thank_[1] ticket_[1] tired_[1] tomorrow_[4] toy_[2] weather_[1]

AWL types: adult_[1] job_[22] professional_[1] topics_[1] traditional_[2] transfer_[1]

OFF types: bakery_[1] baseball_[5] boyfriend_[6] british_[1] career_[3] cd_[1] choo_[4] comedy_[2] concert_[4] curry_[2] cute_[11] delicious_[2] disk_[1] eve_[6] everytime_[1] french_[1] handsome_[3] headache_[1] hobby_[1] homestay_[2] hometown_[1] homework_[3] illumination_[2] ipod_[4] japanese_[5] junior_[4] linguistic_[2] linguistics_[1] movie_[2] movies_[1] nervous_[1] nods_[1] noodle_[1] oclock_[5] okay_[2] pasta_[1] piano_[2] prefecture_[2] radish_[3] romance_[1] skate_[2] skiing_[1] soccer_[2] spicy_[3] talent_[1] tv_[2] unbelievable_[1] vacation_[13] versus_[2] yeah_[29]

References

Bogaards, P., & Laufer-Dvorkin, B. (2004). *Vocabulary in a second language: Selection, acquisition, and testing*. Amsterdam: John Benjamins.

Coady, J., & Huckin, T. N. (1997). *Second language vocabulary acquisition: A rationale for pedagogy*. New York: Cambridge University Press.

Coxhead, A. (1998). *An academic word list*. English Language Institute Occasional Publication Number 18. Wellington: Victoria University of Wellington.

Daulton, F. E. (1999). English loanwords in Japanese: The built-in lexicon. *The Internet TESOL Journal, 5* (1)

Daulton, F. E. (2007). Japan's built-in lexicon of English-based loanwords. *5* (1)

Huebener, T. (1944). The teaching of conversation. *The Modern Language Journal, 28* (8), 655-659.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32* (2), 145-164.

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in a second language. In L. Eubank, M. Sharwood-Smith & L. Selinker (Eds.), *The current state of interlanguage* (pp. 265-272). Amsterdam: Benjamins.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics, 19* (2), 255-271.

Laufer, B. (2005). Lexical frequency profiles: From Monte Carlo to the real world: A response to Meara (2005). *Applied Linguistics, 26* (4), 582-588.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16* (3), 307.

MacWhinney, B. (2000). *The CHILDES project : Tools for analyzing talk* (3rd ed.).

Mahwah, NJ: Lawrence Erlbaum.

McCarthy, M. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.

Meara, P. (2002). The rediscovery of vocabulary. *Second Language Research, 18* (4), 393-407.

Nakatani, Y. (2006). Developing an oral communication strategy inventory. *The Modern Language Journal*, *90* (2), 151-168.

Seward, B. H. (1973). Measuring oral production in EFL. *ELT Journal, XXVIII* (1), 76-80.

Simon-Maeda, A. (1995). Language awareness: Use/misuse of loan-words in the English language in Japan. *The Internet TESL Journal, December* Retrieved from http://iteslj.org/Articles/Maeda-Loanwords.html

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17* (1), 65-83.

West, M. (1953). *A general service list of English words*. Essex: Longman.