Apples – Journal of Applied Language Studies Vol. 8, 3, 2014, 85-96



Frequency Change Patterns across Proficiency Levels in Japanese EFL Learner Speech

Mariko Abe, Chuo University

This study investigated the overall patterns of variation across seven oral proficiency levels of 1,263 Japanese EFL learners and native English speakers. The methodological approach combined a learner corpus, language processing techniques, and multivariate statistical analyses to identify patterns of language use. The largest spoken learner corpus in Japan, the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) Corpus was used for the analysis. This corpus consists of transcribed interviews with 1,281 learners and contains over one million running words of spoken English. The transcriptions were compared with data gathered from 20 native English speakers who performed identical speaking tasks. In the present study, 58 linguistic features (e.g., grammatical features) were used from the original list of 67 linguistic features in Biber's (1988) study. The following research questions were addressed. First, what linguistic features do and do not vary among Japanese EFL learners at different oral proficiency levels and native English speakers? Second, is computer-aided analysis of multiple linguistic features useful for determining which ones characterize particular oral proficiency groups? This study found interesting rising, flat, and falling frequency patterns in how several linguistic features are used in different oral proficiency levels. Some linguistic features (e.g., phrasal coordination and nouns) were frequently used by the learners at a low level. The frequencies of some others (e.g., contraction, pronoun it, and emphatics) increased as oral proficiency increased. The study identified a set of linguistic features that differentiate among second language oral proficiency groups as well as between non-native and native speakers of English.

Keywords: learner speech, oral proficiency, frequency change, multivariate statistical analysis, foreign language learning

1 Introduction

Learner corpora have the potential to clarify our understanding of language development. However, few attempts have been made to provide systematic data about language development that can be applied to the development of language teaching and assessment materials (Pendar & Chapelle, 2008). Despite early work in the field of second language acquisition (SLA), relatively few researchers have been concerned with the development of oral proficiency, and

Corresponding author's email: abe.127@g.chuo-u.ac.jp

ISSN: 1457-9863

Publisher: Centre for Applied Language Studies, University of Jyväskylä

© 2014: The authors http://apples.jyu.fi

few have described language development using multiple linguistic features, such as lexical and grammatical features. The number of targeted linguistic features in previous SLA studies has been limited (Biber, Conrad, & Reppen, 1998). Consequently, it is still not clear which linguistic features characterize different oral proficiency levels, and this has resulted in an insufficient understanding of the characteristics of learner language.

Learner corpora are relatively recent, with the construction of such corpora increasing in the 1990s, including the well-known International Corpus of Learner English (ICLE), which was created as a part of the International Corpus of English (ICE). Some recent studies, such as those by Van Rooy and Terblanche (2009) and Asención-Delaney and Collentine (2011), have overviewed various features of learner language, but studies in which learner performance has been analyzed from multiple aspects are rare. Traditional learner language analyses have focused on a single linguistic item, such as a particular lexical item or grammatical feature; however, this approach is not likely to provide an overview of the steps taken by foreign language learners as they progress through different levels of oral proficiency. In order to fill this research gap, the present study aims to identify multiple linguistic features that discriminate oral proficiency levels.

Practical problems with learner corpus studies may have made it difficult for researchers to provide convincing results. The following deficiencies in using learner corpora for SLA research have been observed (Lu, 2010; Myles, 2005; Pendar & Chapelle, 2008): (a) a large-scale learner corpus containing proficiency level information based on an objective rubric has not been readily available to researchers; (b) a reference corpus is lacking; (c) language processing techniques, such as the automatic detection of relevant linguistic features, have not been used to their full potential; and (d) multivariate statistical analyses for assessing the differences of learner language development have not been used to their full potential. Biber (1988) developed a method to analyze language variation quantitatively using digitized data. He used a language processing technique to extract large amounts of frequency information on specific language features, which was subsequently analyzed using factor analysis to investigate overall patterns of variation between the spoken and written language of native English speakers. This classic study provided new insights into the differences between spoken and written language. Using Biber's (1988) list, which consist of multiple linguistic features, the present study aims to provide a solution to the practical problems associated with learner corpus studies and demonstrate the effectiveness of combining a learner corpus, language processing techniques, and multivariate statistical analyses in learner language studies. In particular, the following research questions were investigated. First, which linguistic features do and do not vary among Japanese EFL learners at different oral proficiency levels and native English speakers? Second, is computer-aided analysis of multiple linguistic features useful for determining which ones characterize particular oral proficiency groups?

2 Methods

2.1 Corpus Data

The spoken data utilized in this analysis were extracted from the National Institute of Information and Communications Technology Japanese Learner English (NICT JLE) Corpus (Izumi et al. 2004) based on the interview protocol data elicited from a 15-minute oral proficiency test. The data set was made up of 325 hours of interviews conducted with 1,281 Japanese EFL learners. When the corpus was originally created, all the learners who provided data for this corpus were informed that their recorded utterances would be used for research, and only those who agreed to this condition took the test. The data that form the NICT JLE Corpus are publicly available, but the SST, the oral English examination that was especially constructed for assessing the oral proficiency of Japanese learners of English, is a commercial English test. There are nine oral proficiency levels in the SST, ranging from level 1 (novice low) to level 9 (advanced), and its evaluation criteria conform to the American Council on the Teaching of Foreign Language Oral Proficiency Interview (ACTFL OPI). Test takers are not provided with any planning time, and the use of reference material is not allowed. The test is divided into the following five stages: (a) warm-up questions for initial assessment, (b) a single picture description task and questions, (c) a role-play task and questions, (d) narrative task with a 4 or 6 sequence of picture, and (e) wind-down questions. In addition to the NICT JLE Corpus, a reference corpus was constructed (Izumi et al. 2004). The data was obtained from 20 native English speakers who performed the same speaking tasks as the learners. From oral proficiency level 4 onwards there is a drop in the number of learners. This is likely to contribute to a narrowing of the range of scores as proficiency increases and n-size decreases. However, one of the most striking distinguishing feature of the proficiency levels is the average length of the spoken text, that is how much learners speak and probably how fluently they speak, with native English speakers saying ten times more than level 3 learners in the same time, and three times as much as level 9 learners (Table 1). This is the reason for the gradually decreasing numbers of learners from level 7 in our data, ending up with only 20 native English speakers. Information about age, nationality, gender and educational level of these native English speakers is not open to the public; however, these data can be used to investigate which vocabulary or grammatical structures the native English speakers used frequently or infrequently in the interview, and they can be used as guidelines for examining how Japanese learners of English overuse, underuse, or avoid using specific linguistic features.

Table 1. Basic Information on Data Size

Level	3	4	5	6	7	8	9	NS	Total
Test takers	222	482	236	130	77	56	40	20	1,263
Tokens	95,352	308,544	204,048	130,678	85,395	68,539	54,394	84,774	1,031,724
Mean tokens	429.51	640.13	864.61	1005.22	1109.03	1223.91	1359.85	4238.70	816.88
Word types	7,987	13,514	10,752	8,836	6,763	6,095	5,207	7,692	-
index	25.87	24.33	23.80	24.44	23.14	23.28	22.33	26.42	-

Note. Level 1 and level 2 test takers were deleted from the data set because interviewers did most of the speaking during the interview.

2.2 Procedures

In the present study, 58 linguistic features were used from the original list of 67 linguistic features in Biber's (1988) study. The following seven items: (a) demonstratives, (b) gerunds, (c) present participial clauses, (d) past participial clauses, (e) present participial WHIZ deletion relatives, (f) sentence relatives, and (g) subordinator-that deletion were not included. This is mainly because of a difference in the software used to annotate part-of-speech tags compared with Biber's (1988) study, and programming problems in automatically extracting these linguistic features. Additionally, type/token ratio and word length were excluded because they involve lexical rather than grammatical information. The part-of-speech information was marked automatically with TreeTagger (Schmid, 1994). The raw frequencies of features were automatically counted using a computer programming language, Perl, which has often been employed for linguistic analysis (Hammond, 2003). The algorithm for the programming script was mostly adopted from Biber (1988), but as the original algorithms include some defects, as pointed out in de Mönnink, Brom and Oostdijk (2003), minor changes that can improve the precision and recall rate were added. The programming scripts were created based on the scripts developed by Murakami (2009). Afterwards, some search formulas were modified to fit the characteristics of learner language.

2.3 Statistical Analysis

Biber (1988) aimed to distinguish between linguistic variation in written and spoken processing modes multi-dimensionally, so he selected a list of linguistic features carefully based on previous research on spoken and written language use. Then he used exploratory factor analysis to identify common factors (i.e., dimensions) that could statistically explain variation between the two processing modes. However, the present study does not aim to interpret variation across different oral proficiency groups based on the linguistic function of each dimension. The aim is to specify linguistic features that can discriminate English oral proficiency levels and to typify proficiency levels. Thus, the linguistic features used in Biber's (1988) study were applied to the analysis of learner language in the present study, but factor analysis was not employed. Other multivariate statistical methods, such as correspondence analysis and cluster analysis, are more powerful tools for investigating similarity among variables (Oakes, 1998; McEnery & Hardie, 2012). First, correspondence analysis was used to reduce the complexity of the data by identifying oral proficiency groups based on various linguistic features. Cluster analysis was then employed to classify large groups into meaningful clusters according to their similarity.

3 Results and Discussion

The results of the clustering are displayed in the dendrogram in Figure 1, which was obtained from the coordination scale resulting from the correspondence analysis. The Ward method on Euclidean distances was used for clustering oral proficiency levels because it is efficient and has been widely used in various applications (Gries, 2009). This method minimizes the variance within and

between groups to maximize the sum of squares of clusters, resulting in the creation of smaller clusters. As shown in Figure 1, there are two main clusters (levels 3 + 4 and all the others). Other than levels 3 + 4, most of the proficiency groups are divided in two subclusters (level 5 + 6 and level 7 - native English speakers). The latter (level 7 - native English speakers) splits into two more subclusters (native English speakers and levels 7 - 9), which divide into two more groups (level 9 and levels 7 + 8). In interpreting the results of the hierarchical structure, the ordinate is supposed to express the heights of the dendrogram. The longer lines of the dendrogram indicate more autonomous clusters, and the shorter lines indicate more similar clusters. Cluster analysis is an exploratory statistical approach, and there is no optimal point for cutting the dendrogram. In general, the number of groups is determined by terminating the dendrogram at a certain height, which is shown in the y-axis, and this height shows the distance between clusters. Therefore, large changes in lines that indicate the distance between clusters were considered for the grouping. In the present study, as shown in a dashed line in Figure 1, the dendrogram was terminated at the height of 0.2.

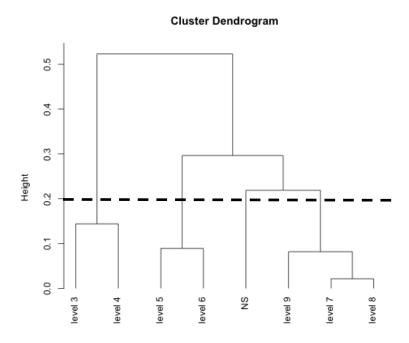


Figure 1. Dendrogram representing the results of the hierarchical cluster analysis using raw frequency data.

Accordingly, the 58 linguistic features from Biber's list are clustered into four large groups: Cluster I (levels 3 and 4), Cluster II (levels 5 and 6), Cluster III (native English speakers), and Cluster IV (levels 7, 8, and 9). Thus, we can conclude that the linguistic features analyzed by Biber (1988) can be used to distinguish the different oral proficiency groups, both between the native and non-native English speakers and also among the non-native speakers of different levels.

3.1 Falling and Rising Frequency Changes

Three linguistic features, (a) phrasal coordination, (b) WH questions, and (c) synthetic negation, displayed similar frequency change patterns (Figure 2). These three linguistic items were frequently used by level 3 learners, but as the learners' oral proficiency increased, the frequency decreased before slightly increasing again. The results showed that these features were most frequently used by level 3 learners and used in relatively high frequency by native English speakers.

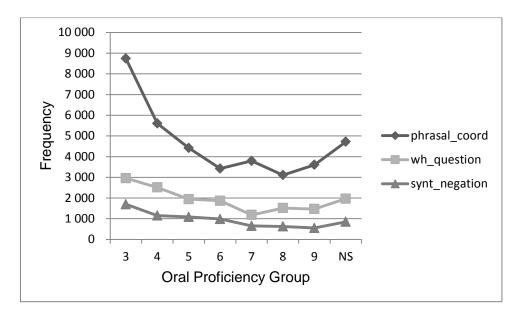


Figure 2. Linguistic features showing falling and rising changes in frequency. *Note*. The frequency counts are normalized to a text length of 1,000,000 words.

In addition, WH questions and synthetic negation were not used as frequently by all of the test takers as phrasal coordination was. Phrasal coordination was used by more than 90% of test takers in all groups. However, more than 60% of the language learners and 100% of the native English speakers used WH questions. In addition, synthetic negations were used by approximately 50% of the language learners and by 95% of the native English speakers. These results indicate that not all language learners use WH questions and synthetic negations as much as native English speakers do. Therefore, we suggest that phrasal coordination may be used by most test takers, whereas WH questions and synthetic negations are used at different rates across the different oral proficiency groups. However, some linguistic forms might reflect the ongoing process, not the end-point, of second language acquisition. For example, the use of synthetic negation (e.g., no answer is good enough for Jones) could occur as a fragment at earlier stages of acquisition and could be interpreted as a construction that appears at a stage prior to full-fledged analytic negation. Consequently, the frequency of synthetic negations needs careful interpretation.

In addition to these examples of frequency changes, other rising and falling frequency changes were also observed. The total frequencies of the linguistic features (a) past tense, (b) private verbs, (c) TO clause, and (d) analytic negation were not used very frequently or they did not display clear rising and falling

3.2 Flat Frequency Changes

The following linguistic features were observed in similar frequencies across the oral proficiency groups: (a) present tense, (b) first person pronouns, (c) attributive adjectives, (d) independent clause coordination, (e) third person pronouns, (f) BE as a main verb, (g) predicative adjectives, and (h) existential THERE (Figure 3). They can be categorized in a group that apparently does not distinguish oral proficiency groups.

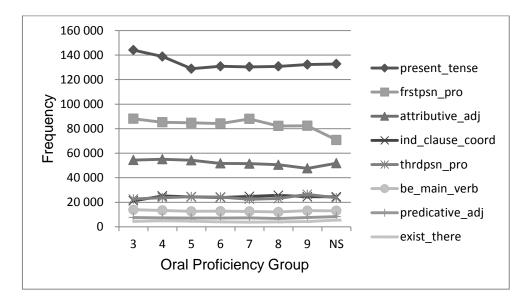


Figure 3. Linguistic features showing flat changes in frequency. *Note.* The frequency counts are normalized to a text length of 1,000,000 words.

The frequency counts of these linguistic features were higher than those of the linguistic features in other clusters; however, in general their frequency changes were insignificant across the oral proficiency groups. Most of the linguistic features were used by more than 90% of the test takers, except for the existential THERE. However, when we observed the range between the minimum and maximum frequency scores of each individual test taker, the differences appeared larger for the native English speakers than for the language learners. Thus, we can assume that native English speakers may have a tendency to use these linguistic features in a wider variety of ways than language learners do. In general, these linguistic features are necessary for constructing basic simple English structures. They can be used for describing pictures in one of the speaking tasks, such as '1st and 3rd personal pronouns + BE as main verbs + predicative adjectives' or 'existential THERE + BE as main verbs + a/an + attributive adjectives.' Thus, these linguistic features were evenly used by most test takers to produce simple basic English sentences, whereas the native English speakers used them in a wider variety of ways than the language learners did.

3.3 Rising Frequency Changes

All the linguistic features shown in Figure 4 were used at higher frequencies by the native English speakers. However, three features, pro-verb *do*, perfect aspect, and stranded prepositions, had low frequency counts compared with the others. A generally rising pattern was apparent for the three most frequently used features, contraction, pronoun *it*, and emphatics, which showed that their frequency increased as oral proficiency rose. Emphatics include items that are used to emphasize ideas (such as *a lot*, *for sure*, and *really*).

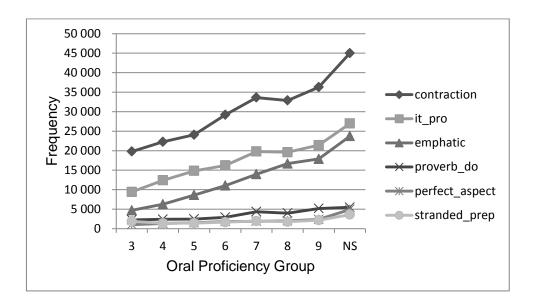


Figure 4. Linguistic features showing rising changes in frequency. *Note.* The frequency counts are normalized to a text length of 1,000,000 words.

Three of the linguistic features that exhibited lower frequencies (i.e., pro-verb *do*, perfect aspect, and stranded prepositions) were used by 100% of the native English speakers, but they were not widely used by the non-native English speakers. However, because the other three linguistic features (i.e., contraction, pronoun *it*, and emphatics) were used at higher frequencies by most of the test takers, we can assume that they reflect the development of learner language.

Another set of linguistic features, (a) demonstrative pronouns, (b) hedges, (c) split auxiliaries, (d) that as a relative subject, (e) that as a relative object, (f) concessive adverbial subordinators, and (g) split infinitives, also showed rising frequency changes. However, their use was not clearly associated with any particular oral proficiency groups. Other linguistic features, including relative pronouns and concessive adverbial subordinators, were used by a fairly high proportion of the native English speakers, but they were low in frequency. The total frequency counts increased as oral proficiency level rose, and advanced learners (level 9) as well as native English speakers used them much more frequently than those at the lower oral proficiency levels. However, the frequencies of relative pronoun usage may be low because that is often deleted in relative clauses. Thus, low frequency does not necessarily indicate a lack of development. Biber's (1988) list of linguistic features did not include the deletion of the relative pronoun.

In addition, another group of features showed rising frequency changes, but they probably could not be used for distinguishing oral proficiency groups because of their low frequency. Some linguistic features in Biber's (1988) study were not frequently used even by the native English speakers, which may present a serious problem for corpus-based studies. Low-frequency items are not well represented in the corpus unless the researcher obtained a huge amount of data, which could skew the results of the statistical analysis. However, even if some variables are underrepresented in the data, the correspondence analysis clustered them into a group. The low frequency features were as follows: (a) conditional adverbial subordinators, (b) other adverbial subordinators, (c) that verb complements, (d) WH relatives in subject position, (e) WH clauses, (f) seem and appear, (g) that adjective complements, and (h) WH relatives with fronted position. The total frequency counts of these linguistic features were fairly low, but most of them increased as the oral proficiency level rose.

The linguistic features related to subordination were (a) conditional adverbial subordinators (e.g., If we don't do anything, nothing will change) and (b) other adverbial subordinators (e.g., you need to provide ways so that people can use money). The frequencies of these different types of adverbial subordinators increased as the oral proficiency level rose. It can be concluded that as the learners' oral proficiency became more developed, they began to state more reasons in order to support their speech, or they learned to propose a conditional sentence to qualify their discussion more precisely. For instance, Norris and Ortega (2009: 563) suggested the use of different linguistic indices that can be expected to distinguish learners at different proficiency levels. They proposed the possibility that coordination can be used as an index for novice learners, subordination for intermediate learners, and complexity via phrasal elaboration (e.g., grammatical metaphor) for advanced learners. Although Norris and Ortega refer to language use in general, we assume that this is true also for oral language. However, because the categories of linguistic features in the present study are wide-ranging, other functions of adverbial subordinators, such as (a) temporal adverbial clauses (e.g., After she went back to work...) and (b) adverbial clauses that have the meaning of purpose (e.g., They stopped working to have a little rest), could be added in future studies for a more detailed analysis as suggested by Biber and Conrad (2009).

Another point regarding grammatical complexity concerns nominal postmodifiers, such as (a) WH relatives in subject position (e.g., a boy who is listening to music), (b) WH relatives in object position (e.g., the person who I don't know well), (c) past participial post nominal (reduced relative) clauses (e.g., lodge made by wood), and (d) WH relatives with fronted preposition (e.g., They have more room in which to construct the house). Among these linguistic features, the frequency of WH relatives in the subject position increased until learners reach level 9, but the frequencies of the other linguistic features did not indicate clear frequency change patterns. A previous study that used the written compositions of Japanese EFL learners' presented interesting results. Negishi (2012) used a checklist of CEFR criterial features to evaluate the written compositions of Japanese EFL learners. The checklist was drawn from the English Profile Program with modifications that were necessary for analyzing the compositions of Japanese EFL learners. He concluded that 42 criterial features effectively discriminated between A2 to B2+ level learners, and the following features were reported to be particularly effective: (a) relatives, (b) past/present participial post nominal clauses, (c) participial constructions, and (d) modal auxiliaries.

However, in the present study, WH relatives in the subject position were the only linguistic features that might have discriminated among oral proficiency levels. Relative pronouns can be considered difficult linguistic items for Japanese EFL learners, but interestingly, the native English speakers were not the most frequent users of WH relatives in the subject position.

3.4 Falling Frequency Changes

The frequency change patterns of nouns are different from the other parts of speech. First, the results showed that the total frequency of nouns was much higher in comparison with the others. Although the total frequency decreased as the proficiency level rose, nouns were the most frequently used linguistic items. On the other hand, the frequencies of two other function-word classes, prepositions and adverbs, increased as the proficiency level increased. However, time adverbs exhibited a falling frequency. A comparison of the different frequency changes of place adverbs showed some frequency change patterns. Both time and place adverbs were frequently used by most test takers in all proficiency groups; however, place adverbs were used much more frequently by the native English speakers, whereas time adverbs were used much more frequently by the non-native speakers of English.

3.5 Determining Linguistic Features

The second research question asked whether the corpus-based multidimensional approach is useful for determining linguistic features that characterize particular oral proficiency groups. The learner corpus in general contains short and erroneous sentences in limited variations of lexical and grammatical use. However, the findings of this study indicate that the response to this question is positive. In addition, we were able to conclude that the linguistic features listed by Biber (1988) could be used to distinguish Japanese EFL learners' oral proficiency levels and differences in linguistic features used by native and nonnative speakers of English.

4 Conclusion

By using correspondence analysis and cluster analysis to analyze the data matrix of 58 linguistic features across different oral proficiency groups, we were able to detect a close relationship between these features and the oral proficiency groups. The study revealed that some linguistic features (Biber, 1988) were frequently used by a particular group, so they may be used as markers to characterize different levels of oral proficiency. For example, phrasal coordination can be used to characterize the spoken performance of Japanese EFL learners at a low level. Moreover, the following linguistic features were used at similar frequencies across the oral proficiency groups: (a) present tense, (b) first person pronouns, (c) attributive adjectives, (d) independent clause coordination, (e) third person pronouns, (f) BE as a main verb, (g) predicative adjectives, and (h) existential THERE. Therefore, they can be categorized as a group of features that apparently does not distinguish oral proficiency groups. In addition to these findings, a generally rising pattern was apparent in three of the linguistic features: (a) contraction, (b) pronoun it, and (c) emphatics. These features were used by most of the test takers, and their frequency clearly increased as English oral proficiency increased. Thus, we can assume that they reflect the development of learner language. The results also showed that the frequency change patterns of nouns differed from those of other parts of speech. The total frequency of nouns was much higher in comparison with the frequency of other items, and it decreased as the proficiency level increased.

In summary, this study found interesting rising, flat, and falling frequency patterns in how several linguistic features are used in different oral proficiency levels. In other words, it identified a set of linguistic features that differentiate among second language oral proficiency groups as well as between non-native and native speakers of English. These features could thus be used to analyze how learner language changes across oral proficiency groups.

References

- Asención-Delaney, Y., & Collentine, J. 2011. A multidimensional analysis of a written L2 Spanish corpus. *Applied Linguistics*, 32 (3), 299-322.
- Biber, D. 1988. Variation across speech and writing. Cambridge: Cambridge University Press.
- Biber, D., & Conrad, S. 2009. Register, genre, style. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S., & Reppen, R. 1998. Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.
- Hammond, M. 2003. Programming for linguists: Perl for language researchers. Oxford: Blackwell.
- de Mönnink, I., Brom, N., & Oostdijk, N. (2003). Using the MF/MD method for automatic text classification. In S. Granger & S. Petch-Tyson, S. (Eds.), Extending the scope of corpus-based research: New applications, new challenges. Amsterdam: Rodopi, 15-
- Izumi, E., Uchimoto, K., & Isahara, H. 2004. 日本人 1200 人の英語スピーキングコーパス [A speaking corpus of 1200 Japanese learners of English]. Tokyo: ALC Press.
- Gries, S. T. 2009. Statistics for linguistics with R: A practical introduction. Berlin: De Gruyter Mouton.
- Lu, X. 2010. What can corpus software reveal about language development? In A. O'Keeffe & M. McCarthy (Eds.), The Routledge handbook of corpus linguistics. Abingdon, England: Routledge, 184-193.
- McEnery, T., & Hardie, A. 2012. Corpus linguistics: Method, theory and practice. Cambridge: Cambridge University Press.
- Murakami, A. 2009. A corpus-based study of English textbooks in Japan and Asian countries: Multidimensional approach. Unpublished master's thesis. Tokyo: Tokyo University of Foreign Studies.
- Myles, F. 2005. Interlanguage corpora and second language acquisition research. Second Language Research, 21 (4), 373-391.
- Negishi, M. 2012. Exploring the possibility of assessing English writing based on the checklist of CEFR criterial features. ARCLE REVIEW, 6, 90-99.
- Norris, J., & Ortega, L. 2009. Organic approach to investigating CAF in instructed SLA. Applied Linguistics, 30, 555-578.
- Oakes, M. 1998. Statistics for corpus linguistics. Edinburgh: Edinburgh University Press.
- Pendar, N., & Chapelle, C. 2008. Investigating the promise of learner corpora: Methodological issues. CALICO Journal, 25 (2), 189-206.
- Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. Revised version of a paper presented at the International Conference on New Methods in Language Processing, Manchester, England.
- Van Rooy, B., & Terblanche, L. 2009. A multi-dimensional analysis of a learner corpus. In A. Renouf & A. Kehoe (Eds.), Corpus linguistics: Refinements and reassessments. Amsterdam: Routledge, 239-254.