報告番号	※　　　　　第　　　　号

主論文の要旨

論文題目	A Data-driven Approach to the Development of Constructions in Japanese EFL Learners (日本人英語学習者のコンストラクションの発達へのデータ駆動型アプローチ)
氏　　名	阿部大輔

論文内容の要旨

The present study conducted analyses of Japanese learners of English as a Foreign Language, attempting to uncover developmental patterns in the acquisition of constructions, or linguistic units comprising multiple parts.

Complexity in writing, an aspect of proficiency, has traditionally been measured through measures based on clauses and T-units, such as clauses per T-unit, clauses per sentence, dependent clause per clause, and other similar measures. However, Biber, Gray, and Poonpon's (2011) work criticized this practice, claiming a lack of evidence that clausal measures and T-units captured proficiency in academic writing. Their argument was that clausal subordination is a feature more common in conversation rather than academic writing, and that the evaluation of academic writing should be performed through phrasal measures. Biber et al. confirmed their claim through a consultation of corpus data, and went on to suggest a developmental order for grammatical structures in learner writing. However, their data came from corpora of native speakers, and they have been criticized for making claims about learner development through the consultation of only native data (Yang, 2013). This inspired a series of studies (Parkinson and Musgrave, 2014; Ansarifar, Pishgadam, and Shahriari, 2018; and Kreyer and Schaub, 2018) trying to provide empirical evidence from learner data to confirm Biber et al.'s claims.

Although these studies found some evidence of Biber et al.'s suggested developmental stages, only part of their hypothesis was confirmed in each study. In addition, even though Ansarifar et al. suggested that "a re-consideration of features among these stages is advised," none of these studies actually suggested a new list of features and stages. A limitation of these studies is that because all of them started out with a predetermined list of features to examine, they did not investigate whether there were any other features that should be included in the list. Furthermore, all of these studies took a time-consuming manual approach in order to extract the

features of interest from their data, making it impractical to use a very large dataset.

The present study attempts to address these issues by using natural language processing tools to automatically extract structures from learner data. The present study first replicates the previous studies in order to compare the results against different learner data. The investigation of predetermined features is referred to as the top-down approach. This study also conducts an investigation through a bottom-up approach, in which all frequently occurring syntactic structures in learner writing are extracted, with the possibility of finding linguistic features not taken into account in the previous studies. Although the previous studies focused on noun phrases, the present study covers sentences, noun phrases, and verb phrases.

The present study uses part of the ETS corpus of Non-native Written English, essays written as part of the TOEFL test, as its learner data. The subset written by Japanese learners of English is used, and three subcorpora of equal word counts are created from this data. These essays are scored on a scale of 1 to 5, and the resulting score is recorded in the corpus as low, medium, or high. The present study investigates the differences between these three score groups in order to find patterns of structures appearing in certain stages of development.

Through a combination of the Stanford Parser, Disco-DOP, and Tregex, the present study adds syntactic tags to learner data and extracts recurring structures from the corpus. The structures are represented in a form called tree fragments, which are parts of sentences represented in tree form. The result is frequency tables of all structures occurring in the data. The present study also examines the frequencies of specific expressions occurring in the corpus.

In order to gain an understanding of the overall distributional patterns of tree fragments, five properties are calculated for each tree fragment: number of nodes, number of terminal nodes, maximum depth, number of lexical nodes, and number of nonlexical nodes. The number of nodes indicates the overall size of the tree; a node could be a syntactic category such as a sentence or noun phrase, a part-of-speech like noun or verb, or a lexical item like *cat* or *is*. The number of terminal nodes indicates the breadth of the tree, or all nodes with no child nodes under them. The maximum depth indicates the deepest part of the syntactic structure, or the distance from the topmost node to the bottommost node. The number of lexical nodes is the number of nodes that are specific lexical items, and the number of nonlexical nodes is the number of nodes that are not lexical items. These measures indicate the overall complexity of a tree fragment.

Through the investigation of the distributional patterns of tree fragments based on the above-mentioned properties, it was discovered that although the lowest group always produced the greatest number of tree fragments regardless of the tree fragment property in sentences and verb phrases, in noun phrases, the frequency ranking of tree fragments reversed depending on the complexity of the tree. For example, the low group produced the greatest number of fragments with 2 and 3 nodes, but for fragments with 4 or more nodes, the high group produced the most fragments. The same pattern could be seen in terminal nodes, depth, and nonlexical nodes, reinforcing Biber et al.'s assertion that complexity in academic writing should not be measured by clausal measures but by phrasal measures.

The investigation of structures revealed that many of the features Biber et al. placed in the higher stages were actually produced at about the same frequency by writers of all score groups, suggesting that these structures were acquired at an earlier stage than assumed by Biber et al. In addition, two structures not included in Biber et al. were found to display a difference between the score groups. Linking verbs, or the copular construction, were found to be frequent among the low score group, indicating that it was acquired at an early stage. On the other hand, noun phrases with determiners increased with score level, indicating that this is a feature that develops late in the acquisition process. The present study compiles these findings and compares them against Biber et al.'s hypothesized developmental stages. A version of developmental stages for Japanese learners of English, supported by empirical evidence from learner data, is presented.

A secondary contribution of the present study is that it demonstrates the effectiveness of automatic parsing and extraction in the investigation of constructions in a corpus. This method can be applied in future research to vastly increase the size of data to be investigated, as well as making the bottom-up approach introduced in this study possible.