論文の内容の要旨

論文題目　　Automatic Question Generation for Language Testing
and its Evaluation Criteria
(語学テストのための自動問題生成とその評価基準)

氏名　　星野　綾子

AQG (Automatic Question Generation) is an emerging research topic, which is gaining attention with the trends of ever increasing demand for specialized educational materials on one side, and the abundant text resources on the Internet on the other side.

In this study, we propose a method for automatically generating Multiple-Choice Fill-In-the-Blank (MC-FIB) questions for English grammar and vocabulary. Under the light of recent advances in test theories (Psychometrics) and NLP (Natural Language Processing), some attempts to AQG for language testing have already been made. However, the evaluation method varies from research to research, making it difficult to compare the effectiveness of different methods. Evaluating and comparing an AQG system is not easy, because the problem is not as trivial as maximizing one index value. The traditionally-used ones are quality (or errorlessness) judged by human experts and discrimination indices measured from the result with student groups. We argue that, in

addition to quality and discrimination, construct validity, and appropriate difficulty should be taken into account.

We have sequently developed three AQG methods and conducted evaluations in terms of the proposed viewpoints. First, we have built a semi-automatic question generator that allows the test author to compose questions just with some clicks on the screen. Secondly, we evaluate randomly-generated questions with a group of students. In terms of discrimination power, our AQG method for grammar questions is as efficient as workbook questions. Evaluation on construct validity shows some evidence that the pattern-generated questions measure intended grammar targets. Finally, we present a CAT (Computer Adaptive Testing) system that administrates automatically generated questions. We have developed a difficulty predictor using machine learning techniques, which can be used for newly generated questions. Evaluation on difficulty adaptivity shows that the predicted difficulty value has more information on the actual correct response rate than the baseline index (sentence length) alone.

Keywords: Language testing, Grammar and vocabulary test, Multiple-Choice Fill-In-the-Blank (MC-FIB), IRT (Item Response Theory), NLP (Natural Language Processing), Machine learning