Identification of Linguistic Features for Classifying L2 Proficiency Levels Using Beginning-level L2 Learner Corpora and Machine Learning Techniques

Yukio TONO
Graduate School of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan
y.tono@tufs.ac.jp

One of the interesting issues in learner corpus research is how to identify features which serve as "criteria" for particular L2 proficiency levels. The term "criterial feature" has been used by the researchers working on the Cambridge Learner Corpus as part of the English Profile Programme (Hawkins and Battery, 2010). Whilst some scholars are skeptical about such profiling research (cf. Hulstijn, 2010), it will surely provide the interesting possibilities of using a large amount of learner output data for extracting features that help classify the learners from various performance perspectives.

This thread of research is closely linked to the application of the Common European Framework of Reference for Languages (CEFR) in learner corpus research. It has been a serious problem that a comparison across different learner corpora is difficult, due to the fact that they do not always share the same corpus design criteria regarding learner proficiency levels. Sometimes classifications were made based upon external criteria such as school grades or age, but sometimes on external exams such as TOEIC or TOEFL, which would make it difficult to compare against the samples without those scores. Therefore, the use of CEFR for classifying the learner corpus data into generic proficiency levels will help make a cross-comparison between different learner corpora and share the results.

The problem here is that there are no specific linguistic features available yet, which have been found to be useful for classifying CEFR levels satisfactorily. In the past few years, various linguistic criteria have been proposed as "criterial", but they need to be refined in such a way that each proposed criterial feature should be evaluated and weighed in terms of usefulness as CEFR-level "classifiers". Then a bundle of criterial features have to be tested and validated to find out which combinations of criterial features work best to predict the CEFR-levels.

In this study, previous criterial features will be briefly reviewed and a bundle of features will be selected for evaluation. Using the samples rated for CEFR-levels from the International Corpus of Crosslinguistic Interlanguage (ICCI) data (Tono, Kawaguchi & Minegishi, 2012), those linguistic features under study were first automatically extracted from the training corpora, and then several different machine-learning algorithms, such as decision tree, support vector machine, and random forest, will be tested to see how classifications are made and which linguistic features are used as predictor variables for the classification, and which learning algorithms works best.

Finally, a set of selected features will be tested over the new test data taken from the ICCI in order to evaluate the performance of those criterial features for classifying essays

according to the CEFR levels. The study has some important implications. It will be useful for developing a system of automatic essay evaluations similar to e-rator by ETS (Monaghan & Beidgeman, 2005), using different heuristics. This will also lead to a new area of L2 learner profiling research, in the sense that it provides more holistic pictures of how acquisition will take place in terms of criterial features and how the findings from such studies will be interpreted in the mainstream of SLA research.

References

Hawkins, J. A. & Buttery, P. (2010). 'Criterial features in learner corpora: theory and illustrations.' *English Profile Journal* 1 (1). DOI: 10.1017/S2041536210000103.

Hulstijn, J. (2010). 'Linking L2 proficiency to L2 acquisition: Opportunities and challenges of profiling research.' In: Bartning, I., Martin, M., & Vedder, I. (eds). *Communicative Proficiency and Linguistic Development: intersections between SLA and language testing research*. European Second Language Acquisition Monograph Series 1: 233-238.

Monaghan, W. & Bridgeman, B. (2005). 'E-rater as a quality control on human scores.' In: *ETS R & D Connections*: 1-4. Available from: http://www.ets.org/Media/Research/pdf/RD_ Connections2.pdf. [31/01/2012].

Tono, Y., Kawaguchi, Y. & Minegishi, M. (eds). (2012). *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: John Benjamins.