E. Arshavskaya

AUTOMATIC PROFILING OF LEARNER CORPORA

This study undertakes a comparison of non-native speakers' (NNS) written language to the academic language of English native speakers (NS). In previous research[1], it was found that EFL[2] learners of French background overuse items of spoken register and underuse items of academic vocabulary in their academic written samples (International Corpus of Learner English (ICLE) database). This study was designed to find out whether ESL and EFL students of various L1 backgrounds also lack knowledge of formal academic vocabulary and instead opt for informal doublets (e.g., *in spite of* vs. *despite*, *till* vs. *until*). In case the study re-confirms the findings of S. Granger and P. Rayson (1998), we can conclude that lack of acquaintance with academic register and prevalence of spoken language items in written samples are characteristic of upper-level English learners regardless of their L1 and constitute a stage in learning L2.

In the first part of this study, MELD (Montclair Electronic Language Database, ESL corpus)[3] and BAWE (corpus of British

[1] *Granger S., Rayson P.* Automatic profiling of learner texts // S. Granger (ed.). Learner English on Computer. New York: Longman, 1998. P. 119–131.

[2] EFL (English as a foreign language) stands for learning and teaching English in countries (e.g., Japan, Russia) where English is not a major language of commerce and education and which students do not usually hear outside their classrooms. ESL (English as a second language) stands for learning and teaching English in countries (e.g., the US, the UK, India) where English is a major language of education and commerce and which students often hear outside their classrooms. – *Brown D.* Teaching by Principles: an interactive approach to language pedagogy. White Plains, NY: Longman, 2001. P. 3.

[3] *Fitzpatrick E., Seegmiller M.S.* Montclair Electronic Language Database (MELD).

Academic Written English)[1] were tagged with the Tree-tagger[2] and the Penn Treebank tagset[3]. Then, part-of-speech (POS) profiles (frequency lists) in the NS (BAWE database) and NNS corpora (MELD) were compared. The differences in POS's use in the NS and the NNS corpora were found to be statistically significant (one sample *t*-test). The comparison of the use of the POS in the two corpora showed that upper-level ESL students predominantly use items of spoken language and rarely make use of academic vocabulary. For example, ESL students underused nouns, whereas English academic texts favor a predominance of nouns, and overused personal pronouns, which are disfavored by English academic register and are characteristic of spoken language. The first study also showed that learners overused closed classes of words (e.g., particles, auxiliaries) and underused open classes of words (e.g., nouns, adjectives). This may signify that upper-level ESL learners' vocabulary is limited (repetitions of words; use of function words instead of content words). The type-token ratio in MELD was low (0,045), too. These data also support the claim for the poverty and lack of variety in the ESL learners' vocabulary.

In the second part of this study, POS profiles of more advanced learners (BAWE ESL corpus) and the same NS corpus (BAWE) were compared. In this case, the POS's had a similar distribution across the two corpora, except for the nouns (which were overused by the EFL learners). The differences in the usage of the POS's in the two corpora proved to be significant (X^2 test). Based on the analysis of the data of the POS frequencies in the NNS (BAWE NNS) and the NS (BAWE

[1] *Nesi H., Gardner S., Thompson P., Wickens P.* The British Academic Written English (BAWE) corpus.

[2] *Schmid H.* Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, September 1994. P. 44–49.

[3] *Marcus M., Santorini B., Marcinkiewicz M.* Building a Large Annotated Corpus of English: The Penn Treebank // Computational Linguistics (Special Issue on Using Large Corpora). 19 (2). 1993. P. 313–330.

NS) corpora, we conclude that writing skills of highly proficient NNS students approach those of the NS's (longer exposure to the target language, higher initial L2 proficiency, etc.).

Thus, the first study re-confirms the speech-like nature of learner writing of upper-level ESL students[1] of different linguistic backgrounds. Since the learners whose writing samples were analyzed come from a number of different backgrounds, this finding (i.e., the speech-like nature of L2 learners' writing) cannot be attributed to L1 transfer. Upper-level ESL/EFL learners of various L1-s lack knowledge and acquaintance with academic vocabulary. It has been shown that general and informal register vocabulary is preferred by the L2 learners to the more abstract and formal registers of lexicon. The low type-token ratio (0,045) also points to the lack of variety in the learners' lexicon. As has been suggested earlier by S. Granger and P. Rayson (1998), these findings have a great potential for the future design of ELT materials.

The study by B. Sardinha and M. Shimazumi (2003)[2] demonstrated that 15-year-old native speakers of British English exhibit similar performance (to the NNS L2 learners) in their written assignments (overuse of spoken and general vocabulary; underuse of academic vocabulary). Academic writing in English is a skill which is teaching-induced and is equally found in novice native and non-native writers[3]. S. Granger and P. Rayson (1998) suggest that exposure to formal writing (e.g., to the editorials of quality newspapers) can serve as a possible remedy. As a result, ESL/EFL learners improve their writing skills and may match those of native speakers (the second part of this study).

Among the possible explanations for the lack of acquaintance of the NNS students with academic style of writing, there are:

[1] *Granger S., Rayson P.* Automatic profiling of learner texts…

[2] *Sardinha B., Shimazumi M.* Schoolchildren writing: a corpus-based analysis. Linguagem & Ensino. Vol. 6 (1). 2003. P. 11–33.

[3] *Granger S., Rayson P.* Automatic profiling of learner texts…

1) predominance of communicative ELT approach which places more emphasis on speech rather than on writing; 2) lack of exposure to good quality academic writing by ESL and especially EFL students.[1]

As a suggestion for further research, it would be interesting to know what L1s contributed to the specific L2 learners' tendencies to under- and/or overuse certain POS's (Both of the NNS corpora (MELD and BAWE NNS) contain essays of students who come from various L1's). This would allow us to see what errors (i.e., the NNS's patterns of under- and overuse of certain POS's, when compared to the NS's) were caused by L1 interference and what errors (i.e., patterns of under- and overuse) could be considered universal.

Also, it would be interesting to analyze the distribution of lemmas across the NNS and the NS corpora. In the current study, most of the analysis concerned the usage of word forms in the NS and the NNS written language. On the other hand, the analysis of lemmas would allow us to see whether the L2 learners' lexicon has or, on the opposite, lacks variety. For example, L2 learners may correctly and frequently use word forms of a certain lexical item (i.e., show acquaintance with inflectional morphology), which would result in a high frequency of a certain POS. However, the type-token ratio for this particular POS may be low. This would signal to the lack of variety in the L2 upper-level learners' lexicon.

Lastly, the reconstructed (for errors) corpus (e.g., MELD) can be used for further analyses of the L2 learners' written interlanguage. The reconstructions of learners' errors can help in making the distinction between word choice and errors clear (i.e., errors will be reconstructed, while word choices will be not). Additional annotation of the NS and the NNS corpora (e.g., syntactic parsing; discourse parsing) will allow us to compare L2 learners' use of specific syntactic constructions (e.g., passive vs. active voice) and of discourse markers to the NS's use of the same features and constructions.

[1] *Granger S., Rayson P.* Automatic profiling of learner texts…