

Automatically Annotated Learner Corpus

*Yu-Yin Hsu, **Charese Smiley, ***Kirsten Todt

*hsuy@indiana.edu, **chsmiley@indiana.edu, ***krtodt@indiana.edu

Linguistics Department at Indiana University Bloomington

With the evolution of computational linguistics, computer-assisted language learning (CALL) has been incorporated into the field of second language teaching, in an attempt to improve and or provide diverse learning environments. Based on pedagogical and computational linguistic purposes, in this project, we focus on online systems for second language writing. We propose to develop a system of data collection that will automatically collect and annotate learner data and provide an extractable database for different research purposes in linguistics.

Current systems in use are designed specifically for curriculum activities. On the other hand, there are also systems that allow online language exchange in an environment of social network communication, such as *Lang-8*. We find that online systems for curriculum fit pedagogical motivations better than a social network, like *Lang-8*, because the design aims at improving specific language skills at the target stage of acquisition, and the quality of feedback is controlled to be adequate to L2 learners. From a computational point of view, to automatically analyze learner language, we need to gather large corpora of learner data that are stored in a way that can help further research. However, feedback and data used in such online curriculums may not be reusable, which is the same problem in *Lang-8*. Repurposing existing systems for research use is complicated by the fact that correction of learner data is not done in a systematic way that would allow an error scheme to be automatically applied. Also, the data structures behind the site are both inaccessible and unsuitable for corpus study.

In this project, aiming specifically at L2 writers, we propose to develop a system of data collection that will 1) cut back on the man-hours needed to collect and annotate learner data, 2) provide an extractable database for different research purposes in linguistics, and 3) offer second language learners a non-threatening environment to receive feedback on their L2 writing. We will focus on issues of copyright and IRB requirements, standardizing error annotation, and the development of user-interface. In accordance with university and research community standards, we will seek to IRB approval of the project with proper implementation in the website. In addition, we propose to design an annotation scheme to be used on the learner data both by users of the site and by the corpus administrators. The corpus will contain both the original L2 texts and associated files produced by other users containing suggested corrections and annotations. An evaluation will be provided for the development of the system and also for the design of pedagogical purposes.

References:

ICLE corpus

Lang8 <http://lang-8.com/>

Language Exchange <http://jones.ling.indiana.edu/~chsmiley>

Moin Moin <http://moinmoin.wikiwikiweb.de>

PmWiki <http://www.pmwiki.org/>

Study on Japanese learners' English writing <http://www.eng.ritsumei.ac.jp/asao/lcorpus/>

Chapelle, Carol A. & Joan Jamieson. 2008. Tips for teaching with CALL: Practical approaches to computer-assisted language learning. White Plains: Pearson Education.

Flege, James Emil, Ocke-Schwen Bohn & Sunyoung Jang. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25.437-70.

Granger, Sylviane. 2003. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal* 20:3. 465-480.

Lightbown, Patsy M. & Nina Spada. 1999. *How Languages are Learned*. Oxford: Oxford University Press.

Levey, Sandra. 2004. Discrimination and Production of English Vowels by Bilingual speakers of Spanish and English. *Perceptual & Motor Skills* 99.445-62.

Suri, Linda & Kathleen F. McCoy. 1993. A Methodology for Developing an Error Taxonomy for a Computer Assisted Language Learning Tool for Second Language Learners. Technical Report 93-16, Department of Computer and Information Sciences, University of Delaware, Newark, DE.