Apples – Journal of Applied Language Studies
Vol. 8, 3, 2014, 1-4

Special issue on

Learner Language, Learner Corpora: From corpus compilation to data analysis

Guest editors:

Jarmo Harri Jantunen, University of Jyväskylä;
Sisko Brunni, University of Oulu &
Marianne Spoelman, University of Oulu

Contents

Corresponding editor's email: jarmo.h.jantunen@jyu.fi
ISSN: 1457-9863
Publisher: Centre for Applied Language Studies, University of Jyväskylä

http://apples.jyu.fi

Editorial

Jarmo Harri Jantunen, University of Jyväskylä; Sisko Brunni, University of Oulu & Marianne Spoelman, University of Oulu

The articles in the present special issue of *Apples* are based on papers presented at the international conference *Learner Language, Learner corpora* (LLLC2012) held in Oulu October 5–6, 2012. The aim of the conference was to bring together researchers interested in second language acquisition and foreign language learning; particularly those working with either or both learner corpora and the learning and teaching of Finno-Ugric (FU) languages as a second or foreign language.

About 80 conference participants came from fifteen different countries, making it a truly international event. Plenary talks were given by Charlotte Gooskens (University of Groningen), Fanny Meunier (Université Catholique de Louvain), and Minna Suni (University of Jyväskylä). The programme included five parallel sessions and about 60 papers and poster presentations. The conference was organized to celebrate the 15th anniversary of the VIRSU project and to introduce the ICLFI project. VIRSU (viro and suomi i.e. Estonian and Finnish as target languages) seeks to create connections between all linguists working with any of the Finno-Ugric (foreign or second) target languages or with the language situations in the countries where Finno-Ugric languages are spoken.

The ICLFI (*The International Corpus of Learner Finnish*) project, in turn, has focused on compiling a multi-mother tongue corpus of Finnish as a foreign language, which represents several CEFR levels. The articles based on non-corpus-related studies represented at LLLC2012 and which address the study on Finno-Ugric languages as a target language, were published in the special issue of *Lähivõrdlusi. Lähivertailuja* 23.

The articles in this issue focus on Learner Corpus Research (LCR). LCR is a relatively new and active field that combines theories and methodologies from second language acquisition, foreign language learning and corpus linguistics. In its first stage, LCR was limited to studies on learner English; however, the spectrum of languages has widened rapidly, which is reflected in the articles found in this issue. Today, the range of learner languages that have been studied is broad; this issue contains articles on the learning of Finnish, Hungarian, German, Swedish and English. However, the articles in this issue do not only report results from pure data analysis but they also discuss corpus compilation, methodology, tools and annotation in cases where the data involve morphologically rich languages, such as Finnish and Hungarian.

The first article in this issue is authored by *Aivars Glaznieks, Lionel Nicolas, Egon Stemle, Andrea Abel* and *Verena Lyding*, who provide an overview of a generic workflow to build written learner corpora for the specific needs of linguists. It introduces how the workflow results from an extensive collaboration between linguists who annotate and use the corpus, and computer linguists who are responsible for providing technical support. The paper addresses linguists' research needs as well as aspects of availability and usability of language technology tools. In addition, it illustrates the relevance of

the suggested workflow using the example of the German "KoKo"corpus (written German as L1 learner language, compiled in South Tyrol/Italy).

The second article of this issue, written by *Ilmari Ivaska*, also concentrates on corpus compilation but has an additional focus on annotation schemes. It introduces the Corpus of Advanced Learner Finnish (LAS2), which has been compiled at the University of Turku since 2007. The article deals with the LAS2 corpus in detail and discusses issues such as its compilation criteria, structure and annotation procedure. In addition, a set of query tools designed within the framework of the corpus project is introduced, and it is explained how these tools can be applied.

Automatic annotation and especially error-tagging of learner data is further discussed in the article by *Péter Durst, Martina Katalin Szabó, Veronika Vincze* and *János Zsibrita*, but from the perspective of Hungarian as a target language. The authors report on the development of automatic tools for the processing and error-tagging of Hungarian learner language. These tools make use of annotation schemes developed on the basis of the particular characteristics of learner language and of Hungarian morphology. Due to its focus on a morphologically rich language, the article provides particularly valuable insights into the complex principles that underlie automatic morphological processing and error-tagging.

The next three articles—by *Marianne Spoelman, Outi Toropainen and Sinikka Lahtinen, and Mariko Abe*—address learner production. Marianne Spoelman's article investigates the use of partitive plural predicatives in texts written by Estonian, German and Dutch learners of Finnish. The research materials were taken from the Estonian, German and Dutch subcorpora of the ICLFI. Both the similarities and differences between the learning of a closely related and non-related target language are addressed in the light of the use of prior linguistic knowledge as a strategy for facilitating foreign language learning.

Both Finnish and Swedish learner languages are in focus in *Outi Toropainen's* and *Sinikka Lahtinen's* article. Here three types of interrogative clauses (*yes/no* questions, *wh*-questions and subordinate interrogative clauses) in semi-formal e-mail messages written by learners of Finnish and of Swedish spoken in Finland are investigated. The frequencies of occurrence of these interrogative clauses were examined across the proficiency levels of the Common European Framework of Reference for languages (CEFR). The study was conducted within the framework of the Topling project, a large research project which aims to provide a linguistic basis for the CEFR.

Finally, the article by *Mariko Abe* analyzes the overall patterns of variation of Japanese learners of English across seven oral proficiency levels. The research materials were selected from the *National Institute of Information and Communications Technology Japanese Learner English* corpus, which is the largest spoken learner corpus in Japan. The study seeks to identify linguistic features on the basis of which oral proficiency groups as well as native and non-native speakers can be distinguished.

We would like to thank all those who contributed to this special issue: the authors, reviewers and *Apples'* editorial staff. We believe that this issue reflects the wide variety of topics and perspectives raised at our conference and that it constitutes a valuable collection of articles, covering different aspects of learner corpus research from corpus compilation to data analysis, with special reference to Finno-Ugric languages.