Lang Resources & Evaluation (2014) 48:1–3
DOI 10.1007/s10579-014-9264-5

SI: RESOURCES FOR LANGUAGE LEARNING

Introduction to the special issue on Resources and Tools for Language Learners

Serge Sharoff · Stefania Spina · Sofie Johansson Kokkinakis

Published online: 20 February 2014

This special issue of *Language Resources and Evaluation* is devoted to Resources and Tools for Language Learners.

The use of language resources in the teaching and learning of a foreign language goes back to the middle of the last century: the *General Service List* published by Michael West in 1953 is one of the first examples of the usefulness of corpus information for language learning. In subsequent decades, corpora and corpus evidence became increasingly exploited for teaching and learning of foreign languages, including annotated frequency lists, corpus-based reference works such as dictionaries, concordances, and grammars, and tools for mining corpora (e.g., procedures for the extraction and analysis of collocations) for linguistic data that can be used in the classroom, CALL programs, and, in more recent years, web-based environments. As a result, a strong link has been forged "between the two previously disparate fields of corpus linguistics and foreign/second language research" (Granger 2002: 4).

In addition to methods and results from corpus linguistics, language learning research has benefited in recent years from the availability of Natural Language Processing (NLP) tools, such as part-of-speech taggers, parsers, and machine learning (ML) techniques, to make raw corpora more useful for language teaching and learning (Borin 2002). In recent years, this research strand has been supported by the development of NLP applications specifically designed for educational purposes; innovative NLP techniques, for example, have been used to develop tools

S. Sharoff (✉)
Centre for Translation Studies, University of Leeds, Leeds, UK
e-mail: S.Sharoff@leeds.ac.uk

S. Spina
Dipartimento di scienze umane e sociali, Università per Stranieri di Perugia, Perugia, Italy

S. J. Kokkinakis
Department of Swedish, University of Gothenburg, Göteborg, Sweden

and algorithms for readability analysis and generate exercises and tools for assessment and test development (see for example Burstein 2003; Brown et al. 2005; Schwarm and Ostendorf 2005; Kilgarriff et al. 2008). As a result, several international conferences and workshops are now held annually, including Eurocall, Calico, Teaching and Language Corpora, NLP-BEA, etc.

Traditionally, there are two main practical applications of corpora and language resources in language learning (Leech 1997). In the first—the so-called "data-driven learning" approach—the resources are intended for the learner, who uses them in CALL applications or in web-based learning programs; in the second, language resources primarily provide a tool for teachers for creating syllabi and teaching materials, or as a source of linguistic information that can be incorporated into language learning tasks (Aston 1997; Römer 2008). The contributions to this special issue mainly belong to this second group: they describe language resources, based on native, parallel or learner corpora intended for use by teachers in classroom activities or other learning environments. The seven papers in this issue cover a broad range of central issues and activities within this area of activity, including error detection, annotation and correction, text simplification, lexicography and language pedagogy.

Handling and analyzing errors made by language learners is an important part of pedagogical education for teachers and teacher educators. This topic is a common denominator for the papers by Tetrault et al., Espunya, Ferraro et al. and Hana et al.

Tetreault et al. present an overview of approaches to automatic detection of errors made by the language learners, and discuss arguments for their reliable annotation by multiple judges via crowdsourcing. Espunya describes a learner translation corpus in her paper, which is a multiple translation corpus of English and Catalan. The corpus is manually annotated for grammatical and usage errors and enables data querying in one or both languages. Ferraro et al. present an approach to automatic detection and correction of collocation errors using a frequency-based approach. They suggest and evaluate three metrics for suggesting corrections, which are based on affinity, lexical context and context features. The paper by Hana et al. on annotation, after describing a 2 million word corpus compiled from texts written by students of Czech as a second or foreign language at all levels of proficiency, presents and discusses a broad-based annotation scheme for noting grammar and usage errors. The paper also presents an evaluation of the consistency of the annotation and explores and discusses the potential to apply automated linguistic annotation to the corpus.

The rest of the papers in the Special Issue addresses other aspects of the research field.

The paper by Saggion and Bott is aimed at developing a text simplification system for Spanish, which is claimed to be the first attempt to identify and quantify the text simplification operations required for this language. It describes a corpus of manually simplified articles aligned with their original versions, which is used to develop a rule-based system for simplification of Spanish texts.

Kilgarriff et al. present a challenging project of designing frequency-based vocabulary lists for nine languages and thirty-six language pairs. To deal with a

range of languages, they developed generic methods for dealing with polysemy, synonymy, as well as with multiword units.

Finally, the pedagogical uses of a parallel corpus are discussed in the paper by Montero Perez et al. The paper outlines two case studies using a Dutch parallel corpus on the comprehensible input and output enhancements in a reading comprehension and in a writing task to demonstrate the pedagogical value of a parallel corpus.

The 33 submissions we have received to this special issue far exceeded our expectations. It is clear that the research field is vivid and is of significant interest to many researchers in our multilingual society. This issue presents a small sample of ongoing research. We would like to thank the authors for their contributions and also for their patience in the editorial process, as well as our reviewers for their careful work in assessing the submissions.

References

Aston, G. (1997). Enriching the learning environment: Corpora in ELT. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 51–64). London: Longman.

Borin, L. (2002). What have you done for me lately? The fickle alignment of NLP and CALL. In *Proceedings of NLP in CALL*. Jyväskylä, Finland.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 819–826), Vancouver, Canada.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. Shermis, & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*, (pp. 113–121). Routledge.

Granger, S. (2002). A Bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: John Benjamins.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of Euralex*.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). London: Longman.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling & K. Merja (Eds.), *Corpus linguistics. An international handbook* (Vol. 1, pp. 112–130). Berlin: Mouton de Gruyter.

Schwarm, S. E., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of ACL*. Ann Arbor, MI.

West, M. (1953). *A general service list of English words*. London: Longman.