Developing an Open-Source Web-Based Exercise Generator for Swedish

Elena Volodina* and Lars Borin

Språkbanken (Swedish Language Bank), Department of Swedish,
University of Gothenburg, Gothenburg, Sweden

Abstract. This paper reports on the ongoing international project *System architecture for ICALL* and the progress made by the Swedish partner. The Swedish team is developing a web-based exercise generator reusing available annotated corpora and lexical resources. Apart from the technical issues like implementation of the user interface and the underlying processing machinery, a number of interesting pedagogical questions need to be solved, e.g., adapting learner-oriented exercises to proficiency levels; selecting authentic examples of an appropriate difficulty level; automatically ranking corpus examples by their quality; providing feedback to the learner; and selecting vocabulary for training domain-specific, academic or general-purpose vocabulary. In this paper we describe what has been done so far, mention the exercise types that can be generated at the moment as well as describe the tasks left for the future.

Keywords: intelligent computer-assisted language learning, ICALL, natural language processing, NLP, language technology, corpora, exercise generator, interoperability.

1. Introduction

Learning languages with the assistance of a computer – computer-assisted language learning (CALL) – has become widespread since the early 1980s. Traditional CALL applications are inflexible; they provide limited exercise types or number of items, along with limited ability to provide feedback, because the exercises are static, i.e., pre-programmed, and the answers pre-stored. In an attempt to remedy this, researchers have turned to the field of Natural Language Processing (NLP). As a result, the interdisciplinary field of Intelligent CALL (ICALL) has emerged over the past 20 years or so.

At present, there are many mature NLP resources and tools potentially available for re-use in ICALL applications for some languages, but this opportunity has so far

* Contact author: elena.volodina@svenska.gu.se

In L. Bradley & S. Thouësny (Eds.), *CALL: Using, Learning, Knowing, EUROCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings* (pp. 307-313).

remained relatively underdeveloped. In the project *System Architecture for ICALL*[*] funded by NordPlus Sprog we are trying to address this issue. The main task in this project is to design and implement an open-source system architecture for ICALL that would:

- Allow the re-use of NLP tools and resources for language learning tasks;
- Allow the addition of new modules on a plug-and-play basis;
- Be language independent and therefore easily adapted to different languages.

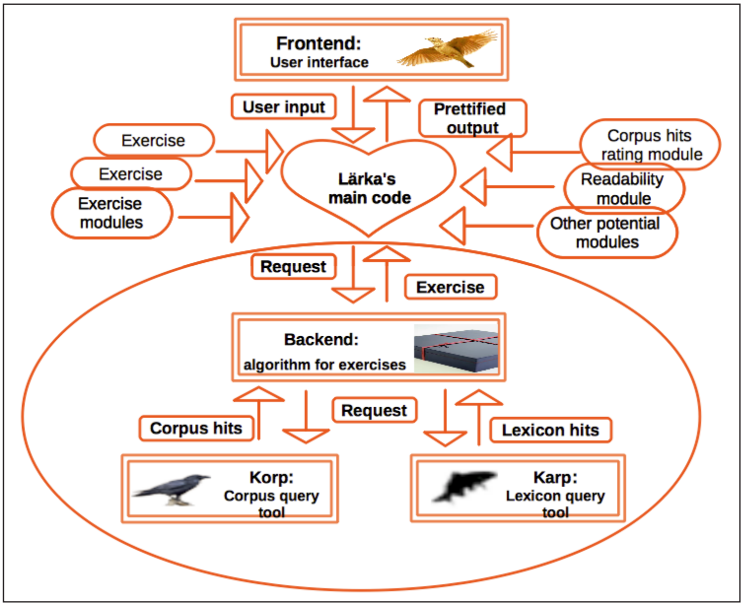
Our system architecture design is such that relevant previous theoretical and applied research results may be added to the system on a plug-and-play basis benefiting language learning and teaching. This calls for cooperation between several fields making ICALL a truly interdisciplinary endeavor. In this project researchers from NLP, linguistics, pedagogy and human-computer interaction (HCI) are working together.

2. An emerging ICALL architecture for Swedish

2.1. Lärka's architecture in a nutshell

A minimal prerequisite for our architecture is an existing infrastructure of interoperable tools and resources, Språkbanken's web-service based infrastructure components for language-resource access.

Figure 1. Lärka's architecture



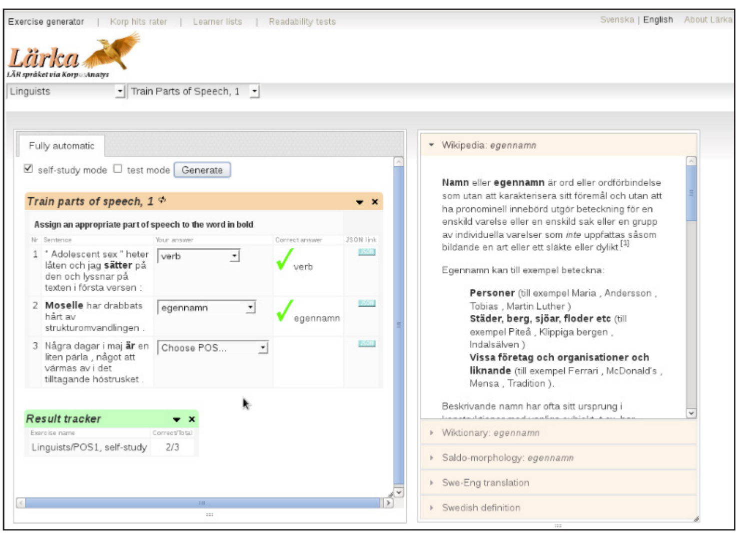
* Participating partners: Reykjavik University, University of Iceland, University of Gothenburg
http://spraakbanken.gu.se/swe/forskning/system-architecture-icall

The application developed to test the architecture is web-based and is called Lärka – "LÄR språket via KorpusAnalys" ('learn language via corpus analysis'; in English *Lark* – "Language Acquisition Reusing Korp"). The four main components of Lärka's architecture are presented in Figure 1:

- *Korp* is Språkbanken's existing web-service based infrastructure for maintaining and searching a constantly growing corpus collection at the moment amounting to about one billion words of Swedish text (Borin, Forsberg, & Roxendal, 2012). The corpora available through Korp contain multiple annotations: lemmatization, compound analysis, part-of-spech (POS) tagging, and syntactic dependency trees;
- *Karp* is the corresponding infrastructure for Språkbanken's collection of lexical resources (Borin, Forsberg, Olsson, & Uppström, 2012);
- The Lärka *backend* is a collection of web services for creating language exercises and selecting distractors. For copyright reasons, the unit used in exercise generation is the sentence. The backend can be used for other applications, for example mobile apps;
- The *frontend* (Figure 2) is the graphical user interface that collects user input and sends requests to Lärka's backend. The design has been inherited from Korp and Karp, so that, for instance, exercise configurations (exercise type, training mode, corpus, level, etc.) can be referenced directly as URLs, saving the user the hassle of always going through the menus on the main webpage.

Each exercise is added as a separate module to the architecture with minimal additions to the user interface code.

Figure 2. Lärka user interface, exercise generator view, self-study mode. POS exercise with reference support window to the right.



2.2. Annotated corpora as a basis for exercises

The exercises are generated using authentic sentences retrieved from two Swedish corpora that have been manually processed, thus ensuring the annotation quality.

SUC is a one-million word corpus of texts from the 1990s, carefully selected to comprise a representative, balanced sample of general-purpose published language, and annotated with lemmas and POS tags (Källgren, Gustafson-Capková, & Hartmann, 2006). The texts have been assigned readability levels using several indices (Volodina, 2010) and the levels are used by Lärka for selection of appropriate sentences for learners of different language proficiency levels.

Talbanken is a manually constructed treebank from the 1970s, containing both written and spoken parts (Einarsson, 1976; Nivre, Nilsson, & Hall, 2006; Teleman, 1974). Currently, the professional prose part of the corpus is used for the exercise generation (about 86,000 words).

2.3. Learning "modes" and feedback

Two exercise modes are available: *self-study* and *test* activities. The s*elf-study mode* offers the learner an opportunity to consider different answers, come back to the previously (incorrectly) answered item and change the answer; the correct answer is not revealed until the user selects it. Every time the user makes some choice, relevant reference material (e.g., Wikipedia articles and dictionary entries) is available to support the learning process (Figure 2 and Figure 3).

In *test mode* the user can answer each item only once. Reference material is not shown to avoid revealing the clues. Eventually one more test mode variant will be added: a timed test when the item should be answered in an assigned period of time (defined by the user). No reference material will be provided in this mode.

A *result tracker* keeps record of correct/incorrect answers.

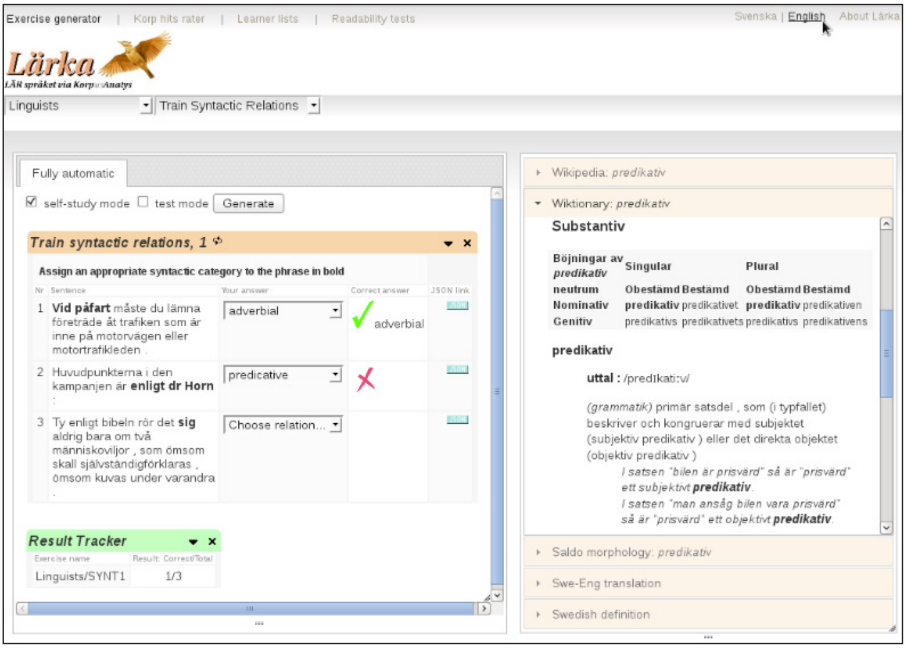
2.4. Exercise types

Currently three exercise types are offered: (1) *POS*; (2) *syntactic relations*; and (3) *multiple-choice vocabulary exercises*.

The *POS* exercises are designed primarily for linguistics students (Figure 2). Here, a random sentence containing a relevant POS is selected from SUC. The target word is presented to the user in bold in its sentence context, and a menu with five potential answers. The distractors are generated dynamically so that two of the distractors are close to the target POS (e.g., *subjunction* or *preposition* for the target POS *conjunction*) and the other two less close (e.g., *determiner* and *pronoun* in the case of *conjunction*). Once the item has been answered a new one is automatically generated.

The *syntactic relation* exercises are also aimed at linguistics students (Figure 3). The design is similar to the POS exercises, but sentences are retrieved from the Talbanken treebank. The distractors are always the same since only seven of the (clause-level) syntactic categories in the corpus are currently used.

Figure 3. Exercise Train syntactic relations with reference support window to the right.

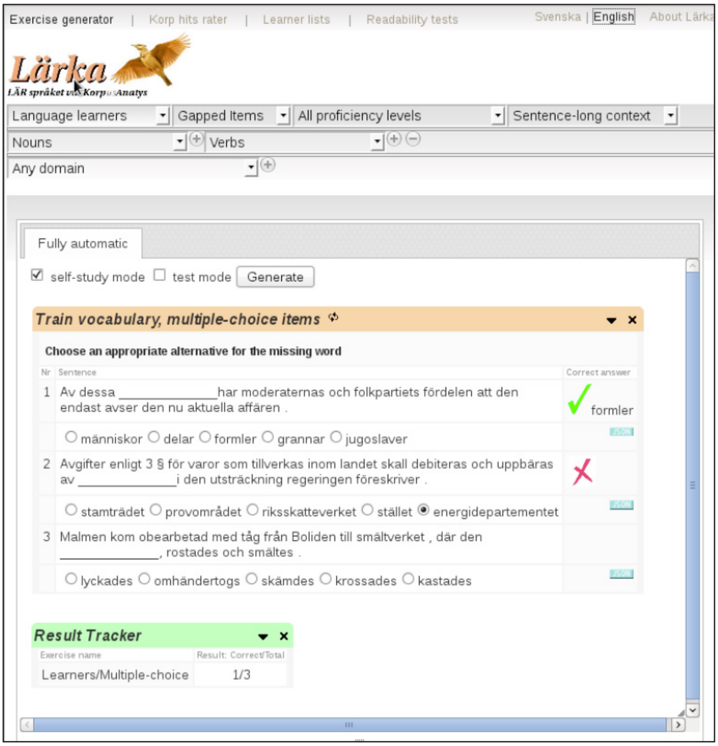


The *multiple-choice vocabulary* exercises (Figure 4) target learners of Swedish and take into consideration sentence difficulty and the desired vocabulary for training. Sentence difficulty level is determined using the LexLIX readability index (Volodina, 2010). The target vocabulary characteristics are chosen by the users, e.g., restricted as to POS, domain, or proficiency level. For this purpose precompiled vocabulary lists are needed, e.g.,:

- Frequency-based word lists with assigned proficiency levels. We are currently using the Swedish Kelly-list (Volodina & Johansson Kokkinakis, 2012) and the Base Vocabulary list (Forsbom, 2006);
- Domain-specific vocabulary lists. At the moment we can use: the academic wordlist (Jansson, Johansson Kokkinakis, Ribeck, & Sköldberg, 2012) and topic vocabulary lists from the Lexin picture series (Lexin, 2006).

Distractors are chosen according to proficiency level or frequency band, and morphosyntactic form. There is, however, an idea to test a more refined approach for the lower proficiency levels where distractors are graded by difficulty level, for example, two of them come from a different part of speech.

Figure 4. Multiple-choice exercise with POS constraints set on the target vocabulary



3. Future plans

During the development of Lärka we have formed a clearer picture of both system requirements and the pedagogical activities we would like to realize. In the near future we plan to add a number of vocabulary training exercises, namely gap cloze and wordbox exercises as well as a diagnostic test for evaluating the learner's vocabulary knowledge level. Additionally, we plan to add a syntactic tree to every sentence; hyperlink all words in a sentence to relevant encyclopedia and lexicon entries; and provide a possibility to save generated items in a number of formats (e.g., QTI (Question and Test Interoperability); IMS (2006)). Further down the road we are planning to add:

- An option of modifying automatically generated exercises by providing user-defined word lists or texts or by providing user-selected distractors;
- A module for ranking corpus hits according to different linguistic features and parameter settings;
- The possibility to test texts for readability using several readability indices;
- The possibility to select and save sub-lists from learner lists of domain or general vocabulary;

- Several new exercise types, e.g., for grammar, word-building, morphology, etc.

Another important issue which we plan to focus on in the future is formal evaluation of Lärka's architecture as well as of the learner activities offered by Lärka.

4. Conclusion

In designing an open-source system architecture for ICALL we want to promote re-use of available mature NLP resources and tools in language learning and teaching. Of course, many aspects of teaching and learning cannot be successfully handled by computers. However, some of the more mechanical aspects of language learning can be successfully implemented – e.g., (some) test item production(s), selection of appropriate corpus examples, analysis of text complexity by proficiency level, feedback generation, etc. – leaving more scope for teachers to develop the more creative aspects of language teaching.

References

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012* (pp. 474-478). Istanbul: ELRA.

Borin, L., Forsberg, M, Olsson, L.-J., & Uppström, J. (2012). The open lexical infrastructure of Språkbanken. *Proceedings of LREC 2012* (pp. 3598-3602). Istanbul, Turkey: ELRA.

Einarsson, J. (1976). Talbanken: Talbankens skriftspråkskonkordans/Talbankens talspråkskonkordans. Lund University.

Forsbom, E. (2006). Deriving a Base Vocabulary Pool from the Stockholm Umeå Corpus. Retrieved from http://stp.lingfil.uu.se/~evafo/resources/baseformmodels/

IMS. (2006). IMS Question and Test Interoperability Overview. Version 2.1 Public Draft (revision 2) Specification. IMS Global Learning Consortium. Retrieved from http://www.imsglobal.org/question/qtiv2p1pd2/imsqti_oviewv2p1pd2.html

Jansson, H., Johansson Kokkinakis, S., Ribeck, J., & Sköldberg, E. (2012). A Swedish Academic Word List: Methods and Data. Forthcoming in *Proceedings of the XV Euralex International Congress*. Oslo: EURALEX.

Lexin. (2006). *Svenska ord med uttal och förklaringar*. Språkrådet.

Källgren, G., Gustafson-Capková, S., & Hartmann, B. (2006). *Manual of the Stockholm Umeå Corpus version 2.0.* Department of Linguistics, Stockholm University.

Nivre, J., Nilsson, J., & Hall, J. (2006). Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In P*roceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)* (pp. 1392-1395). Genoa: ELRA.

Teleman, U. (1974). *Manual för grammatisk beskrivning av talad och skriven svenska*. Lund.

Volodina, E. (2010). *Corpora in Language Classroom: Reusing Stockholm Umeå Corpus in a vocabulary exercise generator*. Saarbrücken: Lambert Academic Publishing.

Volodina, E., & Johansson Kokkinakis, S. (2012). Introducing Swedish Kelly-list, a new lexical e-resource for Swedish. *Proceedings of LREC 2012* (pp. 1040–1046). Istanbul: ELRA.

Published by Research-publishing.net
Dublin, Ireland; Voillans, France
info@research-publishing.net

CALL: Using, Learning, Knowing
EUROCALL Conference, Gothenburg, Sweden
22-25 August 2012, Proceedings
Edited by Linda Bradley and Sylvie Thouësny

Typeset by Research-publishing.net
Cover design: © Raphaël Savina (raphael@savina.net)
Aquarelle reproduced with kind permission from the illustrator: © Sylvi Vigmo (sylvi.vigmo@ped.gu.se)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-03-2 (paperback)
Print on demand (lulu.com)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Bibliothèque Nationale de France - Dépôt légal: décembre 2012.