Lärka: an online platform where language learning meets natural language processing

Ildikó Pilán[1]*, David Alfter*[1]*, Elena Volodina*[1]

[1]Språkbanken, University of Gothenburg, Sweden

{ildiko.pilan, david.alfter, elena.volodina}@gu.se

1. General overview

We present Lärka[1], an Intelligent Computer Assisted Language Learning (ICALL) platform developed at Språkbanken. Lärka is an openly available web-based tool that builds on a variety of existing language resources such as corpora, lexical resources and language technology tools. This makes the platform flexible and a valuable source of additional learning material (e.g. via corpus-based exercises) and a support tool for both teachers and learners of Swedish. Lärka has recently received a new user-interface that is more suitable for different screen sizes. Moreover, the system has also been augmented with new functionalities. These recent additions aim at improving the usability and the usefulness of the platform for pedagogical purposes. Thanks to Lärka's *service-oriented architecture*, most functionalities are also available as web services which can be easily re-used by other applications.

2. Automatic exercise generation

One of the main functionalities of Lärka is the automatic generation of exercises based on real-life language examples from corpora. Exercise generation is aimed at two groups of learners: students of (Swedish) linguistics and learners of Swedish as a second language (L2). The currently available exercises have a multiple-choice format. Each exercise item consists of a sentence containing either a highlighted word or a gap, as well as a list of five answer alternatives out of which one is the correct answer and four are *distractors*, i.e. incorrect options. Students of linguistics can train parts of speech, syntactic relations and semantic roles. Language learners can choose between vocabulary exercises and inflectional exercises. A recent addition to our platform is a simple word-level exercise, *WordGuess*, that takes a step towards gamified learning. WordGuess reimplements the well-known Hangman game mechanism: users are presented with a number of hidden characters and their task is to guess characters contained in the word, which eventually helps them guessing the word itself. Every time the guessed character is not in the word, users receive penalty points. In our learning-oriented version of the game, users can choose to receive clues such as the translation of the word, its definition in Swedish or information about its morphological form. This game is a simple example of reusing information from lexical resources for gamified language learning activities.

3. Corpus example selection

In Lärka, the automatically generated exercises for language learners rely on *HitEx* (*Hitta Exempel* 'find examples'), a tool for selecting and ranking corpus examples. The main purpose of HitEx is to identify sentences from generic corpora which are suitable as exercise items for L2 learners. The suitability of the sentences is determined based on a number of parameters that reflect different linguistic characteristics of the sentences. Through a graphical user interface, it is also possible to perform a sentence search based on parameters customized by the user. The selection criteria include a wide variety of linguistic aspects such as the desired difficulty level based on the CEFR (Common European Framework of Reference for Languages), typicality based on word co-occurrence measures, as well as the absence of anaphoric expressions and sensitive vocabulary (e.g. profanities), just to name a few. Besides its applicability to the language learning domain, HitEx can be also useful for lexicographers for finding dictionary examples that illustrate the meaning and usage of lexical items.

4. Text complexity evaluation

Another functionality, *TextEval*, offers an interface to automatically assess Swedish texts for their degree of complexity according to the CEFR. Texts can be either learner productions (e.g. essays) or texts written by experts as reading material for learners. The machine learning based automatic analysis returns an overall CEFR level for the text, as well as a list of linguistic indicators relevant for measuring text complexity, such as the average length of sentences and tokens, LIX score and nominal ratio. In addition, it is possible to add a color-enhanced highlighting for words per CEFR levels which provides users with a straightforward visual feedback about the lexical complexity of a text. Behind this feature are two word lists, one based on expert-produced texts to reflect *receptive* vocabulary and another based on learner-produced texts representing *productive* vocabulary. In the case of both lists, frequency distributions of lemmas have been mapped to a single CEFR level. For each CEFR level, a darker and a lighter shade of the same color represents productive and receptive vocabulary respectively at the given level.

5. Ongoing work and planned extensions

Besides the activities described in Section 2, the migration of the previous version of our spelling exercises and the addition of new exercise formats are currently under development. In the near future we plan to add a login functionality as well as an infrastructure to log user data. This would enable us to create a valuable resource for modeling learners (e.g. L1-specific errors, learners development over time) and to offer adaptive exercises. We also plan on offering a diagnostic test to assess learners' proficiency levels based on different exercises. Further extensions under development include an annotation interface for learner corpora which facilitates the process of entering metadata about learner essays. We are also investigating the possibility of annotation learner errors through this interface.

[1]https://spraakbanken.gu.se/larka