Modeling Individual Learner Knowledge in a Computer Assisted Language Learning System

Alfter David, Volodina Elena

Språkbanken
University of Gothenburg, Sweden
firstname.lastname@svenska.gu.se

Abstract

We present a project that aims at facilitating language learning by modeling individual learner knowledge in an online language learning platform. At the same time, we can collect useful data about the learner, their progress and their interaction with the platform. The presented project is still at a very early stage, and this paper defines the objectives and proposes a methodology, rather than describing initial results. Implementation and evaluation are future work.

1. Introduction

Traditional language learning platforms present their knowledge in the same form for every learner. However, this is not the most learner friendly solution (Heift, 2007). Indeed, every learner has a different language learning background and prior knowledge, different learning speed, different needs. The most learner friendly solution would be to dedicate a single learning platform to each learner; however, this is not feasible in practice (Heift, 2007).

Intelligent Computer Assisted Language Learning (ICALL) platforms try to address this problem by adapting themselves to the learners, their progress and their specific needs. In order to do so, the system needs information about the user. This information can be obtained in different ways, for example by

1. asking the learner questions
2. asking the learner to complete certain tasks (e.g. write an essay)
3. infer information based on observations while the learner uses the system

The most learner friendly solution would be option 3. Asking questions is a feasible option but it can be tedious for the learner to answer question after question, especially in the beginning when the system does not have any knowledge about the learner. The first question is whether, if, and how, we can infer important information about the student with minimal invasive questioning.

Moreover, every learner has one or more mother tongues that influence their language learning experience. While the mother tongue is a frequent source of errors due to language *interference* (Myles, 2002; Wong and Dras, 2009; Hanafi, 2014), it can also facilitate learning through language *transfer* (Hanafi, 2014). Due to language transfer, learners might be able to use certain words, phrases or constructions that have not been formally introduced yet. Can we, in this case, distinguish between learned knowledge and language transfer?

This leads to the following two research questions:

1. RQ1. (How) can we infer knowledge about a learner with minimal explicit questioning?
2. RQ2. Can we distinguish between learned knowledge and transferred knowledge?

2. Motivation

While there is ongoing research at the international level, Swedish ICALL systems are rare, despite the availability of the necessary resources (Volodina and Borin, 2012). Most language learning platforms base their assumptions of learner levels purely on strictly frequency-based approaches and are thus language-independent. To the best of our knowledge, there are no ICALL platforms that follow a pedagogically aware approach within the context of CEFR. Another motivation could be that we need to know more about learners to help teachers cope with the new wave of immigrants; that this way we also offer a platform that generates language learning materials, and thus supports learning of Swedish in a teacher-free scenario.

Another strength of our project is that through this platform we expect to collect rich information about learners and learning process, data that can eventually be used for various types of research within for example Swedish as a second language or second language assessment.

3. Methodology

This work uses the Common European Framework of Reference (CEFR) (Council of Europe, 2001) as pedagogical framework for evaluating student performance. The CEFR is a framework for language teaching and language evaluation. It defines six levels of proficiency or stages of language acquisition, ranging from A1 (beginner) over A2, B1, B2, C1 to C2 (expert).

The CEFR also defines four skills: reading, writing, listening and speaking, and different learner competences related to these skills. This work focuses on vocabulary in writing and modeling individual vocabulary knowledge.

In order to offer a personalized experience, we first have to design and implement the required models for learner modeling. In addition to the actual learner model, we also need ideal models which represent the target to reach.

For each CEFR level, we would need one ideal model which represents an ideal learner having reached this level. The ideal model defines what a learner should know when reaching a certain level. It is also possible to envision ideal topic models which represent different topics to be learned. Another possibility is to compare a learner model to an ideal native speaker model. Which model to choose under which circumstances largely depends on the learner's goals.

For the actual learner model, we need to select relevant CEFR competences and descriptors from the pool of all available competences which mainly indicate what will be modeled. Additionally, the learner model will include information such as mother tongue and year of birth.

4. Learner Interaction Cycle

The work on learner modeling is based on an existing ICALL platform Lärka (Volodina et al., 2014a). When a learner starts using the ICALL platform, the so-called learner interaction cycle starts. First, learner goals with regard to vocabulary, grammar or topics are set by either the learner through menu settings, by the system using default values or adaptive approaches or by both. In the second step, relevant texts are selected using existing sentence and text selection algorithms which can retrieve sentences or texts for a given CEFR level (Pilán et al., 2013; Pilán et al., 2014). From the selected text, a set of assessment tools is generated. This encompasses both generated exercises as well as the tools required for automatic assessment of the generated exercises. Based on the learner interaction with the ICALL platform, learner knowledge is (re-)assessed and updated. The current goals are then evaluated with regard to whether they have been reached or not. If necessary, new goals are set and the cycle starts anew.

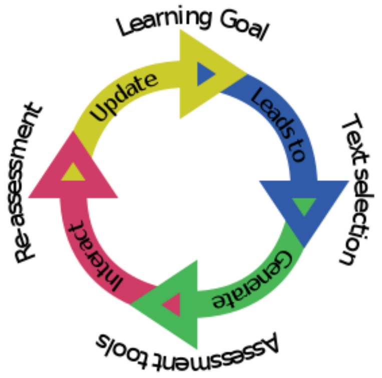


Figure 1: Learner Interaction Cycle

In order to get started, we need to collect learner information, then assess prior learner knowledge either by using a short diagnostic test or by learner self-assessment. Finally, we can start the learner interaction cycle.

5. Lexical Complexity Analysis and Automatic Essay Grading

The first research question is rather abstract and is left as future work at the moment. However, we can get some information about the learner's knowledge by analyzing their written productions as illustrated in the next paragraph.

The second research question can be at least partially addressed by using a lexical complexity analysis. If we look at the vocabulary used productively (actively) by learners, we can draw certain conclusions as to the proficiency level of the student, but also possibly topics of interest. Indeed, if a learner uses a lot of vocabulary of a higher level than their current level, we can conceivably draw one of two different conclusions. If the lexical items of higher levels are mostly function words, it is likely that the learner transferred grammatical structures from their mother tongue. If, on the other hand, most lexical items of higher level are content words, it is not unlikely that the learner knows these words because they are part of a topic that the learner is interested in; however, it is also possible to transfer content words from one's mother tongue.

The proposed lexical complexity analysis works with two resources:

- SVALex
- SweLL list

SVALex (François et al., 2016) is a list of words with their respective frequency distribution over all CEFR levels derived from the COCTAILL corpus (Volodina et al., 2014b), a corpus of text books used in CEFR teaching. Analogously, SweLL list (Llozhi, 2016) is a list of words with their frequency distribution over all CEFR levels derived from the SweLL corpus (Volodina et al., 2016), a corpus of essays written by learners of Swedish as a second language. SVALex covers receptive (passive) vocabulary while SweLL list covers productive (active) vocabulary.

The main problem is defining a "target level" for each word given a distribution. Indeed, the distribution does not contain this information, neither explicitly nor inherently. Figures 2 and 3 show sample distributions respectively for the word *annars* 'otherwise' and the word *fin* 'beautiful'. The mapping from distribution to a CEFR label is currently work in progress.

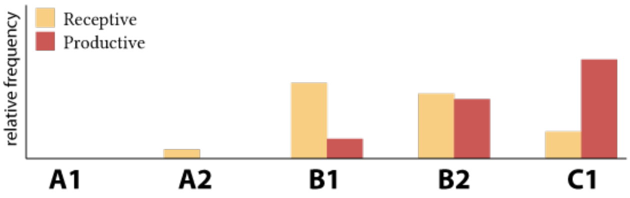


Figure 2: Distribution for *annars* 'otherwise'

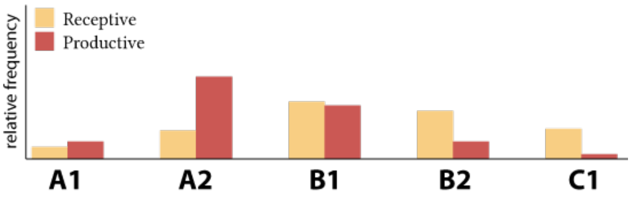


Figure 3: Distribution for *fin* 'beautiful'

Lexical complexity analysis is currently being implemented in an experimental automatic essay grading system for the assessment of essays written by learners of Swedish as a second language which itself will be embedded in the openly accessible experimental language learning platform Lärka (Volodina et al., 2014a). There is also ongoing work on the Lärka platform development, extending the existing platform with new functions such as a database for collecting learner variables, design of (new) exercise types and an option for essay/text assessment. Figure 4 shows a prototype of the essay classification.

References

Council of Europe. 2001. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of LREC 2016*.

Aissa Hanafi. 2014. The Second Language Influence on Foreign Language Learners' Errors: The Case of the French Language for Algerian Students Learning English as a Foreign Language. *European Scientific Journal*.

Trude Heift. 2007. Learner personas in CALL. *Calico Journal*, 25(1):1–10.

Lorena Llozhi. 2016. SweLL list. A list of productive vocabulary generated from second language learners' essays. Master Thesis in Language Technologies. Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg.

Johanne Myles. 2002. Second language writing and research: The writing process and error analysis in student texts. *The Electronic Journal for English as a Second Language*, 6(2):1–20.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013. Automatic Selection of Suitable Sentences for Language Learning Exercises. In *20 Years of EUROCALL: Learning from the Past, Looking to the Future: 2013 EUROCALL Conference Proceedings*, pages 218–225.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 174–184.

Elena Volodina and Lars Borin. 2012. Developing a freely available web-based exercise generator for Swedish. *EuroCALL 2012 Proceedings*.

Elena Volodina, Lars Borin, Hrafn Lofsson, Birna Arnbjörnsdóttir, and Guðmundur Örn Leifsson. 2012a. Waste not; want not: Towards a system architecture for ICALL based on NLP component re-use. In *Proceedings of the SLTC 2012 workshop on NLP for CALL; Lund; 25th October; 2012*. Linköping University Electronic Press.

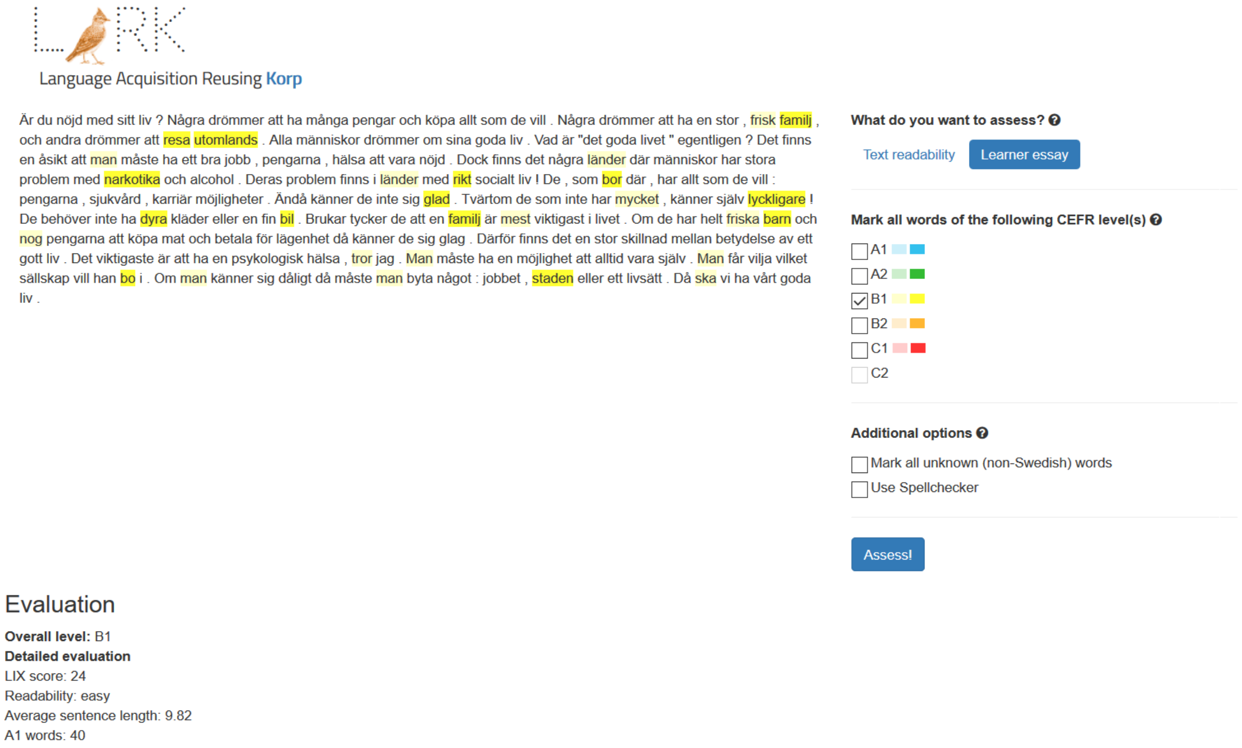
Elena Volodina, Hrafn Loftsson, Birna Arnbjörnsdóttir, Lars Borin, and Guðmundur Örn Leifsson. 2012b. Towards a system architecture for ICALL. In *Proceedings of the 20th International Conference on Computers in Education. Singapore: Asia-Pacific Society for Computers in Education*.

Elena Volodina, Ildikó Pilán, Lars Borin, and Therese Lindström Tiedemann. 2014a. A flexible language learning platform based on language resources and web services. In *LREC*, pages 3973–3978.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014b. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, number 107. Linköping University Electronic Press.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. *arXiv preprint arXiv:1604.06583*.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61.



Language Acquisition Reusing Korp

Är du nöjd med sitt liv ? Några drömmer att ha många pengar och köpa allt som de vill . Några drömmer att ha en stor , frisk familj , och andra drömmer att resa utomlands . Alla människor drömmer om sina goda liv . Vad är "det goda livet " egentligen ? Det finns en åsikt att man måste ha ett bra jobb , pengarna , hälsa att vara nöjd . Dock finns det några länder där människor har stora problem med narkotika och alcohol . Deras problem finns i länder med rikt socialt liv ! De , som bor där , har allt som de vill : pengarna , sjukvård , karriär möjligheter . Ändå känner de inte sig glad . Tvärtom de som inte har mycket , känner själv lyckligare ! De behöver inte ha dyra kläder eller en fin bil . Brukar tycker de att en familj är mest viktigast i livet . Om de har helt friska barn och nog pengarna att köpa mat och betala för lägenhet då känner de sig glag . Därför finns det en stor skillnad mellan betydelse av ett gott liv . Det viktigaste är att ha en psykologisk hälsa , tror jag . Man måste ha en möjlighet att alltid vara själv . Man får vilja vilket sällskap vill han bo i . Om man känner sig dåligt då måste man byta något : jobbet , staden eller ett livsätt . Då ska vi ha vårt goda liv .

What do you want to assess?

Mark all words of the following CEFR level(s)

□ A1 🟦🟦
□ A2 🟩🟩
☑ B1 🟨🟨
□ B2 🟧🟧
□ C1 🟥🟥
□ C2

Additional options

□ Mark all unknown (non-Swedish) words
□ Use Spellchecker

Evaluation

Overall level: B1
Detailed evaluation
LIX score: 24
Readability: easy
Average sentence length: 9.82
A1 words: 40

Figure 4: Essay evaluation tool