Towards a gold standard for Swedish CEFR-based ICALL

Elena Volodina[1], Dijana Pijetlovic[1], Ildiko Pilán[1], Sofie Johansson Kokkinakis[1]

(1) Språkbanken, University of Gothenburg, Box 200, 405 30 Göteborg, Sweden

elena.volodina@svenska.gu.se, guspijdi@student.gu.se, ildiko.pilan@gmail.com, sofie.johansson.kokkinakis@svenska.gu.se

Abstract

In qualitative projects on ICALL (Intelligent Computer-Assisted Language Learning), research and development always go hand in hand: development both depends upon the research results and dictates the research agenda. Likewise, in the development of the Swedish ICALL platform *Lärka*, the practical issues of development have dictated its research agenda. With NLP approaches, sooner or later, the necessity for reliable training data becomes unavoidable. At the moment Lärka's research agenda cannot be addressed without access to reliable training data, so-called “gold standard”. This paper gives an overview of the current state of the Swedish ICALL platform development and related research agenda, and describes the first attempts to collect the reference corpus (“gold standard”) coming from course books used in CEFR-based language teaching.

Keywords: ICALL, CEFR, exercise generator, course book corpus compilation

1 Background

The ICALL platform *Lärka* described in this paper is an open-source web-based application that uses principles of Service-Oriented Architecture (Volodina et al., 2012a; Volodina & Borin, 2012). The platform is divided into several modules: an exercise generator with activities for university students of linguistics and second/foreign language (L2) learners; and modules facilitating different aspects of development and research, at the moment consisting of an experimental sentence readability module and an editor for learner-oriented corpora.

The main focus of Lärka is on L2 learners. This sets certain requirements, first of all, on the use of a pedagogical framework. Among different pedagogical theories and approaches, the Common European Framework of Reference for Languages (CEFR) is one of the most influential. CEFR is a document containing guidelines for harmonization of language teaching and assessment across languages and countries (Council of Europe, 2001). It provides a common metalanguage to talk about objectives, assessment and proficiency levels. Further, it offers a descriptive scheme that can help analyze learner's needs, target communicative competences and define the course curriculum. It is useful for tracking learner progress as well as for designing assessment tests and assigning proficiency levels (Little, 2007, 2011; North, 2007). CEFR defines language competences and skills through "can do" statements at six proficiency levels (A1, A2, B1, B2, C1, C2) which offer flexibility in interpreting them for different languages and target groups. Since the publication of the CEFR guidelines in 2001, a number of countries including Sweden have adopted the system and reorganized language teaching and testing practices to fit into this framework.

Other existent proficiency scales for Swedish language learning include the ones used in SFI (Swedish for immigrants) and SVA (Swedish as a Second Language), both aligned to fit into the CEFR paradigm. SFI, containing levels A, B, C, D correspond to CEFR's A1-/A1, A1/A2, A2/A2+, B1/B1+ respectively according to the recommendations provided by the Swedish National Agency for Education (Skolverket). The language proficiency scale used for SVA, is said to be roughly equivalent to the CEFR level C1 when sva B is reached. Since the CEFR scale combines all the extremes of development of Swedish as L2, and offers interoperability across different countries, we have chosen this scale for our platform.

Ideally, the use of CEFR scales in the context of an ICALL platform should offer a clear-cut possibility to generate exercises and materials adjusted to the proficiency levels. It is, however, a non-trivial task to apply the CEFR descriptors to the practical task of automatic selection of language samples appropriate for different proficiency levels. CEFR's flexibility, being a positive feature on the one hand, has a reverse side. As a number of Second Language Acquisition (SLA) researchers have mentioned, it is non-specific

and therefore it is difficult to associate the different kinds of competences and levels of accuracy that learners would need in order to perform language learning tasks with different CEFR levels (Westhoff, 2007). Milton (2009) says that the lack of objectivity in the CEFR descriptors makes it possible that learners with different amounts and kinds of knowledge can be placed into the same CEFR level; or that performance outweighs competence so that competent but insecure performers can be assigned to a lower CEFR level than they deserve. Among other things, insufficient specifications for vocabulary and grammar competence have been pointed out by Byrnes (2007); Milton (2009); Westhoff (2007); Little (2007, 2011).

Special efforts have been undertaken to interpret CEFR guidelines as sets of Reference Level Descriptions (http://www.coe.int/t/dg4/linguistic/dnr_en.asp) as well as to establish procedures to relate language exams to the CEFR (Council of Europe, 2009), but to the best of our knowledge that has not been done yet for Swedish. Attempts at *aligning texts and tests with CEFR* for a number of other languages are ongoing (e.g. Khalifa et al., 2010; Szabó, 2010; Dávid, 2010) with what could be called a *top-down approach*, i.e. starting from CEFR descriptors and going all the way down to the actual selection of appropriate language samples. We suggest a *bottom-up approach*, where we start from the actual language samples labeled by experienced teachers or coursebook writers for levels, analyze them for linguistic constituents with the help of machine learning approaches and then try to map the identified constituents to the CEFR descriptors. The two approaches should be viewed as complementary of each other.

This is the starting point for our “quest” for data collection, designed to help us interpret CEFR descriptors in a way that can facilitate automatic methods in L2 material generation, among other things: to identify receptive vocabulary scope per level, and to adjust algorithms for sentence readability per proficiency level. Both aspects are described in detail in the following section.

The paper is structured as follows: section 2 reports on the current state of development where the lack of exact interpretation of CEFR scales into linguistic constituents for Swedish has so far hindered implementation of desired exercises or their adjustment to learner proficiency levels. Section 3 describes the compilation of a corpus of CEFR-related course book texts as a way to cope with that obstacle. Section 4 concludes the paper.

2. Current state - in need of a gold standard

Use of NLP for language learning tasks has been pursued in different studies (e.g. Amaral and Meurers, 2011; Amaral et al., 2011; Heift, 2003; Nagata, 2009). Most of the implemented applications generate learning materials, tasks or feedback customized to user interests, needs and proficiency levels. However, the question of automatic classification of authentic language samples (e.g. texts or sentences) into proficiency levels is not always directly addressed. In Meurers et al. (2010) and Knoop & Wilske (2013), the user

finds the texts on the web him-/herself, and the exercise is generated on the basis of that text. In Toole & Heift (2002) this issue is solved indirectly through teachers feeding in sample texts containing examples of learning objective. In Aldabe et al (2006) this issue is ignored and only questions for "high language level" are generated. The question of text classification into levels is directly approached in REAP and Choosito applications (Collins-Thompson & Callan, 2007; Heilman et al., 2007; Francois & Miltsakaki, 2012), elaborating on two major factors: vocabulary frequency and a readability measure based on a selection of linguistic parameters.

2.1 Module for university students of Linguistics

An exercise generator for linguists comes with two exercises: training syntactic relations and training parts of speech (Figure 1).

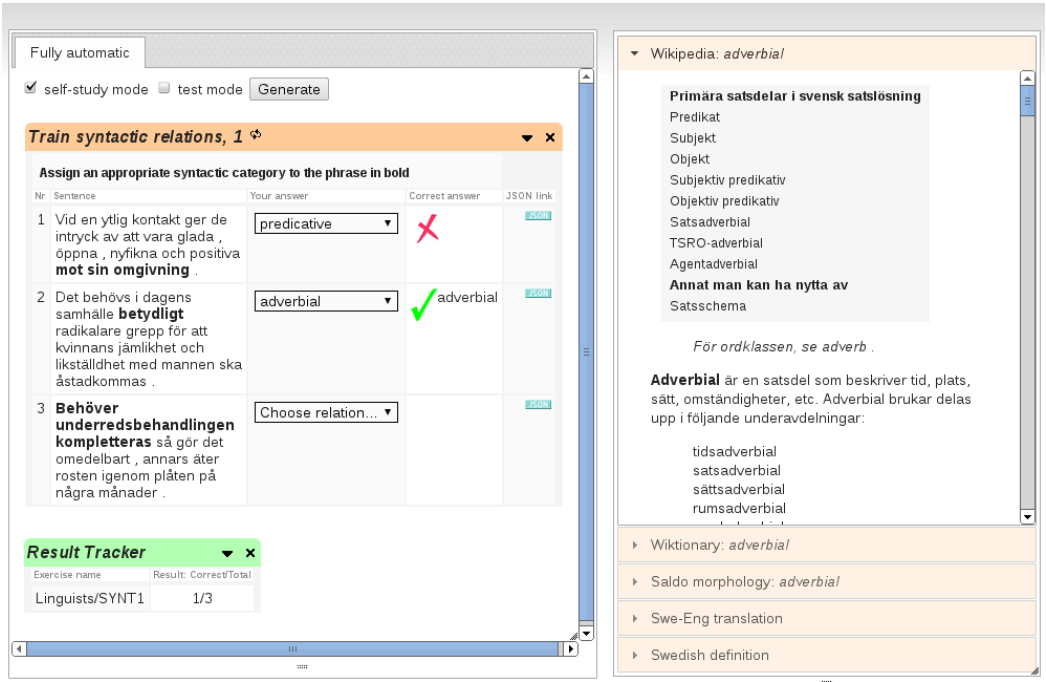

FIGURE 1. EXAMPLE OF AN ITEM FOR TRAINING SYNTACTIC RELATIONS. INTENDED USERS: LINGUISTS

Both exercises use multiple-choice model and are based on sentences randomly selected from several manually checked corpora of Swedish: Stockholm Umeå Corpus (Källgren et.al., 2006), Talbanken (Teleman, 1974; Einarsson, 1976; Nivre et al., 2006) and Läsbart (Heimann Mühlenbock, 2013). The user is offered support in the form of Wikipedia and lexicon entries, as well as feedback in the form of correct-incorrect answers and a result tracker. Once the item is answered, another one is generated.

The system has been tested in real-life setting with students of Linguistics and the first feedback has revealed the general acceptance of the exercises.

However, teachers have expressed their reservation against the use of Wikipedia instead of reference sources of higher quality/reliability. Among other desired improvements a better sentence selection has been mentioned. “Better” sentences should be understood as non-eliptic well-formed simple sentences (as opposed to complex ones). The problem of selection of “appropriate” sentences is described under “Sentence readability” below.

2.2 Multiple-choice vocabulary items for L2 learners

An exercise generator for language learners comprises at the moment multiple choice exercise items for vocabulary training, see Figure 2.

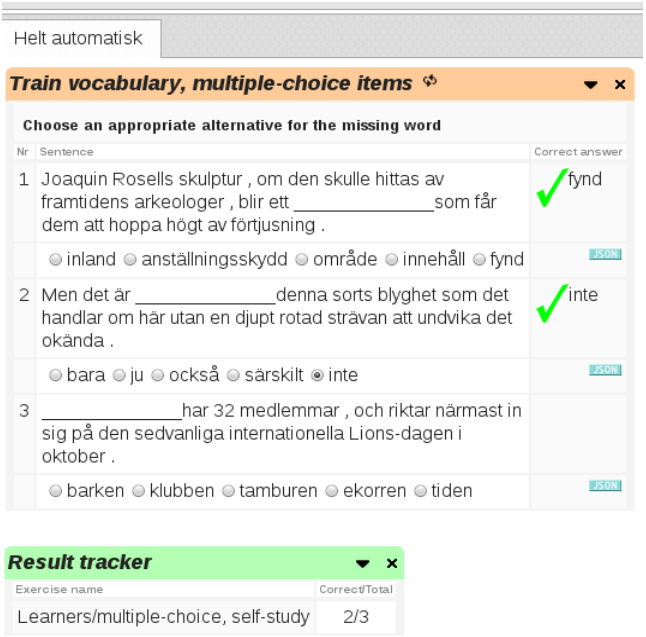

FIGURE 2. MULTIPLE-CHOICE ITEMS FOR LANGUAGE LEARNERS

The target vocabulary for training is at the moment selected randomly from the Swedish Kelly list (Volodina & Johansson Kokkinakis, 2012), a frequency-based vocabulary list for language learners. A sentence containing the target vocabulary is then randomly selected from SUC (Stockholm Umeå Corpus, Källgren et.al., 2006) guided by the principle of maximum sentence length limited to 15 tokens. Distractors to the correct answer are selected based on the principle of the shared frequency band with the correct answer, the same part of speech and shared morpho-syntactic tag.

However, to generate exercise items appropriate at different learner proficiency levels, selection of target vocabulary should be aligned with the CEFR levels. The latter means the need to study the vocabulary used in the CEFR-based courses, both receptively in course books and productively in

written essays, per proficiency level. Addressing this problem without reference data labelled for CEFR levels is however impossible.

Another problem arising in connection with vocabulary training is the appropriateness of the language samples where the target item is used in its context. For copyright reasons, the usual context in Lärka is limited to sentences. Selection of appropriate sentences for language training at different proficiency levels needs a reliable method to classify available sentences by CEFR levels. This, in turn, cannot be studied without an extensive collection of appropriate sentences labelled for proficiency levels, which again points to the need of a corpus of CEFR-related texts.

2.3 Dictation and spelling items for L2 learners

The dictation and spelling items have been recently implemented, but the development is still in progress (Pijetlovic & Volodina, forthcoming).

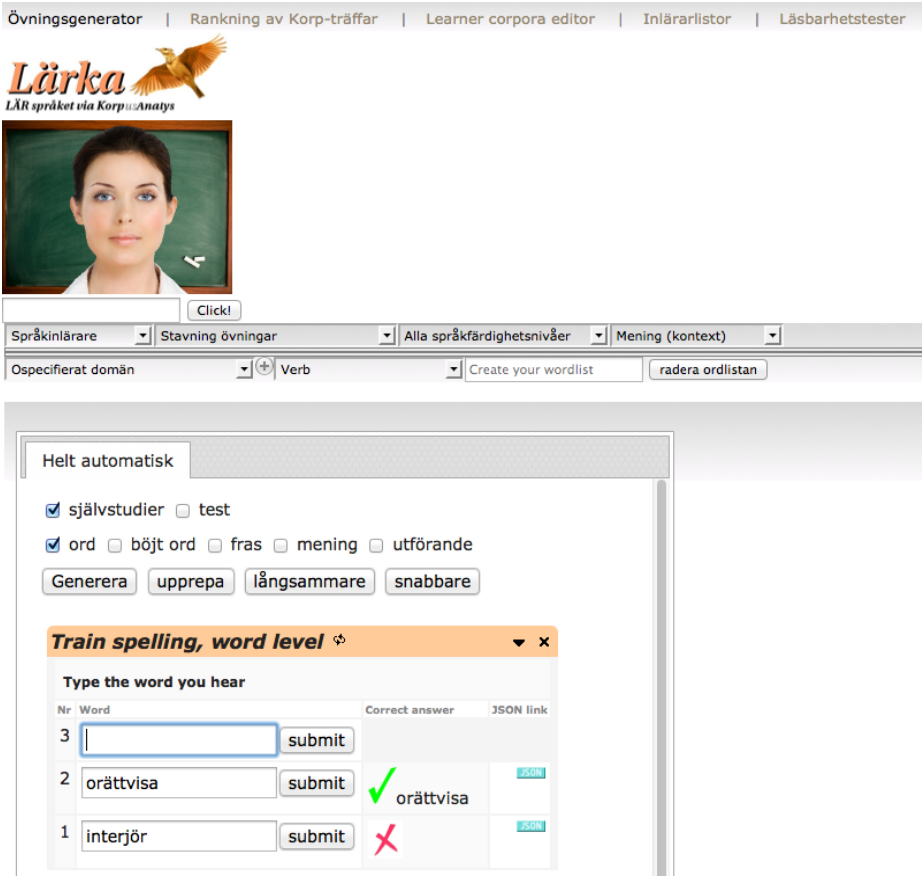

FIGURE 3. DICTATION AND SPELLING ITEM

The goal of this module is to offer web services for automatic generation of spelling exercises using Text-To-Speech technology for Swedish, thus

facilitating training of listening and spelling competences. The exercise is planned to be “adaptive” in the sense that once the users are confident with spelling single words, they are offered the target word in inflected forms, in phrases, and finally in sentences (Figure 3).

Spelling errors can be distinguished between performance-based and competence-based. To account for a more fine-grained distinction between errors, a collection of real-life spelling mistakes needs to be consulted in order to give a useful feedback to the user. Due to the lack of Swedish spelling error corpora, one part of this module involves collecting spelling errors through online dictation&typing exercises with both Swedish native and non-native speakers.

The success of this exercise type depends upon the two factors mentioned before: selection of vocabulary and sentences appropriate for learner level.

2.4 Current research agenda

From the short description above, it is clear that the immediate research agenda contains, among other things, (1) the issue of identifying receptive vocabulary scope per proficiency level and (2) the issue of finding a reliable algorithm for sentence readability assessment. Both issues depend on the availability of reference data, which we are now actively collecting.

2.4.1 Receptive vocabulary scope

According to the CEFR document, there are four main sources of vocabulary that potentially can constitute the vocabulary scope of a CEFR-based course, namely: (1) words typical for the topics required for the learners’ communication (domain-specific vocabulary); (2) vocabulary that is based on lexical-statistical principles of selection (highest frequency words); (3) words randomly coming from texts that are selected as learning material by teachers, and finally (4) words learnt in response to the communicative needs that arise. (Council of Europe, 2001:150-151) The users of CEFR guidelines are encouraged to define *what specific/particular lexical elements the learner might need and how they have been selected.*

To identify the scope of receptive vocabulary for exercise generation needs, we intend to collect a frequency-based vocabulary list from the CEFR-related texts labelled for levels. The lists will be ordered by lemmas and their parts-of-speech as a unique unit in the list. Previous attempts at generating learner-oriented frequency-based word lists have been made in the Kelly project (2009-2011, http://www.kellyproject.eu/), an EU-funded project on building learner-oriented frequency-based monolingual and bilingual word lists for 9 languages intended to be used in a commercial language learning tool (Volodina & Johansson Kokkinakis 2012; Kilgarriff et al., forthcoming; keewords.com). In the Kelly project, target vocabulary has been collected from a large web-corpus of written language used on the web. The basis of the Kelly list is the general-purpose vocabulary, providing the range of both lexical and grammatical elements as specified in the CEFR (Council of

Europe, 2001:110-111). However, during the post-Kelly period we have observed the need for additional modifications: (1) the list needs to be validated against the reading materials used in the CEFR-based courses, to make sure that vocabulary in the list is correctly streamed into CEFR levels; (2) we need to fill in the gaps in relevant vocabulary, for example, missing lexical items like “toothpaste”, “toothbrush”, etc. that clearly need to be present in the learner-oriented vocabulary lists, but do not gain any prominent place in the frequency lists generated from written native speaker corpora. We thus need to analyze which vocabulary should be added, removed or relocated in the list with regard to the CEFR guidelines based on the evidence of materials used in the real-life CEFR-based courses; and (3) we need to look specifically into the domain-specific vocabulary according to the CEFR themes – which words, which levels, how many per level – and evaluate if domain vocabulary should be included into the Kelly list or should be available as a “satellite list” following the implicit indications in the CEFR (Council of Europe, 2001:52-53).

The suggested approach will help us identify (concrete) lexical curriculum for CEFR-based courses in Swedish, both in terms of *what* words and *how many* per level a student of each level should acquire. The resulting list will be used primarily as an instrument for training vocabulary in the Lärka-based exercise generator. Apart from this, the list can be used for testing authentic examples (e.g. texts and sentences) for appropriateness for learners of different proficiency levels; for assessment of language proficiency in L2 learner language production, etc. The crucial prerequisite for this sub-project is access to an *annotated corpus* containing texts labeled for proficiency levels, the gold standard described in the next session.

2.4.2 Sentence readability

The degree to which a text can be understood by a human reader is referred to as its *readability* (Kate et al., 2010). In the second language context, readability corresponds to the extent to which learners are able to understand a text at a certain proficiency level.

Texts and sentences can be mapped to a corresponding level with the use of a measure based on statistical information about different linguistic properties of a text. Traditional readability measures, such as Flesh-Kincaid, Dale-Chall etc. for English, and LIX (Läsbarthetsindex) for Swedish, however, are limited to surface text features such as sentence length and the number of syllables (Heilman et al., 2007; Heimann Mühlenbock, 2013). Moreover, they consider text readability from a first language point of view and focus at the text level, and thus have shortcomings when used on very short passages (Kilgarriff et al., 2008) or when applied to L2 contexts (Beinborn et al., 2012).

Second language teachers and writers of teaching materials often need to make human judgments about readability at both text and sentence level, but recent NLP research started to explore automated techniques for this task, which combine syntactic and lexical information with *machine learning methods*. An important first step in most machine learning-based readability

methods is a sufficient amount of *annotated training data* containing texts labeled with a corresponding level. Then, a number of features, i.e. information and characteristics of the text that one wishes to take into consideration, should be selected (Collins-Thompson & Callan, 2005; Tanaka-Ishii et al., 2010). Finally, these features need to be mapped to a readability level (or score) with a machine learning algorithm. Hybrid approaches combining rule-based, statistic and machine learning methods are also explored in the area of text readability in L2 context (François & Miltsakaki, 2012).

The sentence readability project for Swedish is currently under development (Volodina et al., 2012; Pilán et al., forthcoming). It has arisen in response to the need for a reliable algorithm for classification of sentences into appropriate CEFR-levels in Lärka context.

This project was initially focused on general ranking of corpus hits according to their "appropriateness" (Volodina et al., 2012b). The aim has gradually evolved and eventually crystallized into finding an NLP-based algorithm to predict which lexical, morpho-syntactic and possibly other linguistic elements which students are able to understand at a certain language learning level (Pilán et al., forthcoming).

This project builds upon experimenting with both manually weighted heuristic rules, as well as with machine learning techniques. During the selection of parameters and features not only superficial readability criteria such as sentence and word lengths are taken into consideration, but also deeper linguistic aspects from a second language teaching perspective (part-of-speech, depth of dependencies etc.). The manually set parameters are tested with different thresholds and weights until optimized for a certain CEFR level (see Figure 4). However, to know that the parameter setting is optimal, we need access to experienced teachers who can assess the result (a kind of crowdsourcing), or an open-source collection of sentences labelled for levels to test the prediction accuracy of heuristic rules.

The machine learning part involves supervised techniques to classify the difficulty level of sentences, the training data being a corpus based on second language teaching materials, labeled with CEFR levels, currently available only for B1 and B2 levels. Depending on the outcome of the experiments and users' preferences, the sentence retrieval process could be fully automatic (based only on the trained model), semi-automatic (with a combination of manual parameters and the trained model) or only manual, so that selection of sentences can be fully customized according to specific needs of teachers and students.

The collection of texts labelled for CEFR levels provides, thereby, a number of opportunities to solve the challenges we face. Moreover, the availability of the training data in question labelled for additional text variables, i.e. not only for CEFR levels but also for topics, genres, etc. can facilitate other research projects relevant for ICALL, for example automatic selection of appropriate texts for the target proficiency level, automatic retrieval of topical texts, automatic question generation, to name just a few.

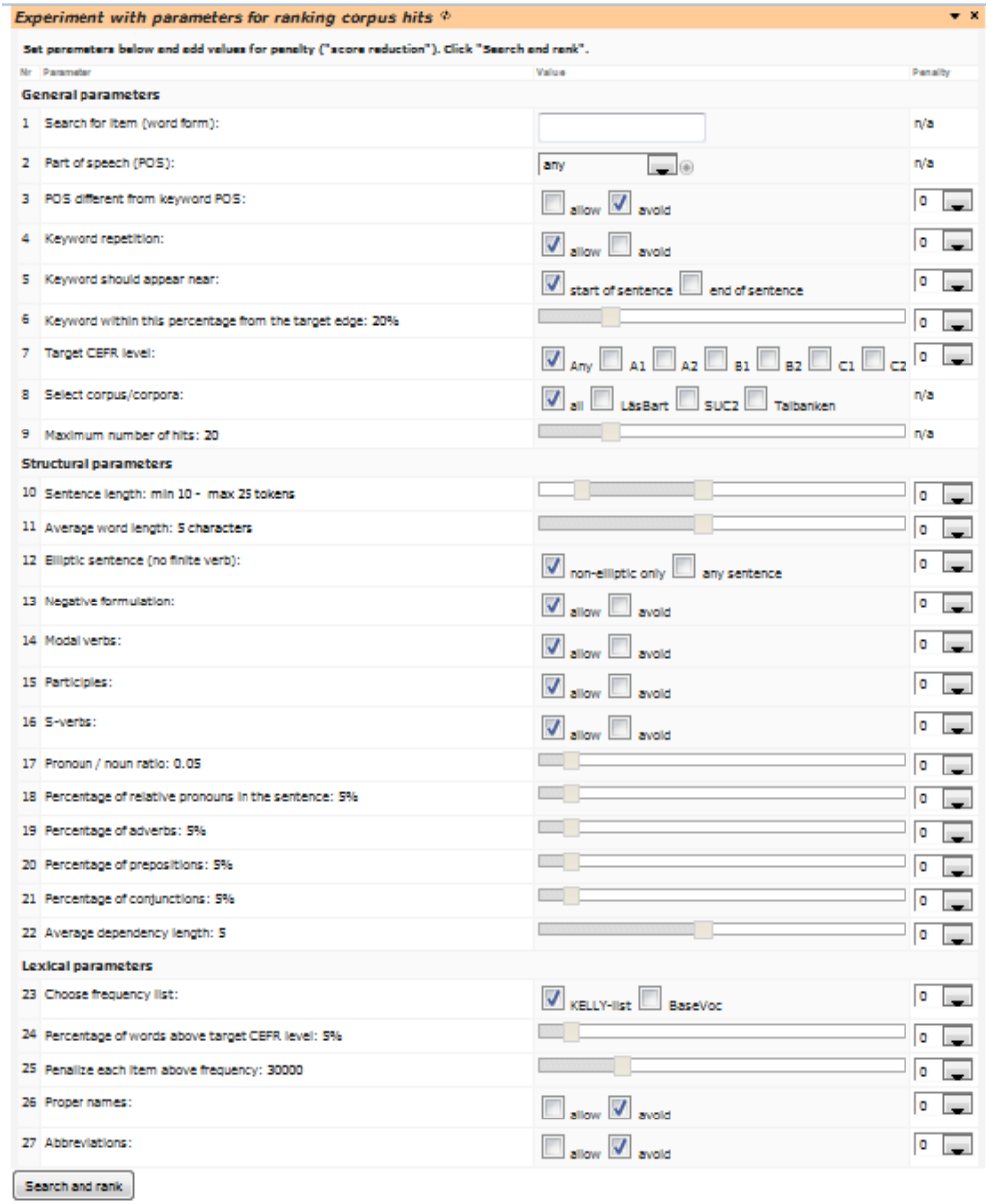

FIGURE 4. LINGUISTIC PARAMETERS FOR SENTENCE READABILITY, HEURISTIC RULES

3. Towards a corpus of CEFR-related course book texts

It is known to be rather controversial to break down CEFR “can-do” statements into concrete constituents, partly due to the “human factor”. Course material producers and teachers often go by their subjective “expert judgements” and intuitions, not necessarily agreeing with each other.

However, we take it for granted that teachers' interpretations of CEFR guidelines, subjective when taken individually, present an objective ground for generalizations and approximations about language complexity and level-wise content, when taken collectively. Therefore, we assume that, given texts used for CEFR-based courses from different authors and publishers, we can perform empirical evidence-based studies of a number of linguistic aspects expected of learners at different levels, for example vocabulary scope, most common grammar per level, text complexity, sentence complexity. Apart from that, we are interested in studying typical linguistic features for texts of different CEFR-based themes (topical domains).

Texts related to language learning fall into two categories: (1) “input” or normative texts provided by course book writers or selected by teachers; and (2) “output” or learner produced texts showing learner performance at the studied level. While learner output texts (not necessarily linked to CEFR levels, though) have been the object of study in different projects for both Swedish (Johansson Kokkinakis & Magnusson, 2011; Hultman & Westman, 1977; Nyström, 2000; Östlund-Stjärnegårdh, 2002) and other languages (Carlsten, 2012; Hawkins & Buttery, 2009), the study of normative course book texts from L2 perspective is rather rarely pursued (Lindberg & Johansson Kokkinakis, 2007, 2009; François & Miltsakaki, 2012). The main (hypothetical) reason for that is absence of accessible digitized data. In the project described in this section we describer our initial efforts at collecting normative texts to fill in the gap and to form the ground for CEFR-based text research for Swedish.

3.1 Collecting corpus materials

To identify relevant course materials, a number of teachers of CEFR-related courses have been interviewed and the relevant publishers have subsequently been contacted for electronic materials. However, texts in electronic format have proven to be rather difficult to obtain. Of all the contacted publishers only *Liber* has shown understanding and provided files for our research. To tackle the problem of lacking texts, we opted for an optical scanning approach subcontracting the relevant digitizing centre. The total amount of course books in pages is 3187; which corresponds to an estimated corpus size of approximately 3 million tokens.

Our pilot level has become B1, with 3 different course books, each containing mixed contents (e.g. half the book B1 level and half the book B2 level; or a part of the book A1/A2, the rest B1), totalling 565 pages.

3.2 Corpus annotation

Annotation of course book texts consists of the following two steps:

1. annotation for CEFR-relevant variables and
2. annotation for linguistic parameters.

We have annotated texts for CEFR-variables using an editor that we developed ourselves. We used Lärka as the basis for the editor. Figure 5

presents the course book editor view: the menu on the left inserts different tags into the text field; the field on the right keeps track of the ids used throughout the file.

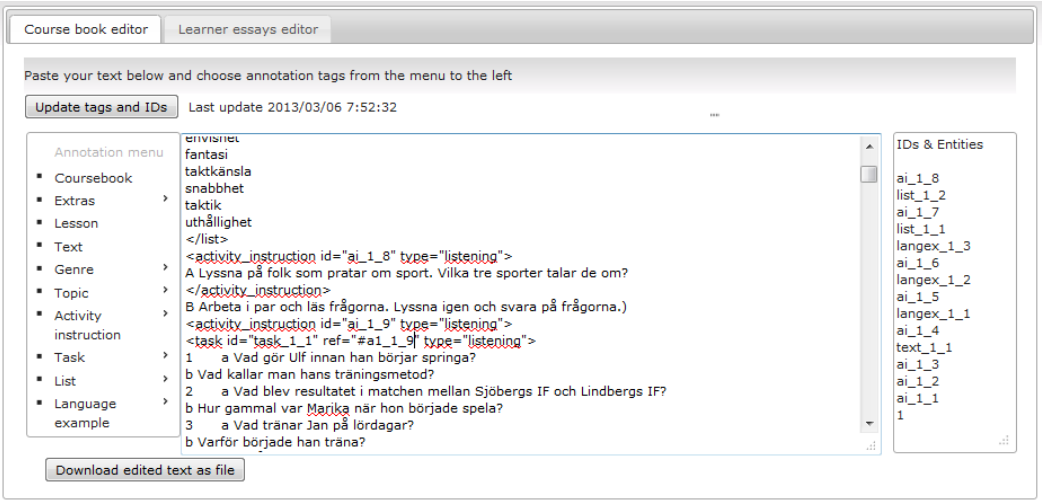

FIGURE 5. COURSE BOOK EDITOR DEVELOPED FOR THIS PROJECT

The taxonomy of text variables gives the key to different empirical and NLP-based studies. In our corpus, the text mass is divided into *Extras* (foreword, contents, acknowledgements, etc.) and *Lessons* (i.e. chapters). *Lessons*, further, contain different types of language and are subdivided into *Texts*, *Activity instructions, Tasks, Lists* and *Language examples*. A more fine-grained division of lesson-related text variables is shown in Figure 6.

Text genres is a modified version of genre families described in Martin & Rose (2008). The scheme over genre families has been extended by some macrofunctions according to the CEFR, e.g. *exposition, exegesis* (Council of Europe, 2001:126); as well as by the genre family marked as "*other*" which contains text types that we could not place in any of the main three families (narration, facts, evaluation). Among the a-typical (compared to Matin&Rose's genre families) text types are *puzzles, rhymes, lyrics, questionnaires, letters*, etc. The genre taxonomy is not final since we expect to encounter other deviating categories during the annotation work.

Topics have been derived from the CEFR document (Council of Europe, 2001:52). As with genres, we expect the list of topics to grow during the annotation period to cover the diversity of the topics in the course books.

Activity instructions usually precede the actual *Tasks* (e.g. exercises or text questions) and contain imperative sentences in the majority of cases. *Lists* provide active vocabulary for training or phrases/sentences to use during some tasks; whereas *Language examples* introduce new grammar or vocabulary patterns, that the learner should focus on and often contain explanations.

The division of the language used in *Lessons* into *Texts* and other categories is made to cater for different types of research that can be performed once the corpus is available. We plan, for example, to study the type of questions on different text genres to generalize about how questions differ in number and contents depending upon the genre and topic of the text, which will influence the question generation engine for that particular text genre.

Text parameters: Genre	Text parameters: Topic	Other types of text in lessons

Genre
• Narration
  • Personal story
  • Fiction
  • Description
  • News article
• Facts
  • Historical facts
  • Biography
  • Autobiography
  • Explanation
  • Instruction
  • Rules
  • Procedures
  • Report
  • Demonstration
• Evaluation
  • Argumentation
  • Exposition
  • Discussion
  • Personal reflection
  • Review
  • Interpretation, exegesis
  • Persuasion
• Other
  • Dialogue
  • Puzzle
  • Rhyme
  • Lyrics
  • Questionnaire
  • Letter
  • Language tip
	Topic
	• Personal identification
	• House and home, environment
	• Daily life
	• Free time, entertainment
	• Travel
	• Relations with other people
	• Health and body care
	• Education
	• Shopping
	• Food and drink
	• Services
	• Places
	• Languages
	• Weather
		Activity instruction
		• Listening
		• Reading
		• Writing
		• Speaking
		• Discussion
		• Grammar exercise
		• Vocabulary exercise
		• Text question
		Task
		• Listening
		• Reading
		• Writing
		• Speaking
		• Discussion
		• Grammar exercise
		• Vocabulary exercise
		• Text question
		• Gaps
		List
		• Vocabulary
		• Grammar
		• Sentences
		Language example
		• Vocabulary
		• Grammar
		• Pronunciation
		• Spelling
		• Writing

FIGURE 6. SUBMENUS OF THE MAIN ANNOTATION MENU FOR TEXT VARIABLES.

Once the *course book editor* is stable, it will be available for use for any other L2 language course book annotation, language independent. Since it is web-based, it can be accessed from anywhere without prior installation.

Annotation for linguistic variables includes annotation for parts of speech (pos), morpho-syntactic information (msd), syntactic relations (ref, dephead, deprel), lemmas, and linking to morphology lexicon (lex, saldo). This is an automated procedure that is used in Korp import pipeline (Borin et al. 2012), Korp being an infrastructure for storing and browsing a large collection of Swedish texts. Example of how a text can look after this annotation is given

in Figure 7. In the near future we plan to build infrastructure in Korp for working with CEFR-related variables.

```
<w pos="DT" msd="DT.UTR.SIN.IND" lemma="|en|" lex="|en..al.1|" saldo="|den..1|en..2|" prefix="|" suffix="|" ref="1" dephead="2" deprel="DT">En</w>
<w pos="NN" msd="NN.UTR.SIN.IND.NOM" lemma="|" lex="|" saldo="|" prefix="|exempel..nn.1|" suffix="|text..nn.1|" ref="2" dephead="3" deprel="SS">exempeltext</w>
```

FIGURE 7. EXAMPLE OF A TEXT ANNOTATED FOR LINGUISTIC VARIABLES

4. Concluding remarks

The problem of sparse data is well known in the area of computational linguistics, especially within machine learning, information extraction and other subfields that require reliable reference and training data, a “gold standard”, i.e. data that perfectly matches the purpose so that the instruments can be trained and fine-tuned on it. A collection of course book texts annotated for CEFR variables presented in this paper provides a unique training dataset for a variety of natural language processing tasks relevant for (but not limited to) ICALL, including topic modelling, genre identification, question generation and automatic classification of texts and sentences by their readability.

Access to such data in pedagogical empirical studies facilitates generalizations and approximations about language use in L2 context. With this project, we lay the ground for further pedagogically relevant studies of CEFR related texts in Swedish. The most important for us, however, is the fact that the access to this corpus is the only way to address the research agenda prompted by the development of the ICALL platform for Swedish.

The corpus based on course book texts cannot be made publicly available due to copyright restrictions. However, once the instruments for level classification and eventually topic categorization are reliable, it will be possible to classify arbitrary texts, e.g. texts available through Språkbanken's corpus infrastructure Korp (Borin et al., 2012) into CEFR levels and thematic domains. Since materials from Korp are digitally available, they will facilitate further studies of CEFR specific linguistic aspects per proficiency level. Text classification into levels and topics is eventually planned to be included into the standard annotation process for Korp for any new text collections.

Acknowledgements

The project on collecting CEFR-related texts has been financed partly by the Swedish Department at the University of Gothenburg (UGOT) and partly by Språkbanken, UGOT. Further, we extend our thanks to the publishing house *Liber* for providing electronic materials for our research.

References

Aldabe, I., Lacalle, M.L.D., Maritxalar, M., Martinez, E., Uria, L. (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In *Intelligent Tutoring Systems* (2006), 584-594

Amaral, L. & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL* 23(1): 4–24.

Amaral, L., Meurers, D. & Ziai, R. (2011). Analyzing learner language: towards a flexible natural language processing architecture for intelligent language tutors. *Computer Assisted Language Learning* 24(1): 1–16.

Beinborn, L., Zesch, T., & Gurevych, I. (2012). Towards fine-grained readability measures for self-directed language learning. In *Electronic Conference Proceedings* (Vol. 80, pp. 11-19).

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

Byrnes H. (2007). Perspectives. *The Modern Language Journal*, 91, iv, p.641–645.

Carlsten, C. (2012). Proficiency Level – a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics,* Volume 33(2), p.161-183

Collins-Thompson, K. & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13). pp. 1448-1462.

Collins-Thompson, K. and Callan, J. (2007). Automatic and Human Scoring of Word Definition Responses. *Proceedings of NAACL HLT 2007,* 476-483. Rochester, NY.

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

Council of Europe. 2009. *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR). A Manual*, Strasbourg: Language Policy Division.

Dávid, G.A. 2010. Linking the general English suite of Euro Examinations to the CEFR: a case study report. In Martyniuk, W. (Ed.) *Aligning Tests with the CEFR*. Cambridge University Press, p.177-203.

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

Francois, T. & Miltsakaki, E. (2012). Do NLP and Machine Learning Improve Traditional Readability Formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Population*,

NAACL

Hawkins, J. A. & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In Taylor, L. & Weir, C. J. (Eds). *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment,* 158-175. Cambridge: Cambridge University Press.

Heift, T. (2003). Multiple learner errors and meaningful feedback: A challenge for ICALL systems. *CALICO Journal*, 20(3), 533–548.

Heilman, M., Collins-Thompson, K., Callan, J. and Eskenazi, M. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of NAACL HLT 2007,* 460-467. Rochester, NY.

Heimann Mühlenbock, K. (2013). *I see what you mean: Assessing readability for specific target groups.* PhD Thesis. Data linguistica, University of Gothenburg.

Hultman, T. G. & Westman, M. (1977). *Gymnasistsvenska*. Lund: Liber Läromedel.

Johansson Kokkinakis, S. & Magnusson, U. (2011). Computer based quantitative methods applied to first and second language student writing. *Young urban Swedish. Variation and change in multilingual settings.*University of Gothenburg, 105-124.

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J. Roukos, S. & Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 546-554). Association for Computational Linguistics.

Khalifa, H., Ffrench, A. & Salamoura, A. 2010. Maintaining alighnment to the CEFR: the FCE case study. In Martyniuk, W. (Ed.) *Aligning Tests with the CEFR*. Cambridge University Press, p.80-101.

Kilgarriff A., Charalabopoulou F., Gavrilidou M., Bondi Johannessen J., Khalil S., Johansson Kokkinakis S., Lew R., Sharoff S., Vadlapudi R, Volodina E. (accepted, LREJ 2013). Corpus-Based Vocabulary lists for Language Learners for Nine Languages. *Language Resources and Evaluation Journal*, special issue.

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

Knoop, S. & Wilske, S. (2013). Automatic Generation of Gap-Filling Vocabulary Exercises for Mobile Learning. 2nd workshop on NLP in Computer-Assisted Language Learning. *Proceedings of the NODALIDA 2013 workshop on NLP for CALL*. Linköping Electronic Conference Proceedings 85.

Källgren, G., Gustafson-Capková, S. and Hartmann, B. (2006). *Manual of the*

Stockholm Umeå Corpus version 2.0. Department of Linguistics, Stockholm University.

Lindberg, I. & Johansson Kokkinakis, S. (2007). *OrdiL - en korpusbaserad kartläggning av ordförrådet i läromedel för grundskolans senare år.* Göteborgs universite

Lindberg, I. & Johansson Kokkinakis, K. (2009). Word Type Grouping in Swedish Secondary School Textbooks - An Inventory of Words from a Second Language Perspective *Multilingualism, Proceedings of the 23rd Scandinavian Conference of Linguistics*. 337-339

Little D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. *The Modern Language Journal* 91, p.645–655.

Little D. (2011). The Common European Framework of Reference for Languages: A research agenda. *Language Teaching,* Vol 44.3, p.381-393. Cambridge University Press 2011.

Martin, J.R. & Rose, D. (2008). *Genre Relations.* Equinox Publishing Ltd.

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V. & Ott, N. (2010. Enhancing Authentic Web Pages for Language Learners. *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT 2010, Los Angeles.*

Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Toronto: Multilingual Matters.

Nagata, N. 2009. Robo-Sensei‘s NLP-based error detection and feed-back generation. *CALICO Journal*, 26(3), 562–579.

Nivre, J., Nilsson, J. and 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)* Genoa: ELRA. 1392-1395.

North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal* 91, p.656–659.

Nyström, C. (2000). *Gymnasisters skrivande. En studie av genre, textstruktur och sammanhang*. Uppsala: Uppsala universitet.

Pijetlovic, D. & Volodina, E. (forthcoming). Developing a Swedish spelling game on an ICALL platform. *Proceedings of EuroCALL 2013*.

Pilán, I., Volodina, E. & Johansson, R. (forthcoming). Automatic selection of suitable sentences for language learning exercises. *Proceedings of EuroCALL 2013*.

Szabó, G. 2010. Relating language examinations to the CEFR: ECL as a case study. In Martyniuk, W. (Ed.) *Aligning Tests with the CEFR*. Cambridge University Press, p.133-144.

Tanaka-Ishii, K., Tezuka, S., & Terada, H. (2010). Sorting texts by readability.*Computational Linguistics*, *36*(2), 203-227.

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

Toole, J. & Heift, T. (2002). Task-Generator: A Portable System for Generating Learning Tasks for Intelligent Language Tutoring Systems. *Proceedings of ED-MEDIA 02, World Conference on Educational Multimedia, Hypermedia & Telecommunications*, Charlottesville, VA: AACE: 1972-1978.

Volodina, E. and Borin, L. (2012). Developing a freely available web-based exercise generator for Swedish. *CALL: Using, Learning, Knowing. EuroCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings.* Eds. Linda Bradley and Sylvie Thouësny. Research-publishing.net, Dublin, Ireland.

Volodina, E., Borin, L., Loftsson, H., Arnbjörnsdóttir, B. & Örn Leifsson, G. (2012a). Waste not, want not: Towards a system architecture for ICALL based on NLP component re-use. Workshop on NLP in Computer-Assisted Language Learning. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 47-58.

Volodina, E., Johansson, R. & Johansson Kokkinakis, S. (2012b). Semi-automatic selection of best corpus examples for Swedish: Initial algorithm evaluation. Workshop on NLP in Computer-Assisted Language Learning. *Proceedings of the SLTC 2012 workshop on NLP for CALL.* Linköping Electronic Conference Proceedings 80: 59–70.

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

Westhoff G. (2007). Challengens and Opportunities of the CEFR for Reimagining Foreign Language Pedagogy. *The Modern Language Journal* 91, p.676–679.

Östlund-Stjärnegårdh, E. (2002). *Godkänd i svenska? Bedömning och analys av gymnasieelevers texter.* Uppsala: Uppsala universitet.