Language Testing and the CEFR: Time for a New Framework?

27–28–29 May 2013, Universiteit Antwerpen, Belgium

Dear author(s)

Congratulations! Your submission for our conference has been accepted as a paper presentation.

Please fill in the fields below, save it as **LT&CEFR_firstname_lastname**.rtf and send it to before **22 March 2013**. Please do not change the formatting and style of this document. The best short paper will be selected as plenary presentation.

CONFERENCE TOPICS
- **Competence and performance**
  What is the link between 'can do' performance statements and areas of linguistic knowledge? To what extent can or should the levels be made more explicit in terms of required vocabulary and grammar?
- **Degree of difficulty of the levels**
  How can we make sure that our examinations are measuring at the CEFR levels we claim they are? What evidence do we have to support our claims?
- **Test purpose**
  Why are we testing? What kind of decisions will be made on the basis of information collected via the test? What will be the consequences of these decisions?
- **Practicality**
  How do we link our tests to the CEFR? How practical, applicable and operational is the CEFR for concrete language testing situations?

Elena Volodina, Sofie Johansson Kokkinakis

University of Gothenburg, Gothenburg, Sweden

elena.volodina@svenska.gu.se, sofie.johansson.kokkinakis@svenska.gu.se

Compiling a corpus of CEFR-related texts

Bio data

Elena Volodina received PhD in Linguistics in Moscow, Russia (1998) and MA in Computational Linguistics in Gothenburg, Sweden (2008). She works as a research engineer at Gothenburg University since 2010, her primary interests being in Intelligent Computer-Assisted Language Learning, text and sentence readability, Corpus Linguistics.

Sofie Johansson Kokkinakis received PhD in Language Technology (2005) and MA in Computational Linguistics, both at the University of Gothenburg, Sweden. She currently works as a Researcher in language technology and the Director of the Institute for Swedish as a Second Language, Department of Swedish, University of Gothenburg. Her interests comprise computer-based lexical and content related analysis of texts (in particular secondary school text books and student written texts), lexical profiling and readability, genre specific texts and academic language use, morpho-syntactic and semantic analysis, computer-based assessment of language skills, corpus linguistics and learner corpora, and intelligent computer-assisted language learning

Abstract

This paper reports on initial efforts to compile a corpus of course book texts used for teaching CEFR-based courses of Swedish to adult immigrants. The research agenda behind compiling such a corpus comprises the study of normative "input" texts that can reveal a number of facts about what is being taught in terms of explicit grammar, receptive vocabulary, text and sentence readability; as well as build insights into linguistic characteristics of normative texts which can help anticipate learner performance in terms of active vocabulary, grammatical competence, etc. in classroom and testing settings.
The CEFR "can-do" statements are known to offer flexibility in interpreting them for different languages and target groups. However, they are nonspecific and therefore it is difficult to associate different kinds of competences and levels of accuracy learners need in order to perform the communicative tasks with the different CEFR levels. To address this problem a systematic study needs to be performed for each individual language, both for "input" normative texts and "output" learner-produced texts. In this project we take the first step to collect and study normative texts for Swedish.
The article describes the process of corpus compilation, annotation scheme of CEFR-relevant parameters, and methods proposed for text analysis, namely statistic and empiric methods, as well as techniques coming from computational linguistics/machine learning.

Short paper

1. Introduction

Since the acceptance of Common European Framework of References for Languages (CEFR) in 2001 (Council of Europe, 2001) many countries inside and outside Europe have abandoned previous practices in language teaching and assessment in favour of the CEFR. The CEFR scale, consisting of 6 proficiency levels, is described intentionally vaguely to cater for the diversity of different languages. As a consequence, there are voices among researchers and educators demanding explicit interpretation of each proficiency level for each individual language in terms of required vocabulary scope, grammatical competence, etc. (Byrnes 2007; Little 2007; Little 2011; Milton 2009; North 2007; Westhoff 2007).

It is known to be rather controversial to break down the 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 approximation*s about language complexity and level-wise content, *when taken collectively*. Thus, we assume that, given texts used for CEFR-based courses, we can perform empiric 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).

Background

Texts related to CEFR-based language learning fall into two categories as shown in figure 1: (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.

Figure 1. Texts in L2 context

The study of learner produced language is a large and active area of research in second language learning (Johansson Kokkinakis & Magnusson, 2011; Hultman & Westman, 1977; Nyström, 2000; Östlund-Stjärnegårdh, 2002). In Sweden, as far as we know, most research in this area is conducted with respect to language development theories, such as "the processability theory" (Pienemann 1998). However, since CEFR is widely spread in everyday practice, there is a need for CEFR-based analysis of learner language as well. Examples of projects devoted to CEFR-based studies of learner-produced language for other languages than Swedish are given under the SLATE research network (Carlsten, 2012; Hawkins & Buttery, 2009;more under <http://www.slate.eu.org/>).

In contrast to research within learner-produced language, we are not aware of any active studies performed on normative texts used in CEFR based courses or on correlation between normative texts and learner production, in spite of the fact that teachers, researchers and language assessors keep expressing the need for formalizing CEFR descriptors in terms of concrete grammar and vocabulary syllabus. In the project described in this paper we aim at collecting normative texts to fill in the gap and to form the ground for that kind of studies for Swedish.

2. Why compile this corpus?

Given the availability of electronic resources of the above-mentioned kind, we can engage in a number of important and interesting from the research point of view studies, eventually useful even outside research circle. For example, using a combination of statistic and empirical methods, as well as methods derived from computational linguistics (e.g. machine learning), we can study features characteristic of different CEFR levels. The possible outputs of such studies comprise (1) an instrument for automatic classification of texts by CEFR level based on text readability; (2) an instrument for automatic classification of sentences by CEFR level based on sentence readability; (3) an instrument for automatic classification of texts into thematic domains based on machine learning approach; (4) receptive vocabulary scope per proficiency level; and (5) receptive grammar scope per level. The main question are, then:

- which linguistic aspects are most important at each particular CEFR level, and why (at sentence and text levels individually); how the identified linguistic aspects match the "can-do" descriptors;

- which linguistic features are characteristic of texts of different thematic domains; and how such texts can be automatically identified;

- which words and how many per proficiency level are important to learn;

- which grammar students are most exposed to during the course of studies.

The studies based on the corpus may help us answer some of the questions often raised in the CEFR-based language testing context, for example, if there is a link between 'can do' performance statements and areas of linguistic knowledge; or to what extent the levels can be made more explicit in terms of required vocabulary and grammar. We view our study as an evidence-based interpretation of vague CEFR descriptors for different levels into concrete linguistic constituents based on expert interpretations of many experienced teachers and course book writiers.

3. Corpus compilation: first experiences

3.1 Interviews with teachers and publishers

Course materials are often copyrighted by publishing houses and cannot be freely obtained, to say nothing of being freely distributed in electronic format. To identify relevant course materials, a number of teachers of CEFR-related courses have been interviewed. It has turned out the materials available in the form of course books only in few cases implicitly mention European framework (Levy Scherrer & Lindemalm, 2008; Levy Scherrer & Lindemalm, 2009; Göransson & Parada, 2010; Göransson et al., 2010; Folkuniversitetets förlang, 2007; Åström, 2011; Åström, 2012; Trevisani, 2011); whereas a number of course books do not provide any indication what level according to the European scale of references they are aimed at, but have been interpreted by teachers as appropriate ones at certain levels (Holm et al., 2001; Ballardini, 2001a, 2001b; Risérus et al., 2002;).

All 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. The total amount of coursebooks in pages is 3187; which corresponds to an estimated size of approximately 3 mln tokens.

3.2 Optical scanning and its challenges

We have subcontracted scanning to a relevant digital centre. Our "pilot" level has become B1, with 3 different coursebooks, each containing mixed contents (e.g. half the book B1 and half the book B2; or a part of the book A1/A2, the rest B1), totalling 565 pages.

During this stage we have encountered a number of challenges. Without getting in to the details of digital document analysis or techniques for optical text recognition, (which is a separate research ares, see for example I*nternational Journal on Digital Analysis and Recognition* <http://www.springer.com/computer/image+processing/journal/10032>) we describe here what we have encountered practically.

First of all, book availability has caused some problems. Since some of the books are rather old, e.g. from 2001, they are neither sold via book stores, nor are they available at the library. In some cases we could find copies from teaching staff, but often they contained scribbles that interfered with scanning.

Second, it is a challenge to scan correctly texts that are diagonally placed, as in figure 2.
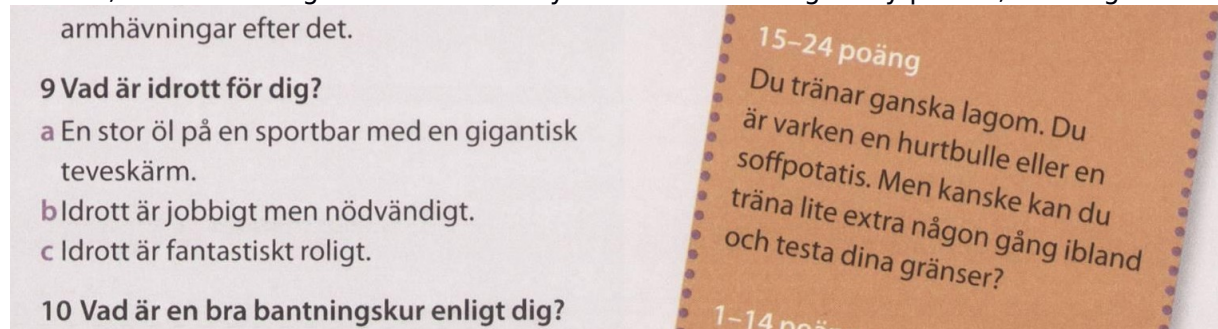
Figure 2. Example of a diagonally placed text.

The extracted text from the scanned document looked like that:
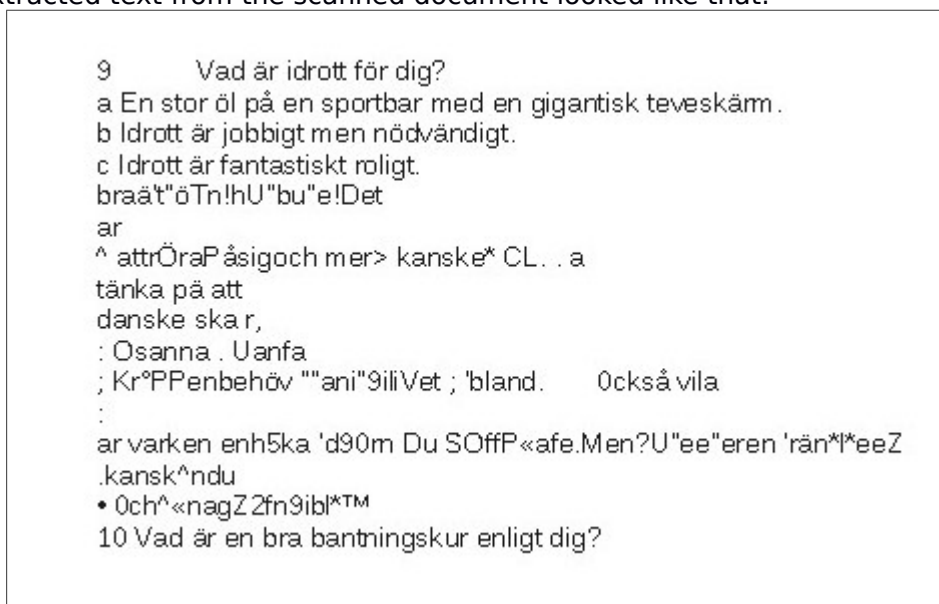
Figure 3. Result of an optical scanning of a diagonally placed text.

Starting with line 5 and till the last line but one (figure 3) there is a lot gibberish. Some of the words or phrases coincide with the phrases in the diagonally placed text but very inconsistently.

Texts given in several blocks or tables (as shown in figure 4) present a problem of texts interfering with each other (figure 5).
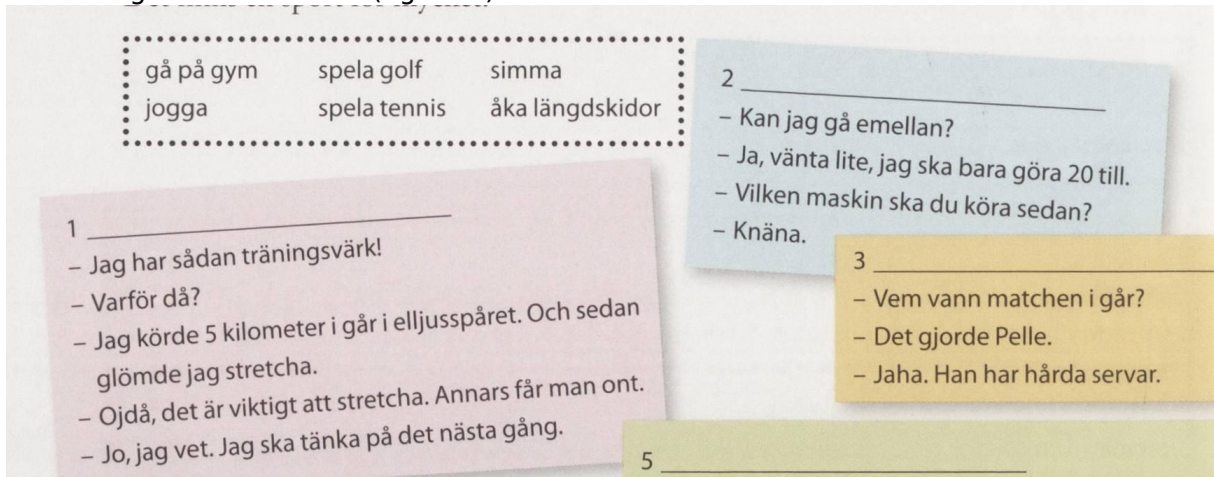

Figure 4. A view of several texts placed in the form of "table" and therefore setting risk to have texts bumping into each other.
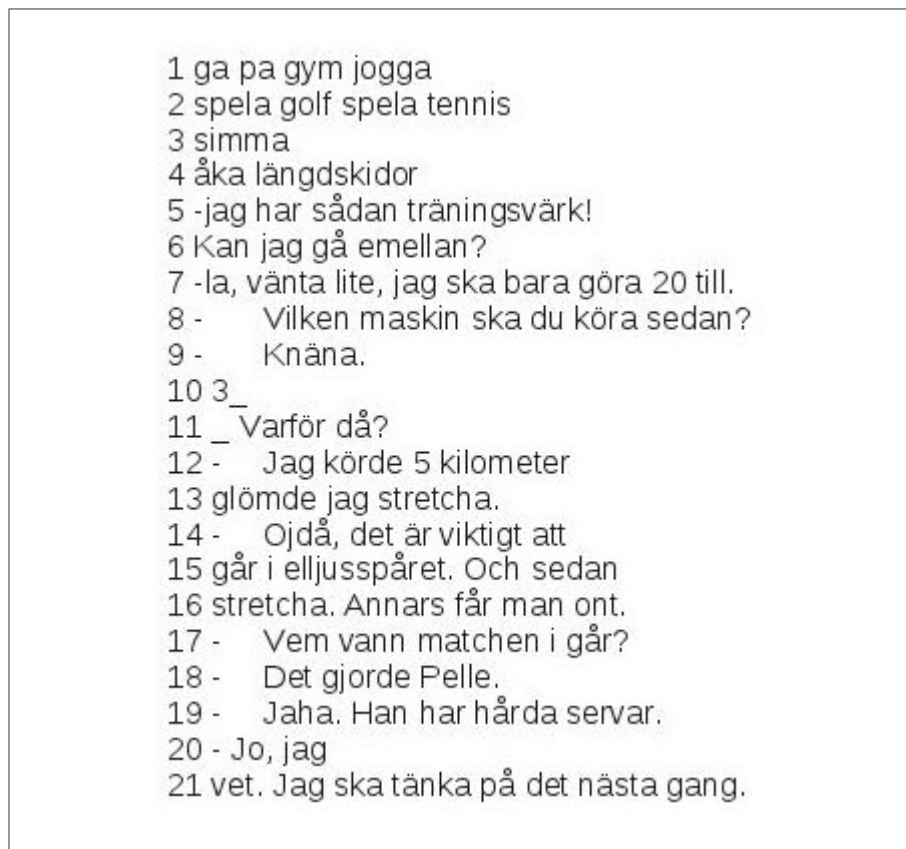

Figure 5. Result of optical scanning of text presented in figure 4.
Line numbers added for easier interpretation.

As can be seen from figure 5, lines 1-4 represent the word list in figure 4; line 5 starts dialogue nr.1, whereas lines 6-9 refer to dialogue nr.2, lines 11-16 continue dialogue nr.1, though in a scrambled order. The correct order should be (given here in line numbers): 11, 12, 15, 13, 14, 16.

We made a decision to ignore texts that haven't been correctly scanned unless it demands little effort to restore the correct text. We have therefore lost a bit of text mass during the post-scanning step.

3.3 Annotation

Coursebook texts annotation consists of two steps:

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

3.3.1 Annotation for CEFR variables

We used Lärka, the ICALL[1] platform for Swedish (Volodina & Borin 2012), as the basis for the editor. Figure 6 presentes the course book editor view:
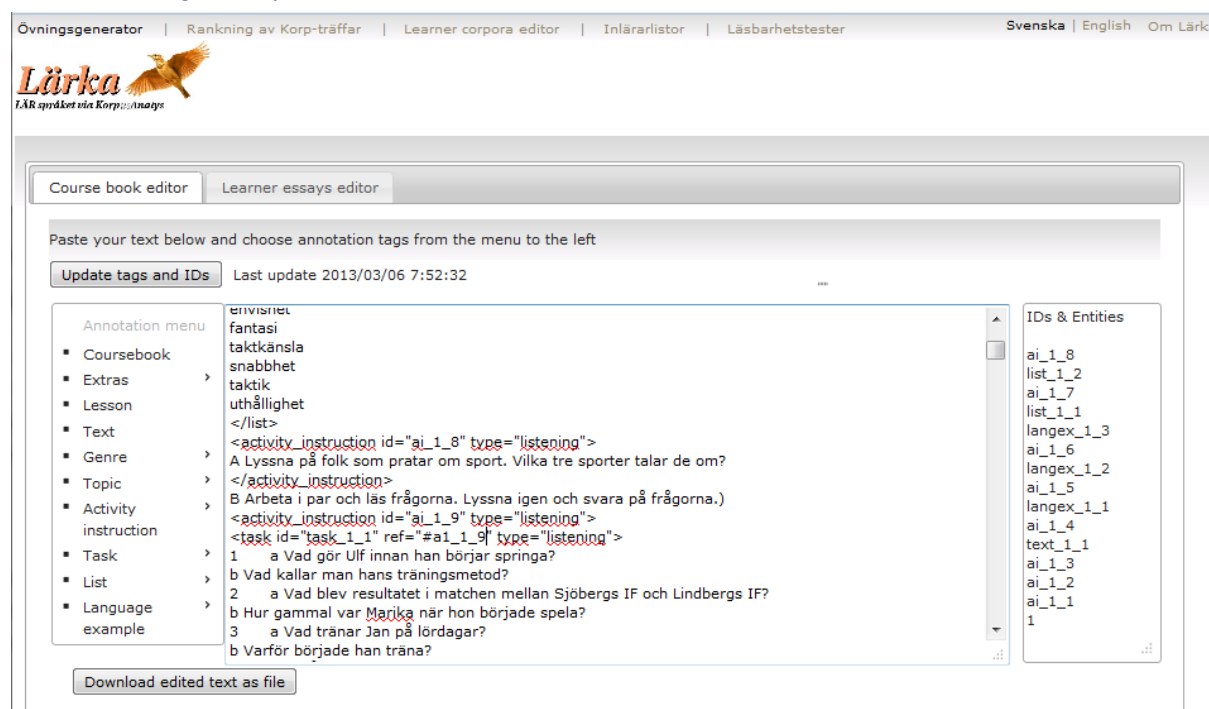
Figure 6. *Course book editor* view in Lärka

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.

The most interesting from CEFR point of view is the taxonomy of text variables. We have divided the text mass in course books into "extras" (foreword, contents, acknowledgements, etc.) and lessons (i.e. chapters). Lessons contain different types of learner-interesting language that we have divided into *texts, activity instructions, tasks, lists* and *language examples*. A more fine-grained division is shown in figure 7.

---

1   ICALL – Intelligent Computer-Assisted Language Learning

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

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 7. Submenus of the main annotation menu
for genre, topic, activity instruction, task, list and language example

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

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.

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

3.3.2 Annotation for linguistic variables

Annotation for linguistic variables includes annotation for parts of speech (*pos* in figure 8), 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[2] import pipeline (Borin et al. 2012). Example of how a text can look after this annotation is given in figure 8.

```
<w pos="PS" msd="PS.UTR.SIN.DEF" lemma="|jag|" lex="|jag..pn.1|" saldo="|jag..1|"
prefix="|" suffix="|" ref="01" dephead="02" deprel="DT">Min</w>

<w pos="NN" msd="NN.UTR.SIN.IND.NOM" lemma="|kompis|" lex="|kompis..nn.1|"
saldo="|kompis..1|" prefix="|komp..nn.1|kompa..vb.1|" suffix="|is..nn.1|" ref="02"
dephead="03" deprel="MS">kompis</w>

<w pos="KN" msd="KN" lemma="|och|" lex="|och..kn.1|" saldo="|och..1|" prefix="|"
suffix="|" ref="03" dephead="" deprel="ROOT">och</w>

<w pos="PN" msd="PN.UTR.SIN.DEF.SUB" lemma="|jag|" lex="|jag..pn.1|" saldo="|
jag..1|" prefix="|" suffix="|" ref="04" dephead="05" deprel="SS">jag</w>

<w pos="VB" msd="VB.PRS.AKT" lemma="|ha|" lex="|ha..vb.1|" saldo="|ha..1|"
prefix="|" suffix="|" ref="05" dephead="03" deprel="MS">har</w>

<w pos="VB" msd="VB.SUP.AKT" lemma="|känna|" lex="|känna..vb.2|känna..vb.1|"
saldo="|känna..2|känna..4|känna..1|känna..3|" prefix="|" suffix="|" ref="06"
dephead="05" deprel="VG">känt</w>
```

Figure 8. Example of annotation for linguistic variables

4 Intended corpus use

Special efforts have been undertaken to interpret CEFR guidelines as sets of Reference Level Descriptions[3] as well as to establish procedures to relate language exams to the CEFR (Council of Europe, 2009; Khalifa et al., 2010; Szabó, 2010; Dávid, 2010; Jones et al., 2010), but to the best of our knowledge that has not yet been done for Swedish.

The availability of electronic resources of the described type opens an opportunity to engage in an *evidence-based interpretation of the CEFR descriptors*. "Evidence-based" in the context of this project is understood as course book materials collected into a linguistically annotated corpus. They present an evidence of conscience expert interpretations of CEFR guidelines into concrete samples of teaching material.

To address the problem of non-specificity of the CEFR descriptors for different levels of language proficiency, a systematic study needs to be performed for each individual language, both for "input" normative texts and "output" learner-produced texts. Attempts at *aligning texts and tests with CEFR* are ongoing (Khalifa et al., 2010; Szabo, 2010; Dávid, 2010; Jones et al., 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 texts/language samples, interpreting the CEFR descriptors on the way. This process consists of four procedures according to the Manual (Council of Europe, 2009): familiarization, specification, standardization, and empiric validation. We suggest a

---

2 Korp – an infrastructure for storing and browsing a large collection of Swedish texts (Borin et al. 2012); www.spraakbanken.gu.se/korp
3 http://www.coe.int/t/dg4/linguistic/dnr_en.asp

bottom-up approach, where we start from the actual language samples labeled for levels, i.e. preselected reading materials for different levels, analyze them for linguistic constituents with the help of machine learning algorithms and then try to map the identified constituents to the CEFR descriptors. The two approaches should be viewed as complementary rather than exclusive of each other.

Once ready, the collection of normative texts introducd in section 3 can be studied internally to generate an instrument that can reliably classify any arbitrary Swedish text by its appropriate CEFR level and domain. Availability of the corpus will also make it possible to identify receptive vocabulary and grammar scope per proficiency level.

Figure 9. Linguistic parameters for sentence and text analysis

The first use of the corpus is planned for an internal project that will help us identify an automatic approach to the readability assessment of Swedish sentences in the L2 context (Pilan et al., forthcoming). More concretely, the aim is to create an algorithm which would try to predict at which language learning level students are able to understand sentences containing certain lexical, syntactical, morphological and other linguistic elements. This approach is a combination of evidence-based empiric methods combined with statistical and machine-learning techniques and leads us to the explicit mapping between required vocabulary, grammar and syntax and the reached CEFR levels; the identified linguistic parameters can be further connected to the level-wise 'can-do' statements.

The linguistic parameters we have selected so far for scrutiny are presented in the left column of figure 9. We initially plan to study A1, A2, B1 and B2 course book texts in contrast to non-restricted texts used for native speakers coming from generic balanced corpora of Swedish. This will show us how the linguistic features in figure 9 are distributed in normative texts of different proficiency levels.

The same type of study is planned for text-long contexts at different levels.

As a further step we intend to collect a corpus of student essays written at different CEFR levels and compare linguistic features used in normative texts, i.e. the ones that learners are expected to cope with receptively when using course books, versus learner-produced texts, showing how these features are reflected in their productive use.

5. Concluding remarks

In this paper we have presented our initial work on compiling and annotating a corpus of CEFR-based course book texts, and outlined the prospects of its usage for CEFR-based pedagogical studies. This kind of data labeled for CEFr levels, topical themes, etc. is critical for pedagogical empirical studies like the ones proposed above since it facilitates conclusions, 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.

Acknowledgements

References

Ballardini, K., Stjärnlöf, S. & Viberg, Å. (2001a). *Nya Mål: svenska som andraspråk. 1.* Natur & Kultur.

Ballardini, K., Stjärnlöf, S. & Viberg, Å. (2001b). *Nya Mål 2 Lärobok.* Natur & Kultur.

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.

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.

Folkuniversitetets förlag. (2007). *På svenska! 3 lärobook.* Folkuniversitetets förlag.

Göransson, U. & Parada, M. (2010). *På svenska! 1 lärobok.* Folkuniversitetets förlag.

Göransson, U., Helander Annika & Parada, M. (2010). *På svenska! 2 lärobok.* Folkuniversitetets förlag.

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.

Holm, B., Nyborg, R. & Pettersson N-O. (2001). *Svenska utifrån: Lärobok i svenska.* Svenska institutet.

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

Khalifa, H., ffrench, A. And 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.

Johansson Kokkinakis, S. and 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.

Jones, N., Ashton, K. and Walker, T. (2010). Asset Languages: a case study of piloting the CEFR Manual. In Martyniuk, W. (Ed.) *Aligning Tests with the CEFR*. Cambridge University Press, p.227-246.

Levy Scherrer, P. & Lindemalm, K. (2008). *Rivstart: A1+A2 Textbok*. Natur & Kultur.

Levy Scherrer, P. & Lindemalm, K. (2009). *Rivstart: B1+B2 Textbok*. Natur & Kultur.

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.

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

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.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. Amsterdam: John Benjamins.

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

Risérus, H., Sandahl, Å. & Stjärnlöf, S. (2002). *Nya mål 3.* Natur & Kultur.

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.

Trevisani, M. (2011). *Avancera skriv: skrivövningar*. Liber.

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

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

Åström, M. (2011). *Språkporten Bas.* Studentlitteratur AB.

Åström, M. (2012). *Språkporten 1, 2, 3 Textbok med webbdel: Svenska som andraspråk.* Studentlitteratur AB.

Östlund-Stjärnegårdh, E. (2002). *Godkänd i svenska? Bedömning och analys av gymnasieelevers texter.* Uppsala: Uppsala universitet. Carlsten, C. (2012). Proficiency Level – a Fuzzy Variable in Computer Learner Corpora. *Applied Linguistics,* Volume 33(2), p.161-183