# You get what you annotate:

# a pedagogically annotated corpus of coursebooks for Swedish as a Second Language

*Elena Volodina[1], Ildikó Pilán[1], Stian Rødven Eide[2], Hannes Heidarsson[3]*

(1) Swedish Language Bank, Department of Swedish, University of Gothenburg, Sweden
(2) Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden
(3) Department of Swedish, University of Gothenburg, Sweden

elena.volodina@svenska.gu.se, ildiko.pilan@svenska.gu.se,
stian@fripost.org, hannes.heidarsson@live.se

ABSTRACT

We present the COCTAILL corpus, containing over 700.000 tokens of Swedish texts from 12 coursebooks aimed at second/foreign language (L2) learning. Each text in the corpus is labelled with a proficiency level according to the CEFR proficiency scale. Genres, topics, associated activities, vocabulary lists and other types of information are annotated in the coursebooks to facilitate Second Language Acquisition (SLA)-aware studies and experiments aimed at Intelligent Computer-Assisted Language Learning (ICALL). Linguistic annotation in the form of parts-of-speech (POS; e.g. nouns, verbs), base forms (lemmas) and syntactic relations (e.g. subject, object) has been also added to the corpus.

In the article we describe our annotation scheme and the editor we have developed for the content mark-up of the coursebooks, including the taxonomy of pedagogical activities and linguistic skills. Inter-annotator agreement has been computed and reported on a subset of the corpus. Surprisingly, we have not found any other examples of pedagogically marked-up corpora based on L2 coursebooks to draw on existing experiences. Hence, our work may be viewed as “groping in the darkness” and eventually a starting point for  others.

The paper also presents our first quantitative exploration of the corpus where we focus on textually and pedagogically annotated features of the coursebooks to exemplify what types of studies can be performed using the presented annotation scheme. We explore trends shown in use of topics and genres over proficiency levels and compare pedagogical focus of exercises across levels.

The final section of the paper summarises the potential this corpus holds for research within SLA and various ICALL tasks.

KEYWORDS: L2 coursebook corpus, annotation scheme, CEFR proficiency levels, SLA-aware ICALL, inter-annotator agreement

Elena Volodina, Ildikó Pilán, Stian Rødven Eide and Hannes Heidarsson 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. *Proceedings of the third workshop on NLP for computer-assisted language learning.* NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107: 128–144.

# 1 Background

## 1.1 Corpora in CALL and ICALL

Corpora have become a useful and often central component in Computer-Assisted Language Learning (CALL) applications and especially in Intelligent CALL, i.e. CALL based on Natural Language Processing and Speech Technologies. Primarily, corpora of two types are being employed in such applications: native speaker (NS) corpora (e.g. Vajjala & Meurers, 2013) and corpora consisting of L2 learner production, such as essays (e.g. Hancke & Meurers, 2013). In both cases variation can be observed in the mode of language, i.e. written vs spoken language. NS corpora are primarily used for automatic selection and generation of learning materials (e.g. Volodina et al., 2014), while L2 learner corpora are used for development of different types of grammar and writing support (e.g. Attali & Burstein, 2006).

However, a number of tasks that need to be modelled for the automatic generation of L2 materials, such as text readability classification for the automatic selection of appropriate texts, depend on access to a special type of language which cannot be classified as *typical* NS or L2 learner language in the full sense of this word. NS corpora are unable to provide a reliable basis for modelling for instance text difficulty at the beginner or lower intermediate levels, since NS corpora exhibit a mixture of easy and complex linguistic phenomena, such as vocabulary, grammar, sentences, texts. L2 corpora, on the other hand, contain errors and hence cannot be used to model the language that L2 learners should be exposed to. However, reading and coursebook materials used for L2 courses can – hypothetically – be used as a subset of NS language that is appropriate for modelling L2 learner levels, for example for identifying texts understandable at each of the proficiency levels.

Corpora of coursebook (CB) texts is no novelty in itself, see Meunier & Gouverneur (2009) for an overview. A number of recent projects dealing with collection and annotation of coursebooks indicate a rise in interest in textbook analysis for various applied and theoretical studies (e.g. Gamson et al., 2013). However, CB corpora research has dominated the area of Second Language Acquisition (mainly English as a Foreign Language, EFL) to a larger extent than ICALL-driven research. L2 researchers usually pursue a narrowly defined aim, e.g. teaching of grammar/vocabulary in EFL coursebooks (Anping, 2005) or teaching phraseology at advanced EFL levels (Meunier & Gouverneur, 2007). To our knowledge, there are very few electronic CB corpora that have been compiled (e.g. Römer, 2006), with numerous studies carried out using paper copies of CBs (e.g. Reda, 2003). Systematic studies of textbooks from different angles (textual, pedagogic, didactic, linguistic) have so far been outside of research focus, which partly depends upon the lack of richly annotated electronic CB corpora.

## 1.2 CEFR and L2 coursebook corpora

The corpus described in this article is an electronic collection of textbooks used for teaching of L2 Swedish at CEFR-based courses. CEFR – Common European Framework of Reference for Languages (COE, 2001) – is an influential cross-national initiative that aims at providing language course syllabuses and assessment according to the same model of proficiency levels. CEFR contains 6 levels - A1, A2, B1, B2, C1, C2 – where A1 is the beginner level and C2 is the full proficiency level.

Our interest towards studying CEFR descriptors has resulted from the lack of systematic description of the CEFR levels for Swedish in concrete linguistic terms that could be useful for ICALL applications. The CEFR descriptors, that are intentionally very general to cover different languages, provide very vague guidelines on e.g. text complexity, vocabulary and grammar scope, as can be seen from Figure 1. Subject to interpretation would be: how short should "short pieces of information" and "short written passages" be? What does "collate" mean? What is meant by "in a simple fashion"?

> *Can collate short pieces of information from several sources and summarise them for somebody else. Can paraphrase short written passages in a simple fashion, using the original text wording and ordering.*

FIGURE 1. CEFR descriptor for B1, for ability to process text. (COE, 2001:96).

Our assumption is that the necessary basis for interpretation of (a part of) the CEFR descriptors can be obtained from texts used for practical teaching, e.g. coursebooks. A corpus of CB texts linked to the CEFR levels can, firstly, facilitate pedagogical text studies which would help (1) establish a relationship between how texts selected for reading influence productive writing skills, and thus facilitate SLA research; (2) break down CEFR descriptors into concrete linguistic constituents based on the evidence of the corpus of "input" (i.e. normative) texts - thus attempting at the standardization of CEFR descriptors. Secondly, from the ICALL perspective, CEFR-linked CB corpus can provide basis for comprehensive analysis of normative language that students at CEFR courses are being exposed to. This would, among other things, entail studies of vocabulary and grammar scopes per level; text and sentence readability experiments. Depending on the type of annotation, other studies might also be possible, for instance investigation of development in genre features and use of topics; change in type and format of exercises across levels; shifts in the focus on language skills across levels. Besides, experiments on topic modelling, automatic genre identification, analysis of text questions and text question generation, etc. could also become feasible.

However evident the value of such data for ICALL and SLA might seem, there are very few attempts undertaken to compile corpora of (CEFR-based) coursebooks. François (2011) describes the only known to us CB corpus of CEFR-based texts stretching over all levels of proficiency. The main aim with François' corpus is to use it for NLP-based CALL applications for L2 French. The corpus consists of 21 coursebooks distributed over the 6 proficiency levels, see Table 1:

| | **A1** | **A2** | **B1** | **B2** | **C1** | **C2** | **Total** |
|---|---|---|---|---|---|---|---|
| **Nr textbooks** | 10 | 8 | 8 | 4 | 3 | 3 | 36 |
| **Nr texts** | 452 | 478 | 681 | 198 | 184 | 49 | 2042 |

TABLE 1. Overview over the French CB corpus (François, 2011)

All CBs have been published after 2001, have an explicit link to the CEFR levels of proficiency and are aimed at general L2 French (as opposed to French for specific purposes). After scanning, only reading materials (i.e. texts properly) have been extracted, leaving aside exercises, lists, instructions, etc. found in the coursebooks. Texts have been labelled with the proficiency level of the (chapter of the) book where texts came from, and assigned a genre (e.g. dialogue, recipe, poem) and linguistic annotation (POS, lemmas). The corpus compiled by François has up to date been used for readability studies of L2 French texts and for extraction of a graded lexicon aimed at L2 learners of French (François, 2011; François et al., 2014).

## 2 COCTAILL: collection and annotation

Work on COCTAILL (**C**orpus **o**f **C**EFR-based **T**extbooks **a**s **I**nput for **L**earner **L**evels' modelling) was initiated in 2013 and has been funded partly by the Department of Swedish at the University of Gothenburg (UGOT), and partly by the Center for Language Technology, UGOT. The process of corpus compilation consisted of several stages, shortly presented in Figure 2 below:

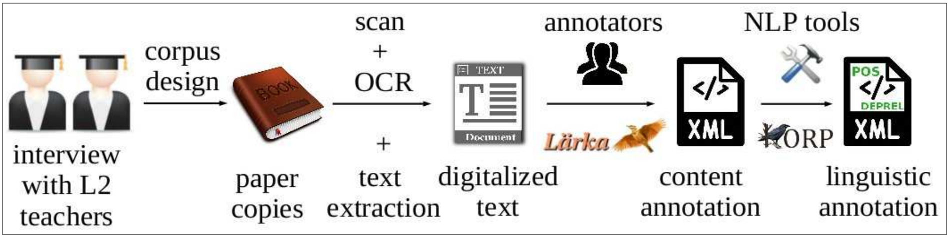


FIGURE 2. Overview of the CEFR-corpus creation

- *Interviews with L2 teachers.* To identify candidate coursebooks, we have carried out interviews with teachers engaged in CEFR-based courses as well as studied course plans for such courses. Altogether, 7 teachers at different levels, schools and institutions have agreed to have an interview. A number of CBs have been named as being used at more than one level. In such cases, to decide the border between levels, we organized a CB workshop where trained teachers discussed such coursebooks with each other and suggested division.

- *Corpus structuring & purchase of coursebooks.* Books that have been suggested by at least two teachers have been selected as core material. We have aimed at a balanced representation at each level with respect to the number of coursebooks per level. However, very few courses are offered at C1 and none at C2 levels that we know of, so the number of coursebooks at these levels differ from the others: 2 titles at C1 and none at C2, see section 2.1 for an overview of the corpus structure. Before books were purchased, we explored the possibility of getting electronic versions from the publishers, but only the publishing house *Liber* was willing to cooperate. However, the titles that Liber could provide have been named by only one teacher, and consequently have not been included into the final corpus.

- *Optical scanning & extraction of raw text*. Once the books were purchased, optical scanning was ordered from an outside contractor. PDF alongside XML files were delivered as resulting output data. Raw text extracted from the XML files was used as the input for the next stages.

- *Implementation of a coursebook editor*. At this stage we defined a taxonomy of textual and pedagogical features for annotation, as well as the format of the output data. Previously, no richly (pedagogically) annotated L2 coursebook corpora have been compiled. Therefore, there were no available editing tools to reuse. After experimenting with XML editors and DTD schemas, we have opted to develop our own editor as described in subsection 2.3.

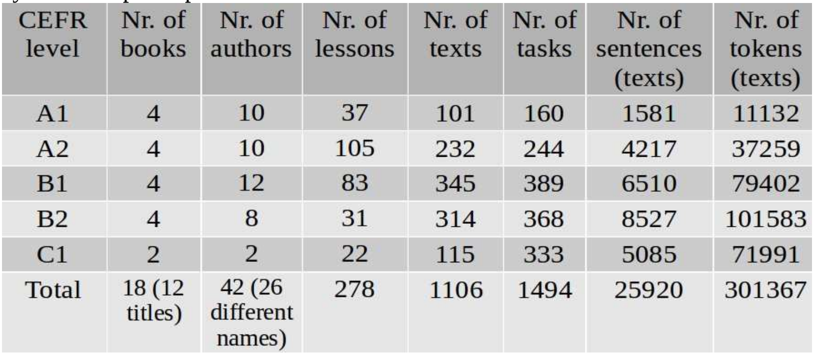
- *Annotation for pedagogical and textual features* involved manual work. Altogether, four people have been involved in the content annotation. Initial annotation of the first two CBs was performed to test the editor and to establish an acceptable taxonomy of textual and pedagogical variables, see section 2.2. In the next round, one more annotator was trained, and as a result, a number of revisions were suggested to improve the taxonomy of pedagogical and textual features. The introduced changes led to a necessity to revise the two initially annotated coursebooks. By the end of this round, annotation guidelines have been produced. Finally, two more annotators have been trained. This stage was concluded by an inter-annotator agreement experiments, which entailed revisions to the annotation guidelines and highlighted the need of another round of revision of the already annotated books, as described in section 2.4.

- *Linguistic annotation* in the form of parts-of-speech, syntactic relations and lemmas has been automatically added using Korp web services (Borin et al., 2012). Whereas annotation of text passages and activity instructions holds good quality, we would need to assess annotation quality of all other types of information. The reason for that is the fact that tasks, lists, and language examples have an unpredictable structure – often incomplete sentences, or lists of mixed linguistic units, which tends to get a very low-level accuracy when it comes to e.g. parts of speech and dependency annotation.

- *Release of the corpus*. Unfortunately, the corpus as a whole cannot be made freely available for download for copyright reasons, however, it is browsable for research purposes via Korp (Borin et al., 2012) with password protection. Besides, parts of the corpus in the form of a bag of sentences (as opposed to connected texts) for each proficiency level are released as downloadable data[1].

## 2.1 Corpus overview

The COCTAILL consists of 12 coursebooks, 5 of which are used at more than one level. The corpus is balanced in the number of coursebooks per level (4 titles/level), except level C1 (2 titles/level). C2 level is not included in this corpus since it represents full language proficiency when learners “can understand with ease virtually everything heard or read” (COE, 2001:24), hence, from the point of view of linguistic modelling it corresponds to regular NS language. The summary of the corpus is presented in Table 2.



| CEFR level | Nr. of books | Nr. of authors | Nr. of lessons | Nr. of texts | Nr. of tasks | Nr. of sentences (texts) | Nr. of tokens (texts) |
|---|---|---|---|---|---|---|---|
| A1 | 4 | 10 | 37 | 101 | 160 | 1581 | 11132 |
| A2 | 4 | 10 | 105 | 232 | 244 | 4217 | 37259 |
| B1 | 4 | 12 | 83 | 345 | 389 | 6510 | 79402 |
| B2 | 4 | 8 | 31 | 314 | 368 | 8527 | 101583 |
| C1 | 2 | 2 | 22 | 115 | 333 | 5085 | 71991 |
| Total | 18 (12 titles) | 42 (26 different names) | 278 | 1106 | 1494 | 25920 | 301367 |

TABLE 2. Overview of the Swedish CEFR corpus

[1] Contact Elena Volodina <elena.volodina@svenska.gu.se> or Ildikó Pilán <ildiko.pilan@svenska.gu.se> to get access to the files.

The COCTAILL comprises a total of 708 589 tokens, about half of which belong to texts, the rest to activity instructions, tasks, lists and language examples. The columns “Nr. of sentences (texts)” and “Nr. of tokens (texts)” refer to sentences in texts only, other elements were excluded from these counts since they often contain smaller linguistic units than a full sentence. The amount of tasks in the corpus (a total of 1494) outnumbers the number of texts (1106). The largest amount of material in terms of texts and tasks is available for B1 and B2 levels.

The values in Table 2 are meant primarily to give an idea of the size of the corpus, rather than present data from which generalizations about the CEFR levels can be made, since authors' choice varied to a great extent as far as the division into lessons and the number of texts and tasks included per level are concerned.

## 2.2 Coursebook content annotation

An overview over the taxonomy of textual and pedagogical annotation is provided in Figure 3. XML elements are shown on the left with their corresponding attributes on the right:

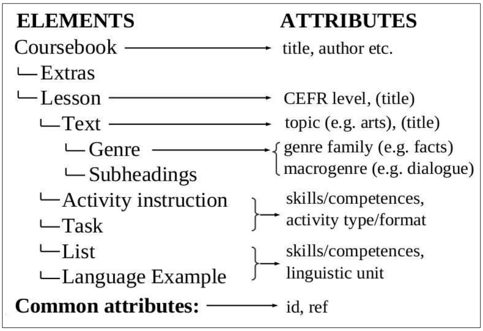


FIGURE 3. Overview over the textual and pedagogical annotation: XML elements and their attributes

Structurally, each coursebook is divided into extras (contents, foreword, copyright note, etc) and lessons (chapters). The running text in each lesson has been manually split into texts aimed at reading comprehension and other types of information typical of coursebooks, such as activity instructions, tasks, lists and language examples, whereby reading comprehension materials have been annotated for *textual features* (section 2.2.1), and the rest of information for *pedagogically relevant features* (section 2.2.2).

### 2.2.1 Textual annotation

By *textual* annotation we understand mark-up of text passages for *topics* and *genres*.

We have listed 28 text *topics* (Table 3) which follow the CEFR guidelines (COE, 2001) in the first place, with modifications introduced as a result of our practical work on the first coursebooks (Volodina & Johansson Kokkinakis, 2013).

In general, we followed the recommendation to opt for a broader topic, e.g. if a text is about a political crisis in some country, including military actions, Politics and power would probably be the best choice. In most cases, more than one topic has been applicable, in which case two or more topics have been assigned. In case there were no topics that corresponded to the text, we considered adding new ones, see Table 3 for the alphabetic list of the topics we have been using so far.

- Animals
- Arts
- Clothes & appearances
- Crime & punishment
- Culture & traditions
- Daily life
- Economy
- Education
- Family & relatives
- Famous people
- Food & drink
- Free time, entertainment
- Greetings/introductions
- Health & body care
- House & home, environment
- Jobs & professions
- Languages
- Personal identification
- Places
- Politics & power
- Relations with other people
- Religion, myths & legends
- Science & technology
- Services
- Shopping
- Sports
- Travel
- Weather & nature

TABLE 3. List of topics

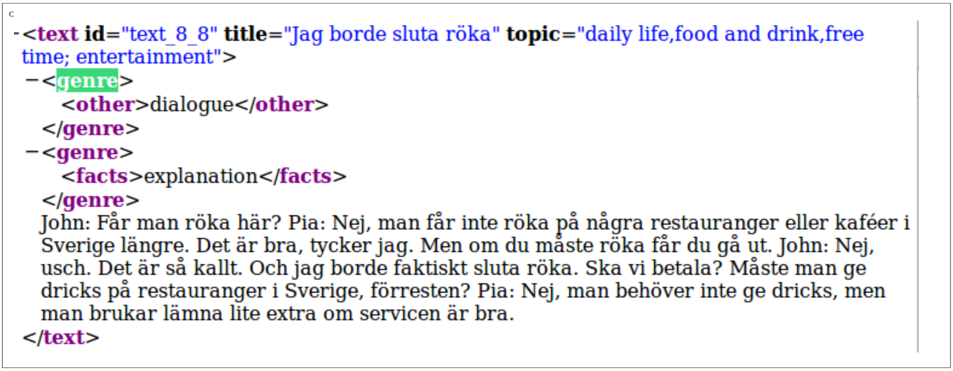
The taxonomy of *genre families* is comprised of four elements: Narration, Facts, Evaluation and Other, following the taxonomy described in Johansson and Sandell Ring (2010) with slight modifications as a result of the work on the first annotated coursebooks (Volodina and Johansson Kokkinakis, 2013). Such a modification is the addition of the genre family Other which contains text genres (e.g. puzzle) that were difficult to place into the other three Narration, Facts or Evaluation families. Further subdivision of genre families into macrogenres is shown in Table 4.

| Narration | Facts | Evaluation | Other |
|---|---|---|---|
| Description<br>Fiction<br>News article<br>Personal story | Autobiography<br>Biography<br>Demonstration<br>Explanation<br>Facts<br>Geographical facts<br>Historical facts<br>Instruction<br>Procedures<br>Report<br>Rules | Advertisement<br>Argumentation<br>Discussion<br>Exposition<br>Interpretation, exegesis<br>Personal reflection<br>Persuasion<br>Review | Anecdote, joke<br>Dialogue<br>Language tip<br>Letter<br>Lyrics<br>Notice, short message<br>Puzzle<br>Questionnaire<br>Quotation<br>Recipe<br>Rhyme |

TABLE 4. List of genre families and macrogenres

It can be discussed whether some of the Other macrogenres can be moved to any of the other three genre families (e.g. Anecdotes to the Narration family).

In a lot of cases, where there were no clear-cut genres, a combination of genres became an optimal solution, see Figure 4.



```
-<text id="text_8_8" title="Jag borde sluta röka" topic="daily life,food and drink,free
 time; entertainment">
  -<genre>
     <other>dialogue</other>
   </genre>
  -<genre>
     <facts>explanation</facts>
   </genre>
   John: Får man röka här? Pia: Nej, man får inte röka på några restauranger eller kaféer i
   Sverige längre. Det är bra, tycker jag. Men om du måste röka får du gå ut. John: Nej,
   usch. Det är så kallt. Och jag borde faktiskt sluta röka. Ska vi betala? Måste man ge
   dricks på restauranger i Sverige, förresten? Pia: Nej, man behöver inte ge dricks, men
   man brukar lämna lite extra om servicen är bra.
 </text>
```

FIGURE 4. An example of textual annotation, text at level A1

### 2.2.2 Pedagogical annotation of coursebooks

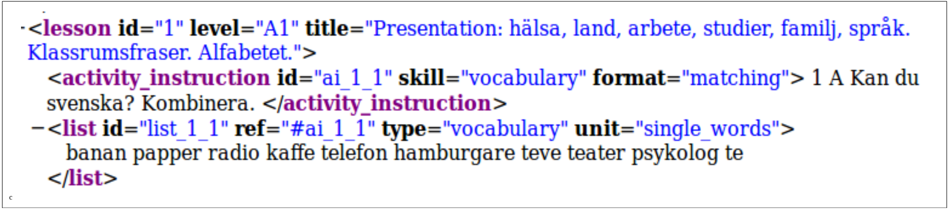
*Pedagogical* annotation in this corpus is understood as mark-up assigned to all types of information found in coursebook lessons except texts used for reading comprehension. All books are structured by lessons (i.e. chapters in coursebooks), which are assigned a proficiency level, which then applies to all texts and activities in the lesson. The taxonomy of the pedagogical mark-up within each lesson is presented by lists, language examples, tasks and activity instructions.

| Activity instructions, Tasks, Language examples, Lists | Activity instructions, Tasks | Activity instructions, Tasks | Language examples, Lists |
|---|---|---|---|
| **Target skills:**<br>Listening<br>Reading<br>Writing<br>Speaking<br><br>**Target competences:**<br>Grammar<br>Pronunciation<br>Spelling<br>Vocabulary | **Activity types:**<br>Brainstorming<br>Composition/essay writing<br>Dialogue/interview<br>Dictation<br>Discussion<br>Error correction<br>Form manipulation<br>Information search<br>Monologue<br>Pre-reading<br>Question answering<br>Reading aloud<br>Role-playing<br>Summary<br>Text questions<br>Translation | **Activity formats:**<br>Category identification<br>Category substitution<br>Free/short answers<br>Free writing<br>Gaps<br>Matching<br>Multiple choice<br>Narration, retelling, presentation<br>Reordering/ Restructuring<br>Sorting<br>True-false/Yes-no<br>Wordbank | **Linguistic units:**<br>Characters<br>Dialogues<br>Full sentences<br>Incomplete sentences<br>Numbers<br>Phrases<br>Question-answer<br>Single words<br>Texts/examples of text writing |

TABLE 5. Overview over the taxonomy of the pedagogical mark-up

Further, each of the pedagogically-relevant elements is associated with the target `skills/competences` (e.g. reading) they are aimed at. `Lists` and `language examples` are assigned `linguistic units` (e.g. single words), and all `tasks` and `activity instructions` are associated with `format` and `type` of exercises (e.g. gaps), see Table 5 for an overview. In the terms of the output XML data, the table headings represent XML elements, the text in bold corresponds to XML attributes, and the running text stands for a set of attribute values.

An example of pedagogical annotation follows below (Figure 5)

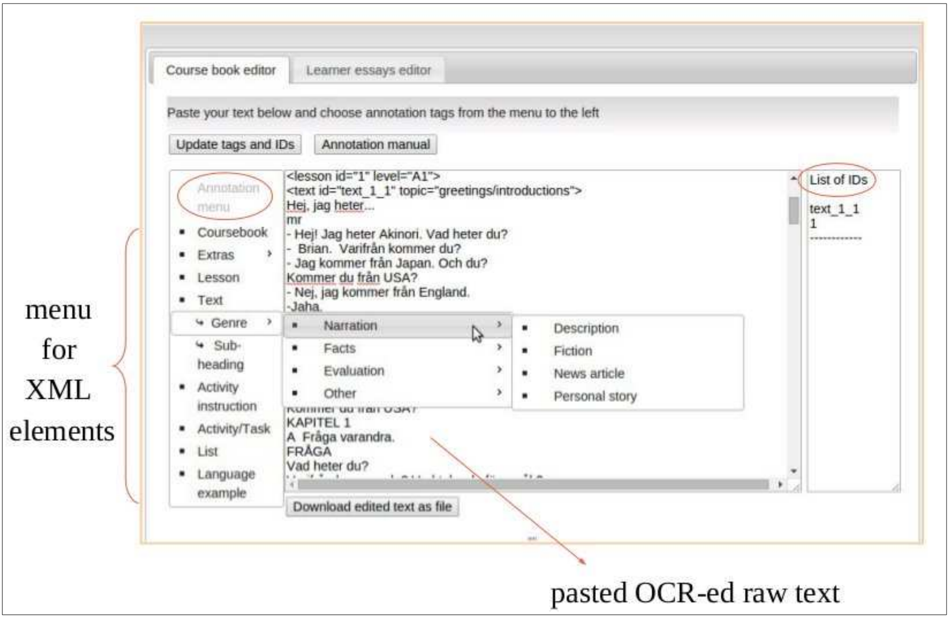


```
-<lesson id="1" level="A1" title="Presentation: hälsa, land, arbete, studier, familj, språk.
 Klassrumsfraser. Alfabetet.">
   <activity_instruction id="ai_1_1" skill="vocabulary" format="matching"> 1 A Kan du
   svenska? Kombinera. </activity_instruction>
  -<list id="list_1_1" ref="#ai_1_1" type="vocabulary" unit="single_words">
     banan papper radio kaffe telefon hamburgare teve teater psykolog te
   </list>
```

FIGURE 5. An example of pedagogical annotation, level A1

## 2.3 Online coursebok editor

To simplify the process of inserting XML-annotation into the OCR-ed raw texts, an online coursebook editor has been developed early in the project (Volodina & Johansson Kokkinakis, 2013).[2]



```
<lesson id="1" level="A1">
<text id="text_1_1" topic="greetings/introductions">
Hej, jag heter...
mr
- Hej! Jag heter Akinori. Vad heter du?
- Brian. Varifrån kommer du?
- Jag kommer från Japan. Och du?
Kommer du från USA?
- Nej, jag kommer från England.
-Jaha.
Kommer du från USA?
KAPITEL 1
A Fråga varandra.
FRÅGA
Vad heter du?
```

FIGURE 6. The online corpus editor.

[2] http://spraakbanken.gu.se/larka/larka_cefr_editor.html

The annotation scheme for content annotation described in section 2.2 has been implemented in the form of user-friendly menus (Figure 6, on the left). In the centre (Figure 6) is an editable text area where text for annotation is pasted, and on the right is a field for an overview of all inserted `IDs`. Link to annotation guidelines and an option of downloading the annotated text as a file are also offered.

Each menu element is accompanied with a pop-up dialogue, which prompts what information should be added, for example `IDs`, or `references` to previously used `IDs`, or `titles`. For categories where lists of options exist, such as `topics`, `genres` or `skills`, options are offered as multi-select drop-down menus. Besides, there are sub-menus for inserting `subheadings` and `extra` information, such as text author, source of information, etc. Each new inserted XML element closes the previously opened one, except in cases of `lessons`, `extras`, `genres` and `subheadings`.

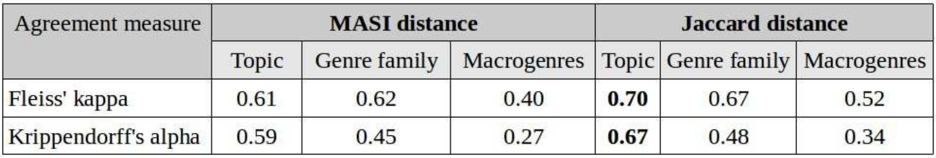
Meta-information about each coursebook is collected once before the annotation of the rest of the book starts, and includes `title`, `author`, `publication year`, `publisher`, `ISBN`.

The editor is language independent, freely accessible over internet and can be easily reusable in other L2 coursebook annotation projects.

## 2.4 Text-level annotation: inter-annotator agreement

*Inter-annotator agreement* is the degree of agreement among annotators about assigning categories to the same objects (Artstein and Poesio, 2008). Our intention with the inter-annotator agreement experiment was to estimate the quality of the text-level (textual) annotation on the one hand, and to detect categories causing large number of disagreements and inconsistencies, on the other.

We have investigated randomly chosen parts of the CEFR corpus, targeting at least one chapter (lesson) per level. The controlled subset of the corpus comprised 21630 tokens at the five proficiency levels, divided between 32 texts and a number of accompanying coursebook activities. Our focus has been on texts: text topics, genre families and macrogenres. Three annotators have been involved in this experiment with knowledge of linguistics, language teaching and computational linguistics.



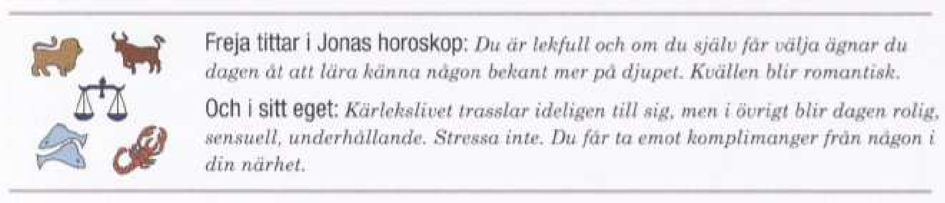
| Agreement measure | MASI distance | | | Jaccard distance | | |
|---|---|---|---|---|---|---|
| | Topic | Genre family | Macrogenres | Topic | Genre family | Macrogenres |
| Fleiss' kappa | 0.61 | 0.62 | 0.40 | **0.70** | 0.67 | 0.52 |
| Krippendorff's alpha | 0.59 | 0.45 | 0.27 | **0.67** | 0.48 | 0.34 |

TABLE 6. Results of the inter-annotator agreement for topics, genre families and macrogenres

We report inter-annotator agreement in terms of Fleiss' multi-kappa (Davies and Fleiss, 1982) and Krippendorff's alpha (Krippendorff, 1980) being that the task involved multiple (i.e. three or more) annotators. Both measures take into account chance agreement (Artstein and Poesio, 2008). Each annotator could assign more than one category to each text object, i.e. multiple topics out of 28 possible ones, multiple genre families out of 4 choices and multiple macrogenres out of 34 options, therefore, we used distance measures that would calculate the dissimilarity between sets of multiple values. We considered both Jaccard's distance metric (Jaccard, 1908) and MASI (Measuring Agreement on Set-valued Items; Passonneau, 2006) when calculating agreement with the previously mentioned measures. Both metrics are based on the union and the intersection between sets, MASI including also an additional term, M, which equals 1 if the sets are identical, 2/3 in case of subsumption, and 1/3 if there is at least one element in common between the two sets (Passonneau, 2006). For both the distance[3] and the agreement[4] measures the NLTK Python module has been used (Bird, 2006). Results are shown in Table 6.

Fleiss' kappa within the range between 0.61-0.80 means substantial agreement, which given our type of annotation is a very encouraging result. However, the original results for Fleiss' kappa were lower than the ones reported in Table 6 (e.g. Fleiss' kappa for topics 0.52 with Jaccard distance and 0.37 with MASI). The reason for that proved to lie in the fact that some of the texts had substantial difference in the number of assigned values, with the intersection being a good common ground. This has led us to the conclusion that we should set a maximum number of values that may be assigned to each text object. To simulate that, we have calculated inter-annotator agreement based on the intersection of values (i.e. considering only values that were common between at least two of the three annotators, leaving out the ones that have been assigned only once, except when only one label was provided), as reported in the table above. The results have improved substantially. Following this experiment, in the near future, a revision of the corpus annotation is planned where we will consider reducing the number of assigned topics to a maximum of 3 and macrogenres – to a maximum of 2.

To exemplify cases with different interpretations, look at Figure 7 where a text with a horoscope is given in the original language and translated into English in Figure 8.



Freja tittar i Jonas horoskop: *Du är lekfull och om du själv får välja ägnar du dagen åt att lära känna någon bekant mer på djupet. Kvällen blir romantisk.*

Och i sitt eget: *Kärlekslivet trasslar ideligen till sig, men i övrigt blir dagen rolig, sensuell, underhållande. Stressa inte. Du får ta emot komplimanger från någon i din närhet.*

FIGURE 7. Text on horoscope, level B2.

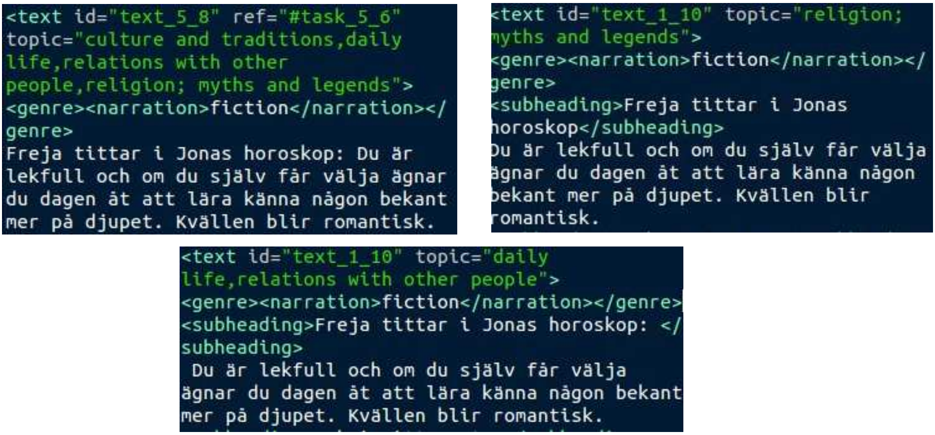
***Freja looks into Jonas's horoscope**: You are playful, and if you can choose, you'd spend the day getting to know better somebody you are acquainted with. The evening will be romantic.*
***And then into her own**: The love life is a mess, but otherwise, the day will be funny, sensual and entertaining. Don't work yourself up. You will receive compliments from somebody in your surrounding.*

FIGURE 8. Translation of the text into English

Figure 9 provides the three annotations that have been provided to this text.

The first annotator assigned 4 topics: (1) culture and traditions, (2) daily life, (3) relations with other people, (4) religion; myth and legends. The second annotator assigned topic (4), whereas the third annotation assigned topics (2) and (3).

[3] http://www.nltk.org/_modules/nltk/metrics/distance.html
[4] http://www.nltk.org/_modules/nltk/metrics/agreement.html



```
<text id="text_5_8" ref="#task_5_6"
topic="culture and traditions,daily
life,relations with other
people,religion; myths and legends">
<genre><narration>fiction</narration></
genre>
Freja tittar i Jonas horoskop: Du är
lekfull och om du själv får välja ägnar
du dagen åt att lära känna någon bekant
mer på djupet. Kvällen blir romantisk.
```

```
<text id="text_1_10" topic="religion;
myths and legends">
<genre><narration>fiction</narration></
genre>
<subheading>Freja tittar i Jonas
horoskop</subheading>
Du är lekfull och om du själv får välja
ägnar du dagen åt att lära känna någon
bekant mer på djupet. Kvällen blir
romantisk.
```

```
<text id="text_1_10" topic="daily
life,relations with other people">
<genre><narration>fiction</narration></genre>
<subheading>Freja tittar i Jonas horoskop: </
subheading>
 Du är lekfull och om du själv får välja
ägnar du dagen åt att lära känna någon bekant
mer på djupet. Kvällen blir romantisk.
```

FIGURE 9. Annotation of the text for topics and genres

For the experiments we used triples of values (annotator-code, text-code, list of assigned values), in Table 7 shown with the original set-up in the first column, and with an intersection set-up in the second column.

| Original experiment | Intersection-based experiment |
|---|---|
| • (ann1, text_5_8, [1,2,3,4])<br>• (ann2, text_5_8, [4])<br>• (ann3, text_5_8, [2,3]) | • (ann1, text_5_8, [2,3,4])<br>• (ann2, text_5_8, [4])<br>• (ann3, text_5_8, [2,3]) |

TABLE 7. Original versus "intersection"-based triples

As can be seen, the value "1" has been removed from the list of assigned values from annotator 1, since this value has not been used by any other annotator. We can see here that annotator 1 has agreed with both annotators 2 and 3, whereas there was no agreement between annotator 2 and 3.

Summarizing the results of the experiment on inter-annotator agreement, we can say that categories causing a lot of disagreement proved to be the difference in number of assigned values, rather that the values themselves, which is the reason for planned revisions in the annotation guidelines and in the annotated files. However, the experiment has also shown that the annotation is reliable and can be used for experiments as it is, in the sense that among the multiple values there has always been a central overlap between different annotators. Non-overlapping topics and genres can be considered peripheral adding an extra value to text characteristics.

## 3 Initial quantitative explorations of the COCTAILL

We carried out an initial quantitative analysis of the corpus observing variables such as text genres, topics as well as skills and competences targeted by tasks at each CEFR level.

Texts showed a substantial variation both in genre and in topics across proficiency levels. About half of the texts were dialogues at A1 level, but this amount steadily decreased at each CEFR level, C1 level coursebooks containing barely any. Factual texts were presented at all levels, but at higher proficiency levels they were almost twice as common. The percentage of dialogues and factual texts at each level is presented in Figure 10.

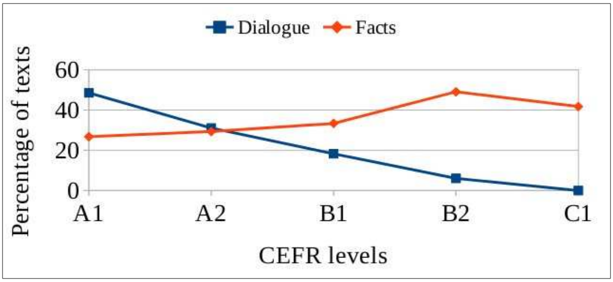


FIGURE 10. Percentage of dialogues and factual texts per CEFR level

Not only genres, but also certain topics showed large difference in distribution at different CEFR levels, as Figure 11 below shows.

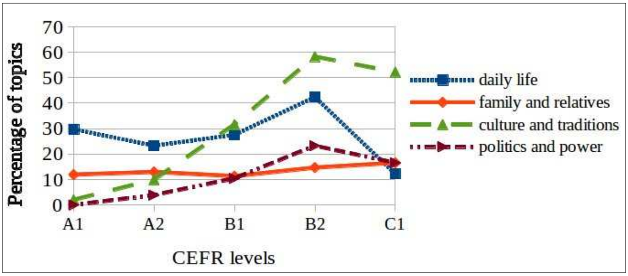


FIGURE 11. Percentage of text topics per CEFR level

The topics “culture and tradition” and “politics and power” are either not present or appear to a very limited extent at A1 level, but at higher proficiency levels their proportion increases substantially. The topic of “daily life”, although appears at all CEFR levels, seems to be less common at C1 level. Interestingly, the percentage of texts focusing on “family and relatives” remains the same across all levels. Such topics would be particularly suitable for the analysis of how linguistic complexity changes at different proficiency levels within the same topic.

Further, we retrieved some quantitative data from a more pedagogical perspective aiming at tracing how the proportion of skills and competences targeted by tasks change at various levels. This information is presented in Figure 12.

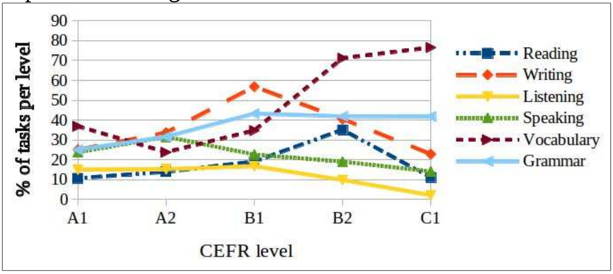


FIGURE 12. Target skills and competences per CEFR level

At A1 and A2 levels, the focus is primarily on the productive skills of speaking and writing, each of which accounted for about one fourth of the exercises at this level. Tasks involving the receptive skills of reading and listening are about 10% less frequent at this initial stage. The corpus also shows a shift in focus from oral language use to the written one at B1 and B2 levels. More than half of the tasks are writing exercises at B1 level, and the highest percentage of reading tasks (35%) appears at B2 level. The proportion of grammar exercises increases until B1 level, then it keeps its rather dominant presence (about 40%) at all further stages. Vocabulary teaching is a primary target skill of tasks at A1 level, but less so at A2 level, whilst from intermediate (B1) level on, vocabulary exercises dominate the items proposed for students, which is especially obvious at C1, which supports Singleton's (1995) hypothesis that vocabulary doesn't have a critical period at which it should be taught or learnt.

Another interesting piece of statistics we have looked at is average sentence length per CEFR level (Figure 13). Numbers have been calculated upon sentences retrieved from texts aimed at reading comprehension.

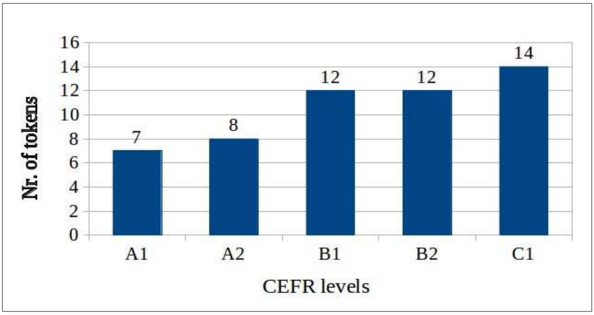


FIGURE 13. Average sentence length per CEFR level.

The graph shows that sentence length grows steadily from lower levels to more advanced ones, the largest increase being observed between A2 and B1 with no difference between B1 and B2. The most feasible explanation for the less drastic increase in sentence length starting from B1 is that texts at the higher levels contain a broader mixture of sentence types – both typical of the level itself, and of all the lower levels, e.g. B2 texts hypothetically contain sentences typical of levels A1, A2, B1 and B2. The sentence length typical of the lower levels would in that case influence the calculations of the average length at B2. Another potential explanation might be connected to the number of texts of certain genres: to take one example, dialogues that tend to contain very short sentences, dominate at A1 and A2 levels and decrease in number from B1.

These numbers show some similarity in the tendency of increase to the reported average sentence length in the L2 French corpus (François, 2011) , as shown in Table 8:

| A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|
| 9,1 | 14,54 | 16,85 | 18,6 | 19,36 | 21,43 |

TABLE 8. Average sentence length in L2 French corpus ( François, 2011:359)

There is a steady increase in the average sentence length over the levels in both languages. However, there is a larger increase between A1 and A2 in L2 French coursebooks, and more moderate growth between the rest of the levels. Differences in the average values between the two languages can be accounted for by linguistic characteristics of the two language, by differences in tokenization and segmentation tools, as well as by the variety of text genres present in the two corpora. In general, this piece of statistics raises interesting questions about linguistic complexity of each proficiency level and asks for deeper investigations of the problem.

## 4 Concluding remarks

We have presented our work on COCTAILL, a corpus of L2 coursebooks, richly annotated for textual, pedagogical and linguistic variables. The corpus is innovative in a number of ways: there are no other existing electronic corpora that have pedagogical annotation alongside proficiency level-labelling, textual annotation, and linguistic annotation covering all the spectrum of proficiency levels interesting for linguistic modelling of learner levels. We pioneered in the development of a taxonomy of pedagogical variables for L2 coursebook annotation, which up-to-date remains the only one we are aware of. Besides, unlike a number of other coursebook projects, where only reading materials are selected or only a subset of CB language is analysed, we present a possibility to study coursebooks in their entirety with important implications for correlating proficiency levels, L2 input as well as various pedagogical and textual variables, such as target skills and competences. COCTAILL is available for browsing with password protection and is downloadable as a bag of sentences labelled with coursebook levels.

In the future, we plan several iterations on the improvement of COCTAILL content annotation. This will include the revision and a potential decrease in the number of assigned topics and macrogenres. Besides, the topic and genre taxonomy may need to be revised to contain fewer, but more general categories, i.e. going from a more detailed taxonomy to one with broader categories.

Certain parameters have yet been outside the inter-annotator agreement experiment. In future we plan to focus on

(1) activity instructions and tasks, where we will calculate agreement in assigning target skills and exercise formats; and

(2) lists and language examples, where the main focus will be on the annotation of target skills and linguistic units

We can foresee that results of the inter-annotator experiments will yield another round of annotation revision.

Availability of the corpus opens prospects to engage in numerous SLA-aware ICALL-relevant studies, such as CEFR profiling, vocabulary and grammar profiling, studies on sentence and text readability, question generation, automatic genre identification, automatic topic modelling – to name just a few potential directions of research.

The taxonomy of textual and pedagogical variables present in COCTAILL provides the key to various empirical studies of coursebooks, which can help critically assess and reflect on the relation between coursebook design and SLA research. Pedagogically annotated coursebook corpora such as COCTAILL, have a potential to become a crystallized form of what should be taught, at which level and in which format, which is crucial for various ICALL tasks, such as material generation. We expect that these insights, implemented into ICALL applications, will facilitate generation of pedagogically appropriate learning materials. To put it simply, you get what you annotate.

## References

Anping He. (2005). Corpus-Based Evaluation of ELT textbooks. Paper presented at the joint *conference of the American Association of Applied Corpus Linguistics and the International Computer Archive of Modern and Medieval English*, 12-15 May 2005, University of Michigan.

Artstein Ron & Massimo Poesio. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4): 555-596.

Attali Yigal & Jill Burstein. (2006). Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Bird Steven. (2006). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions, pp. 69-72.

Borin Lars, Markus Forsberg & Johan Roxendal. (2012). Korp – the corpus infrastructure of Språkbanken. *Proceedings of LREC 2012*. Istanbul: ELRA. 474–478.

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

Davies Mark & Joseph L. Fleiss. (1982). Measuring agreement for multinomial data. *Biometrics*, 38(4): 1047–1051.

François Thomas. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*, Ph.D. Thesis, Université Catholique de Louvain. Thesis Supervisors : Cédrick Fairon and Anne Catherine Simon.

François Thomas, Nuria Gala, Patrick Watrin & Cédrick Fairon. (2014). FLELex: a graded lexical resource for French foreign learners. In the 9th *International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland, 26-31 May.

Gamson David A., Lu Xiaofei, & Eckert Sarah Anne. (2013). Challenging the research base of the common core state standards: A historical reanalysis of text complexity. *Educational Researcher*, 42(7):381-391.

Jaccard Paul. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 44: 223-270.

Hancke Julia & Detmar Meurers. (2013). Exploring CEFR classification for German based on rich linguistic modeling. *Learner Corpus Research 2013, Book of Abstracts*. pp. 54-56. Bergen, Norway.

Johansson Britt & Anniqa Sandell Ring. (2010). *Låt språket bära: genrepedagogiken i praktiken.* Hallgren och Fallgren, Stockholm.

Krippendorff Klaus. (1980). *Content Analysis: An Introduction to Its Methodology*, chapter 12. Sage, Beverly Hills, CA.

Meunier Fanny & Gouverneur Céline. (2007). The treatment of phraseology in ELT textbooks, In: *Corpora in the Foreign Language Classroom. Selected papers from the Sixth International Conference on Teaching and Language Corpora* (TaLC6), University of Granada, 4-7 July 2004, Encarnación H., Quereda L. and Santana J. ed(s), Amsterdamm & New York, Rodopi, Language and Computers Series 61, p. 119-139.

Meunier Fanny & Gouverneur Céline. (2009). New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material, In: *Corpora and Language Teaching*, Aijmer, K. ed(s), Amsterdam & Philadelphia, Benjamins, p. 179-201

Passonneau Rebecca J. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of LREC,* Genoa, pp. 831–836.

Reda Ghsoon. (2003). English Coursebooks: Prototype Textsts and Basic Vocabulary Norms. *ELT Journal* 57(3): 260-268.

Römer Ute. (2006). Looking at *Looking*: Functions and Contexts of Progressives in Spoken English and 'School' English. In: Renouf, Antoinette & Andrew Kehoe (eds.). *The Changing Face of Corpus Linguistics.* Papers from the 24th International Conference on English Language Research on Computerized Corpora (ICAME 24). Amsterdam: Rodopi. p.231-242.

Singleton David. (1995). *Introduction: A Critical Look at the Critical Period in Second Language Acquisition Research*, In Singleton D. & Lengyel, Z. (Eds.), The Age Factor in Second Language Acquisition (1-29). Avon: Multilingual Matters, Ltd.

Vajjala Sowmya & Detmar Meurers. (2013). On The Applicability of Readability Models to Web Texts. *Proceedings of the Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, ACL 2013

Volodina Elena, Ildikó Pilán, Lars Borin, & Therese Lindström Tiedemann. (2014). A flexible language learning platform based on language resources and web services. *Proceedings of LREC 2014, Reykjavik, Iceland.*

Volodina Elena & Sofie Johansson Kokkinakis. (2013). Compiling a corpus of CEFR-related texts. *Proceedings of the Language Testing and CEFR conference*, Antwerpen, Belgium, May 27-29, 2013.