Annotation guide

for the *Course book editor* in Lärka

Version 1.0

Elena Volodina, Ildikó Pilán

Språkbanken, Department of Swedish
University of Gothenburg

elena.volodina@svenska.gu.se
ildiko.pilan@svenska.gu.se

June 2014

Table of Contents

The following document is a manual for the *Course book editor*, an annotation tool for course-book texts used in language teaching, available through the online platform Lärka[1] (Volodina et al. 2014). The annotation aims at describing the structure of the text in such books, identifying chapters and their sub-parts such as activity instructions, tasks and proposed readings. The future use of the annotated books includes, but is not limited to, different corpus-based linguistic or pedagogical research examining, among others, the activity types and the language used.

1 A quick introduction to XML

The format used during the annotation is the Extensible Markup Language (XML) a standard data format readable by both human and machines.
An XML document has two main components, the *markup* and the *content*. The markup consists of two parts, a start- and an end-*tag*. Tags are delimited by the < and > symbols. The unit beginning with a start-tag and closing with and end-tag is referred to as *element*. Elements may have *attributes* (name/value pairs) describing a specific property. An example of an XML element is shown below. For additional details see the World Wide Web Consortium's site[2] or the Wikipedia article[3] on XML.

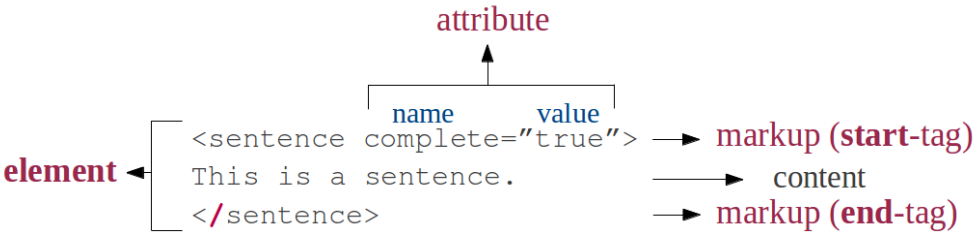


Figure 1. Key XML concepts

You can link one element to another through *referencing* to express a connection between them. This is done with the use of the *ref* attribute and the # symbol. See the example below:

```
<text id="text_1">
This is the first text. This is the first
text.
</text>
<text id="text_2" ref="#text_1">
This is the second text. This is the
second text.
</text>
```

Figure 2. Referencing

In the figure above, each text is assigned a unique ID (e.g. *text_1*) and the second element's *ref* attribute uses this ID as value to refer to the first text. Examples of what kind of coursebook elements should be linked to each other are listed in *section 2.2.2, point 13* below.

1 http://spraakbanken.gu.se/larka
2 http://www.w3.org/TR/REC-xml/
3 http://en.wikipedia.org/wiki/XML

2 The annotation process

2.1 Before you annotate

2.1.1 Lärka and the annotation page

To access the annotation tool, open Lärka and click on the 'Learner corpora editor' link on the top of the main page. Alternatively, you can open the following link directly:

http://spraakbanken.gu.se/larka/larka_cefr_editor.html

The first tab displayed on this page is 'Course book editor', the tool you will use for the annotation. On the left side of the page, you can find the *Update tags and IDS* button (for updating the list of IDs used), *Annotation manual* button (to open a pdf file containing the current guidelines) and the *Annotation menu* (containing all the possible elements for the annotation). There is a text box in the middle of the page, this is where you will paste your text to annotate. On the right of that, the *List of IDs* shows the IDs already used (IDs must be unique, i.e. used only once, except when used in references). Below the text box you can find a '*Download edited file as text file*' button. Use can use it to save your work while and after annotating. We recommend, though, that you *copy and paste the annotated text into the local file*, instead, since this is faster.

2.1.2 Preparatory tasks before the annotation

Read through the annotation manual. Try to get a hold of the whole annotation process, the single elements and read through the guidelines in section 2.2 and detailed description of each element in the annotation menu in section 3.
Look at how the chapters are structured in the textbook you will annotate, try to identify a pattern / logic behind it. You can also try to pen-and-paper annotate the first chapter of your book.

2.2 While annotating

2.2.1 First steps

1. Copy the plain text from one of the files that you will work with.
2. Paste the entire raw text in the text box in the middle of the page where the 'Paste your text here' appears. If you continue working on a partly annotated document, click on `Update tags and IDs` to ensure correct automatic suggestions to your IDs.
3. Move with the cursor to the beginning of the text and click on the '`Coursebook`' button in the menu on the left side of the page.
4. A series of dialogue boxes will open asking for generic information about the book (title, author, ISBN etc.). Fill in the information and click on '`OK`'.

2.2.2 Adding new elements

For each part you want to annotate in the text, follow these steps:

1. Click with the cursor to the beginning of the **relevant part of the text.**
2. Select the relevant **tag name from the Annotation menu**. (You will notice that as soon as you have entered the next xml element, the first one is automatically closed (e.g. `</text>`). This is true of all elements except "`lesson`". The previous lesson is closed only when the new lesson is added.)
3. Fill the **dialogue boxes** asking for specific information.
4. First, you'll need to **assign an ID** to your element. (Consecutive ID numbers are suggested to you based on your previously used IDs.)
5. Then you can provide a **reference** task for the part you're annotating. Type the ID of an XML element used (e.g. `task_1_2`). See point 13 below in this list on referencing for further information.
6. For some elements, a **title** can also be provided.
7. Once you fill in the required information, the opening XML tag and the attributes will be added in the text box automatically.
8. Remove any **OCR junk** not belonging to the text (series of strange symbols, page numbers, structural references etc.). Feel free to **correct OCR errors** on the basis of a comparison with the original book / text.
9. Move to the beginning of the next part to annotate, and insert the relevant tag (same as step 2). The closing tag for your previous XML element will appear automatically once you continue with the annotation of the next piece of text. Double-check that the tag is correct, in case, change it manually.
10. If you need to **insert a tag before / between** already annotated parts, you will need some manual editing. A closing XML tag for your last inserted XML element will appear preceding the newly added element, make sure that:
    1. you **add** a closing tag manually <u>after</u> your last added element before adding the new element in preceeding parts of the text
    2. **delete** the automatically created closing tag from <u>before</u> your new element.
    3. <u>Update tags and IDs</u> to ensure correct tag flow in the next annotation steps

11. In case you see some **error** in the information you have filled, correct it directly in the xml-tags. Alternatively, delete the last inserted xml-tag(s), both the previously closed one, and the opening one, update tags and IDs, and start over.
12. Repeat steps 1-5 for the whole book. After the last lesson, insert the </lesson> tag manually.
13. When **referencing**, **the rule of thumb** is to reference **from** a **new element to** an **already used ID** (i.e. backwards in the text). For example, a book chapter (lesson) starts with a task "Look at the pictures and match them with the words in the list below". After the task comes a list of words. You first assign an ID to the task/activity instruction (e.g. ID="ai_1"; you then mark the list of words (ID="list_1") as a list referencing the task (ref="#ai_1"). The chaining of interrelated activities and elements always goes from the new elements to the already added elements. In cases where the new element can be linked to more than one previous element, link to the closest relevant one unless there is a more relevant element several steps away. Reference activity instructions (what and how to do + examples) to a text / wordlist that the activity is based on and reference (link) tasks to instructions (the activity / exercise itself). In case of task without instructions, use a reference for the text / list which it builds on, if any.

NB! Save your work regularly!
Be careful not to close accidentally the tab in your browser.

*TIPS:

- Try to set your browser so that it asked for a confirmation message before closing the window.
- Spell check: with Microsoft Word or LibreOffice Writer set to Swedish can help you see OCR (spelling) errors in the text

2.2.3. Additional guidelines

- Try to be as **consistent** as possible in your decisions during annotation.
- In case of decisions which might not be straightforward, **make a note** and indicate the relevant page. Send these notes together with the annotated file.
- Examples on how to perform a task should be included into activity instructions
- When adding IDs: use a notation with the number of lesson, underscore, running number of the element: e.g. *langex_1_1* (for language example, lesson 1, language example nr. 1)
- **Pad the gaps** in tasks with 6 underscores: _____
- **Mark dialogues** either with a name (initial) followed by a colon symbol (e.g. *Daniel:* or *D:*) or by a dash symbol (e.g. - *Jag är Daniel*.)
- You can sometimes make smaller changes in the order of pieces of text (e.g. activity instruction) to avoid interrupting bigger units of text / task. (eg. Pg 94 in På svenska)
- for a **well-formed XML** structure:

    1. Do not leave unannotated text (without assigned XML tags) between tags. Do not jump to other parts of the text, work consequently.
    2. Attribute values cannot contain double quote symbols (“), substitute them with single quotes (') if quotes are needed. For example, if the text title is *How to “appeal” to the court*, you should write *How to 'appeal' to the court* in the dialogue menu, replacing double quotes with single ones. Double quotes are automatically added in the beginning and the end of each attribute, e.g, ID=”list_1”

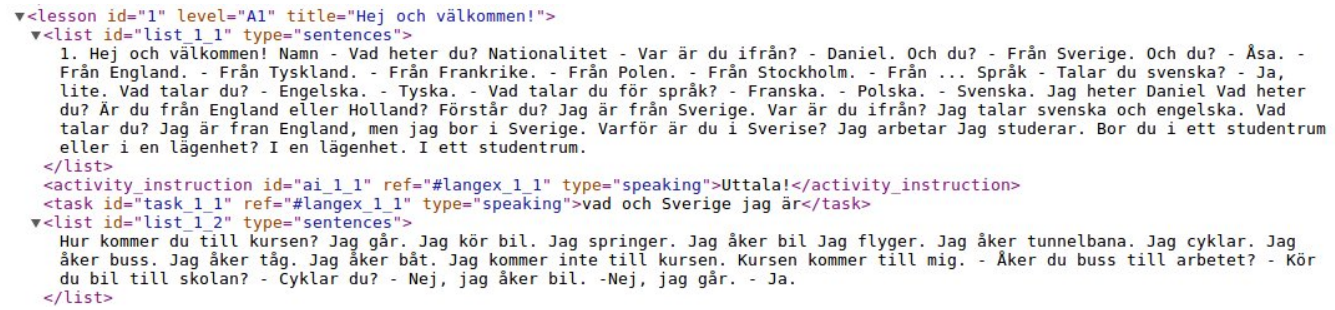
2.2.4 Troubleshooting

- Sometimes a genre / topic cannot be added, only an empty line appears. Click on the `Update tags and IDs` button. If that doesn't solve the problem, add the elements manually (you can check the previously automatically inserted genre / topic tags to see how the element should look like).

2.3 After annotation

2.3.1 Fixing XML-specific errors

When you're finished annotating the book, you may need to check if all XML tags are correctly placed. You can use your browser for this purpose. Right-click on your XML file, select 'Open with' and choose the browser you want to use. If you made any mistakes, a red box with a specific error message will be displayed at the top of the page in the browser. Go to the line number indicated to find the problematic part. (Problems may be caused also by open tags long before or after the indicated line.) Continue removing the errors until no more error messages are left and your document is displayed as a document tree with XML tags highlighted in different colors (see figure below).



Examples for types of errors:

a) Opening and ending tag mismatch:

```
<language_example id="langex_16_1" ref="#text_16_2" type="phrases">
Gott nytt år!
Glad påsk!
</task>
```

b) unclosed tag (raises the same error: 'Opening and ending tag mismatch:')

```
<list id="list_16_1" ref="#text_16_1" type="vocabulary">
16.  Helger och högtidsdagar
Nyårsafton
```

c) two opening / or closing tags:

```
<topic>personal identification<topic>
```

d) a number of special xml symbols may cause problems if they are used incorrectly. For example, angle brackets (> and <) and an &-symbol in the running text (e.g. not within the xml-tag) should be replaced according to the following principle:

>	>
<	<
&	&

2.3.2 Reporting

Finally - Note down the uncertainties, inconsistencies or suggestions for changes during your work, referencing page numbers in the course books or element ids in your files. We would like to know what kind of improvements/changes are necessary.
We would also like to know your opinion about the process of working on the course book texts using our tool. Please, write down your comments in some informal way. Thank you!

3. Annotation menu. HOW-TO in detail.

Overall hierarchical structure of the elements

Annotation menu contains 8 elements/tags (see sections 3.1-3.9 for detailed information):

Coursebook
Extras
Lesson
Text
-> Genres
Activity instruction
Activity/task
List
Language example

3.1 Coursebook

This menu element collects all bibliographical information about the coursebook: title, authors, year, publisher, ISBN. You will get several pop-up menus, each asking you for some specific information Please, enter all the information thoroughly so that it can be referenced later. After you have completed filling in the menus, you will see something like this in the text field:

```
<coursebook>
<title>Title</title>
<author>Author</author>
<year>Year</year>
<publisher>Publisher</publisher>
<isbn>ISBN</isbn>

Your text here

</coursebook>
```

Automatically, the end tag is added to the end of the text file.
In case of multiple authors, fill in only the first author in the pop-up menu. Then, copy and paste the “author” element as many times as you need, and fill in the second, third, etc. authors, e.g.

```
title>Title</title>
<author>Author1</author>
<author>Author2</author>
<author>Author3</author>
...
<year>Year</year>
```

3.2 Extras

Extras is used to mark the text that is of no immediate interest, but since it is present in the book, it is kept here. You will find "`Table of contents`" here, and "`Extra info`" (for anything else except table of contents). The information that ends up here, need not be edited.

3.3 Lesson

Each new chapter in a course book is called "Lesson" in this editor. Clicking on the "`Lesson`" you will be asked about the proficiency level that this chapter describes. The proficiency level is assigned to the book chapters either by the coursebook writers, or by a council of teachers actively working with the CEFR-based courses, i.e. there will be no need for you to make judgements about the levels. You choose the appropriate level from a drop-down menu in the dialog window, and click `ok`. To get a picture over the levels, look at the following link:

http://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

Next, you will be asked about the lesson id. Automatically, the id will be assigned as a running number of the chapter/lesson. Keep it this way unless there is a strong reason not to.
Further, you will fill in the title of the lesson, which usually coincides with the title in the table of contents. On clicking OK, you will see an added xml-tag, e.g.

```
<lesson id="1" level="B1" title="Some title">
```

and an added ID in the list of IDs, e.g.:

```
List of IDs

1
------------
```

3.4 Text

Step 1. When assigning a text-tag to a passage in the course book, you will be asked for text ID. An automatically suggested ID follows the principle of "text"- lesson ID – running text number. For example, if the lesson-ID is "1", and the text you mark is the third one in this lesson, the ID will be

“text_1_3”. Keep to the automatic suggestions unless there are strong reasons not to.

Step 2. Next, you will be asked if there is any other activity that the text should be referenced to. The general principle is that the references should be to the previous text elements. For example, if before the text there was a list of words that are trained in the text (e.g. <list id=” list_1_2” >), provide the reference to that list ID, i.e.<text id=” text_1_3” ref=” #list_1_2”>. You will be able to see the list of all IDs in the field to the right of the text field. If there is no element that the text takes reference to, leave this field empty.

Step 3. Further, you will be asked about the title of the text. If there is a title, please fill it in; otherwise leave it empty. Please delete the title from the text once you typed the title as XML attribute value.

Step 4. Now you will be offered to select a topic for the text from a list of topics. A general recommendation is to stick to a broader topic, e.g. if it 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 may be applicable, so you may need to mark two or more topics. In case there is no topic that corresponds to the text, please, note that and report to us to introduce changes. Meanwhile, select any topic, and then go in and replace the irrelevant topic with the relevant one manually. The 28 topics available in the list, follow CEFR guidelines (COE, 2001) in the first place, with modifications introduced as a result of our practical work on the previous annotation experience (Volodina & Johansson Kokkinakis, 2013). Here is the list of the topics available at the moment:

- Animals
- Arts
- Clothes & appearances
- Crime & punishment
- Culture & traditions
- Daily life
- Economy
- Education
- Family & relatives
- Famous people
- Food & drink
- Free time, entertainmant
- 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

As a result of your choices you will see an xml tag added to the text field, e.g.:

```
<text id="text_1_1" topic="animals">  -----
```
OBS! No reference in this case

and the list of IDs will be populated with a new ID (added always on the top).

```
List of IDs

text_1_1
1
------------
```

3.5 Genre

A submenu to the `text` element that you will have to fill is on the `genre` of the text. Click on the submenu (in the list to the left). There are four elementary genres to choose from: `narration, facts, evaluation, other`, following the taxonomy of elementary genre families as described in Johansson and Sandell Ring (2010) slightly modified by us during the work on the first annotated books (Volodina and Johansson Kokkinakis, 2013). The genre family of "`Other`" appeared as a result of this work, and contains text genres that were difficult to place into `narration, facts` or `evaluation` families.

Further subdivision of genre families into macrogenres is as follows:

```
Narration:
      Description
      Fiction
      News article
      Personal story

Facts:
      Autobiography
      Biography
      Demonstration
      Explanation
      Facts
      Geographical facts
      Historical facts
      Instruction
      Procedures
      Report
      Rules
```

Evaluation:

- Advertisement
- Argumentation
- Discussion
- Exposition
- Interpretation, exegesis
- Personal reflection
- Persuasion
- Review

Other:

- Anecdote, joke
- Dialogue
- Language tip
- Letter
- Lyrics
- Notice, short message
- Puzzle
- Questionnaire
- Quotation
- Recipe
- Rhyme

It can be discussed whether some of the "other" macrogenres can be moved to the groups above (e.g. "anecdotes" to the narration family). We welcome your suggestions about such modifications.

In a lot of cases, there will be no clear-cut genre. In that case a combination of genres will be an optimal solution. Add several genres if this is what you see necessary by clicking several times at the genre-element in the menu and selecting the appropriate macrogenre, e.g.:

```
<text id="text_1_1" topic="famous people">
<genre><other>dialogue</other></genre>
<genre><facts>biography</facts></genre>
```

3.6 Sub-heading

Sub-heading is used when within the same text there are several mini-texts with separate sub-headings. Those sub-headings should be then kept within <subheading> tags, i.e. deleted moverd from the text to subheading-tags, e.g.

<text id="text_0_1" title="Insändare" topic="crime and punishment">
<genre><other>letter</other></genre>
<subheading>Insändare 1</subheading>
Bla-bla-bla
<subheading>Svar till insändare 1</subheading>

Bla-bla-bla
</text>

3.7 Activity instruction

Activity instruction is the text that usually precedes the task students are supposed to do, e.g.:

```
Välj bland orden i rutan och skriv in rätt form av orden i meningarna
```

To separate these bits of text from other text types, we annotate them as activity instructions.

Step 1. You will be asked to assign an ID. As long as possible keep to the automatically suggested IDs e.g. `ai_1_1`.

Step 2. Where applicable, assign a reference. Remember the rule of thumb that the referencing happens from the actual element to one of the previous elements. The same ID (e.g. for a text) can be referenced from several other course book elements, e.g. lists, language examples, tasks, etc.

Step 3. You will be asked to select target skills/competences that are trained in this activity. The possible choices are:

```
Skills:
      Listening
      Reading
      Writing
      Speaking

Competences:
      Grammar
      Pronunciation
      Spelling
      Vocabulary
```

Mark as many as necessary. It is very rare that only one skill or competence is targeted.

Step 4. In this step you will receive a dialogue asking you for relevant formats/activity types. Mark the relevant ones from the drop-down menu. Note down if something you would like to mark is absent and report to us.

```
Activity types:
      Brainstorming
      Composition/essay writing
      Dialogue/interview
      Dictation
      Discussion
```

```
        Error correction
        Form manipulation
        Information search
        Monologue
        Pre-reading
        Question answering
        Reading aloud
        Role-playing
        Summary
        Text questions
        Translation

Formats:
        Category identification
        Category substitution
        Free/short answers
        Free writing
        Gaps
        Matching
        Multiple choice
        Narration, retelling, presentation
        Reordering/Restructuring
        Sorting
        True-false/Yes-no
        Wordbank
```

After you have filled in all the information and made all the choices, you will receive a tag that would look something like this:

```
        <activity_instruction id="ai_1_1" skill="listening,reading,vocabulary" format="brainstorming,
        discussion,pre-reading">
```

3.8 Activity/Task

This part replicates the previous one (3.6 Activity instruction), with the only difference that here you mark the task itself (e.g. exercise), separating it from the instruction.

```
        </activity_instruction>
        <task id="task_1_1" ref="#ai_1_1" skill="reading,grammar" format="discussion,question
        /answers">
```

3.9 List

You will be marking lists in cases where lists of vocabulary items, phrases, grammatical points, etc. will be provided. They can appear on the side of the pages, but often as a part of activity instructions. We are interested in having them marked as lists separated from instructions.

The process of list annotation is very similar to the activity instructions (3.6) and activity/task (3.7), including assigning of id, reference, and skills/competences. Besides, you will be asked to add a title if relevant, and a linguistic unit. The following linguistic units are available from a menu:

```
Linguistic units:
      Characters
      Dialogues
      Full sentences
      Incomplete sentences
      Numbers
      Phrases
      Question-answer
      Single words
      Texts/examples of text writing
```

More than one category can be selected if there is a mixture of units, e.g.:

```
      </task>
      <list id="list_1_1" ref=” task_1_1”  type="" title="Hos läkare"
      skill="speaking,vocabulary" unit="full_sentences,incomplete_sentences">
```

3.10 Language example

Language example dialogue menus replicate the `list` dialogues. However, they often are seen as examples since they have a different pedagogical aim. Often they introduce a new (grammatical) topic, e.g.

```
Verb i presens = -r/-er
talar
arbetar
bor
-r efter vokal

heter
kommer
-er efter konsonant
```

or

```
Relativt pronomen där
Jag vet ett ställe där det brukar finnas många kantareller
```

Though this is not always obvious, try to identify whether this language information can be referenced to some previous activity, e.g. text or task.

References

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

Britt Johansson & Anniqa Sandell Ring (2010). Låt språket bära. Genrepedagogiken i praktiken. Hallgren och Fallgren

Elena Volodina & 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. [pdf, p.248-259]
<http://webh01.ua.ac.be/linguapolis/LT-CEFR2013/Conference%20Proceedings%20%27Language%20Testing%20in%20Europe%20-%20Time%20for%20a%20New%20Framework%27.pdf>

Elena Volodina, 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. <http://www.lrec-conf.org/proceedings/lrec2014/index.html>

Appendix. Genre families (Johansson & Sandell Ring, 2010:24-25)