Transc&Anno transcription and annotation tool user guide

Contents

1 Preface

Transc&Anno is a browser based transcription and annotation tool. It works with *Mozilla Firefox*.

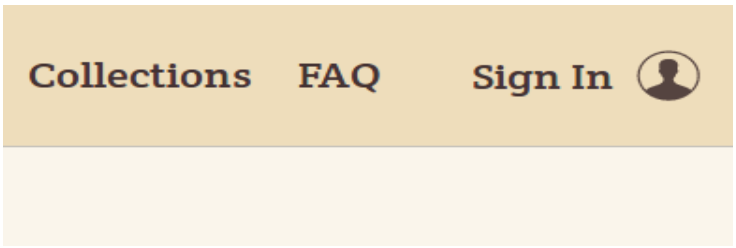
Transc&Anno is useful, if one has text documents in image format and wants to transcribe and annotate them. The transcription and annotation result will be saved to a database in XML format.

Here are the instructions to follow in order to use ***Transc&Anno***:

1
You have to go to the URL https://kommul.eurac.edu/transcanno/

2 Signing in

You have to sign in (top right corner of the page).



There are 4 types of user rights: administrator, collection owner, transcriber and guest.

The collection owner is a person who creates a collection: he uploads the scanned text to be transcribed and defines the annotation tags.

The transcriber transcribes and annotates the texts.

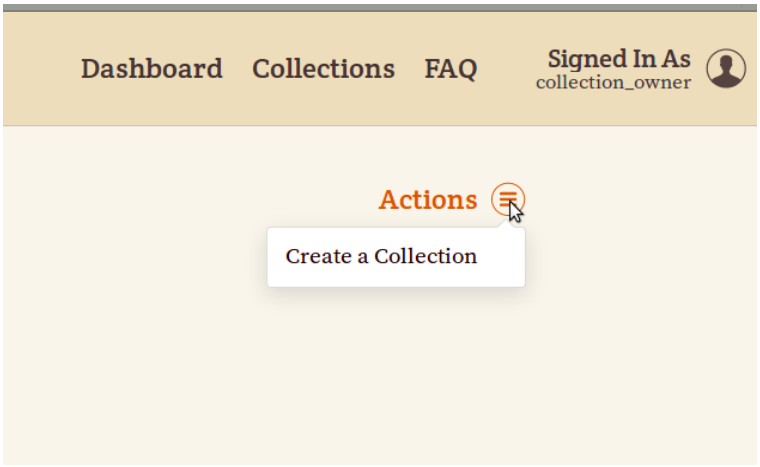
The guest is only allowed to make a few transcriptions.

The collection owner receives his identification details from the administrator.

In order to be able to **create a collection**, you need to have **collection owner rights**. Any account can be given collection owner rights by another collection owner. This collection owner should go to the *Settings* tab of a collection and add an account to the list of collection owners (on the right side of the page). From now on, the account has collection owner rights and can create a new collection.

When ***Transc&Anno*** is deployed to a server, 1 collection owner account is created. Its credentials are saved in the password database of the LT group.

The following 9 steps, from **3** to **12**, have to be executed by the collection owner in order to set up a new collection.



3 Starting a Project

In order to start a new transcription project, you have to click on *Start A Project*. In the top right corner of the page you will see an *Actions* menu. Click on *Create a Collection*.

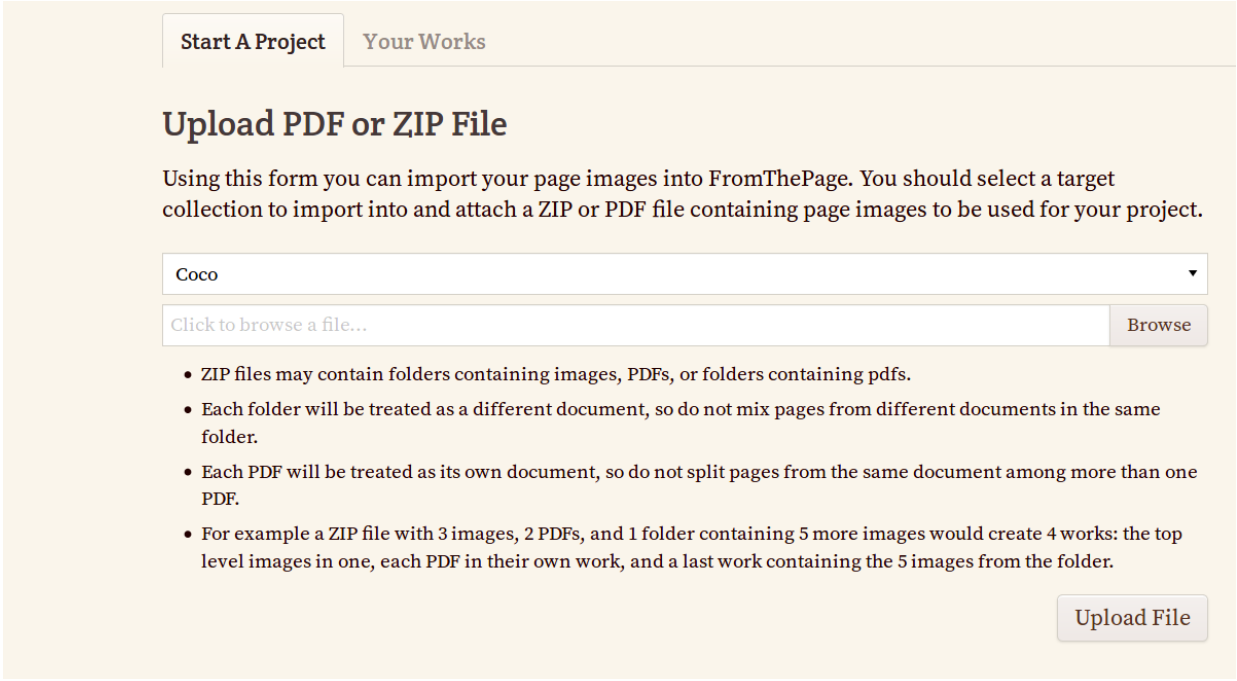
4 Defining User Access to a Collection

By default, a new collection is public. It means that anyone can see its contents and transcribe its

pages. In order to limitate access to a collection, you have to make it private. Go to the *Settings* tab of the collection and on the right side you will see a *Collection Privacy* title. Under this title there is a button *Make Collection Private*. Press this button. From now on only collection owners have access to this collection. Collection owners can be added via a menu underneath, in the *Collection Owners* section. The collection owners you will add will be able to transcribe pages, but they won't be able to change the collection's annotation categories and settings. Only the creator of a private collection (its principle owner) will be able to change the collection's annotation categories and settings.

5 Uploading Text Images

Once you have created a collection, you can upload scanned files to be transcribed.



Upload a ZIP archive containing images and their plain text transcriptions

You may also upload an already transcribed collection and use ***Transc&Anno*** only to annotate it. In that case, you need to prepare a ZIP archive and inside this archive as many folders as there are transcribed documents. Each of these folders will contain one PDF document that can consist of any number of pages. Next to this PDF document, in the same folder, there will be as many plain text documents as there are pages in the PDF. Each plain text document will contain the transcription of one page in plain text format and its name will be as follows: [page_number].txt

Example of a files hierarchy that can be uploaded to ***Transc&Anno***:

Corpus/

butterflies/

Swallowtail.pdf

1.txt

2.txt

birds/

Doves.pdf

1.txt

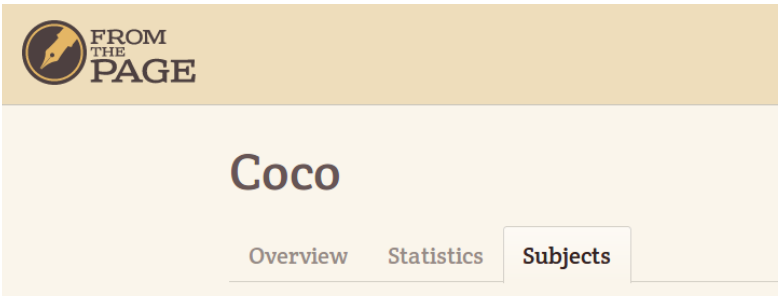
This example shows a corpus consisting of 2 works. The first work contains 2 pages and the second work contains 1 page.

Upload a ZIP archive downloaded from Transc&Anno and containing page images and their transcriptions

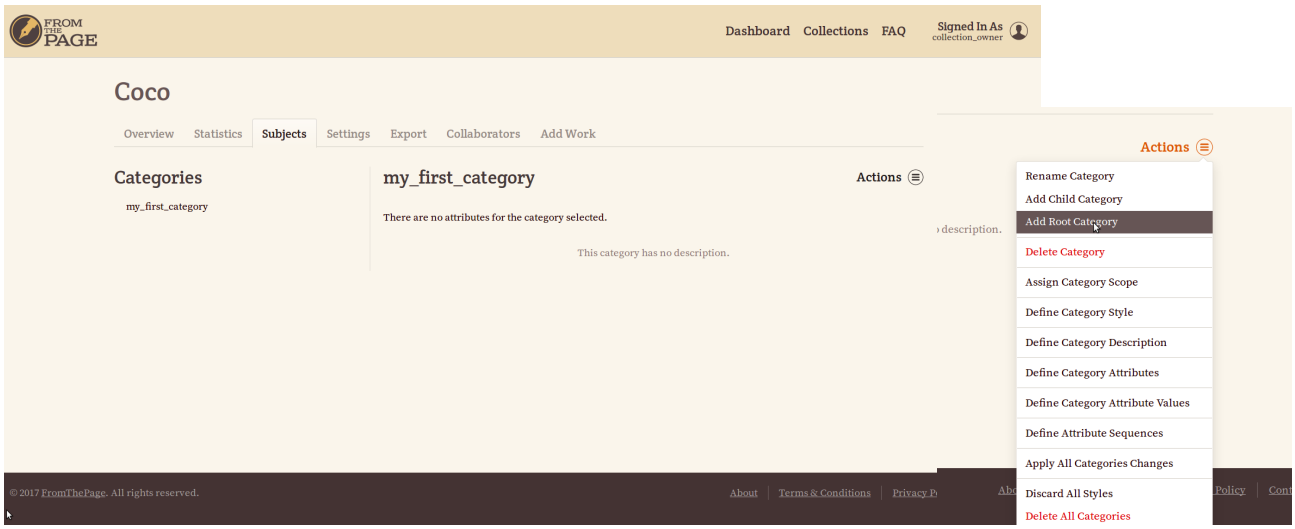
A third option is to upload a collection already transcribed on ***Transc&Anno*** and downloaded with the button *Export All Works Uploadable* on the *Export* tab. Be careful: this will create works and transcriptions of their pages, but won't create annotation categories, neither normal, nor header categories. You have to create them yourself, or, if the collection you are importing is on the same instance of ***Transc&Anno***, import the whole annotation system from the other collection. It won't upload page versions either.

6 Defining Tagging Categories

You also have to define the set of transcription tags. Go to the *Subjects* tab and create the first category.



In the top right corner of the page you will see an *Actions* menu that will let you add new categories and define their styles and types. You have to create **root categories**.



When you are creating a category, you have to decide weather it is an annotation category or a header category. Annotation categories are used to annotate the transcription text. Header categories are used to give additional information about the whole text, for example, the name of the author, the place where the text was collected etc. If you want to create a header category, you have to

check the corresponding checkbox in the category creation menu. However, it is possible to do it later via the menu *Is a Header Category* in the same *Actions* drop down menu.

Header categories cannot have any attributes. They also cannot have a style or a scope. For these reasons the corresponding menus are automatically deactivated for header categories. However, you have to define header categories' values: allow free user input or predefine a set of possible values or both. This is done via the menus *Is a Header Category* and *Define Header Category Values*.

Note that before finalising a transcription, the transcriber will be obliged to fill in the values of **all the header categories** you defined for this collection.

For each annotation (non-header) category you have to fill in all the forms listed in the *Actions* drop down menu and in the **same order as** they are listed in the **menu**:

7 Assigning Category Scope

The *Assign Category Scope* menu allows you to decide if the category will be usable only in the ***simple mode***, only in the ***advanced mode*** or in both modes.

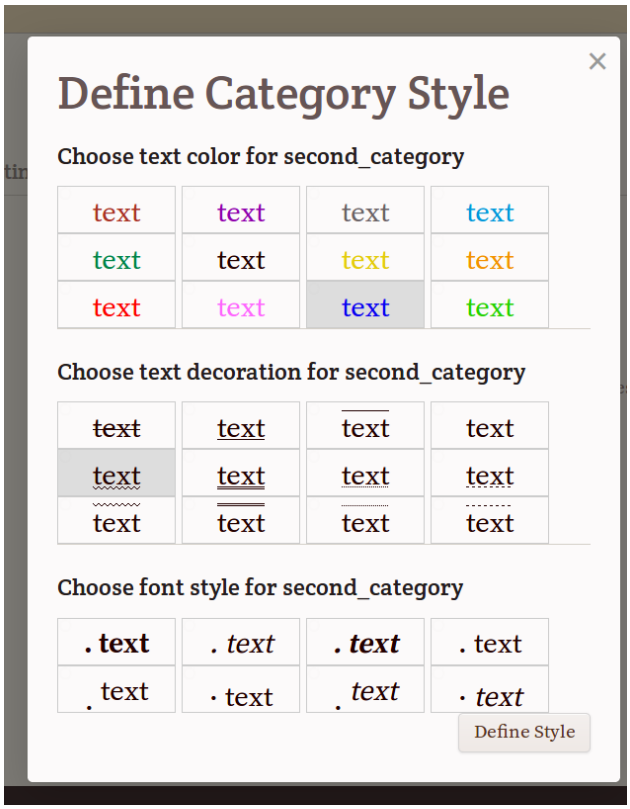
The ***simple mode*** allows you to add a number of attributes to each category and all of these attributes will receive a value during annotation. None of them will be skipped.

The ***advanced mode*** offers the possibility to choose a number of compulsory attributes for each category, all of which will receive a value, like in the simple mode. The advanced mode also offers you a possibility to fill additional attributes in case some other attribute was given a certain value.

> *Example:*
> For example, we may need to make detailed descriptions of word forms. We may need to describe all the verb forms of a text by defining their tense, voice, person etc. In this case, the value of an attribute may allow or not allow the existence of another attribute. Let's take the example of an **adjective**. In German a singular **adjective** has 3 declension paradigms, a masculine, a feminine and a neutral, while a plural **adjective** has only one declension paradigm. Therefore, if the value of the number attribute is *singular*, the annotator needs to define the gender, but if the value of the number attribute is *plural*, he doesn't need to define the gender. This use case demonstrates the necessity to define complex relations between attributes and their values.
>
> Here is an example of a sequence of attributes:
>
> *part of speech -> verb -> tense (**present indefinite** -> person)*
>
> *adjective -> number (**sg** -> gender)*
>
> *noun -> gender*
> *number*

8 Defining Category Style

The *Define Category Style* menu will let you choose the colour and decoration of the tagged text.



It is **not compulsory** to define a category's style, but it is **highly recommended**. If no particular style is defined, the annotator won't know if a certain piece of text has been tagged or not.

Tags can overlap. This should be taken into consideration while choosing categories' styles. For example, if one wants to tag parts of speech and syntactic roles, he could use different colors for different parts of speech and different types of underlining for different syntactic roles.

Buttons corresponding to styles that have already been chosen for categories of this collection appear in gray.

After choosing the style it is important to push the button *Define style*. Otherwise, all the menu information will be lost.

9 Defining Category Description

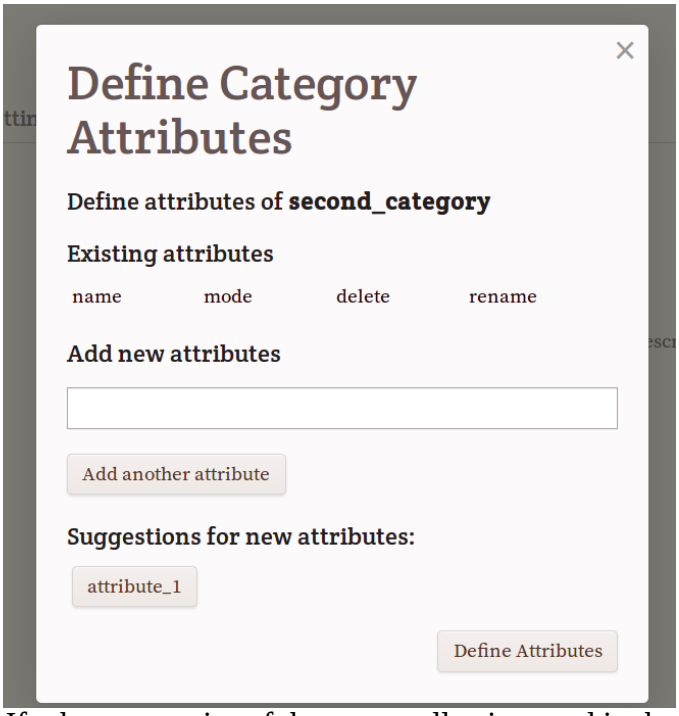
The *Define Category Description* menu allows you to provide a description of a category and its attributes, anything you want the transcriber to know. The description will appear on the *Subjects* page next to the category name.

10 Defining Category Attributes

The *Define Category Attributes* menu will let you define different attributes of the same category. You can either add attributes to a category or not, it's **not compulsory**.

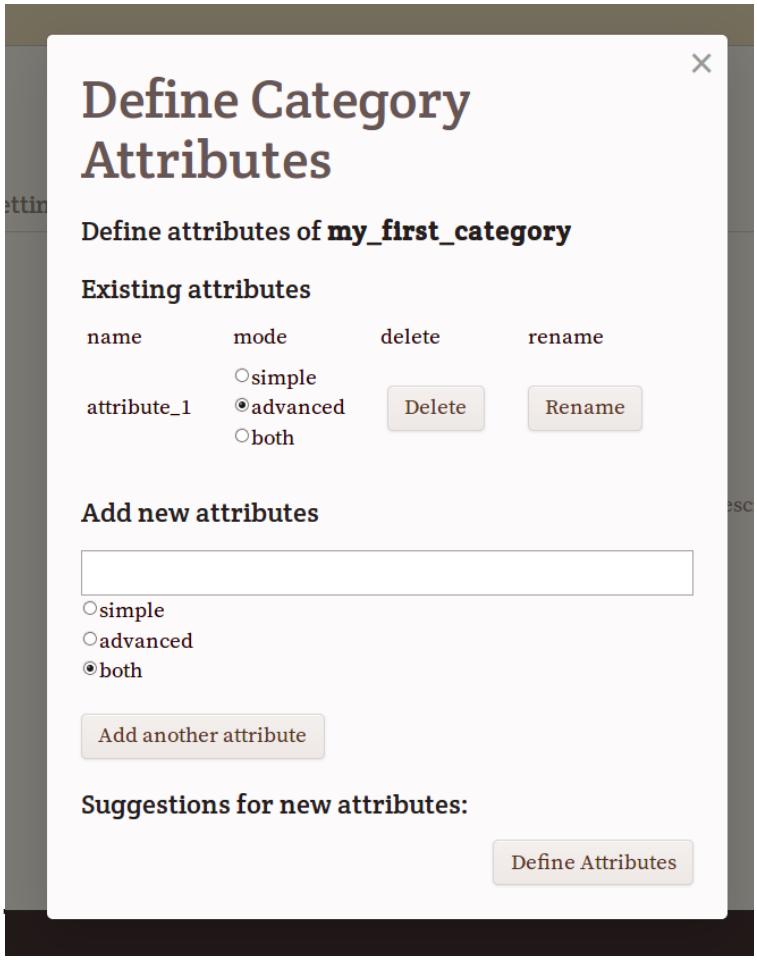
The distinction between categories and attributes depends on the project's needs. One and the same concept can be defined as a category or as an attribute of a category. For example, lets imagine we have to tag prepositions. If prepositions are the only part of speech we are interested in, we won't define an annotation category « preposition », because it would be unique. We would rather define categories referring to the prepositions' meaning, for example, « spacial », « temporal » etc.

However, if we are not going to tag exclusively prepositions, we will define a category « preposition » and it will have an attribute « meaning ».



If other categories of the same collection used in the same scope already have some attributes, those attributes will appear at the bottom of the dialog window as **suggestions for attributes** of the current category. If you push the button with the suggested attribute name, it will appear in the existing attributes' list and will become one of the attributes of this category.

If a category is used in both modes, it's possible to restrict its attribute usage to only one mode.



If the category is used in only one mode, its attributes will also be used only in the same mode as the category.

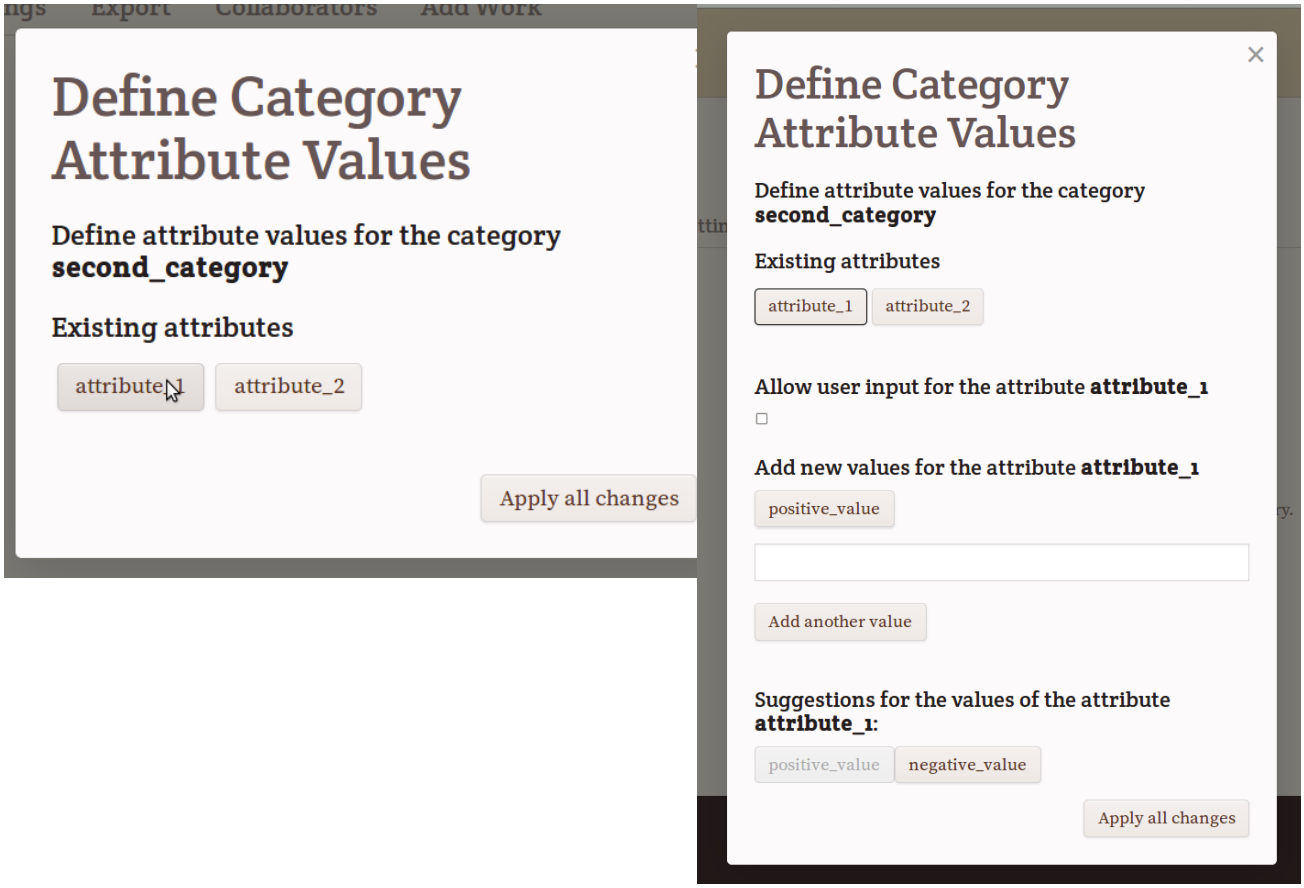
This menu also allows to **delete** and **rename** attributes.

After making all the necessary changes you have to confirm them by pushing the *Define Attributes* button. Otherwise, all the changes will be lost.

11 Defining Category Attribute Values

The *Define Category Attribute Values* menu will let you define values for the attributes you have just defined.

In order to define an attribute's values, you have to click on the button with the attribute's name.



A menu will appear.

You can either make a list of possible values or let the annotator type a new value while transcribing or both. In order to let the annotator type a value while transcribing, you have to check the *Allow user input* check box.

Each attribute must have a possibility to receive a value. Otherwise, the tool will show an error message and the annotator won't be able to use this category.

If **other categories** of the same collection have the **same attributes** as the current category, values defined for these attributes for other categories of the same collection will appear as **suggestions** for values of the attributes of the current category.

> *Example:*
> Lets imagine you need to tag errors and distinguish syntactic and orthographic errors. In this case you need to create a root category **error** and define an attribute type. Then you need to create 2 possible values for it: *synt* and *ort*.
>
> If you also want the annotator to give the correct form, you have to introduce a special attribute for that (for example, correct_form), and check the check box *allow user input*.
> Here is an example outcome:
> *<error-id28 tagcode="1496064202555" type="ort" correct_form="these" class="medium-error-id28">thees</error-id28>*

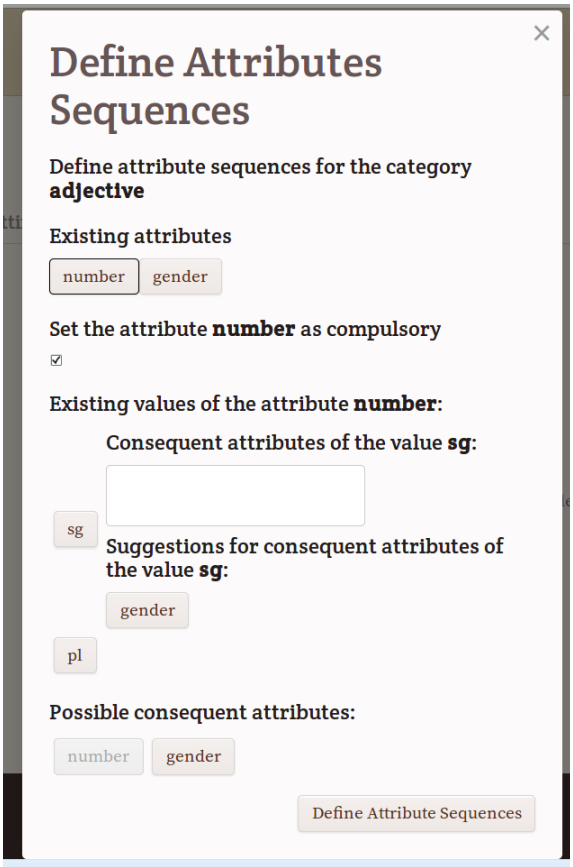
12 Defining Attributes' Sequences

If you have made a category available in the **advanced mode**, you have to define attribute sequences using the *Define Attribute Sequences* menu.

After clicking on the attribute's name you will see a menu letting you define attributes that the transcriber will have to fill if he chooses certain values of the current attribute (image on page 7). First, you have to decide which attributes will be compulsory. At least one attribute must be compulsory. Otherwise, this category tags will have **no attributes at all in the advanced mode**. The choice of compulsory attributes is made by checking the *Set the attribute X as compulsory* check box. Each of the predefined values of the chosen attribute is listed below (*sg* and *pl* on the picture). By clicking on one of them, we open a dialog that allows us to define "consequent" attributes of this value. All the possible attributes are listed at the bottom of the dialog window (*Possible consequent attributes*). These are attributes of the same category. Just under the white area where the consequent attributes will be listed there may be a list of **suggestions for consequent attributes** of the chosen value, coming from the same values of attributes of other categories of the same collection. If other categories of the same collection don't have attributes with the same values or these values don't have consequent attributes, the suggestions list will be empty.

You have to be very attentive **not to create infinite loops**. For example, if a *singular* value of the number attribute induces the gender attribute, a *neutral* value of the gender attribute induces a case attribute and the *Nominative* value of a case attribute induces a number attribute, the annotator will never be able to finish the annotation. The program doesn't verify the validity of the sequences and won't warn you, if you create an infinite loop.

After defining all the necessary sequences for this category, you have to push the *Define Attribute Sequences* button. Otherwise, all the defined sequences will be lost.



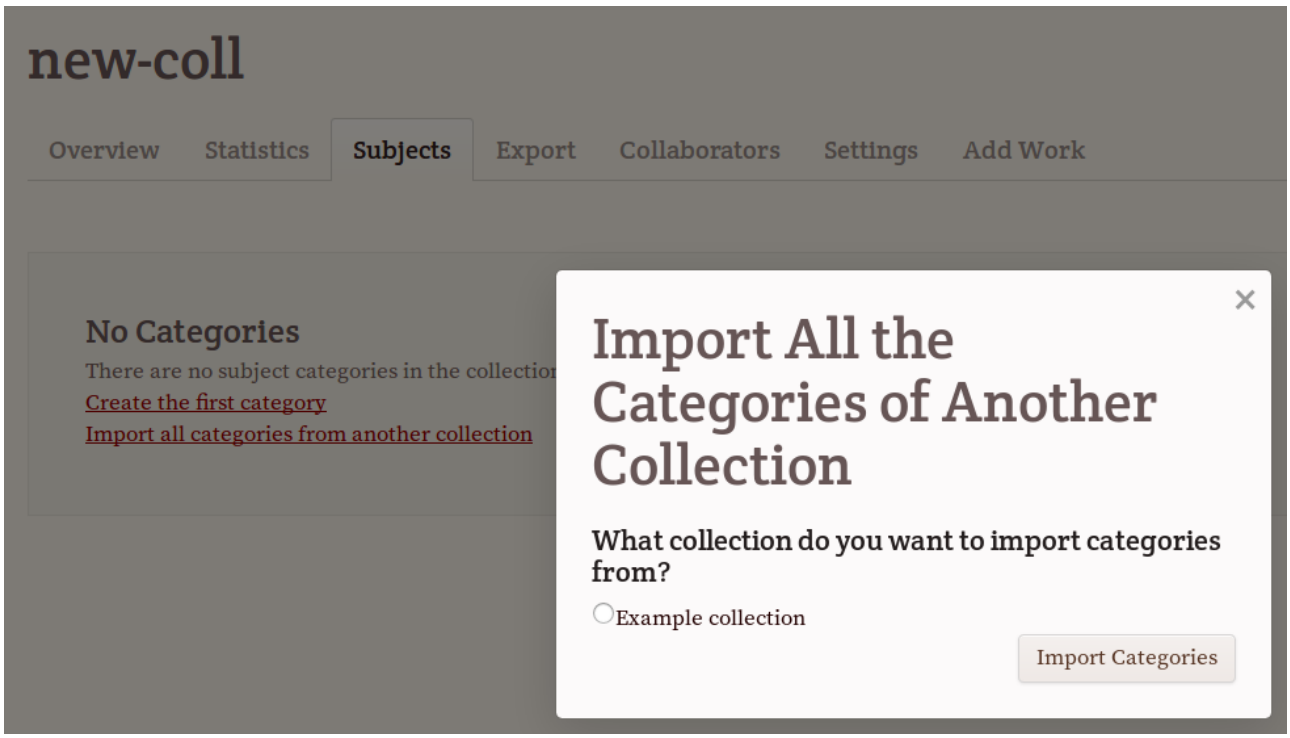
Before starting the transcription process, transcribers can get acquainted with the tagging categories by visiting the *Subjects* page of the collection where all the categories, their descriptions and their attributes are visible.

13 Importing Annotation Categories, Attributes and Values from Another Collection

You can import the whole annotation system from another collection of the same instance of ***Transc&Anno***. However, you cannot import it from another instance of ***Transc&Anno*** deployed on another server.

You can only do it when the new collection has no categories.

Go to the Subjects menu and click on Import all categories from another collection. Then choose the collection you want to import from and click on the button *Import Categories*.



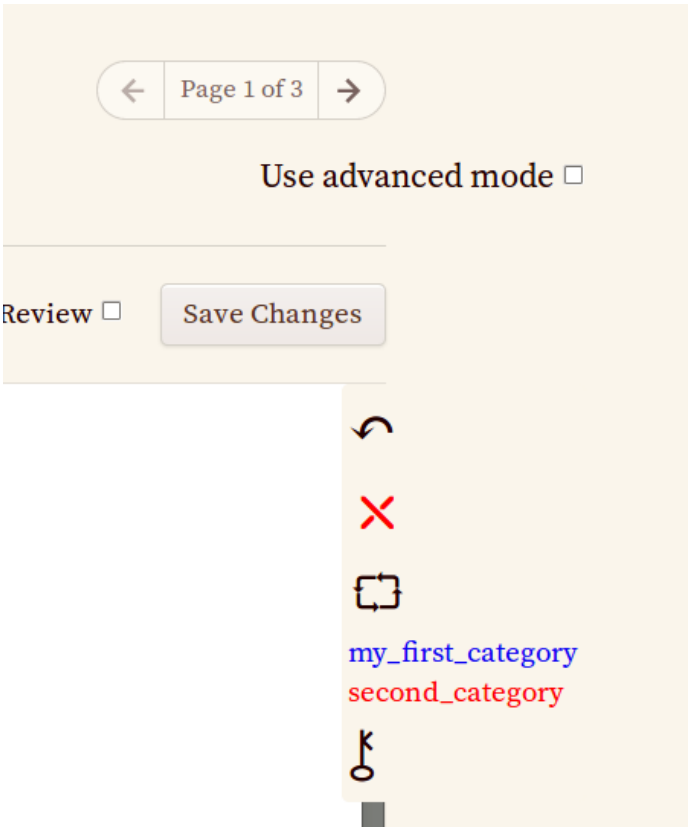
When all the categories from the other collection have been imported, with their descriptions, styles, attributes and values, you can modify them. The changes you make to the annotation system of the new collection will have no incidence on the annotation system of the collection you imported it from.

14 Transcribing

Now that all the categories, their attributes and values have been defined, the transcription process can start.

In order to start transcribing, go to the *Works* tab and click on a work, then choose the page you want to transcribe. You will see the scanned page and an empty space for the transcription next to it. On the right side you will see a list of categories you have just defined.

The annotator has to **choose the annotation mode**. It is the simple mode by default, but he can also choose the advanced mode by checking a check box at the top right corner of the window.



The simple mode lets the annotator use categories having the simple mode scope, the advanced mode lets the annotator use categories having the advanced mode scope.

Both modes allow the use of categories having the “both” flag. However, the list of attributes that have to be filled may not be the same. As you remember, attributes of such categories can themselves receive a scope: **simple**, **advanced** or **both**. Attributes having received the “simple” flag are only usable in the simple mode, attributes having received the “advanced” flag are only usable in the advanced mode and attributes having received the “both” flag are usable in both modes. In the simple mode all the attributes defined for this category have to be given a value during annotation, while in the **advanced** mode it depends on the defined **sequences**. For example, if no attribute is defined as **compulsory**, no attributes will be proposed during annotation.

There are 2 ways of inserting tags:

1 First insert a tag and then type the text inside.
2 First type the text, then highlight a portion of it with a mouse and tag it.

The tag can be added either using hot keys or using buttons on the right side of the screen.

Here is the list of hot keys:

Alt+C : insert a tag or tag a highlighted piece of text

When you press Alt+C, a pop up menu will appear offering a choice of categories available in this transcription mode. You can choose one of them with the mouse, with an arrow key or by typing the first letters of the category name. Once you have chosen the category, if the latter has attributes, each of them will appear accompanied by a list of possible values or a text input field where you will be able to type a new value. When all the attributes will have their values assigned, the tag will be inserted into the text.

Alt+X: get the cursor out of the current tag in order to continue typing outside the tag

Alt+R: make a pop up menu disappear. In this case no tagging is performed.

Alt+N : delete a tag from the portion of the text where the cursor is

A menu listing all the tags present at the cursor position will appear, the user has to choose the one he wants to delete.

This functionality is also available through the “cross” looking button on the right side of the screen.

Alt+M : modify a tag from the portion of the text where the cursor is

A menu listing all the tags present at the cursor position will appear, the user has to choose the one he wants to modify. Once the tag chosen, its attributes and there values will appear. The user will be able to change the values of these attributes. He won’t be able to add or delete attributes. In order to do this, he should delete the tag and insert another one in its place. And if the category configuration defined by the collection owner will suggest other attributes as consequences of the newly chosen values (in the advanced mode), the attributes finally present in the tag will be different.

If the user modifies attribute values of a certain tag, the program doesn’t check if they are compatible with each other according to the system of sequences of attributes and values defined by the collection owner (in the advanced mode). It means that in case of annotation errors the result may be slightly different from the one expected by the collection owner. The annotation quality is the responsibility of the annotator.

The tag modification functionality is also available through the “electric circuit” looking button on the right side of the screen.

Alt+, : insert a German lower double quote

Alt+. : insert a German upper double quote

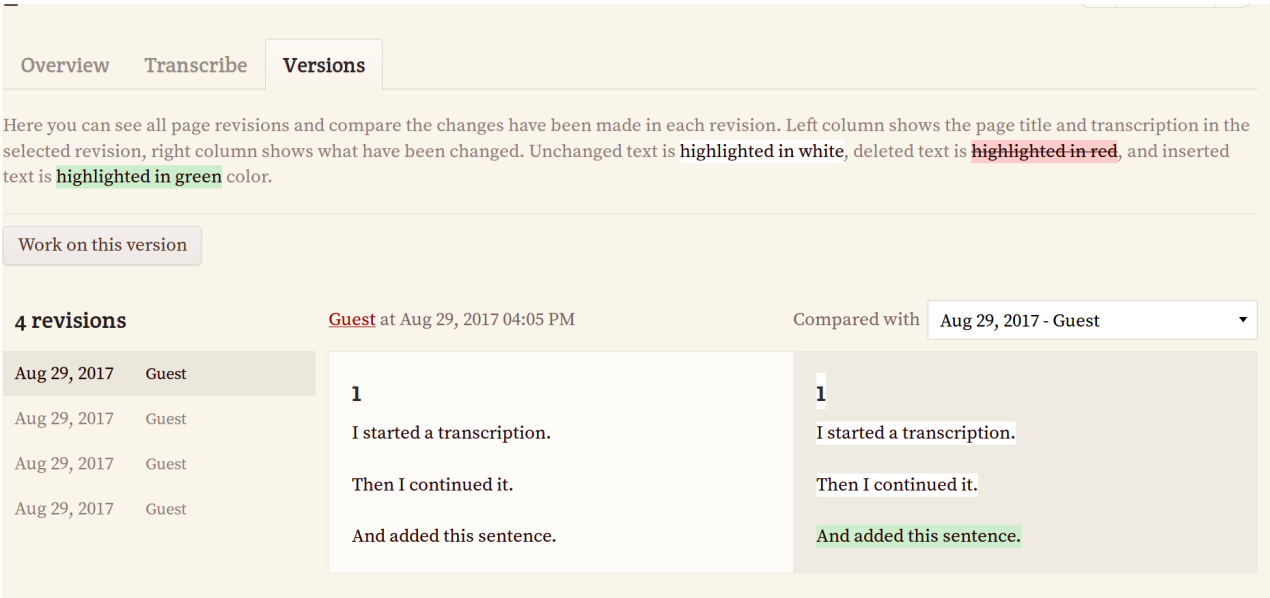
All these **hot keys may be modified**. In order to change them, you have to push the “key” looking button in the right part of the window.

15 Saving Transcriptions

Transcriptions are automatically saved every 3 minutes. However, you have the possibility to change the transcription saving frequency using a sand clock looking menu on the right side of the screen. After changing the transcription saving frequency you have to reload the page so that the new value be taken into account. Note that if you decide to use another browser, you will have to repeat the procedure, otherwise, it will be the default value. The custom transcription saving frequency value is saved in the browser cookies.

If there are errors in the XML code of the page (for example, due to accidental mouse mismanipulation), an error message will appear and the page won’t be saved. If you see an error message, it means that there is a mistake in the work you have done during the last 3 minutes (or another period of time if you have redefined it). In that case or if you notice that you have made

some other mistakes and want to go back to a previous version of the page, you can go to the Versions tab next to the *Overview* and *Transcribe* tabs on the transcription page. This page allows you to compare different versions of the same transcription and choose the one you want to work on.



By pushing the *Work on this version* button on the left side of the screen you define the chosen version as the main one. Now you can go back to the Transcribe tab and continue transcription. You will be working on the version you have just chosen.

Despite of the automatic saving, when you have finished transcribing, you have to push the button *Save changes* to be sure that all the changes have been saved.

When you are sure that everything is right in your transcription and you will never have to go back to a previous version, you should push the button *Transcription finished* and leave the transcription page. The *Transcription finished* button deletes all the previous versions of the page and only leaves the last one. The fact that you leave the *Transcription* tab insures that no more automatic saving will be made and no more page versions will be registered. It is important to **delete the previous versions**, because in this way you liberate precious space on the computer hosting the transcriptions database. Thank you in advance for pushing the *Transcription finished button* every time you have finalised a transcription!

This button also allows you to verify that you have filled in the values of **all the header categories**. If you have forgotten at least one of them, you will be notified.

The values of the header categories are saved to the text of the transcription into the <textinfoheader> tag as shown in the example below.

Here is an example of a text tagged with ***Transc&Anno***:

```
<?xml version='1.0' encoding='UTF-8'?>
    <page>
        <textinfoheader><author_id>2</author_id><collected_in>Bozen</collected_in></textinfoheader>
            <greeting_id8 class='medium-greeting_id8' mode='1' tagcode='1501593576726'>
                    <adjective_id6 class='medium-adjective_id6' mode='1'
            tagcode='1501595124365'>Dear</adjective_id6>
```

```
                Reader
            </greeting_id8>,
            <main_clause_id13 class='medium-main_clause_id13' mode='1' tagcode='1501596400575'>
                <subject_id11 class='medium-subject_id11' mode='1'
            tagcode='1501595065526'>I</subject_id11>
                <predicate_id12 class='medium-predicate_id12' mode='1'
            tagcode='1501595071144'>
                    <verb_id10 class='medium-verb_id10' mode='1' person='1'
            tagcode='1501595105142' tense='present ind'>am</verb_id10>
                    glad
                </predicate_id12>
            </main_clause_id13>
            <subordinate_clause_id14 class='medium-subordinate_clause_id14' mode='1' tagcode='1501596407114'>
                you
                <verb_id10 class='medium-verb_id10' mode='1' person='2' tagcode='1501595080621' tense='past
    ind'>read</verb_id10>
                this tutorial.
            </subordinate_clause_id14>
            <br/>
            <closing_id9 class='medium-closing_id9' mode='1' tagcode='1501593710728'>
                <adjective_id6 class='medium-adjective_id6' mode='1'
            tagcode='1501595133332'>Best</adjective_id6>
                regards
            </closing_id9>
</page>
```


Dear Reader,

I am glad you read this tutorial.

Best regards

16 Exporting Transcriptions

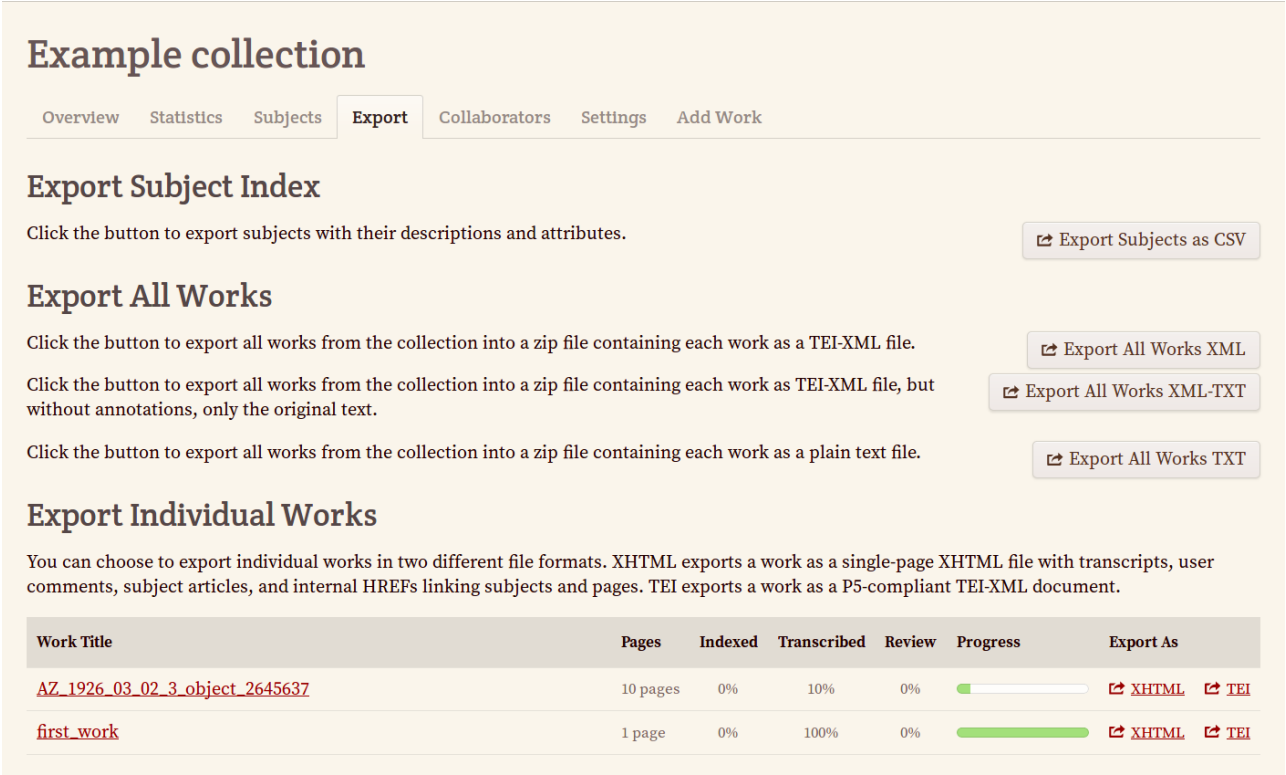
The transcription result can be exported either in a variety of formats. In order to do so you need to go to the *Export* page of your collection.

The whole collection can be exported into an archive containing 1 file for each Work. Each work can be saved in:

- TEI-XML containing all the annotations
- XML containing only the original text of the transcriptions, without annotations
- TXT containing the original text of the transcriptions in plain text format, without annotations
- XHTML format into a .zip folder or in TEI-XML format.

Each work can also be downloaded separately in TEI-XML or XHTML format.

As for the annotation categories, they can be downloaded as a CSV file that will contain each category's name, description and list of attributes. This CSV file **cannot** be used to upload annotation categories back into ***Transc&Anno***.



All the people that can see the collection can export its content, not only the collection owner.

17 Searching the Collection for a Word or Number

The search menu on the right side of the screen allows you to search the collection content for words or numbers. You can type a word or number in the search field and on the left side you will find only the pages containing the string you were looking for. By default only the text content of the transcription will be searched. If the search string is part of an annotation category name or value appearing on a page, it won't be considered as a match. If you want the search string to be looked for also in the annotation tags, you need to check the box *"search in transcriptions including tags"*.

By default, if the search string is part of a word, the page containing this word will be selected. If you want to only look for whole words, you can check the corresponding checkbox.

Example:

Text:

<greeting type="informal">Hi!</greeting> I am waiting in the garden. <closing type="informal">See you.</closing>

Search strings:

garden: will always match

gar: won't match with the *"whole word"* checkbox checked

informal: will only match with the *"search in transcriptions including tags"* checkbox checked

type: will only match with the *"search in transcriptions including tags"* checkbox checked

greeting: will only match with the *"search in transcriptions including tags"* checkbox checked

There is also a possibility to search in the values of header categories. To do that, you need to check the header categories in the values of which you want to find the string. However, it is impossible to search at the same time in the transcription **in**cluding tags and in the values of header categories. On the other hand, it is possible to search at the same time in the transcription **ex**cluding tags and in the values of header categories.

The search is **case insensitive**. It means that the search strings **boy** and **BOY** will give the same result.

Regular expressions can be used in the search field, with some limitations. Regular expressions allow to widen the meaning of a search string, to make it less precise and more permissive, and therefore to find more pages in the same search.

Here come some examples of regular expressions:

(boy|girl) => looking for pages containing the string "boy" OR the string "girl"

bo. => looking for pages containing the string "boy", or "boa", or "bot", or "bob", or "bo1" etc. : 3 letter words starting with "bo". The point character replaces any character.

However, ^ and $ symbols defining the beginning and the end of a string cannot be used. For example, the expression **^(boy|girl)$** will give no result.