Lesson 6: Exercises

Learning goals

- Starting, saving and continuing annotation work (e.g., transcription of a conversation) in ELAN
- Orthographic transcription in ELAN
- Exporting a conversation transcript from ELAN into a plain text file
- *Types* of annotation tiers in ELAN
- Tokenizing a transcribed annotation tier in ELAN
- Performing queries from a number of annotated files in ELAN

Exercises

1. If you do not already have ELAN on your computer, download and install it (https://archive.mpi.nl/tla/elan).

   Note: If you are using a computer maintained by the university, you may need to ask your IT support person to install the program. On computers maintained by the University of Helsinki, ELAN should also be available as a package that can be installed directly from the Software Center (see instructions for Windows).

2. Open ELAN.

3. Watch the video tutorial on how to annotate speech with ELAN: https://aoe.fi/#/materiaali/1637

   If you like, you can try repeating the same things with the same material while watching the video or afterwards. The video uses the files **reitti_a-siipeen.wav** and **reitti_a-siipeen.mp4** that are included in the openly available *Route to A-wing corpus*. You can download the audio and video files from Kielipankki – the Language Bank of Finland. (The corpus is officially called *Route to A wing Corpus, Downloadable Version* and you can find the metadata about it via the persistent identifier urn:nbn:fi:lb-2020112929).

   Don't worry if you do not have a video in your own language, the purpose here is just to see what can be done in ELAN and to practise the basic workflow with some speech material.

   You have already practised annotating in Praat. This work becomes a lot easier when you can simultaneously watch the video, doesn't it?

   Practice just enough so you know how to add and edit annotations and how to create new tiers as needed. Try saving the file and make sure you can open it again in ELAN (including both audio and video). Also try to export the annotation file to text format with the **Export As: Traditional Transcript Text…** command or to Praat TextGrid format with the **Export As: Praat TextGrid** command.

   Let's move on then!

4. Select a WAV sound file on your computer or download one (you can use for instance one of the files in the story North Wind and the Sun you downloaded before). The sound file should contain at least one complete sentence, preferably several. Start annotating it in ELAN:

   Create a new annotation project (**File: New…**), and add just one media file in the project. Save the annotation file (remember to use the file extension ***.eaf***).

5. To start with, it is enough to create one single annotation tier. But we should change the name of the annotation tier called *default*, to make it a bit more descriptive. Select **Tier: Change Tier Attributes…** Rename the tier as *M1-clause* (or something similar, depending on the type of speaker and the content of the tier) and click **Change** to accept the new tier name.

6. Delineate and transcribe all sentences in the audio file to this annotation tier. You can locate the sentence boundaries roughly or more precisely according to your own interest.

   Remember to save the annotation project!

7. Once you've got the story, or at least most of it, written down, let's try linking some annotation tiers in ELAN.

   - In ELAN, each annotation tier can be ***typed***: the tier can be equipped with information about the type of information described by the annotation units it contains. For example, the same file can have multiple annotation tiers, some containing the orthographic transcript of the speech of different participants in the conversation, other tiers can contain grammatical descriptions of individual words, some tiers may be used for indicating gaze, i.e., to what direction a participant is looking at each moment, some tiers may contain a description of background noise, etc. Types can be used for grouping different annotation tiers together or even for creating annotation hierarchies.

   - Typing is especially valuable for searches. Once an entire corpus with multiple annotation files has been annotated in a consistent manner, searches can be made of the entire corpus and a single search can be applied at once to specific types of tiers within the entire corpus.

   - In ELAN it is also possible to use so-called ***controlled vocabularies*** (*CV*) for specific types of annotation tiers, in which case only a closed set of labels are allowed in the annotations contained in the tier. This can be useful in specific application areas, e.g., if you want to annotate the part-of-speech of words, categorize voice quality, or label hand movements with a particular system. When a controlled vocabulary is enabled, ELAN automatically monitors the annotation tiers and makes sure that the user can only enter one of the labels that are allowed.

   - In the default project in ELAN, only one "*Tier type*" is used for annotation tiers. The default type is labeled as *default-lt*. To make the default type more descriptive, select **Type: Change Tier Type…** The tier annotated so far probably contains units resembling a clause or a sentence, so you could change the name to "clause", for example. Click **Change** to accept the change.

- If you go and view the properties of the annotation tier you previously annotated (**Tier: Change Tier Attributes…**), the *Tier Type* should now display "*clause*" if the change in the type was successful.

8. Next, let's try to create a new type for annotation tiers where we wish to annotate the individual word tokens that are included in the annotated clauses or sentences.
   - Select **Type: Add New Tier Type…**
   - Type something like "*word"* as the ***Type Name.***
   - We also want the *word*-type annotation units to occur consistently within the boundaries of *clause*-type units. You can specify this relationship between annotation tiers by selecting ***Included In*** under ***Stereotype***.
   - Finally, click **Add.**

   Extra tip:
   If you are going to annotate, for example, the part-of-speech of each word to a separate annotation tier, you should create your own type for part-of-speech. It is done otherwise as above, but under ***Stereotype*** you need to select ***Symbolic Association*** (which means that each annotation of the "parent tier" corresponds to exactly one annotation in the "child tier"). Of course, you could also name the tier type as "*part of speech*" or "*POS*".

   Since a well-defined set of labels (e.g., verb, noun, adjective, numeral, etc.) are used to denote part of speech, a controlled vocabulary can be defined in advance. If you like, you can try creating one in **Edit: Edit Controlled Vocabularies…** First enter a name (***CV Name***) for the vocabulary and click **Add**. The titles in the vocabulary and their descriptions are then added one at a time at the bottom of the window. The new vocabulary must then also be applied to the selected tier type by selecting **Change Tier Type: Use Controlled Vocabulary**. After this, ELAN will help you to select one of the labels you defined by displaying a drop-down list each time you edit a part of speech annotation.

9. Next we eill automatically **tokenize** a tier, i.e., we will segment all the annotations in the tier called *M1-clause* into individual words, by creating a new annotation tier with the type ***word*** (which we just defined):

   - Select **Tier: Tokenize Tier…**
   - For *Source tier (parent tier)* you need to select the original tier called *M1-clause.*
   - Click on the button **Create New Tier…**, to the right of *Destination tier.*
   - A new tier is created where the individual words will be inserted. You can call the tier "M1-word", for example. Under Parent Tier, select the original tier *M1-clause*. Under **Tier Type**, select "word". Finally, click Add and then Close to close the window. A new tier called *M1-word* will appear in the background, but it is still empty.
   - Proceed with the tokenization by clicking Start in the **Tokenize Tier** window.
   - The word units corresponding to the words in the original annotation tier will automatically appear in the new annotation tier. Isn't this convenient! (Of course, word boundaries will be automatically marked at even intervals, and so they will not match the correct time points in the audio file. In case you need precise word boundaries for your study, you will need to manually check and move the boundaries.)

- Make sure to save the file again at this point!

10. If you want to annotate the part-of-speech of each word to its own tier and you already created a *tier type* for this property (see above), you can now create a tier for the part-of-speech of the words read by speaker M1:
    - Select **Tier: Add New Tier…**
    - Name the tier as *M1-POS, for example*.
    - Select the previously created word tier as the parent tier for the new part-of-speech tier: *Parent Tier: M1-word*.
    - Select *Tier Type: part-of-speech* (or *POS*).
    - Accept by clicking **Add.**
    - Annotation units created in the part of speech tiers will now automatically follow the boundaries of their corresponding word units, i.e., the boundaries cannot be moved independently in the POS tier.

11. What kind of annotation tiers would you need for your own studies? What kind of relationships between them could be defined in ELAN (by using tier types)?

12. What if you wanted to search the material you annotated? There are several search functions in ELAN's **Search** menu:
    - **Find (and Replace)…,** search in the annotation file that is currently open
    - **Search Multiple eaf…**, simple string search from eaf files on your local computer or from directories
    - **Structured Search Multiple eaf…**, advanced search from eaf files or directories on your local computer, defining the search criteria according to annotations in multiple tiers

    Note. To begin with **multiple eaf** operations, you first need to specify the search domain where all the annotation files in EAF format will be queried. For the search domain, you can pick a specific set of individual *eaf* files and/or entire directories on your computer. The search domain can be saved and used again later, i.e., this allows you to specify complete speech corpora for your personal use.

13. Try searching the files you annotated in ELAN by applying the different search commands. If you need more specific instructions, see the ELAN manual.

    When defining the criteria for searching for words or similar units, you can also apply so-called regular expressions. Regular expressions can be used to define, for example, the beginning or end of the annotated string, characters that occur or recur within a single annotation, characters that must not appear in the target annotations, etc. A list of all operators for regular expressions supported by ELAN can be found here: http://www.mpi.nl/corpus /html/elan/apa.html

    Suggested searches:

    - Look for words whose annotated form contains a double vowel character, e.g., *aa*.
    - Look for words that appear at the beginning (or end) of a sentence (or other units you delineated in your annotations).

    - Tip: You can mark the beginning of an annotated string in regular expressions with ^ and the end with **$**. The word boundary is marked as **\b**.
    - Any word (in English or in Finnish) could possibly be expressed as the following kind of regular expression:
      \b[a-zåäöA-ZÅÄÖ]*\b
      = any string of letters a-ö or A-Ö that occurs between word boundaries. (The asterisk means that there can be any number of the characters mentioned within the square brackets.)
- For example, if you have annotated parts of speech, try searching for all word units in token type annotation tiers that contain a **V** (verb) in the part of speech tier and begin by the letter **i**.

14. Now you have an idea of what can be done with ELAN and how "ordinary" annotations can be **enriched**.
    - Finally, consider for what purpose ELAN is most useful and where you think Praat is better.
    - Which do you think you will need more in the future, Praat or ELAN?
    - Can you use both programs in parallel? If you want, you can try importing a Praat TextGrid annotation file to ELAN by selecting **Import: Praat TextGrid File…**

Congratulations on your hard work - I hope you found it useful!