Guidelines for annotation

KDK-SZTAKI MILAB PROJECT

Set of subject headings

- We currently have a hierarchical list of 220 subject headings, compiled first from the translation of the ELSST thesaurus, then from the 600+ word list selected from that, and then from the 242 words narrowed down from that, with additions that were not included in the extended glossaries
- Each word has a clearly marked place in the hierarchy AND each word is only in one place, whether it is a lower, middle or upper level term
- Anything from the 600+ list that didn't make it into the shortened list has been replaced by something from the 220 list
- In compiling the vocabulary we have tried to keep an overview, highlighting some of the key terms for our collections, e.g. terms related to the Holocaust (but generally speaking, only basic terms will fit in a list of this length)

Clarification of the purpose of the annotation

What is the purpose of the annotation?

- to associate subject headings with sections that describe well the socially relevant content they contain - the primary aim here is not to teach the machine
- and to assign keywords to the selected subject headings that refer well to the subject headings - this is teaching the machine
- treat each passage as if it were a completely new, stand-alone text (as strange as it may sound) - technically, this is how the machine can interpret it.

I.e. if family has been mentioned many times before and is mentioned again in a particular passage, then we will also assign the subject heading there (this is important, by the way, because in the end, the priority subject headings the machine gives to texts may also depend on how many passages a given subject heading is assigned to - so it makes a difference whether someone talks about family once or continuously)

Label Studio

- Task: assign subject headings to each interview passage and select keywords from the text for the latter
    - Section: the interviews can be read in sections and annotated
    - Subject heading: an element of the fixed subject list (KDK-Thesaurus) that describes an important unit of content of a given section
    - Keyword: a phrase specifically (literally) mentioned in the given section, selected by the annotator, which refers to an important unit of content and can be linked to the selected subject heading

- **The aim is to target content that is salient and socially relevant** (e.g. if someone mentions something once but it does not seem important, it is not a subject heading, even if it is an "important topic", e.g. he reads a book about the Holocaust and mentions it, but otherwise talks about something else)
- In all cases, everyone should choose the most precise word/expression possible, both subject headings and keywords:
    - this may mean that, if you find a better connection as you move on in the text, you may want to replace the previous one or add the new one as well
    - avoid interpretative correspondences, which in practice means that if e.g. escape/fleeing is mentioned by the interviewee and the context reveals that it is escape from prison, then we must choose the most important, most precise keyword to which the subject headings is assigned, e.g. in the example above, prison would be ideal, not escape (as escape can be from many places/situations)

Guidelines:

- a maximum of 5 subject headings should be associated with a section (can be 0)
- a subject heading should be accompanied by a maximum of 4 keywords, but more can be used if justified
- a keyword can be a specific word or a specific word combination (2, maximum 3 consecutive words), half sentences should not be selected, AI cannot work with them
- if there is really no good keyword, but it is important for the content of the section to specify a specific subject heading, then no keyword is needed
- if the keyword to be marked consists of several elements (e.g. a parting verb, a multi-word conjunction...) and these elements are not in direct succession, either do not mark it a keyword, just assign the subject heading, or mark the elements of the keyword separately and add the same subject heading to each of them, but avoid such cases, try to find another keyword for the subject heading
- a keyword can be linked to more than one subject heading if it is logical to do so
- it is also useful to associate a keyword with a subject heading that is exactly identical (subject heading=assimilation / keyword=assimilation - that's OK)

The annotation process is a method - steps:

1. assigning keywords to content that seems important while reading the section
2. delete on the fly / at the end of the passage the keywords that were not really good and replace them with better ones
3. reviewing the keywords at the end of the section, deleting keywords that are not important (content that is not important)
4. reconsidering the content of the section (what is important content) -> whether a keyword should be taken out / whether there is content for which there is no

keyword yet (if there is, then search for it, if no good one can be found, then only subject heading should be added, no keyword)

5. select subject headings from the list that describe well the contents identified
6. if there are more than 5 subject headings, decide whether more subject headings are justified (justified if the passage is longer than average) -> if justified, more subject headings can be used, but the number must be communicated to the co-annotator / if not justified, select the 5 most important subject headings
7. finalize subject headings
8. pairing of keywords and subject headings
9. check - make sure all pairing has been done
10. if there is an extreme case (e.g.: there is important content for which there is a subject heading but no good keyword, etc., etc.), then this should be indicated in a separate document

Technical notes on coding:

- to find the right subject heading while annotating, you might want to use the original excel sometimes, it shows all the elements of the hierarchy expanded, it's easier to see, search in it... - can be useful!
- collect specific extreme cases in a separate document, e.g.: there is no good keyword for a given subject heading, or the elements of the keyword are far apart, but this is the only one that could be associated with a selected subject heading... - to see what such cases are, and maybe come up with a better solution than the current one
- if a subject heading is very much missing from the list, we collect suggestions (for possible future improvement of the subject heading list).
- submit button:
    - you should finalize a section with the submit button when it is ready
    - although the update button can still be used to change the annotation, clicking submit gives SZTAKI the freedom to export the subject headings and keywords for the section and start working with them
- when the two annotators have finished their joint interview:
    - let SZTAKI know (e-mail), who will export, compare and feedback whether a third coder is needed
    - the comparison will be made on the basis of subject headings only
    - if a third annotator is not needed, the two original annotators will produce the final annotation (gold standard), from which the machine will learn
    - if a third annotator is needed, a third annotator is assigned and decides on the gold standard

Technical notes for Label Studio:

- if you click on a keyword that you have already selected, the associated subject heading appears in a green box directly above the text - you can probably see what you have assigned to it better here than in the right-hand bar
- if a lower level subject heading is selected, only that will appear in the selected list, but in the exported version (which is what SZTAKI will be working with), the subject heading above that word will also be there, as a subject heading for that section