NRK-Quiz-QA Guidelines

In this document we describe the annotation process for quiz data provided by NRK, the Norwegian national broadcaster. The data presented consist of an ID tied to the original quiz. Each question has an ID, taken from the original data, a date from when the quiz was published, a question, and a set of options, which varies in number. The IDs reflect the original quiz ID together with a question index. Questions also have a fact associated with them. There is also a column indicating which of the options are correct, and a column for different labels. Multiple correct answers can be possible, in which case they are all listed, separated by a comma. The example below shows a question with only two possible answer options.

Below each question there will be a new row with the same ID followed by a "b". This is where the annotators write potential corrections. The main task for this set of questions is to fixing various issues in the original Norwegian data.

id	date	question	option A	option B	fact	correct answer	label
1.15018510-14	15.05.2020	Treff midt i blinken? Treffer midt i blinken?: Kva er rett?	Johannes Thingnes Bø treffer som regel midt i blinken.	Johannes Thingnes Bø treff som regel midt i blinken.	Å treffe/a – treffer – trefte – har treft Er du i tvil på korleis verbet skal bøyast, finn du svaret i ordboka.	A	
1.15018510-14b							

Past events or states

One issue with the Quiz data is that they might have questions related to the specific time when they were published. For example, if a question asks about who holds a certain position, or what the state of something is, then the questions must be changed to fit the time. This might also be relevant for answer options.

For example, in March 2023, the following question was published:

Rettssak: Det florer klipp med Gwyneth Paltrow i ein rettssal – kvifor er ho der?

This could be changed to:

Rettssak: Det florerte klipp med Gwyneth Paltrow i ein rettssal i 2023– kvifor var ho der?

Where the year was added, and the verbs were made into the past tense.

Media-reliant questions

In some cases a question cannot be answered without access to a photo originally provided along with the quiz, or perhaps a sound file or video. Since these questions will not be provided with images, it is important to mark these.

Noise

Some questions or alternatives might contain remaining html tags, urls that cannot be used or references to media etc. These should be removed so that only the question or options themselves remain.

Irrelevant Information and Comments

In some cases, the questions contain comments or descriptions that guide the user through the quiz, or that provide information that is not strictly necessary. This extra text should be removed. Look at the two sentences below:

1) **Var den lett? Vi tar en til.:** Han skal prøve....overbevise folket
2) **Kan flere se sola?:** Kan du se midnattssola også om du befinner deg litt sør for den nordlige
polarsirkelen?

The first two short sentences in 1) contain a question to the user about the difficulty of the previous question, and then a comment about the flow of the quiz itself. In 2) The first sentence is just a prelude to the actual question. These can be removed. The cleaned version of the question is then written in the row below.