RefCo Checklist for the Corpus Reviewer

This document provides the list of quality criteria for evaluating a corpus submitted to RefCo. These quality criteria are grouped in thematic modules. This list of quality criteria comes with the RefCo_CheckList_report.ods file for actually reporting the answers to the list of questions present in this file. As it is part of a quality process, these reports are meant to be given back to the Corpus Submitter as well as the RefCo Certification Entity, so that the Corpus Submitter can improve the corpus and that the RefCo Certification Entity can keep track of its activities.

Reviewing the Corpus Documentation

To review the Corpus and its Corpus Documentation, the reviewer should check all the following elements listed in the sections below. Each title in this section refers to a RefCo module that a corpus has to pass. For each of these modules, after a short introduction describing the module, a question is associated to each item that has to be checked and reported in the RefCo_CheckList_report.ods file.

The first step of the reviewing process is to rename the report file as follows for facilitating the accountability:

YYYYMMDD_CorpusName_RefCo-Report.ods

There is a distinction made between errors and warnings in the RefCo Certification process, as the certification can also contribute to improve the quality of the corpus. Warnings are indicated in the report list by using the tag ***warning***. Contrary to an actual error, a warning does not prevent the corpus to be prevented, but are here as part of the discussion between the Certification Entity and the Corpus Submitter to improve the quality of the corpora.

To report an error or a warning, if for a reviewing step, you have to answer "no" to the question, it means that the item being reviewed fails and that you should report either an *error* or a *warning* in the **Results** column corresponding item in the Report file. In the **Comments** column, please indicate reformulate the question that was used to review the item. Else, if the point currently reviewed passes, simply indicate *pass* and do not fill the **Comments** column. At the end of the report process, please indicate the number of errors in the **Certification Process** tab.

Certification Process

This module contains information useful for certification process the corpus.

Report Summary: List of the title of all the checks that failed here.
Corpus File Checksum: Not to be reviewed, if missing, report to the RefCo Certification Entity.
Date of the Review: When was the review performed? (YYYY-MM-DD)

Functional Tests[1]

A first important quality is whether the corpus is their functional. This module checks whether the files in the corpus are readable, properly formatted and complete, that is, for instance, that an annotation files do not refer to a recording not actually in the corpus.

- **Misc:**
  - Are there errors while *opening the file*?[2]
  - Are there *missing files*? This is particularly relevant as some annotation software required files for managing the project.
  - Are there multiple versions of some files?
- **Annotation:**

1 Not directly part of RefCo, could be tested upstream in QUEST, but some tests are specific to RefCo's usages.
2 These errors are often due to the author moving the files from one folder to another for sending the corpus. It's also a behaviour of ELAN that could be improved by relying more on relative paths.

- Are the annotation files in the corpus valid[3]?
- Are the annotation tiers actually containing annotation units?
- Are the annotations displayed and the characters inside them displayed?

- **Recordings**:
  - Can you play the recordings with an external player?Can you play the recordings with the annotation software integrated player?
- **Documentations**:
  - Are the documentation files readable?
  - and are the documentation files valid?

Corpus Design

This module is about checking the design choices made by the Corpus Creator when the corpus was created.

- **Corpus Structure:**
  - Does the corpus follow the indications given in the Reference manual? If not, does the corpus follow a folder structure that is explicit?
- **Naming Conventions:**
  - Can we easily *match the annotation files with their recordings*?
  - Warning: Are the *file names in the corpus following a convention* (camel case, ISO 8601: YYYYMMDD, corpus specific)?
  - Warning: Does the file name conventions help *interpret their content*?
  - Warning: Is the name *conventions usable*? That is, does it help or, on the opposite, interfere working with them (too short names, without semantic significance, do not help distinguish the files in the corpus)?
  - Warning: Are the *file names conventions are coherent*? Does it help interpret existing subsets of the corpus? That is retrieving the speech genres (interview, narratives, events, etc.) associated with the annotated text?

Corpus Documentation

The Certification module is dedicated to the review of the Having the RefCo Corpus Documentation is part of the quality criteria for the cross-linguistic reusability of a corpus. For a corpus to be validated by Quest RefCo, the Corpus Submitter has to fill a corpus documentation which the RefCo Reviewer will review, report on it and if it passes, validate.

Overview

This module checks whether the metadata of a corpus specifies the license and if the corpus is available on an archive.

- **Corpus Title:**

3 Please use file validators for checking the actual validity and portability of the files (facile.cines.fr/, ELAN software can be use for eaf files).

- Is there a title given to the dataset? does it inform on its content?

- **Target Languages(s):**
  - Are the languages specified using a proper Glottocodes or ISO 639-3 code?
- **Archive:**
  - Is the *corpus archived*?
  - Is there an *online access?*
  - Is how to request *access to the corpus specified*?
- **Corpus Persistent Identifier:**
  - Is the URL given valid?
  - Can you actually access the corpus submitted by the author using this address?
- **License Annotation Files:**
  - Is the *license of the annotation files specified*?
  - Is the *access open* or is it limited?
- **License Recording Files:**
  - Is the *license of the recordings specified*?
  - Warning: Is the *access open* or is it limited?
- **Corpus Submitter Name:**
  - Is the name properly indicated?
- **Corpus Submitter Contact:**
  - Is the email address valid[4]?
- **Corpus Submitter Institution:**
  - Can you find the institution using the name given?

Certification

Corpus ID and Corpus Documentation's Version are not to be checked as they are provided by the RefCo Certification Entity, so they should be already indicated. Please report to the Certification Entity if it is not the case.

Quantitative Summary

- **Number of Sessions:**
  - Is the Number of sessions specified?
  - Can you understand how are the sessions indicated in the corpus using the file names? That is there a semantic way of naming the files so that you can regroup their according to their sessions?
  - Total number of (we still have to figure out which numbers we want).

4 Please send actually an email to check this step.

- **Annotation Strategies:**
    - Are the translation languages indicated using Glottocode or ISO 639-3? CorpusComposition

CorpusComposition

- **Sessions:**
    - Warning: Is the naming convention used for the session column semantically informing?
- **Files:**
    - Error: Are the file names indicated corresponding to actual files in the submitted corpus?
    - Are the extensions of the file indicated?
    - Warning: Are only the files given in the column present? If not, please ensure that there files are relevant for the dataset.
- **Speakers:**
    - Is the speaker's name indicated?
- **SpeakerAges:**
    - Did the Corpus Submitter follow the conventions for approximate age?

AnnotationTiers

This section should provide details about the annotation tiers used in the corpus.

- **Names:**
    - Is the name indicated corresponding to a tier name used in the corpus?
- **Functions:**
    - Is the function given expliciting the name?
- **SegmentationStrategies:**
    - Warning: Does the segmentation strategy indicated correspond to what has been done in the corpus?
- **Languages:**
    - Are the languages specified using a proper Glottocodes or ISO 639-3 code?
- **Speakers:**
    - Is the name given matching with a name in the given in the CorpusComposition Speakers column?
- **MorphemeDistinction:**
    - Is the MorphemeDistinction specified for tiers which deal with glossing?

Transcription

- **Graphemes:**
    - Is the grapheme actually used in the corpus?
    - Are all the graphemes documented?

- **LinguisticValues:**
  - Did the Corpus Creator provide only the linguistic value?(The linguistic value should not come with interpretation marks, like // or [].)
- **Linguistic Conventions:**
  - Warning: Can you find a reference expliciting the linguistic convention given by the Corpus Submitter?

Glosses

- **Glosses:**
  - Warning: Is the gloss used at least only once in the corpus?
- **LGR:**
  - Warning: Is it actually a gloss that is part of the LGR?
- **Meanings:**
  - Is the explanation helping to understand the character's usage?
  - Warning: Is the explanation used only once?
- **Comments:**
  - Warning: Is the comment helpful?
- **Tiers:**
  - Does the tier indicated correspond to an actual tier documented in the AnnotationTier section?

Punctuations

- ***Characters:***
  - Is the indicated character used as a punctuation mark in the corpus?
- ***Meanings:***
  - Is the explanation helping to understand the character's usage?
- ***Comments:***
  - Warning: Is the comment helpful?
- ***Tiers:***
  - Does the tier indicated correspond to an actual tier documented in the AnnotationTier section?
- **Functions:**
  - Error: Did the corpus submitter tried to use the controlled vocabulary before creating their own words?

CorpusOpenDescription

- ***Keywords:***
  - Is the keyword not already taken by the Corpus Documentation?

- ***Descriptions:***
    - Is the description explaining the keyword and its associated values given by the Corpus Submitter?

Glossary

- **Terms:**
    - Warning: Is the term in the corpus?
- **Descriptions:**
    - Warning: Are the explanations given helping to understand the meaning of the term?

General Review

- **Corpus Coherency**:
    - Does the RefCo Corpus Documentation describe adequately the content of the annotation files?
- **Corpus Consistency**:
    - Does the annotation files in the submitted corpus all follow the same annotation conventions?
- **Translation:**
    - Warning: If there is a translation, is there at least a translation per annotation unit ?
- **Annotations:**
    - Is there a morphological annotation layer describing each morpheme of each annotation units ?