RefCo Corpus Documentation

Reference Manual

Date: 2021.07.09

Document Version: 0.6

Authors: Aznar Jocelyn, Seifart Frank

1 Introduction....2
1.A Creating the Dataset for RefCo....2
1.A.a Organizing the Submission Package....2
1.A.b Naming the DataSet....2
1.A.c Naming the Corpus Documentation....3
2 Filling the Corpus Documentation Template....3
2.A.a General Conventions....3
2.A.b Counting the Number of Words or Glosses Using ELAN....3
2.B Definitions of Corpus Documentation's items....4
2.B.a Metadata at the Corpus Level: Overview....4
2.B.b Metadata at the Text and Annotation Levels....5
Corpus Composition....5
Annotation Tiers....6
Transcription....7
Glosses....7
Punctuations....7
CorpusOpenDescription....8
Glossary....8

1 Introduction

This Reference Guide provides explanations for creating a language documentation dataset, including information on how to fill in each entry of the Corpus Documentation. The Corpus Documentation is a template document that should be filled in order to provide the information required for reusing the submitted dataset but which cannot be retrieved or inferred from the corpus itself. The Corpus Documentation should be submitted together with the corpus to the RefCo Certification Entity as it will be used both for checking the dataset by the RefCo Entity and by potential RefCo reusers.

1.A Creating the Dataset for RefCo

When submitting a dataset, the Dataset Submitter should follow certain conventions regarding the file names and the folders. In this section, we provide a description of these conventions.

1.A.a Organizing the Submission Package

The number of files in a corpus can vary from one project to another, but having the same folder structure from one project to another helps dataset reusers to find the information they are looking for. Unless the submitted corpus follows some particular conventions devised specifically for the corpus and its problematic, RefCo recommends Dataset Submitter to respect the following conventions for organizing the folder containing the corpus files.

All the documents related to the corpus should be contained in a unique folder which should respect the Naming conventions indicated in .

This main folder should contain three subfolders: Annotations, Metadata, and Recordings. The Annotation folder should contain only annotation files. The Metadata folder should contain only metadata information, such as the Corpus Documentation and the Recording folder should only contain the audio or video files of which there is the corresponding annotation file in the Annotations folder.

Do not include multiple versions of the same file, old files, or incomplete files. The submitted files are the files that will be evaluated for the certification and potentially reused by other researchers, so they should be submitted only if they are considered coherent and complete.

1.A.b Naming the DataSet

The file should be named as follows:

<DateOfSubmission>_<GlottoCode>_<CorpusSubmitterName>_<CorpusName>.zip

For instance, a valid corpus name would be:

2021-10-22_nisvai123_JocelynAznar_NisvaiCorpusOfNarratives.zip

Each tag in the file name should follow the CamelCase convention (the first letter should be in capital letters as well)[1].

The <DateOfSubmission> tag should follow the format: YYYYMMDD.

1See https://en.wikipedia.org/wiki/Camel_case for a description of this convention.

The <GlottoCode> tag should be taken from https://glottolog.org/, if the language is not in the glottolog database, indicate a name.

The <CorpusSubmitterName> tag should be made the Corpus Submitter's first name first, and then last name.

The <CorpusName> tag should be a title describing the object of the corpus.

1.A.c Naming the Corpus Documentation

The Corpus Documentation file should be named using the following pattern:

CorpusDocumentation_<GlottoCode>_<CorpusSubmitterName>_<CorpusName>.zip

For instance, a valid name would be:

CorpusDocumentation_nisvai123_JocelynAznar_NisvaiCorpusOfNarratives.ods

2 Filling the Corpus Documentation Template

This section gives the necessary indications to fill the Corpus Documentation file.

2.A.a General Conventions

These General Conventions are valid for all Corpus Documentation entries:

- To indicate multiple values in a cell, please use the coma "," to separate the values.
- To indicate an approximate value, please use the tilde "~" before the figure. For instance, the age of someone who is approximately thirty years old should be given as "~30".

In the context of filling the Corpus Documentation, a gloss is a term used to describe a segment of the corpus, that is a morpheme, a word, a sentence or any kind of sequence to which a linguist wants with a description. Glosses are often abbreviated and can be specific to a particular description, thus they should be documented.

2.A.b Counting the Number of Words or Glosses Using ELAN

How the number of words in the corpus is counted will again depend on the file formats. In an ELAN corpus, there are at least two solutions, which might still have to be adapted depending on how your corpus was annotated.

One possibility is to copy and paste the content of the relevant tier in a Word processing software and count the words using that software. To do so, to select Window>Annotation Window, go to the Text tab, chose the relevant tier and copy its content into a text file. Repeat this operation for all the text of your corpus.

The second possibility is to use ELAN's export function (see https://archive.mpi.nl/forums/t/word-count-in-elan/1141). The multiple file export “List of Words” (File->Export Multiple Files As->List of Words), with the option “count occurrences” selected, creates a two-column text file (tab-delimited) with all unique words and the number of occurrences of each word. Opened in a spreadsheet application, the number of rows represents the number of unique words, the sum of the numeric values in the second column should be the total number of words.

Using a regular expression (e.g. for word boundaries) in the “Structured Search in Multiple eaf’s” might also be used to render the word count.

2.B Definitions of Corpus Documentation's items

This section explains every item that can be found in the Corpus Documentation and describes how to fill them. It is designed to be very practical, the list should be used as a dictionary.

2.B.a Metadata at the Corpus Level: Overview

The first section is the **Overview** section, there are four different subsections: ***Corpus Information***, ***Certification***, ***Quantitative Summary*** and ***Annotation Strategies***.

The items in the ***Corpus Information*** section provide metadata regarding the whole corpus.

The **Corpus Title** must be a title given by the Corpus Creator to the dataset. This title should help reusers to have an idea of its content. **Example:** The language documentation of Navajo, a corpus of Nisvai narratives.

The **Subject Language(s)** item corresponds to the main languages documented by the corpus. In addition to the language name, the Corpus Submitter needs to specify a language identification code, preferably Glottocode (https://glottolog.org/), alternatively ISO 639-3 (https://iso639-3.sil.org/). **Example:** (Nisvai,nisva1234)

Archive contains the name of the archive or repository entity hosting the corpus. Please provide the full name, not the abbreviation. **Example:** Pacific and Regional Archive for Digital Sources in Endangered Cultures

Corpus Persistent Identifier indicate if possible the Persistent IDentifier (a PID, which can be an handle or DOI) referring to the corpus's access point. If the corpus cannot be associated with a PID, please indicate an URL that links towards it. **Example:** https://hdl.handle.net/11403/sldr000783/v2

Annotation Files License must correspond to the license the Corpus Submitter associates with the corpus annotations. Specifying different licenses on the file level is not supported. **Example:** CC-BY-ND

Recording File Licence must correspond to the license the Corpus Submitter associates with the recordings in the corpus. Specifying different licenses on the file level is not supported. **Example:** CC-BY-NC

Corpus Creator Name is the name of the person who submits the corpus to the Certification Entity. Please indicate here the name you want to be associated with the corpus when the latter will be cited by future reusers. The recommended practice is to give first the given name (or first name), middle name and then surname (or family name). **Example:** Gabija Eglėkutė

Corpus Creator Contact is the email address the Corpus Submitter wants to be contacted with by future corpus reusers. **Example:** g.eglekute@univ-ama.rt

Corpus Creator Institution is the complete name of the institution to which the Corpus Submitter is associated at the time of application. You can indicate the acronym in parentheses after if relevant. **Example:** Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)

The **Certification** section provides metadata relevant for the QUEST certification process. These are provided by the Certification Entity, so the Corpus Submitter does not have to fill in anything in this section.

The **Corpus ID** is a unique identifier within the QUEST project that should be provided by the Certification Entity after the Corpus Submitter submits the corpus. **Example:** 7df98c045f31007a9ec6dc4824c7cd6b3141568390cc039b88f80021135f02df8a5580837a33d26f859e7fafe026a42615fdc8faa6a969dc8b5629e0b8305ea2

The **Corpus Documentation's Version** indicates which version of the Corpus Documentation's template this document has been made from. Using this information, the Corpus Re-user can know what features to expect from a RefCo corpus. This information is provided with the Corpus Documentation template, the Corpus Submitter should not have to fill it. **Example:** 2

The ***Quantitative Summary*** contains information about the number of annotations of a corpus to allow potential Corpus Re-users to assess whether the size of the corpus matches their requirements.

Number of sessions, where "session" typically corresponds to one text that has typically been recorded during one recording session, and typically consists of one audio recording and one corresponding annotation file (exceptionally more than one of either or both). **Example:** 42

Total number of transcribed words should indicate the number of words contained in the tier associated with the transcription.

Total number of morphologically analysed words should indicate the number of morphologically analysed orthographic words in the corpus, i.e. morphologically segmented and annotated with morpheme glosses, and optionally part-of speech tags.

The **Annotation Strategies** section provides information on the corpus and the choices made by the Corpus Creator when they created the annotations.

Translation Language(s) specifies the language(s) used for translating the Target Language(s). In addition to the language names, the Corpus Submitter needs to specify a language identification code, preferably Glottocode (https://glottolog.org/), alternatively ISO 639-3 (https://iso639-3.sil.org/). **Example:** Bislama,bisl1239

2.B.b Metadata at the Text and Annotation Levels

Corpus Composition

Corpus Composition lists of all the files present in the Submitted Corpus. It requires the Corpus Submitter to provide a minimum list of metadata for each session. The Corpus Submitter can submit additional metadata in the **CorpusOpenDescription** section.

Sessions must contain a unique identifier within the Submitted Corpus that identify a group of files as a coherent linguistic event. The unique identifier should be as transparent as possible on the semantic level, and not a very abstruse symbol. A **Session** typically corresponds to a text that has been recorded during a recording session, and associates the recording(s) to their corresponding annotation file (exceptionally more than one of either or both). But this is also particularly relevant when multiple files are part of the same text or interview. **Recommended values:** the sessions' names should provide some contextual information about the recording, such as speech genre, recording situation or any metadata relevant within the corpus. **Examples:** interview_1, story_of_Naharuelc, speech-event_1

Files must specify which files, including their file extension, belonging to the same session. Each entry should comprise at least one audio file and one annotation file. **Examples:** T1_part1.wav,

T1_part2.wav, T1_annotations.eaf, interview_1a.mp3, interview_1b.mp3, interview_1c.mp3, interview_1.eaf

Speakers must contain names for speaker or speakers identified by a corpus-unique alphanumeric ID. Please use a code if you have to anonymize the name, this code can be a name that connotes the same social characteristics if relevant. **Examples:** Alfred Junior Wash, john1,

Gender must be an English gender equivalent term that you would use to translate how the speaker refers to themselves, using either their personal pronoun or the lexical term they use as an indication of their identity. **Examples:** If the speaker uses he or his, you should indicate male, she or her: female, they or their: other. This list is only indicative and can be adapted depending on the project's context.

SpeakerAges must contain the speakers' age at the time of the recording. If there are multiple speakers, the values should be matching the order given in the **Speakers** columns. If the exact age is not known, precede the age with the character tilde "~" to indicate that it is an approximate age. **Examples:** ~30, 59.

PlacesOfRecording must indicate the name of the place where the session was recorded. The type of toponyms retained for indicating where the recording took place (city, street, country, island) should be decided by the Corpus Submitter according to the purpose of the documentation project. This field is not for situating the corpus within the world but to provide documentary information about the recordings. **Example:** Tokyo, Delhi, Tempelhof, Malekula, 18 place de la Concorde.

DatesOfRecording must indicate the date when the recorded was made. It should be provided following the ISO 8601 standard: YYYYMMDD. If the date is not certain, please indicate only the year and the month, or only the year.

SpeechGenres must identify the genre of the session. We provide here suggestions for genre categories, but others can be used. **Examples:** traditional narrative, personal narrative, conversation, stimulus-based (including the stimulus name e.g. Pear Story).

Annotation Tiers

The **Annotation Tiers** section must provide a description of all the different types of tiers used in the corpus.

Names, each entry of the column must specify the name of a tier as it appears in the annotation files. **Examples:** tx, ft, mb, POS, Trans

Functions must explicit the purpose of the annotation tier. **Recommended values:** Transcription Reference, Note, Part-of-speech, Morpheme glossing, Morpheme segmentation, Free Translation **Acceptable:** Any short description of maximum 5 words.

SegmentationStrategies must contain an explanation specify what is the main unit of segmentation associated with the tier. **Examples:** Intonation units, interpausal units, breath groups, clauses, words, morphemes.

Languages must indicate the main language used in the tier by providing a prose language name (that name must still be present in the list of names provided by the Glottocode or the ISO 639-3[2] convention). If the name of your language is not in these lists, indicate only the name that the

2 See https://glottolog.org/glottolog/language for Glottocodes or https://iso639-3.sil.org/code_tables/download_tables for the ISO codes.

speakers use to refer to their language. **Recommended values:** Mandarin Chinese, English French, German, Indonesian, Portuguese, Russian, Spanish

Morpheme Distinction This field must be filled in for tiers containing morphological glosses to document whether the glossing uses upper vs. lower case to distinguished lexeme glosses from grammatical glosses, as described by the Leipzig Glossing Rules, or another strategy that should be explained. Specify "UpperGramLowerLex" if grammatical morphemes are systematically glossed using upper case while lexical morphemes are glossed using lower case, "None" if the conventions are not used or "Not relevant" is this does not concern this particular tier. If another convention is used, please describe it and provide details about your annotation strategy in the **CorpusOpenDescription** tab.

Transcription

In this section, the Corpus Submitter should specify the linguistic values of the graphemes used for transcribing the annotated text.

Graphemes should list all characters (including di- and trigraphs) that are part of a doubly articulated system used to transcribe the speech in the transcription tier. If the graphemes used are the same as their counterpart in the linguistic value, simply copy them in both columns. **Examples:** ʧ,a,ɓ,

LinguisticValues column indicates the value, typically phonetic or phonological, used in the transcription tier. If a single linguistic value is transcribed by multiple graphemes, please create different entries for each value. If a grapheme in the transcription has multiple values, it is possible to simple indicate them using a comma. **Examples:** ʧ,ɒ, ,b_<,

LinguisticConventions should specify the transcription conventions used to interpret the linguistic values associated to the graphemes. **Examples:** IPA, X_SAMPA,

Glosses

This section primarily serves to explain abbreviations used in morpheme glosses, and also in other tiers, such as part-or-speech tiers. Extensive use of the **Comments** column should be made to explain potentially partial coverage. If a speaker uses an abbreviation so that it is part of the transcription tier, the best practice is to document it using the glossary. Glosses and conventions that are following the Leipzig Glossing Rules do not have to be documented.

Glosses, each entry should contain an abbreviation used in the corpus, acronyms used by the speakers can also be given here. In theory, all the abbreviations used in the corpus should be present in the column. The best practice would be to create different entries if a tag is used both in capital letters and in small case. **Example:** OBLI

Meanings should provide a linguistic description of the abbreviation. **Example:** oblique

Comments is a place open for any additional information the Corpus Submitter wishes to submit.

Tiers, indicate the name of the tier the entry is associated with. **Example:** mb

Punctuations

The **Punctuations** section is dedicated to the description of how punctuation signs were employed within the documented corpus. Each different usage of a punctuation mark or set of punctuation marks according to its annotation tier, the Corpus Submitter should create an entry.

Characters, each entry should contain a form, that is the punctuation mark or set of punctuation marks, described, in the column. **Example:** -

Meanings, the corpus submitter should describe how the characters should be interpreted. The description should be has extensive as necessary. **Example:** The character indicates an affix split between two morphemes.

Functions describe using an open controlled vocabulary, on which level the punctuation character interacts with the description. If the values in the recommended set are not matching your usage, please add another. Use only a short expression to quality the function. **Recommended values:** morpheme break, prosodic cue, external events, unsolved interpretation, notes.

Comments is a place to provide additional (contextual) information about the characters. **Example:** If there is nothing transcribed next to the dash character, it means that it has can be used to indicate a zero morpheme as well that should be glossed in the mb tier.

Tiers, the Corpus Submitter should specify in which annotation tier the characters as used with this specific meaning. Note that if the same punctuation mark is used twice with different meanings, feedback should be given to the Corpus Submitter to correct this ambiguity. **Example:** tx

CorpusOpenDescription

This tab allows Corpus Submitter to provide their own original metadata at the file level, that is in the **Corpus Composition** tab. To do so, a Corpus Submitter should indicate the name of the category they want to add after the existing column by creating an entry in the CorpusOpenDescription tab and put the name in the Keywords column. The Corpus Submitter should then give a description of the entries that will be they will fill in this new column. It is by combination of a keyword and a description that they can then create their own metadata.

Keywords For each original metadata, the Corpus Submitter should create an entry in the tab, specifying the metadata name in. That name should not contain spaces or punctuation marks. **Example:** age_group

Descriptions Each keyword must then be associated with a text in the column that defines the word. Reference to an external text can be provided here but a usable definition should still be given so that a corpus user does not have to access the reference to understand its usage. **Example:** age_group refers to the age as a social characteristic. It can take three different values: "elder", "adult" and "child".

Glossary

The last section of the Corpus Documentation, the **Glossary**, is optional. Any term that is used in the translation, glossing, or other supporting documents that are not widely known (e.g. because it refers to local traditions, local natural environment, to a very specific terminology, etc.) can be documented using the glossary. The Corpus Submitter should create an entry for each term, providing its form and a description of the term.

Terms, the Corpus Submitter should provide the form of the term as it appears in the corpus. **Example:** nahubkac

Descriptions column should provide as much information as the Corpus Submitter feels is relevant for describing the term. **Example:** The term refers to the emotion of emptiness someone feels when, after having invited their relatives for an event, everyone leaves.