RefCo and its Checker: Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Process

Herbert Lange, Jocelyn Aznar

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2721–2729
Marseille, 20-25 June 2022

RefCo and its Checker: Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Process

Herbert Lange, Jocelyn Aznar
Universität Hamburg, ZAS Berlin,
herbert.lange@uni-hamburg.de, aznar@leibniz-zas.de

Abstract

The QUEST (QUality ESTablished) project aims at ensuring the reusability of audio-visual datasets (Wamprechtshammer et al., 2022) by devising quality criteria and curating processes. RefCo (Reference Corpora) is an initiative within QUEST in collaboration with DoReCo (Documentation Reference Corpus, Paschen et al. (2020)) focusing on language documentation projects. Previously, Aznar and Seifart (2020) introduced a set of quality criteria dedicated to documenting fieldwork corpora. Based on these criteria, we establish a semi-automatic review process for existing and work-in-progress corpora, in particular for language documentation. The goal is to improve the quality of a corpus by increasing its reusability. A central part of this process is a template for machine-readable corpus documentation and automatic data verification based on this documentation. In addition to the documentation and automatic verification, the process involves a human review and potentially results in a RefCo certification of the corpus. For each of these steps, we provide guidelines and manuals. We describe the evaluation process in detail, highlight the current limits for automatic evaluation and how the manual review is organized accordingly.

Keywords: QUEST, reusability, quality checking, language resources, oral language, annotated corpora, language documentation

1. Introduction

With the decline of languages around the world (Hale et al., 1992), documenting endangered and minority languages became one of the current goals of linguistic research. The aim is to mitigate the loss of linguistic data as much as possible by creating archived material of existing endangered or minority languages for which we would otherwise have no language data.

A consequent development is the inception of language documentation as a linguistic discipline, with its objectives and own methodology: recording and conserving linguistic data for its future reuse. This discipline emerged from the descriptive tradition and adopted many of its practices and recommendations, in particular from the Boasian school (Himmelmann, 1998; Michaelis, 2014). But archiving data is not sufficient to ensure its usefulness: a specific curation process must be devised to accommodate for each intended audience, for instance linguists or the speech community from which the corpus originated. One approach to make the effort of language documentation pay most is by adhering to the FAIR principles (Wilkinson and others, 2016): **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability. However, adhering to these principles is not always straightforward and there are many ways of achieving them. Our goal is to improve existing and future corpora and their reuse by providing a process that helps corpus submitters[1] to refine their data.

Our method is threefold: documentation of a corpus, evaluation of the corpus against its documentation, and finally discussion of the assessment results between the corpus submitter and a reviewer in order to improve, and ideally, certify the corpus. During the corpus documentation step, by asking for relevant information, we ensure that some of the FAIR criteria are followed and thus guarantee that all corpora curated in our proposed way adhere to them. This step consists in the manual filling of a digital spreadsheet by the submitter. Spreadsheets provide a portable interface with which many field linguists are already familiar. The RefCo spreadsheet is accompanied by extensive documentation guiding the corpus submitter through the process[2].

For the second step, the resulting corpus documentation is parsed by an automatic evaluation procedure. The checker creates a report based on both the corpus documentation and the corpus content itself. This report is intended for the corpus submitter to help improve the data as well as for the reviewer in the next step. The last step of the RefCo process is the dialogue and evaluation involving a human reviewer. This reviewer should be provided by a certification entity implementing the RefCo process. This ultimate step is not mandatory, as the first two steps can be executed independently by a corpus submitter to amend their work before submitting it to a RefCo entity, such as an archive. Using the RefCo reviewing checklist, the automatically generated report and a reporting form devised by us, the reviewer highlights which aspects of the corpus or its documentation require intervention from the corpus submitter. A dialogue is then engaged between the reviewer and the corpus submitter until the corpus meets the RefCo quality criteria by passing the automatic evaluation as well as the manual review.

To allow for a wide-spread use of our proposed method and simplify its adaptation to other contexts, all code

[1] We focus on corpus submitters. They can either be the corpus creator or a person curating someone else's corpus for submission to an archive.

[2] See Aznar and Seifart (2022) for the set of documents forming the RefCo Toolkit.

and written documentation will be available under free and open source licences.

This article is structured in the following way: in Section 2 we put our work into context. Section 3 describes our proposed corpus review process in more detail: the three steps in Section 3.1 – 3.3. The current state of evaluation is described in Section 4, These parts are followed by a general discussion in Section 5 as well as suggestions for future extension in Section 6 and finally concluded in Section 7.

2. Background and Related Work

There are currently multiple initiatives to promote quality criteria and metadata, which aim at improving language documentation corpora's reusability. The software Lameta (Hatton et al., 2021), promoted by the Endangered Languages Documentation Programme (ELDP) and Centre of Excellence for the Dynamics of Language (CoEDL), is developed to assist field linguists in having a systematic description of their documentation project files' metadata. In November 2021, a workshop was organized within the Groupement de recherche "Linguistique Informatique, Formelle et de Terrain" (GDR LIFT, Research Association on "Computational, Formal, and Fieldwork linguistics")[3] to help linguists structuring and describing their language documentation corpora in order to deposit them on Cocoon (Michaud et al., 2016).

As Babinski and Bowern (2021) reported, when current quality criteria are devised for archiving material, the focus is rarely on actual reuse of this material. The RefCo quality criteria and certification process aims at exactly this: enabling and improving corpus data for reuse within linguistics and beyond.

Improving data reusability, by providing data with an open-access licence and having citable resources, contributes to reproducibility (Drummond, 2009) as well as accountability (Berez-Kroeker et al., 2018). When doing language documentation, the fieldwork cannot be repeated under exactly the same conditions, even if it is possible in theory to organize another field trip to produce similar data and try to reproduce similar results. Organizing such a field trip is not a simple task and requires the coordination of obtaining funding, having a fieldwork linguist available and negotiating the fieldwork with the speech community. The situation is even more challenging when working with endangered languages, as further investigations might not be possible in the near future, thus preventing obtaining similar data for reproducing the results. Given these difficulties, it is of utmost importance to ensure that the digital recordings and annotations are available and can be reused by other researchers or for scientific accountability.

[3]See `https://gdr-lift.loria.fr/datathon-2021-bilan/`, for a report regarding this event, in French.

It is a vital endeavour for long term reusability of corpora to ensure that the terminology and typographical conventions used for identifying and glossing corpus segments remain understandable. Furthermore, these two practices must be implemented consistently within a given corpus. It would otherwise increase the difficulties one has to face when reusing corpora created by other linguists, and also when doing studies assisted by automatic processing. There exist already different conventions and references regarding terminology and annotation practices. For instance, the Leipzig Glossing Rules (LGR) (Haspelmath et al., 2015) promote a set of conventions for marking morphological segments as well as glossing abbreviations for some common descriptive terms. These rules are now quite widespread among linguists but cover only a limited range of the conventions needed for a language documentation corpus. Thus, linguists often have to come up with their own glossing and typographical conventions, even if they built them on top of existing standards (see for instance (Mettouchi et al., 2015) with LGR). Another difficulty arises as to how to understand the descriptive terms represented by the glosses, whose definitions, for the same terms, may vary from one linguistic paradigm to another. As summarized by Chiarcos et al. (2020), there exist numerous propositions for standardizing linguistic terminology or rendering it interoperable, such as the GOLD ontology (Farrar and Lewis, 2007), a community-driven standard for descriptive terminology that aims at creating interoperable descriptions from one corpus to another, or the Universal Dependency (UD) annotation framework (de Marneffe et al., 2021; Zeman et al., 2021), which was used for 217 treebanks for 122 languages (Version 2.9, released November 15, 2021). Unfortunately, as the creators acknowledge themselves, UD only aims at defining a common core suitable to describe many languages in a uniform way. The fine-grained and language-specific features required for language documentation currently cannot be expressed in this framework without a loss of valuable information.

3. The RefCo Certification

The RefCo process, as we describe it in this paper, fills an important gap: it provides a sound framework to ensure the reusability of corpora. All involved parties, institutions such as archives as well as corpus creators or maintainers, are provided with guides, documentations, and checklists. An automatic evaluation step provides immediate feedback to improve both the corpus and the documentation. The results of the automatic evaluation also narrow down potential issues in the final dialogue step with human reviewers. To motivate linguists to submit corpora and engage in the certification, the process has been devised in a way that provides immediate feedback and suggestions to improve the overall quality of the submitted corpus while minimizing extra burdens on the researcher.

The certification tries as much as possible to remain agnostic regarding linguistic theories, their terminology, and the kinds of corpus data it should be used on. RefCo aims to be compatible with already existing corpora, regardless of their theoretical background or size. Thus, we do not require linguists to use any specific terminology or conventions, and the corpus documentation provides a template to describe their own conventions.

As both von Prince and Nordhoff (2020) and Babinski and Bowern (2021) show, from one corpus to another, from one linguist to another, that different conventions can be used at every level of a project, including directories, files, and annotation tiers. To ensure that potential reusers will be able to understand the corpus structure and glosses used by the linguists who created the language documentation corpus, the RefCo approach[4] requires linguists to document their corpus structure and glosses. This documentation is essential for the RefCo process. Two quality criteria are at the foundation of the RefCo evaluation: consistency and coherency of a given corpus relative to its documentation.

The consistency of a corpus means that it should contain only the files that it is supposed to. As reported by Babinski and Bowern (2021) and as one of the authors observed in the context of DoReCo, many language documentation corpora contain files in multiple versions or which are not relevant for the corpus documentation. Likewise, the inconsistency of the file naming conventions makes it often difficult for reusers to identify which files belong to the same recording session.

The coherency quality criterion corresponds to the idea that a corpus and its description should match each other's content. For instance, all the glosses described in the corpus documentation should actually occur in the annotation files in order to be considered coherent with its documentation. And accordingly, all the glosses documented should be described in the documentation. Checking gloss descriptions or their accuracy to describe a given morpheme is currently out of reach. The latter, in particular, requires knowledge of the language being described.

The RefCo certification process is intended as such:

1. Corpus submitter prepares the corpus for submission: reads the RefCo reference manual which is part of the RefCo toolkit, creates the corpus documentation, and applies the criteria to their corpus.

2. Submitter uses the RefCo checker before applying for a certification, in order to solve as many issues as possible. This step can be useful while creating a corpus to improve the quality of the data and to ensure the coherency of the documentation

3. After the corpus data is submitted, the results of the automatic evaluation have to be reviewed manually.

Once the manual review is done, a potential certification entity can either directly certify the corpus or deliver feedback to the corpus submitter about issues to be resolved.

3.1. The Documentation Step

The first step of our three-step RefCo process involves the documentation of the corpus by its creator or submitter using our template spreadsheet and following our guideline document.

The spreadsheet has to be submitted as part of the metadata information associated with the corpus, with the corpus data (that is annotation and recording files). Because the spreadsheet format is currently the only format implemented by the RefCo checker, it is required by the evaluation[5].

The corpus documentation spreadsheet is designed to comply with many of the FAIR principles[6], in particular:

F1 (Meta)data are assigned a globally unique and persistent identifier

F3 Metadata clearly and explicitly include the identifier of the data they describe

A1 (Meta)data are retrievable by their identifier using a standardized communications protocol

R1 (Meta)data are richly described with a plurality of accurate and relevant attributes

- **R1.1** (Meta)data are released with a clear and accessible data usage licence
- **R1.2** (Meta)data are associated with detailed provenance
- **R1.3** (Meta)data meet domain-relevant community standards

Our approach expands on these generic principles for research data and mostly focuses on the reusability criteria, the most relevant aspect of language documentation data. The corpus documentation template is divided into three main parts, spread over eight tabs in the spreadsheet document. The first section, corresponding to the tab *Overview*, is related to generic metadata and general properties of the corpus. The set of metadata retained, such as corpus creator, where the data can

[4]Initially devised by von Prince as part of the collaboration between QUEST and DoReCo, it was redesigned by the authors in order to use machine-readable formats and to include automatic processing.

[5]There exists some overlaps between the Refco corpus documentation and other metadata formats such as IMDI or CMDI.

[6]`https://www.go-fair.org/fair-principles/`

be accessed, and its licence, allows following the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group, 2014). This metadata is not unique to the RefCo criteria, but is usually also expressed in a common metadata format, such as IMDI or CMDI. Thus, some redundancy in metadata information is necessary for the current RefCo checking process.

One strict requirement is a persistent identifier for the corpus data, owing to the findability principles. Other information included is the languages involved, as well as some statistics such as number of sessions and total word count. A screenshot depicting this first tab is shown in Figure 1.

The second part, which comprises the *CorpusComposition* and the *AnnotationTiers* tabs in the spreadsheet, documents the structure of the corpus and recording metadata. RefCo focuses on corpus data created from speech and assumes that the speech acts are grouped in recording sessions, documented in *CorpusComposition*. For each of these sessions, two kinds of information are relevant. Firstly, all relevant files have to be referenced. This includes, if available, recording files as well as annotation files, but can also include additional files, e.g., ones used for elicitation strategies. Secondly, for each session, all relevant information about the recording session, most importantly speaker information as well as recording location and date, have to be specified. The tab *AnnotationTiers* describes the structure of the annotations, listing all annotation tiers and giving all necessary information about them. Corpus creators can name their annotation tiers according to their own strategies, but should associate them with specific functions. These functions can be chosen either from a predefined list or defined by the user themself. It is by using one of the predefined functions that the RefCo checker will infer the appropriate checks to perform. The segmentation strategy defines the granularity of information units, e.g., morphemes for glossing or paragraphs for textual description. For each tier, the languages used have to be documented as well.

The third part of the corpus documentation includes the *Transcription*, *Glosses* and *Punctuations* tabs. These three tabs describe the actual content of annotation files in the corpus. This content will be used by the RefCo checker when performing the different coherency checks, to ensure that whatever is annotated in the corpus matches its documentation. The function column in the *AnnotationTiers* tab specifies which tiers contain transcription or glosses using a set of predefined functions. Each grapheme is associated with a form and a linguistic value, and convention from which they originate, such as the International Phonetic Alphabet (IPA) or X-SAMPA. Glosses refer to grammatical concepts, usually as abbreviations in uppercase letters. Following the Leipzig Glossing Rules (Haspelmath et al., 2015) is possible, but not mandatory. All gloss abbreviations have to be defined, and the corpus submitter can add optional comments to clarify their usage. Similarly, an entry must be created for each meaning associated with a punctuation mark. Because they can be used differently in various tiers, the corpus submitter must specify for each punctuation mark in which tier they are used. As they can have varying usages, each punctuation mark entry should be described by a function, such as prosodic cue, gloss separator or morpheme break.

Two more tabs are included in the documentation spreadsheet: a *Glossary* for linguists to document the lexical glosses that they could not translate, and a tab called *CorpusOpenDescription* as a space for linguists who defined their own metadata to describe the events they recorded. They are only relevant for human readers and are not automatically evaluated.

3.2. The Automatic Evaluation Step

After the corpus submitter documented all relevant aspects of the corpus in the previous step, we provide an automatic testing procedure that verifies information given in the corpus documentation and ensures coherent and consistent corpus data. Our automatic checker creates a report with the results of various checks. An example can be seen in Figure 2. Report items can have one of three levels: *Critical*, *Warning*, or *Correct*. Where possible, the items contain additional helpful information such as the location of the problem as well as a suggestion on how to fix the issues. *Critical* errors have to be fixed by the person in charge of the corpus. *Warnings* mark issues that are not systematically problematic, but which still require attention, either by the corpus submitter or the reviewer, in order to be sure that the issues raised will not negatively impact the corpus' reusability. Finally, *Correct* items give some additional feedback about the corpus, but do not require any intervention.

The automatic validation is possible because the corpus documentation is machine-readable and uses the OpenDocument format. This means that the spreadsheet can be read as an XML file and the information can be extracted using standard techniques such as XPath expressions. Even though the XML markup of the spreadsheet is only structural and does not directly represent the intended semantics of the corpus documentation, it is still possible to identify and extract all relevant information.

The RefCo checker[7] is developed as part of the Corpus Services, initially developed at the Hamburger Zentrum für Sprachkorpora (HZSK) (Hedeland and Ferger, 2020). The Corpus Services are a generic processing framework for corpus data. The checker is currently able to process corpus data in the ELAN file format, but can be extended to any file format as long as it is possible to extract annotation tiers and the annotations contained in these tiers.

[7] https://gitlab.rrz.uni-hamburg.de/bba1792/corpus-services/-/blob/develop/doc/README.RefCo.md

	A	B	C
1	**Corpus Information**		
2	**Corpus Title**	Corpus de narrations nisvaies	
3	**Subject Language(s)**	nisv1234	
4	**Archive**	ORTOLANG	
5	**Corpus Persistent Identifier**	https://hdl.handle.net/11403/sldr000783	
6	**Annotation Files Licence**	CC-BY-NC-ND	
7	**Recording Files Licence**	CC-BY-NC-ND	
8	**Corpus Creator Name**	Jocelyn Aznar	
9	**Corpus Creator Contact**	contact@jocelynaznar.eu	
10	**Corpus Creator Institution**	ZAS	
11	**Certification**		
12		**Information**	**Notes**
13	**Corpus Documentation's Version**	2	
14	**Quantiative Summary**		
15	**Number of sessions**	12	
16	**Total number of transcribed words**	28730	
17	**Total number of morphologically analyzed words**	29970	
18	**Annotation Strategies**		
19		**Information**	**Notes**
20	**Translation language(s)**	French	

Figure 1: The general part in the *Overview* tab of the RefCo documentation. It includes relevant information to meet FAIR criteria, such as unique identifier and licensing information for the corpus

ID	Type	Function	Filename:line.column	Error	Fix
0	Warning	RefcoChecker	CorpusDocumentation.ods	Corpus composition: File does not exist: T1_15-12-2013_Levetbao_Aven_WaetMasta_1089.wav	Check the file reference in the documentation and remove the reference to the file if it is removed intentionally
1	Warning	RefcoChecker	CorpusDocumentation.ods	Corpus composition: Files are not documented: T1_15-12-2013_Aven_Levetbao_WaetMasta_1089.wav	Check the file reference in the documentation and add the references to the files if they should be included or delete unused files
2	Warning	RefcoChecker	CorpusDocumentation.ods	Annotation Tiers: potential custom tier detected: Textualité with tier function [text plan of the narrative]	Check if custom tier function is intended or change tier function
3	Warning	RefcoChecker	CorpusDocumentation.ods	Annotation Tiers: language is neither a Glottolog, a ISO-639-3 language code nor otherwise known: Nisvai	Use a valid language code
4	Correct	RefcoChecker	T1_15-12-2013_Levetbao_Aven_Waet-Masta_1089.eaf	Corpus data: More than 99 percent of transcription characters are valid. Valid: 8189 Invalid: 0 Percentage: 100.0	Documentation can be improved but no fix necessary
5	Warning	RefcoChecker	T1_15-12-2013_Levetbao_Aven_Waet-Masta_1089.eaf: Tier:Morphologie.Segment:a411, Time:02:31.115-02:34.546	Invalid morpheme in token: IRR.NEG in IRR.NEG=2SG=dire	Add gloss to documentation or check for typo
6	Warning	RefcoChecker	T1_15-12-2013_Levetbao_Aven_Waet-Masta_1089.eaf: Tier:Morphologie.Segment:a1308, Time:05:22.584-05:25.524	Invalid morpheme in token: IRR.NEG in IRR.NEG=1SG=blesser	Add gloss to documentation or check for typo
7	Correct	RefcoChecker	T1_15-12-2013_Levetbao_Aven_Waet-Masta_1089.eaf	Corpus data: More than 70 percent of tokens are valid gloss morphemes. Valid: 2193 Invalid: 2 Percentage valid: 99.9	Documentation can be improved but no fix necessary
8	Critical	RefcoChecker	T1_15-12-2013_Levetbao_Aven_Waet-Masta_1089.eaf	No annotated text found in one of the expected tiers: Morphologie, gl	Check the tier documentation to make sure that your morphology tiers are covered

Showing 1 to 9 of 9 entries

Previous 1 Next

Figure 2: Sample from the RefCo checker report displaying the three different error levels (*Correct* in green, *Warning* in yellow, and *Critical* in red). It shows several consistency issues such as undocumented files and coherency issues such as undocumented glosses.

The report generated during the automatic evaluation determines the next step. If the report contains critical problems, the corpus submitter has to fix the issues before they can progress towards the certification. However, if the report contains only items of the *Warning* or *Correct* level, the corpus and the report are handed to a human reviewer who decides about the certification of the submitted data.

3.2.1. Checking the Corpus Documentation

The first step of the automatic validation checks the information given in the corpus documentation itself. The corpus description is read into a data structure and, where possible, automatically checked. Directly after reading the spreadsheet, some obvious checks are executed:

- most fields in the documentation are mandatory, missing mandatory fields cause critical errors,
- URLs are resolved to make sure that the linked resource is available, invalid URLs cause warnings,
- dates are checked to follow the ISO 8601 standard, invalid dates cause warnings,
- number fields are checked for valid numbers, invalid values cause warnings.

In addition, more specific checks are implemented in order to either verify the information given in the corpus documentation or guarantee the consistency of the documentation itself:

- There are three ways languages encountered in the documentation can be verified. For languages represented by an ISO-639-3 language code, the list of all defined language codes can be searched. When using Glottocodes, a Glottolog URL can be constructed and resolved to check if the code is

valid. Alternatively, for a few most commonly expected translation languages, as specified in the guidelines, the language name can be used directly. Unknown languages generate a warning.

- The documented number of transcription and annotation tokens is compared to the result of a simple counting procedure in the checker. If the numbers are off by more than a certain factor, currently 10 percent, a warning is created.
- The number of sessions in *Overview* has to match the sessions declared in the *CorpusComposition* tab of the corpus documentation.
- All files listed in the documentation have to be present, and missing documented files cause errors. In addition to checking the presence of documented files, undocumented files in the corpus are identified and cause a warning as well. These checks are necessary to improve the consistency of the corpus.
- Tier functions used in the documentation can either be from a set of tier functions suggested in the guidelines or user defined. If the function is not one of the suggested ones, a warning is added, which should be ignored if the tier is indeed of a user-defined type. These warnings help in catching common mistakes such as typos in tier names as well as inconsistencies in tier naming in the corpus. In addition to checking the tier documentation, all tiers are extracted from the corpus data and compared to the documented tiers. All tiers missing from the documentation are added as warnings.
- The tiers specified for each punctuation and gloss abbreviation are checked to be among the documented tiers in the *AnnotationTiers* tab, resulting in warnings otherwise.

Checking other information, such as names of people and organizations, or email addresses, is outside the scope of the automatic tests. Only the presence of the information is checked automatically, not its validity. This kind of information has to be checked in the manual review step.

3.2.2. Checking the Corpus Data

In the previous section, we presented automatic validation, focusing on the corpus documentation itself and checking the corpus files for consistency. This section focuses on the coherency of the corpus documentation, i.e., to make sure that the corpus documentation and the corpus data are matching each other. The relevant parts of the corpus are the transcription and morphology tiers.

To check the coherency of the transcriptions in a corpus, all transcribed texts are extracted and validated using the information given in the corpus documentation. The documentation contains a list of valid characters for a tier consisting of both transcription graphemes and punctuation marks. Each token, a sequence of characters surrounded by delimiters such as spaces, must consist only of valid characters. If a token contains invalid characters, it is not coherent with the documentation. For each incoherent token, a warning is created. Also, after checking a tier, statistics about the ratio between coherent tokens and incoherent tokens are created. If the ratio of incoherent tokens is above a certain threshold, a warning is added to the report, otherwise the ratio is reported as a *Correct* item (See item 4 in Figure 2).

After checking the transcription data for coherency, the automatic validation moves on to checking the morphology annotations. The morphological annotation can follow a wide range of annotation schemas. A common example would be annotation following the Leipzig glossing. Here lexical and morphological annotation are combined, and the lexical part is separated from an abbreviation encoding the morphological information by special characters. Furthermore, there is a common convention to encode the lexical parts in lowercase while morpheme gloss abbreviations are written in uppercase letters.

To evaluate the coherency of the morphological annotations, all non-lexical morpheme glosses are extracted from a tier and compared to the documented gloss abbreviations. Two cases lead to warnings: either a gloss abbreviation is not documented, or a gloss abbreviation never occurs in the corpus. Additionally, after finishing the validation of a tier, the ratio between coherent and incoherent morpheme glosses is reported, potentially causing a warning if too many glosses are not coherent in respect to the documentation.

3.3. A Dialogue Between the Reviewer and the Corpus Submitter

The final, and optional, step of the RefCo process is the official certification by a RefCo certification entity. In order to make the process sensible, the process has been devised as a dialogue between the applicant and a human reviewer. The purpose is to be able to certify that the corpus follows RefCo's quality criteria. By this stage, the corpus submitter should have already prepared their corpus and do a self-evaluation using the RefCo checker.

Once the results of the automatic validation are satisfactory, the dialogue with the reviewer to certify the corpus can begin. The reviewer will go through the review list provided, which has to be used in combination with the RefCo report form. The review list specifies all checks, either manual or automatic, a corpus has to pass. The reviewer has to report the result of each check using the report form, which will then be given back to the applicant. The purpose of the report form is for both the certification entity and the corpus submitter to be able to keep track of each issue that remains to

be solved.
The review tasks are grouped into four different sections:

- The *Certification Process* contains information regarding the certification process itself,
- the *Functional tests* section deals with issues regarding the files in the corpus,
- the section *Corpus Design* is about ensuring the consistency of the folder and file naming conventions used by the corpus submitter.
- the *Corpus Documentation* section provides additional verification that should be done manually where an automatic evaluation is not possible.

After finalizing the manual review, the reviewer shares their reports with both the certification entity and the corpus submitter. If issues persist, the corpus submitter has the opportunity to update the documentation and the corpus based on the content of the review form before resubmission.
As soon as there are no remaining serious issues, the certification entity bestows a RefCo certificate upon the corpus. In the case that the certification entity is an archive, the archive can ingest the corpus and mark it as RefCo certified.

4. Evaluating the Process

For a proper evaluation, the RefCo process, specifically the automatic checker, has to reach a more mature state. Until then, we can only test the method ourselves with data available to us. We started from a pre-existing corpus and created the documentation for it. Using the RefCo process, we improved both the corpus and the documentation based on the feedback provided by the automatically generated report. The resulting corpus and documentation have been reviewed by us and can now be seen as a gold standard. Based on this gold standard, the code of the automatic validator is tested using modified "flawed" corpora. They trigger procedures that validate our checker and ensure that the expected output is generated. In parallel, an extensive test suite based on unit tests has been developed. That way, we use both synthetic and authentic data in testing our code.
The RefCo criteria have been devised thanks to interviews conducted by von Prince with linguists working in the field of comparative research (Aznar and Seifart, 2020). The RefCo process presented in the paper is based on these criteria. To add further experimental evidence, we plan to develop a user study to show that our claim holds: our process provides researchers with a tool to improve their data for future reuse with acceptable overhead. In this evaluation, we plan to include linguists from various fields, giving us the opportunity to experiment with a wide range of data.

5. Discussion

Since this work is still in the early stages, there are several points open for discussion. Some of these points we want to present here.
For a start, we decided to use a spreadsheet as the user interface, which might seem like a surprising decision. Several other choices, such as a web or standalone application comparable to Lameta would be possible. But, by using the spreadsheet format, we immediately have a machine-readable format at hand. Furthermore, building on top of a spreadsheet saves us the trouble of hosting infrastructure and provides the user with a familiar interface. One point of RefCo is to provide a reference implementation for quality criteria. As the field of language documentation is evolving, this reference should be the object of constant discussion. Having an interface with which fieldwork linguists are familiar and can engage is a step towards facilitating discussions over the quality criteria. As an editable format, linguists will be able to provide their own implementation of the corpus documentation for discussing quality criteria, without having to code a software interface. As RefCo aims to be a collaborative reference for improving the quality of fieldwork corpora, facilitating the discussions and the negotiation around the quality criteria, and the standards of the fieldwork community are of utmost importance. However, adding other interfaces at a later point would be possible, as long as they are compatible with the documentation format presented here[8].
The spreadsheet interface will be kept as our reference implementation, for facilitating the discussion among linguists, while there might be multiple existing implementations.
Our claim to be theory agnostic could be seen contradictory with our multiple mentions of the Leipzig Glossing Rules as it is only one way among other possibilities for annotating language data. But the RefCo corpus documentation as it is, is compatible with other annotation frameworks, such as Universal Dependencies (de Marneffe et al., 2021), i.e., Universal POS tags and Universal features, or the Stuttgart-Tübingen tag set (STTS) (Schiller et al., 1999) that is used for instance by TreeTagger[9]. A corpus using one of these tag sets can be documented with RefCo corpus documentation as well. One could create a new tier for a different annotation scheme such as STTS and document all valid tags as glosses for this tier.
In general, it would reduce the effort to create the documentation if we allow the user to define their transcription or glossing scheme by just naming them, e.g., IPA for transcriptions or any of the glossing schemes named above. However, this would also reduce the flexibility of the RefCo toolkit by introducing a dependence to-

[8] A compatible XML and JSON schema is available as part of the RefCo toolkit (Aznar and Seifart, 2022)

[9] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/STTS-Tagset.pdf

wards a limited amount of transcription or glossing theories. This is where a balance between flexibility and ease of use is a constant challenge, in order to avoid introducing biases in favour of certain theories and social perspectives. Even though we try to avoid biases, we could not prevent the contradiction of trying to document language diversity while promoting the use of "central languages" (Calvet, 2006), a very limited set of less than 10 languages occurring in most translation exchanges, for glossing and translation tiers.

Another issue is the question of how to get the required infrastructure to offer a proper RefCo certification. This would require some kind of certification entity, which we currently cannot provide ourselves. Instead, we would have to find ways to collaborate with existing units, e.g., archives and libraries that are interested in certifying the data submitted to them according to the RefCo principles. However, even without this kind of certification entity, the initial steps of the certification process can still be used to improve the data quality and reusability without resulting in a proper certificate.

6. Future Work

There are plenty of plausible minor or major additions to the RefCo process and the automatic evaluation that would help to achieve our goal: consistent, well-documented corpora usable in linguistic research and beyond.

A major future addition to the current work will be additional evaluation of our tool and process in the form of a user study, as described in Section 4.

Once a full evaluation has been performed, the next step would be to partner up with data archives to integrate the RefCo process into their data deposition workflow, either as a standalone process or as part of a larger set of quality assurance methods such as the full set of methods developed within the QUEST project (Wamprechtshammmer and Arestau, 2021; Wamprechtshammer et al., 2022).

A way which has not been explored yet for improving the quality of corpora would be using statistical or machine-learning methods in RefCo. They could help either improve the data itself, e.g., by detecting typos, or verifying the documented information by, e.g., automatically detecting transcription languages. However, the limited amount of data available and the requirement of high reliability could be limiting factors, and the integration of such methods would currently go beyond the scope of the project.

7. Conclusion

We strongly believe that this work fills an important gap. We do not claim to be the first to implement a machine-readable corpus documentation. People already used spreadsheets to document their work, and some of these people even attempted to establish a standard for machine-readable corpus documentation. What sets us apart is that we do not only present a new standard based on current metadata recommendations, instead we propose a complete, semi-automatic, review and quality assurance process around a flexible documentation format. The suggested process, ideally resulting in proper certification of research data, helps to improve both the data itself and its reusability. Our approach ensures the harmonization between the actual annotations in the corpus and the transcription and glossing conventions intended by the corpus creator. We try to avoid theory-specific biases and imagine the application of the RefCo process in many linguistic areas, not just in language documentation. To guarantee easy access to the RefCo toolkit, we release all documents and code under free licences.

8. Acknowledgements

Thanks for early discussions on RefCo that happened within GDR LIFT. The QUEST project is funded by the German Federal Ministry of Education and Research (BMBF) (06/2019–05/2022).

9. Bibliographic References

Aznar, J. and Seifart, F. (2020). RefCo: An initiative to develop a set of quality criteria for fieldwork corpora. In Thierry Poibeau, et al., editors, *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT)*, pages 95–101, Montrouge, France. CNRS.

Aznar, J. and Seifart, F. (2022). The RefCo toolkit. Technical report, Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS). `https://zenodo.org/record/6470807`.

Babinski, S. and Bowern, C. (2021). Contemporary digital linguistics and the archive: An urgent review. `https://www.youtube.com/watch?v=aC06qrr1gcY`.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., and Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1):1–18.

Calvet, L.-J. (2006). *Towards an Ecology of World Languages*. Polity Press.

Chiarcos, C., Fäth, C., and Abromeit, F. (2020). Annotation Interoperability for the Post-ISOCat Era. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, page 10.

Data Citation Synthesis Group. (2014). Joint Declaration of Data Citation Principles. Technical report, Force11.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, pages 1–54.

Drummond, C. (2009). Replicability is not Reproducibility: Nor is it Good Science. *Proc. of the Eval-*

uation Methods for Machine Learning Workshop at the 26 th ICML, page 4.

Farrar, S. and Lewis, W. D. (2007). The GOLD Community of Practice: An infrastructure for linguistic data on the Web. *Language Resources and Evaluation*, 41(1):45–60.

Hale, K., Krauss, M., Watahomigie, L. J., Yamamoto, A. Y., Craig, C., Jeanne, L. M., and England, N. C. (1992). Endangered Languages. *Language*, 68(1).

Haspelmath, M., Bickel, B., and Comrie, B. (2015). Leipzig Glossing Rules : Conventions for interlinear morpheme-by-morpheme glosses. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf, last accessed 2021-01-03.

Hatton, J., Holton, G., Seyfeddinipur, M., and Thieberger, N. (2021). Lameta. https://github.com/onset/laMETA/releases.

Hedeland, H. and Ferger, A. (2020). Towards Continuous Quality Control for Spoken Language Corpora. *International Journal of Digital Curation*, 15(1):13.

Himmelmann, N. P. (1998). Documentary and Descriptive Linguistics. *Linguistics*, 36(1):161–196.

Amina Mettouchi, et al., editors. (2015). *Corpus-Based Studies of Lesser-described Languages: The CorpAfroAs Corpus of Spoken AfroAsiatic Languages*, volume 68 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company.

Michaelis, M. H. a. S. M. (2014). Annotated corpora of small languages as refereed publications: A vision. https://dlc.hypotheses.org/691, last accessed 2022-01-03.

Michaud, A., Guillaume, S., Jacques, G., Mac, D.-K., Jacobson, M., Pham, T.-H., and Deo, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation: la collection pangloss et la collection auco. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 1 : JEP*, pages 155–163.

Paschen, L., Delafontaine, F., Draxler, C., Fuchs, S., Stave, M., and Seifart, F. (2020). Building a Time-Aligned Cross-Linguistic Reference Corpus from Language Documentation Data (DoReCo). *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 12:2657–2666.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS (Kleines und großes Tagset). Technical report, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart and Seminar für Sprachwissenschaften, Universität Tübingen. https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf, accessed 2022-01-03.

von Prince, K. and Nordhoff, S. (2020). An Empirical Evaluation of Annotation Practices in Corpora from Language Documentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2778–2787, Marseille, France. European Language Resources Association.

Wamprechtshammer, A., Arestau, E., Aznar, J., Hedeland, H., Isard, A., Khait, I., Lange, H., Majka, N., Rau, F., and Schwiertz, G. (2022). QUEST: Guidelines and specifications for the assessment of audiovisual, annotated language data. Technical report, QUEST project. fortcoming.

Wamprechtshammmer, A. and Arestau, E. (2021). Generische und disziplinspezifische Zugänge zur Qualität audiovisueller, annotierter Sprachdaten im BMBF-Projekt QUEST. In Patrick Helling, et al., editors, *FORGE 2021 - Forschungsdaten in den Geisteswissenschaften: MAPPING THE LANDSCAPE - Geisteswissenschaftliches Forschungsdatenmanagement zwischen lokalen und globalen, generischen und spezifischen Lösungen. Konferenzabstracts*. Publisher: Zenodo.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Zeman, D., Nivre, J., et al. (2021). Universal dependencies 2.9. http://hdl.handle.net/11234/1-4611. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.