RefCo: An initiative to develop a set of quality criteria for fieldwork corpora

Jocelyn Aznar, Frank Seifart

RefCo: An initiative to develop a set of quality criteria for fieldwork corpora

Jocelyn Aznar[1], Frank Seifart[1]
(1) ZAS, Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)
Schützenstr. 18, 10117 Berlin, Allemagne
aznar@leibniz-zas.de, seifart@leibniz-zas.de

RÉSUMÉ

RefCo: une initiative pour développer un ensemble de critères de qualité pour les corpus de terrain.
RefCo est une initiative issue du projet QUEST dont l'objet est de s'assurer de la qualité des corpus issus de terrains en linguistique. L'objectif est d'établir des critères pour s'assurer que les corpus soient réutilisables, en particulier dans le cadre de recherches comparatives. L'initiative porte à la fois sur les corpus existants, par la mise en place d'un label de qualité, et sur les futurs corpus, par l'élaboration de recommandations et des guides de bonnes pratiques. Le label RefCo est décomposable en deux volets : les métadonnées pour la comparaison linguistique et la documentation du corpus. Enfin, un système de production de citations est également développé afin de faciliter la citation des corpus.

ABSTRACT

RefCo is an initiative of the QUEST project which aims at ensuring the quality of corpora from linguistic fieldwork. The objective is to establish criteria for reusable corpora, especially for comparative linguistic research. The initiative addresses both existing corpora, through the implementation of a quality label, and future corpora, through guidelines for best practices. RefCo's label is currentely composed of two panels: the metadata for cross-linguistic corpora, and the Corpus Documentation. Besides, a system producing citations is also developed in order to facilitate corpus citation.

MOTS-CLÉS : réutilisation, archivage, normalisation, critères de qualité, comparaison linguistique, langues orales, documentation linguistique, corpus de terrain

KEYWORDS: reusability, archive, standards, quality criteria, cross-linguistics, oral languages, linguistic documentation, fieldwork corpora.

1 Introduction

At least since Hale et al.'s (1992) acknowledgment of language endangerment and extinction as a scientific issue, many small languages around the world have been documented in one way or another. These documentations are potentially valuable sources of knowledge about human language and linguistic diversity. But, as the number of language documentations increased, the difficulties for reusing those materials are also becoming more apparent (e.g., Thieberger et al. 2016). In fact, of the vast materials held in repositories such as the Endangered Languages Archive (ELAR), The Language Archive (TLA), or PANGLOSS, only little has been used in linguistics research so far, especially by linguists other than the corpus creators themselves.

To enhance the reusability of such data, the QUEST project (https://cutt.ly/quest-project), funded by the German Ministry of Science from 2019 to 2022, proposes to develop, promote and validate good practices and standards for language documentation corpora through consultation with the research community. The QUEST consortium, as a whole, deals with different types of linguistic data, including multimodal and multilingual data. The project develops generic quality criteria as well as specific recommendations applicable for certain re-use scenarios. These quality criteria will be the basis for a data quality label, which QUEST develops as a mid-term goal, and as guidelines for ongoing and future data collection. Furthermore, QUEST develops a Web interface as a tool to make the QUEST label testing process accessible to a wider audience (Arkhangelskiy, Hedeland, and Riaposov 2020).

The QUEST component RefCo (Reference Corpora) targets research carried out on fieldwork corpora[1], especially comparative research, as one of the QUEST re-use scenarios. By 'fieldwork corpora' we mean sets of monological narratives, both audio and audiovisual, that are typically collected in the context of language documentation projects. These projects could aim at cultural preservation for multiple users, including speech communities, and thus might contain additional data not relevant for RefCo like song recordings. RefCo's objective is to ensure that such fieldwork corpora can be re-used, especially in cross-linguistic studies.

As a first step, QUEST RefCo carries out consultation with stakeholder in the field, including fieldworkers and researchers who carry out comparative research using fieldwork corpora (e.g., Bender et al. 2013; Haig & Schnell 2016; Mettouchi, Frajzyngier & Chanard 2017; Stave et al. 2020). On the basis of that, it develops a set of quality criteria for re-usability which will be implemented as (i) guidelines for fieldworkers, and (ii) as basis for a certificate of cross-linguistic re-usability that can be awarded to corpora based on an evaluation carried out by an expert committee. RefCo thus aims to be relevant to both already existing corpora as well as future documentation projects. Additionally, QUEST RefCo carries out case studies to test the re-usability of data that meets the proposed criteria. As with other QUEST components, RefCo criteria are conceived as standards that are currently accepted by the relevant research community: they can and must be amended and changed when standards of the field evolve. The aim is to define a minimum set of quality criteria, not an extensive one. RefCo criteria focus primarily on the quality of annotations and of metadata, both at the level of individual files and at the level of corpora and to improve the findability of the resources. To achieve this objective, QUEST policy is to promote and facilitate the application of standardized good practices on a dataset, not to enforce the use of as many as possible metadata or annotations. QUEST criteria align standards for fieldwork corpora with internationally accepted standards for research data, including the FAIR principles (Wilkinson et al. 2016), Dublin Core Terms, Schema.org, FOAF, SKOS as well as OLAC (Bird and Simons 2001), a metadata specification dedicated to language resources. To further enhance data reuse, RefCo, as part of the QUEST label, provides bibliographical references which follows the Austin principles (Berez-Kroeker et al. 2018), to ensure that corpus creators are properly credited for their work.

[1] In this article, we interchangeably use the terms "dataset" and "corpus" to refer to the data submitted to RefCo. Within QUEST, a "dataset" is a more abstract concept that allows encompassing the various deposit scenario: a field linguist submitting a corpus on an oral language, a set of legacy recordings, a teacher providing annotated documents. In the case of RefCo, both terms are equivalent as the deposit are mostly fieldwork corpora.

2 RefCo Curation Standards

The RefCo component is a subproject of the QUEST initiative. As such, it inherits the quality criteria defined within the QUEST project for assessing the submitted datasets. In this section, we will focus solely on the quality criteria and process associated with fieldwork corpora and thus RefCo.

2.1 Metadata for Cross-Linguistic Corpora

The Quality Standards for Audiovisual Corpora developed by the QUEST initiative focus primarily on a comprehensive set of metadata for linguistic datasets. Within this set, the RefCo subcomponent defines the 'RefCo module', a set of basic metadata that is currently considered minimal standard by research community of field workers and linguists carrying out comparative research on fieldwork data (e.g. Bowern 2008: 47–62 ; Thieberger and Berez 2011: 105-109; Meakins, Green, and Turpin 2018: 73-78). This includes Glottocode language identification codes for the languages documented as well as the language in which translation and glosses are provided, date and location of recording, speaker age and sex, and that a license for re-use is specified. RefCo explicitly allows for approximate, rather than exact time and age information to bridge archivists' desire for complete and precise metadata with the realities the corpus creators face during fieldwork.

A corpus submitted to RefCo has to be licenced with a Creative Common licence which enables the scientific re-use of the corpus[2]. This is a requirement for corpora that are intended to be used in cross-linguistic studies.

The metadata are to be specified at two levels: the corpus, called dataset in QUEST, which is the entity corresponding to a whole coherent submission by a data creator, typically on one language. A dataset is composed of datapackages, also called sessions or bundles, an abstract entity typically referring in the case of the RefCo to one monological narrative, including media and annotation files. The abstraction of datapackage allows to handle case of a texts distributed over various files, or a file containing various texts.

As many research questions require a minimal amount of data, the number of words that have been transcribed and translated, and of the number of words that are morphologically annotated, have to be specified at the dataset level. We are aware that the number of words is a rough estimate of corpus size, given differences in morphological type and in definitions of words, but we consider that this is a useful approximation.

2.2 Documenting a Corpus and its Conventions

Annotating data with interlinear glossing is a standard practice for linguists working on oral languages that was popularized by Boas (1922). The process is now assisted by using software like ELAN, EXMARaLDA or Anvil and has been the object of some conventions, in particual the Leipzig Glossing Rules. Still, there are considereable variations in the interlinear glossing practices of linguists, as apparent in corpora we are currently processing for RefCo. Therefore, RefCo requires that the information relevant for re-use but that cannot be gleaned which explicitely from the corpus

[2] It includes thus the six derivated licenses associated with Creative Commons, including the non commercial and non derivative onr which will still allow comparative researches to be made. Other Open or Free licenses will be evaluated on the demand of a corpus submitter.

itself or the metadata to be provided in a set of separate documents that we call "Corpus Documentation"[3]. The redaction of this Corpus Documentation involves both the corpus creator and the RefCo component into checking the coherency of the annotation. Crucially, RefCo does not requires the use of one of the other glossing convention, but explicit description of the specific glossing conventions used, including, e.g., abbreviations used in glossing and punctuation.

The Corpus Documentation must start with an ***Overview*** section wich provides information about the types of files present in the corpus and their format. It also asks about the number of items present (texts, transcription units, tokens, glosses and POS tags). The following section, ***Annotation levels***, specifies whether all the files in the corpus respect or not the same conventions, the authors has to provide information concerning the tiers and their names, the way they were segmented and in which language they are written. QUEST requires identification of the (anonymous) speaker IDs. Regarding the ***Transcription***, the author of the corpus has to provide first a table associating each grapheme with their phonological value. If employed, it is important as well to specify all the particular strategies used to transcribe noise, cough, laugh or other paralinguistical speech. The language used for the ***Translation*** has to be specified as well as whether the translation was provided by a native speaker of the destination language, a more vehicular language. If relevant, a glossary explaining the untranslated words in the corpus should be provided. The ***Morphemes*** section is dedicated to the description of choices made to handle morpheme boundaries and non-linear morphology, that is if Leipzip Glossing Rules were applied and which morpheme separators were used. In the ***Glosses***, if the punctuation was used in that tier, its meaning has to be described. The grammatical abbreviations are explicited here as well. If the corpus contains a layer dedicated to the annotation of Part-of-Speech **POS**, their meaning has to be provided here. Finally, if the corpus providers find it relevent, in the **Other** section, they can provide additional information and comments.

3 Benefits of using RefCo

Submitting a corpus to RefCo is, for the corpus creator, a consequent effort. Still, we believe that the RefCo label provides substantial benefits to the corpus submitter and the other different stakeholder in the field, which will incentivize the submission of data sets.

3.1 For the Corpus Submitter

One of the first incentise comes from the submission process itself. As the QUEST RefCo supports field workers by guiding them through consistenty and completeness checks of their metadata and annotations, the quality of their work is be positively affected.

Making the corpus available to the public through the QUEST labelling process adds to the accountability of the fieldworker's research by allowing for replication of results (Riesberg 1998). Publication of datasets on which results are based is increasingly viewed as important and enforced by publication outlets in linguistics, following practices in other sciences. It facilitates fair recognition of the efforts invested by the fieldworker by facilitating proper use and citation of the data. It also facilitates future re-use of the data by the fieldworker herself, as it implies fully and consistently processing and properly archiving data.

[3] The Corpus Documentation and its template, which explains how to write the document, were designed by Kilu von Prince.

3.2 Funding organizations

From a funding organization's perspective, RefCo provides a yardstick for the success of a project that involves data collection. First, there is a growing consensus that the public has certain access rights regarding the re-use of public funded researches and their results (Wilkinson et al. 2016). Existing private funding schemes for fieldwork projects, like ELDP, have implemented schemes to enforce best practices for ELDP-funded fieldworkers and require their grantees to implement those, or the funding would not be resumed (Holton & Seyfeddinipur 2018).

3.3 To the Linguistic Community

As it has been stated many times, every language offers a unique window into the puzzle of the human intellectual capacity. In this sense, making fieldwork data on under-described languages accessible contributes to basing linguistic theorizing on a greater sample than the relatively few, well-described languages that most current theories rely on (Hale et al. 1992; Anand, Chung & Wagers 2015; Norcliffe, Harris & Jaeger 2015). One of the primary concern of QUEST is to render linguistic corpora available to other linguists as well as the scientific community in general. The RefCo QUEST criteria focus on information are of particular relevance to re-use for typology. Howeverr, consistently described, coherent, and accessible datasets resulting form RefCo labelling process will render these accessible also for anthropologists and social scientist in general.

3.4 The Speech Communty

Finally, for a speech community which have been subject of a documentation project labelled by RefCo, the process ensures its sustainability by making them findable and accessible. It raises the visibility of the language contributes at the conservation of the cultural heritage, in particular in case of languages that are endangered of becoming extinct, as a whole or with respect to certain traditional genres (Mosel 2006).

4 Conclusion

The QUEST project is an implementation of the current standards and good practices coming from the archiving and linguistics communities. As multiplying metadata and recommendations regarding language documentation corpora adds to the difficulties encountered by corpus creator during the production of their data, the perspective we adopted is to facilitate their application by data creator. To do so, we provide guidelines for describing the corpus The quality criteria and metedata we have seen in this article associated with RefCo, a subproject of QUEST dedicated to the problematic of corpora for cross-linguistics. The position adopted by RefCo is to ease the production of a corpus first by limiting to the minimum the quality criteria and metadata the dataset creators have to provide, second by accompanying them during that application by providing support and guidelines.

References

Anand, Pranav, Sandra Chung & Matthew Wagers. 2015. *Widening the Net: Challenges for Gathering Linguistic Data in the Digital Age*. Paper submitted to the National Science Foundation as part of its SBE 2020 planning activity. https://www.nsf.gov/sbe/sbe_2020/2020_pdfs/Wagers_Matthew_121.pdf.

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey & Fei Xia. 2013. Towards Creating Precision Grammars from Interlinear Glossed Text: Inferring Large-Scale Typological Properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 74–83. Sofia, Bulgaria: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-2710.

Bender, Emily M., Michael Wayne Goodman, Joshua Crowgey & Fei Xia. 2013. Towards Creating Precision Grammars from Interlinear Glossed Text: Inferring Large-Scale Typological Properties. *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 74–83. Sofia, Bulgaria: Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-2710.

Berez-Kroeker, Andrea L, Lauren Gawne, Susan Smythe Kung, Barbara F Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. 'Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field'. *De Gruyter*, Linguistics, 56 (1): 18. https://doi.org/10.1515/ling-2017-0032.

Bowern, Claire. 2008. *Linguistic Fieldwork - A Practical Guide*. New York: Palgrave MacMillan.

Haig, Geoffrey & Stefan Schnell. 2016. The discourse basis of ergativity revisited. *Language* 92(3). 591–618. doi:10.1353/lan.2016.0049.

Hale, Ken, Michael Krauss, Lucille J. Watahomigie, Akira Y. Yamamoto, Colette Craig, LaVerne Masayesva Jeanne & Nora C. England. 1992. Endangered languages. *Language* 68(1). 1–42. https://doi.org/10.1353/lan.1992.0052.

Holton, Gary & Mandana Seyfeddinipur. 2018. Reflections on funding to support documentary linguistics. In Bradley McDonnell, Andrea L. Berez-Kroeker & Gary Holton (eds.), *Reflections on Language Documentation 20 Years After Himmelmann 1998* (Language Documentation & Conservation Special Publication 15), 100–109. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10125/24812 (17 December, 2019).

Meakins, Felicity, Jennifer Green & Myfany Turpin. 2018. *Understanding Linguistic Fieldwork*. London, New York: Routledge.

Mettouchi, Amina, Zygmunt Frajzyngier & Christian Chanard (eds.). 2017. *Corpus-based cross-linguistic studies on Predication (CorTypo)*. http://cortypo.huma-num.fr/Publication (3 November, 2018).

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of Language Documentation*, 67–85. Berlin: Mouton de Gruyter.

Norcliffe, Elisabeth, Alice C. Harris & T. Florian Jaeger. 2015. Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances. *Language, Cognition and Neuroscience* 30(9). 1009–1032. https://doi.org/10.1080/23273798.2015.1080373.

Stave, Matthew, Ludger Paschen, François Pellegrino & Frank Seifart. 2020. Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. To appear in *Linguistic Vanguard*.

Thieberger, Nicholas & Andrea L. Berez. 2011. Linguistic Data Management. In Nicholas Thieberger (ed.), *The Oxford Handbook of Linguistic Fieldwork*, 90–118. Oxford, New York: Oxford University Press.

Thieberger, Nick, Anna Margetts, Stephen Morey & Simon Musgrave. 2016. Assessing annotated corpora as research output. Australian Journal of Linguistics 36(1). 1–21. https://doi.org/10.1080/07268602.2016.1109428.

Vasile, Aurelia, Séverine Guillaume, Mourad Aouini & Alexis Michaud. 2020. *Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger*. I2D - Information, données & documents. https://halshs.archives-ouvertes.fr/halshs-02870206.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1). 1–9. https://doi.org/10.1038/sdata.2016.18.