RefCo and its Checker

Improving Language Documentation Corpora's Reusability Through a Semi-Automatic Review Process

Herbert Lange and Jocelyn Aznar
herbert.lange@uni-hamburg.de, aznar@leibniz-zas.de

13th Conference on Language Resources and Evaluation, 20-25 June 2022

QuESt

UHH Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

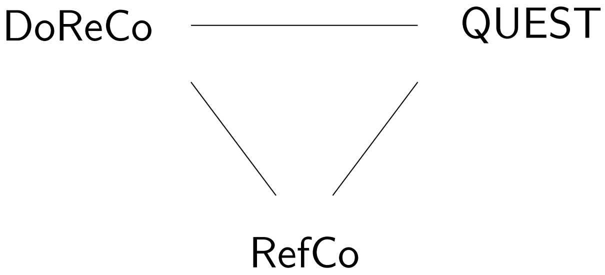
ZAS

Motivations

Why?

- Improve Reusability (FAIR)
- Mitigate the loss of linguistic data as much as possible
- Documenting and amending a corpus is a tedious task

Context

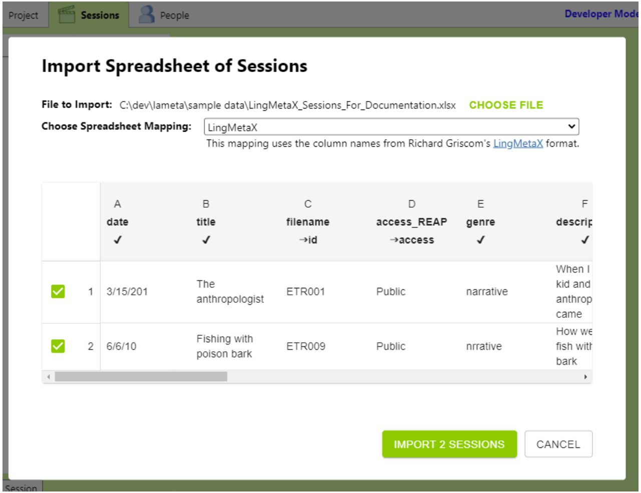


DoReCo

QUEST

RefCo

Initiatives to improve data reusability



Import Spreadsheet of Sessions

File to Import: C:\dev\lameta\sample data\LingMetaX_Sessions_For_Documentation.xlsx

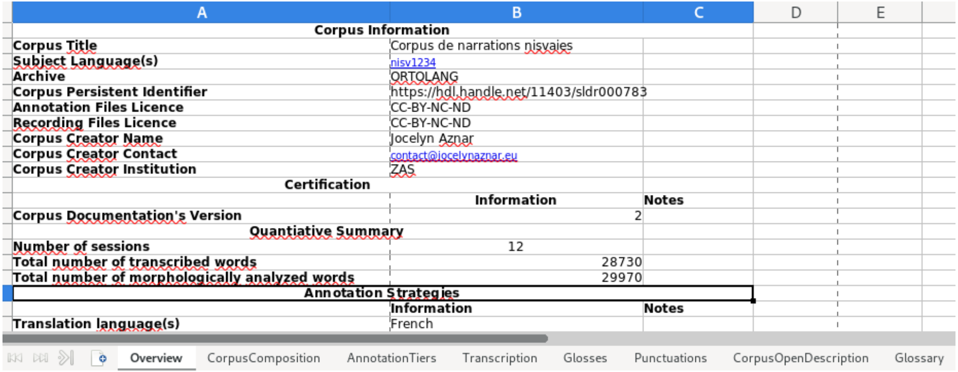
Choose Spreadsheet Mapping: LingMetaX

This mapping uses the column names from Richard Griscom's LingMetaX format.

		A date ✓	B title ✓	C filename →id	D access_REAP →access	E genre ✓	F descrip ✓
☑	1	3/15/201	The anthropologist	ETR001	Public	narrative	When I kid and anthrop came
☑	2	6/6/10	Fishing with poison bark	ETR009	Public	nrrative	How we fish wit bark

Figure: Source: https://github.com/onset/laMETA/releases

The RefCo toolkit



A	B	C
Corpus Information		
Corpus Title	Corpus de narrations nisvaies	
Subject Language(s)	nisv1234	
Archive	ORTOLANG	
Corpus Persistent Identifier	https://hdl.handle.net/11403/sldr000783	
Annotation Files Licence	CC-BY-NC-ND	
Recording Files Licence	CC-BY-NC-ND	
Corpus Creator Name	Jocelyn Aznar	
Corpus Creator Contact	contact@jocelynaznar.eu	
Corpus Creator Institution	ZAS	
Certification		
	Information	Notes
Corpus Documentation's Version	2	
Quantiative Summary		
Number of sessions	12	
Total number of transcribed words	28730	
Total number of morphologically analyzed words	29970	
Annotation Strategies		
	Information	Notes
Translation language(s)	French	

Figure: The overview tab of the RefCo corpus documentation

Contains a reference manual, a checklist for the reviewer, a report form and a documentation template spreadsheet (as well as compatible JSON/XML schemas).

Our proposition for the semi-automatic evaluation

A theory agnostic approach:

The documentation should contain a description of the terms and symbols used to annotate the corpus.

Two main principles:

Consistency: One documentation for the corpus, only files relevant to the corpus.

Coherency: The documentation exactly matches the content, the complete content is described by the documentation.

Annotation Conventions

Example: *na=tog* : ‘*I want*’

- Stuttgart-Tübingen tag set (STTS)
  VMFIN
- Universal Dependencies (UD)
  VERB `VerbForm=Fin,Mood=Ind,Voice=Act,Number=Sing,`
  `Person=1, [...]`
- Leipzig Glossing rules (LGR)
  1SG=*want*

Interlinear Glossing / Leipzig Glossing Rules

(1) Avyn, from the Corpus of Nisvai narratives[1]

E	*nabuqao*	*ili*	*ga=prag*	*pal*	*hni*	*mog*	*bul*	*na=tog*
INTER	fence	DET	3SG=do	ASP.C	OBJ	ASP.F	COO.OP	1SG=want

dara=mai	*van*	*druan*	*ahnao*	*syh*	*dara=roc*	*nyn...*
1INCL=come	ASP.R	COO.NM	OBJ.1SG	COO.CSQ	1INCL=follow	DEI.R

‘Eh, that’s it, I’m done with the fence and I’d like that we go together to follow the...’

[1] Following https://hdl.handle.net/11403/sldr000783/v2

Interlinear Glossing / Leipzig Glossing Rules

(1) Avyn, from the Corpus of Nisvai narratives[1]

E	*nabuqao*	*ili*	*ga=prag*	*pal*	*hni*	*mog*	*bul*	*na=tog*
INTER	fence	DET	3SG=do	ASP.C	OBJ	ASP.F	COO.OP	1SG=want

dara=mai	*van*	*druan*	*ahnao*	*syh*	*dara=roc*	*nyn...*
1INCL=come	ASP.R	COO.NM	OBJ.1SG	COO.CSQ	1INCL=follow	DEI.R

‘Eh, that’s it, I’m done with the fence and I’d like that we go together to follow the...’

Table: Documentation

OBJ	direct object
1SG	first person singular

[1] Following `https://hdl.handle.net/11403/sldr000783/v2`

Interlinear Glossing / Leipzig Glossing Rules

(1) Avyn, from the Corpus of Nisvai narratives[1]

E	*nabuqao*	*ili*	*ga=prag*	*pal*	*hni*	*mog*	*bul*	*na=tog*
INTER	fence	DET	3SG=do	ASP.C	OBJ	ASP.F	COO.OP	1SG=want

dara=mai	*van*	*druan*	*ahnao*	*syh*	*dara=roc*	*nyn...*
1INCL=come	ASP.R	COO.NM	OBJ.1SG	COO.CSQ	1INCL=follow	DEI.R

‘Eh, that’s it, I’m done with the fence and I’d like that we go together to follow the...’

Table: Documentation

OBJ	direct object
1SG	first person singular
ASP.C	durative
ASP.F	completive
ART.C	article common name

[1] Following https://hdl.handle.net/11403/sldr000783/v2

Automatic Checks[2]

Consistency

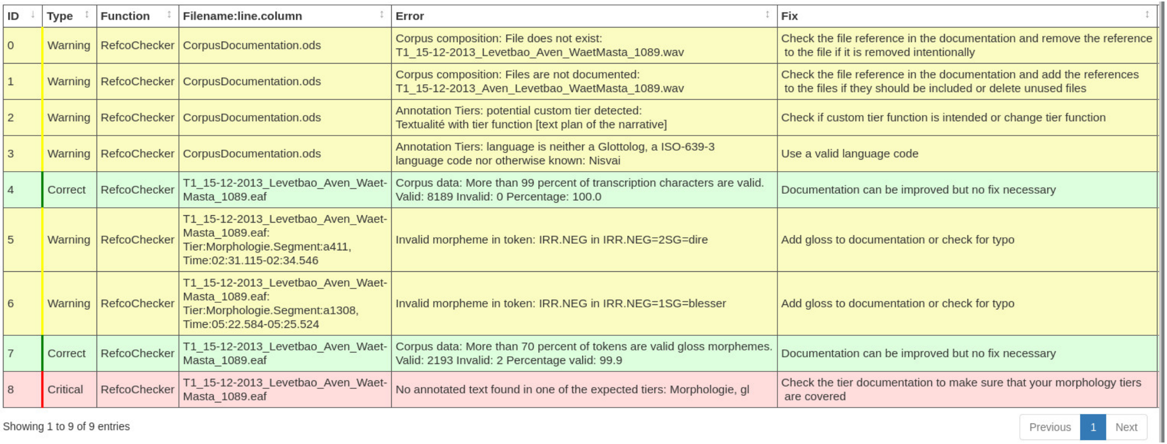
Files present vs. files documented

Coherency

- Coherent documentation of tiers, i.e. all tiers are documented and only documented tiers exist
- Coherent documentation of transcription convention
- Coherent documentation of annotation conventions

[2] Part of the Hamburger Zentrum für Sprachkorpora (HZSK) Corpus Services

Reporting the issues

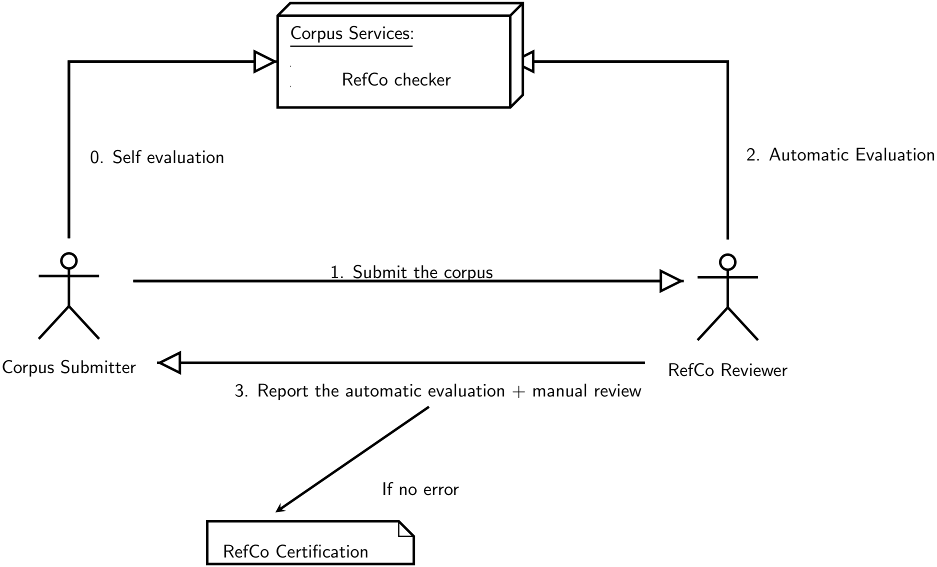


ID	Type	Function	Filename:line.column	Error	Fix
0	Warning	RefcoChecker	CorpusDocumentation.ods	Corpus composition: File does not exist: T1_15-12-2013_Levetbao_Aven_WaetMasta_1089.wav	Check the file reference in the documentation and remove the reference to the file if it is removed intentionally
1	Warning	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

Figure: a report pointing out the errors and problems in the corpus

The certification dialog



Discussion: Spreadsheet as a versatile interface

- as a familiar user interface for field linguists,
- as an interface to discuss the quality criteria.

Contacts

For further questions & comments, please feel free to contact us:

- Herbert Lange : `herbert.lange@uni-hamburg.de`
- Jocelyn Aznar : `aznar@leibniz-zas.de`

RefCo toolkit

RefCo checker