CLD^2

Language Documentation meets NLP for Revitalising Endangered Languages

Roberto Zariquiey, Arturo Oncevay, Javier Vera

PUCP (Peru) U of Edinburgh (UK) PUCV (Chile)

TABLE OF CONTENTS

01 Introduction

02 Language Documentation, Language Revitalization, and CL/NLP

03 NLP and Endangered Language Documentation

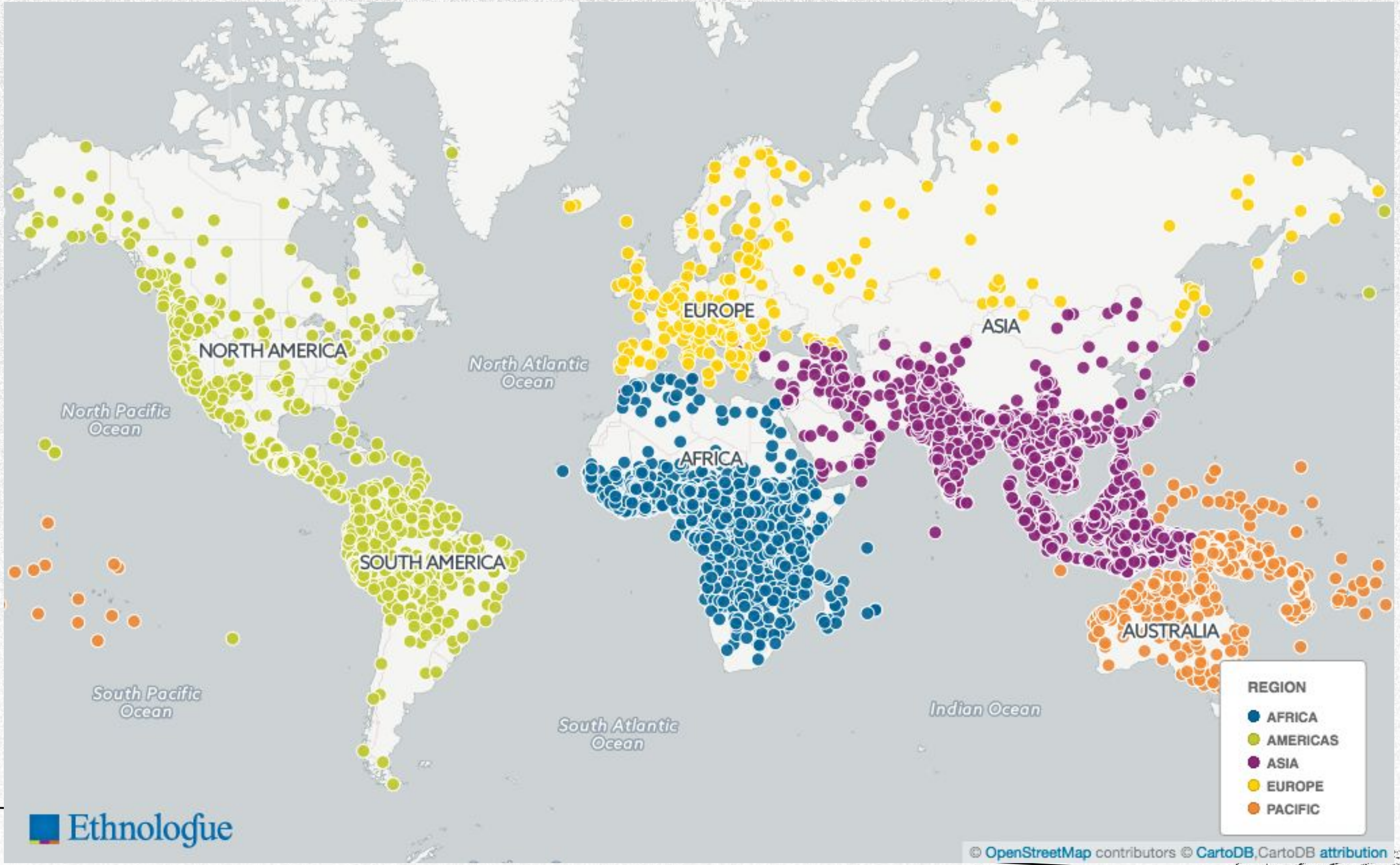
04 CLD^2: Computational Language Documentation and Development

05 Conclusions (+ Extra!)

01

Introduction

Introduction



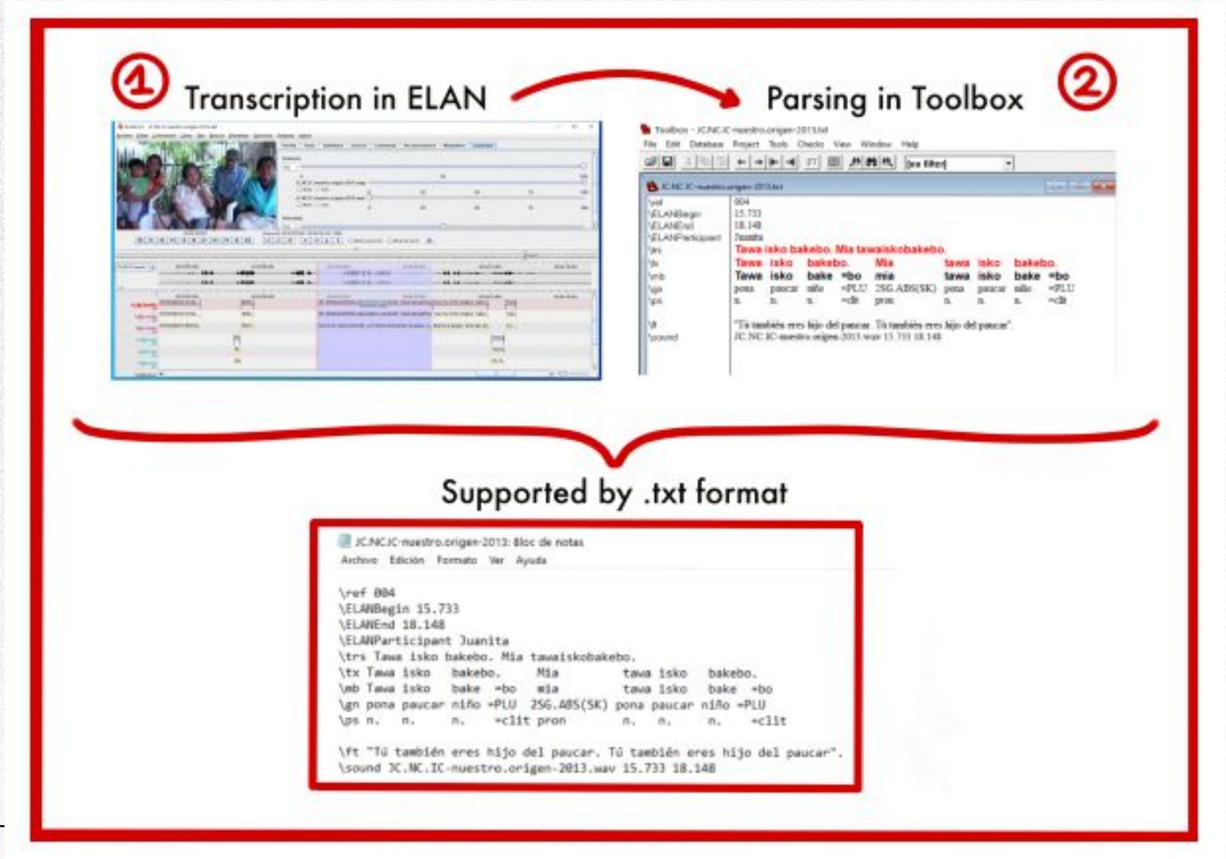
Introduction

Documentary Linguistics attempts to produce permanent **records** of the linguistic and cultural practices of the most threatened speech communities.

Audio and video recordings
Lexical and text collections
etc.

The Endangered Languages Archive

Introduction

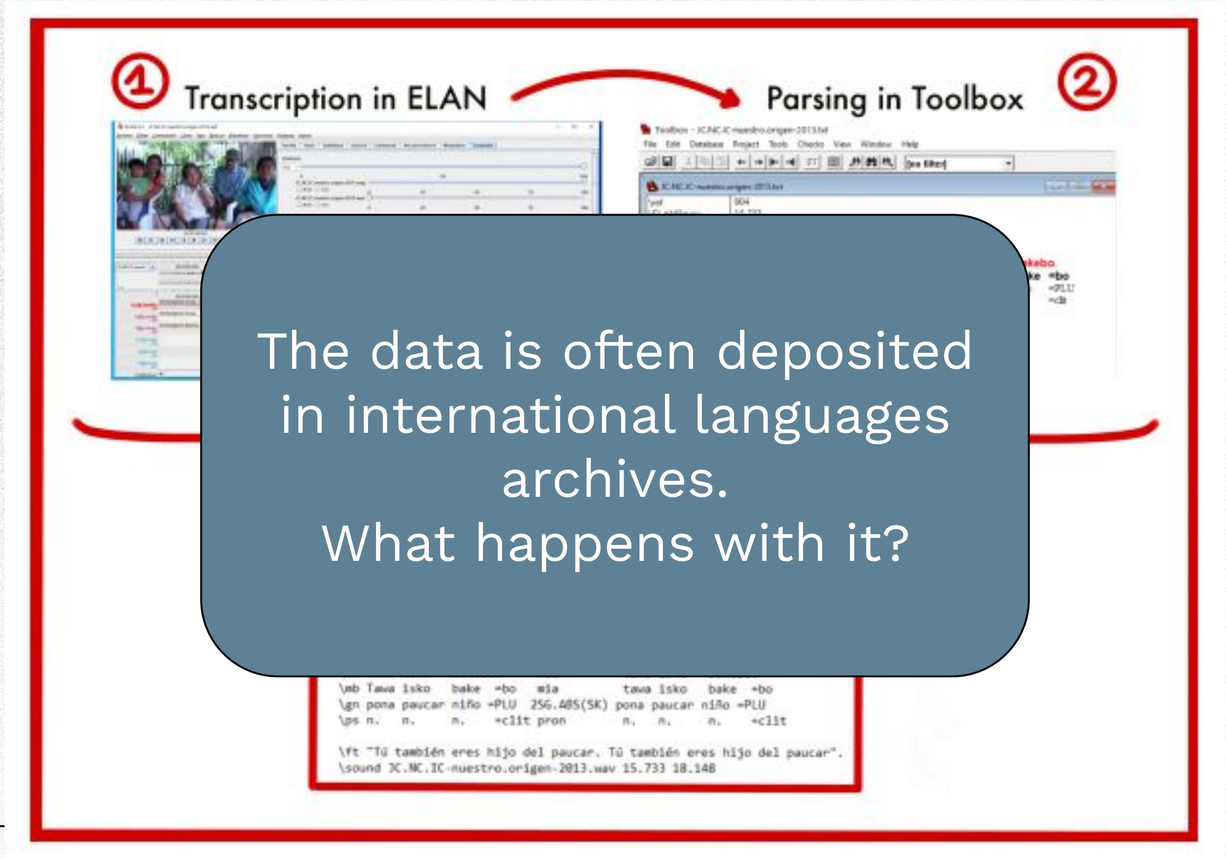

1
Transcription in ELAN
Parsing in Toolbox
2
Toolbox - JC.NC.IC-nuestro.origen-2013.txt
File Edit Database Project Tools Checks View Window Help
JC.NC.IC-nuestro.origen-2013: Bloc de notas
Archivo Edición Formato Ver Ayuda

Supported by .txt format

```
\ref 004
\ELANBegin 15.733
\ELANEnd 18.148
\ELANParticipant Juanita
\trs Tawa isko bakebo. Mia tawaiskobakebo.
\tx Tawa isko   bakebo.     Mia          tawa isko   bakebo.
\mb Tawa isko   bake  =bo   mia          tawa isko   bake  =bo
\gn pona paucar niño =PLU  2SG.ABS(SK) pona paucar niño =PLU
\ps n.   n.     n.    =clit pron         n.   n.     n.    =clit

\ft "Tú también eres hijo del paucar. Tú también eres hijo del paucar".
\sound JC.NC.IC-nuestro.origen-2013.wav 15.733 18.148
```

Introduction


1
Transcription in ELAN
Parsing in Toolbox
2
Toolbox - JC.NC.IC-nuestro.origen-2013.txt
File Edit Database Project Tools Checks View Window Help
JC.NC.IC-nuestro.origen-2013.txt

```
\mb Tawa isko   bake  =bo   mia          tawa isko   bake  =bo
\gn pona paucar niño  =PLU  2SG.ABS(SK) pona paucar niño =PLU
\ps n.   n.     n.    =clit pron         n.   n.     n.    =clit

\ft "Tú también eres hijo del paucar. Tú también eres hijo del paucar".
\sound JC.NC.IC-nuestro.origen-2013.wav 15.733 18.148
```

The data is often deposited in international languages archives.

What happens with it?

CLD^2

In this paper, we reflect on the necessity of increasing the interactions between **Documentary Linguistics** and **NLP**

Similar to Levow et al., (2017), van Esch et al., (2019), among others.

02

Language Documentation and Language Revitalization

What is language documentation

- Aims to create permanent records of the linguistic and cultural practices of the most threatened speech communities.
- It is a long-term and time-consuming task that may take several years and requires considerable funding.

SIL.org

Language documentation and revitalization

What are the expectations for LD? Are they met?

Can language revitalization be supported by LD?

How is CL/NLP helping?

Language Documentation and CL/NLP

The interaction has mostly focused in tools and how to support the documentation process, e.g.:

- Daan van Esch, Ben Foley, and Nay San. 2019. **Future directions in technological support for language documentation.** In Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages.
- Antonios Anastasopoulos, Christopher Cox, Graham Neubig, and Hilaria Cruz. 2020. **Endangered languages meet Modern NLP.** In Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts, pages 39–45, Barcelona, Spain (Online). International Committee for Computational Linguistics

03

NLP and Endangered Language Documentation

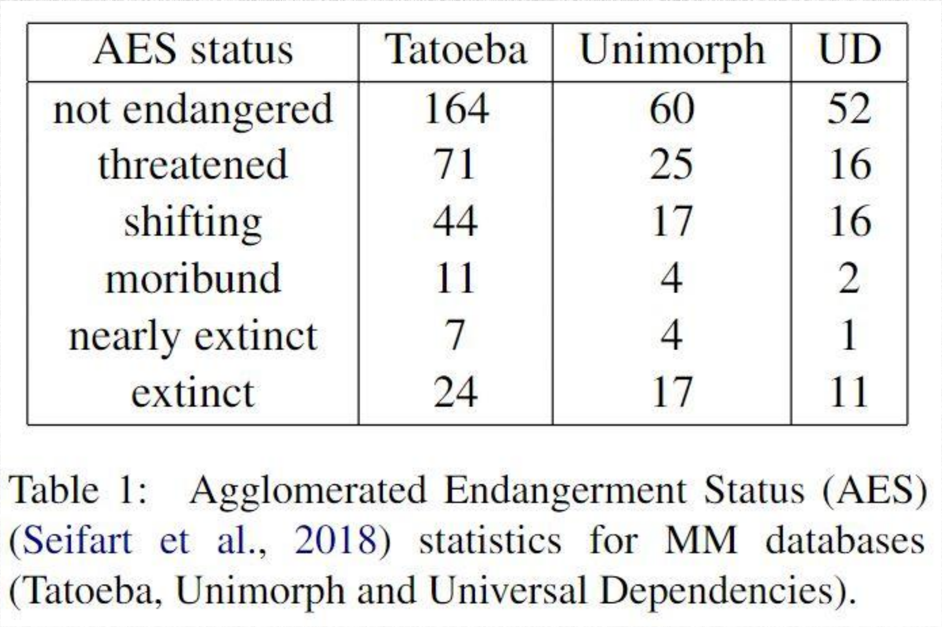
Has NLP taken advantage of the outputs of the documentation projects, especially for endangered languages?

- **Data**:
  - ACL Anthology
  - Language inventory of massive multilingual datasets in NLP research (MM): Unimorph, Universal Dependencies and Tatoeba
  - The Endangered Languages Archive: ELAR

Also, we work with Glottolog 4.4 (extended inventory of world's languages) and the Agglomerated Endangerment Status (AES) (vitality status)

- **Processing**
  - We identified all the publications in the ACL Anthology whose title or abstract explicitly includes the name of a language
  - A similar procedure for ELAR: from all the 570 projects, we identified 307 language matches with Glottolog (only 286 matches with geographical info)
  - MM datasets: ISO codes or languages are matched with Glottolog

Results

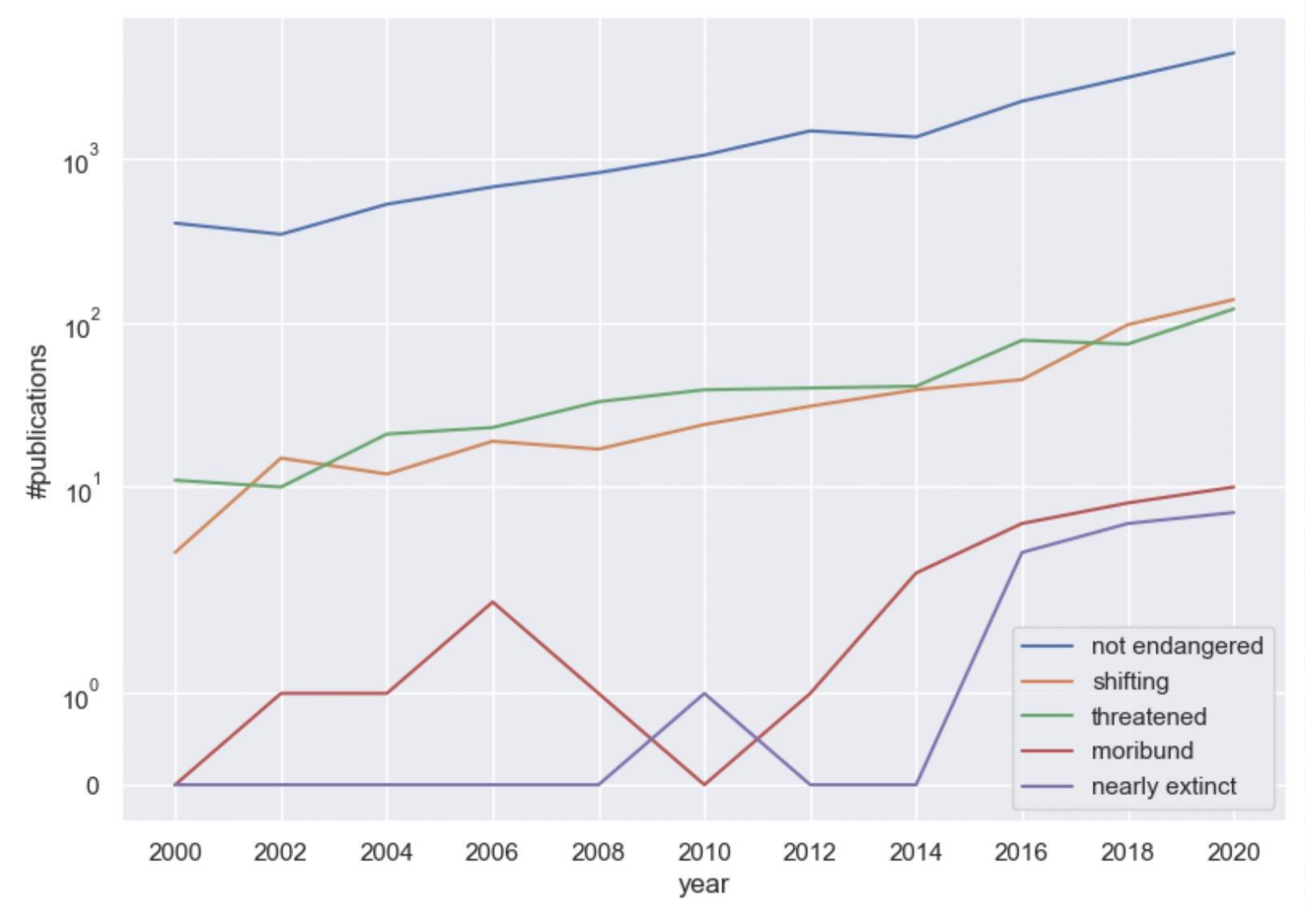


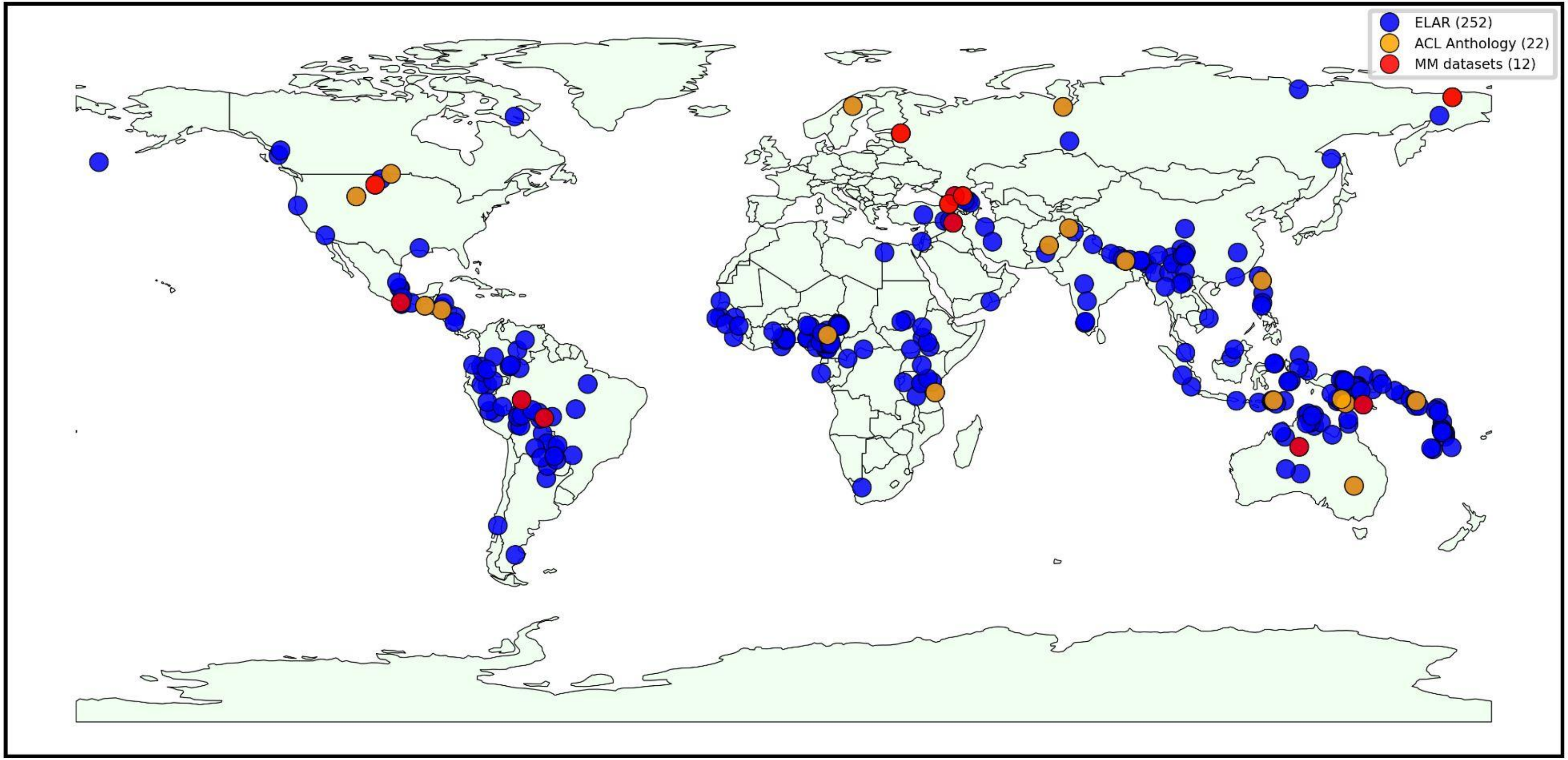
AES status	Tatoeba	Unimorph	UD
not endangered	164	60	52
threatened	71	25	16
shifting	44	17	16
moribund	11	4	2
nearly extinct	7	4	1
extinct	24	17	11

Table 1: Agglomerated Endangerment Status (AES) (Seifart et al., 2018) statistics for MM databases (Tatoeba, Unimorph and Universal Dependencies).

Results

#publications in the ACL Anthology that mentions a language in their title or abstract.





Very low overlapping between ACL Anthology, ELAR and MM datasets

Discussion

- Accessing databases of a wide sample of endangered languages would be beneficial for the NLP agenda. However, this has not been the priority. **Why?**
  - Visibility
    - Language documentation databases are mostly known in the linguistic community
  - Accessibility
    - Partial access to language documentation databases
  - Readability
    - Language documentation outputs are not processed for immediate NLP projects

04

CLD^2: Computational Language Documentation and Development

Can documentary projects consider more “NLP-friendly”* outcomes?

A basic protocol, for instance:

1. Monolingual and parallel corpora
2. A set of annotations in universal frameworks well known in linguistic typology (UniMorph, Universal Dependencies)
3. The annotation scheme (if applicable)

*Audio data could also be output in more friendly formats for speech-related tasks.

How about a basic language toolkit?

- "**Computarisation**" (Berment, 2022) of a language could support revitalisation efforts.
- This does not mean that all language documentation projects must include a large technology development (it is expensive already).
- But, linguistic database could be **multipurpose**: to easier the development of language technologies.

How about a basic language toolkit*?

1. Morphological tools: Morphological information is already crafted in documentation, how can we make it “easier to read” for NLP?
2. Spell-checker: Dictionary or frequency-based approaches are approachables outputs.
3. Syntactic parser (e.g. using the UD schema): It is not a main focus on current LD, but could be relevant for linguistic typology research.
4. POS and NER taggers: There is information that could be adapted from the glosses.

*Without forgetting to ask to the communities what could be relevant to them

05

Conclusions (+ extra)

Conclusion

CLD^2 calls for an **enrichment of language documentation projects** by means of **incorporating** components, outcomes and methods from **NLP research**, as a strategy to promote the computarisation and revitalisation of minority languages.

Most of the interactions between LD and NLP/CL have mostly focused on software and the documentation process. There should be a better interaction from the two ways:

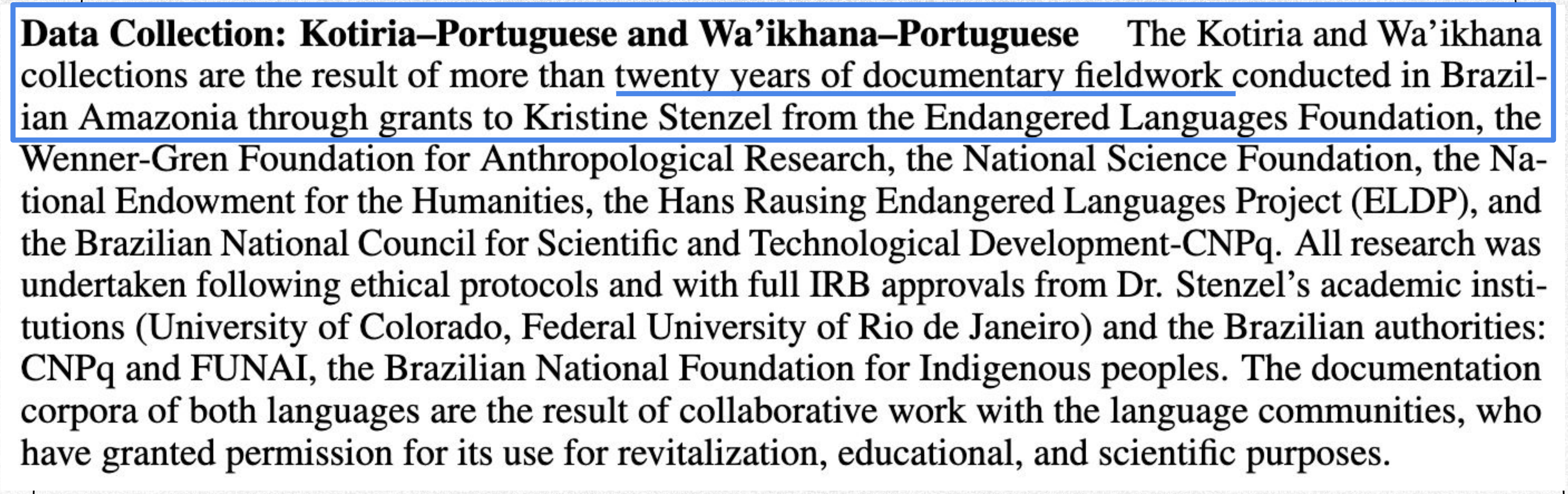
1. The NLP community could pay more attention to the documentation databases
2. Field linguists could consider to make their outputs more NLP-friendly

To appear @ LREC 2022

Building an Endangered Language Resource in the Classroom: Universal Dependencies for Kakataibo

Roberto Zariquiey[ρ], Claudia Alvarado[ρ], Ximena Echevarria[ρ], Luisa Gomez[ρ],
Rosa Gonzales[ρ] , Mariana Illescas[ρ] , Sabina Oporto[ρ],
Frederic Blum[β], Arturo Oncevay[ϵ], Javier Vera[υ]
[ρ]Pontificia Universidad Católica del Perú, Peru
[β]Humboldt-Universität zu Berlin and Leibniz-Zentrum Allgemeine Sprachwissenschaft
[ϵ]University of Edinburgh, Scotland
[υ]Pontificia Universidad Católica de Valparaíso, Chile
{rzariquiey, claudia.alvarado, ximena.echevarria, luisa.gomez, a20175617, m.illescasb, sabina.oporto}@pucp.edu.pe
frederic.blum@hu-berlin.de, a.oncevay@ed.ac.uk, javier.vera@pucv.cl

AmericasNLP @ NeurIPS 2022
Speech-to-text translation shared task



Data Collection: Kotiria–Portuguese and Wa'ikhana–Portuguese The Kotiria and Wa'ikhana collections are the result of more than twenty years of documentary fieldwork conducted in Brazilian Amazonia through grants to Kristine Stenzel from the Endangered Languages Foundation, the Wenner-Gren Foundation for Anthropological Research, the National Science Foundation, the National Endowment for the Humanities, the Hans Rausing Endangered Languages Project (ELDP), and the Brazilian National Council for Scientific and Technological Development-CNPq. All research was undertaken following ethical protocols and with full IRB approvals from Dr. Stenzel's academic institutions (University of Colorado, Federal University of Rio de Janeiro) and the Brazilian authorities: CNPq and FUNAI, the Brazilian National Foundation for Indigenous peoples. The documentation corpora of both languages are the result of collaborative work with the language communities, who have granted permission for its use for revitalization, educational, and scientific purposes.

CLD^2

Language Documentation meets NLP for Revitalising Endangered Languages

Roberto Zariquiey, Arturo Oncevay, Javier Vera

PUCP (Peru) U of Edinburgh (UK) PUCV (Chile)