DoReCo:
Exploring phonetic lengthening and information rate

ANR-DFG 03/2019-02/2022

Matthew Stave, Frank Seifart, François Pellegrino
DDL-CNRS Lyon

DoReCo project overview

- Three main objectives
  1. Create a time-aligned reference corpus of spoken language for over 50 languages (DoReCo = Language Documentation Reference Corpus)
     a. Phonemically time-align the corpus
  2. Perform analyses on the corpus
  3. Make resources available to scientific community

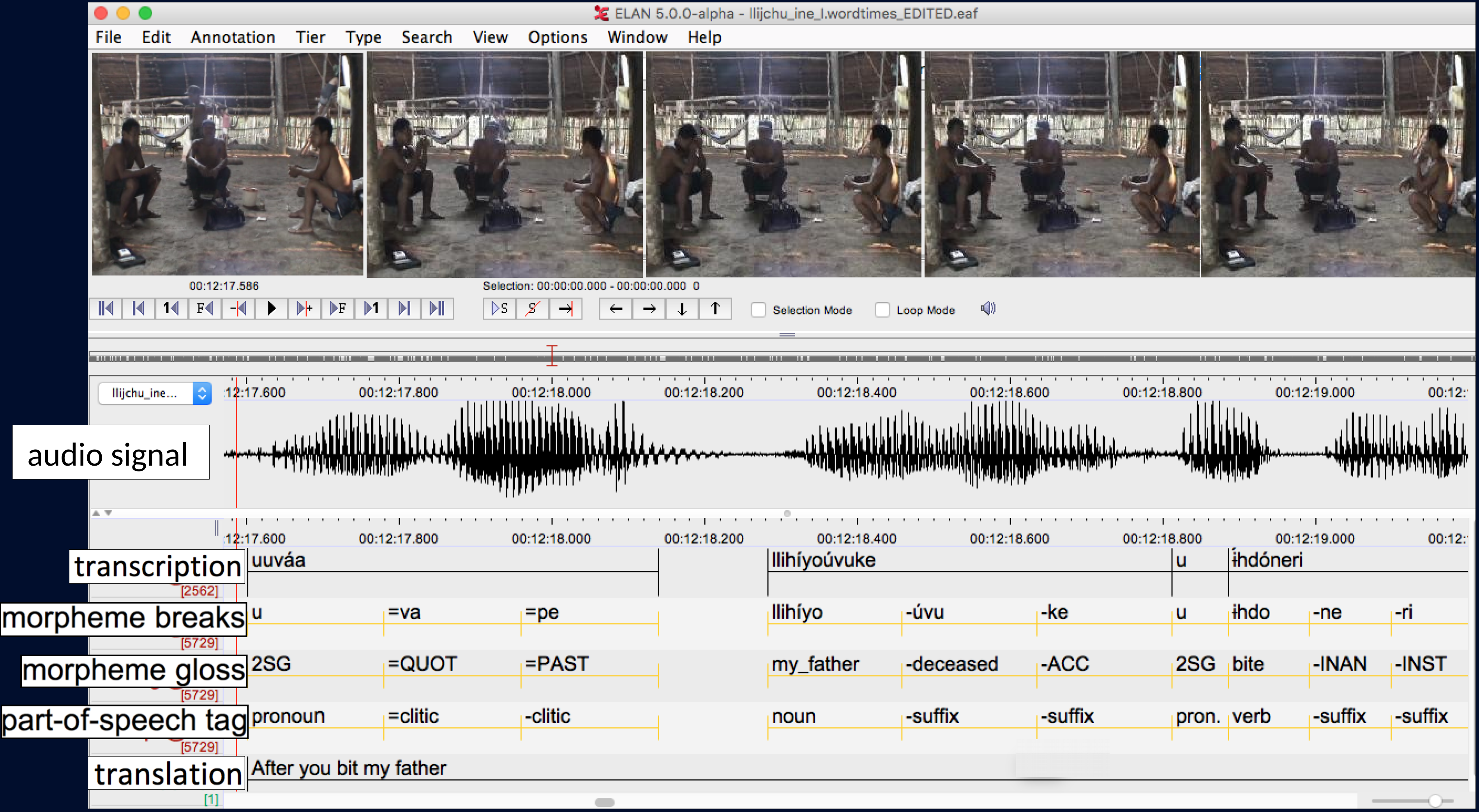
1. Corpus creation

- Broad cross-linguistic coverage

- > 10,000 words per language

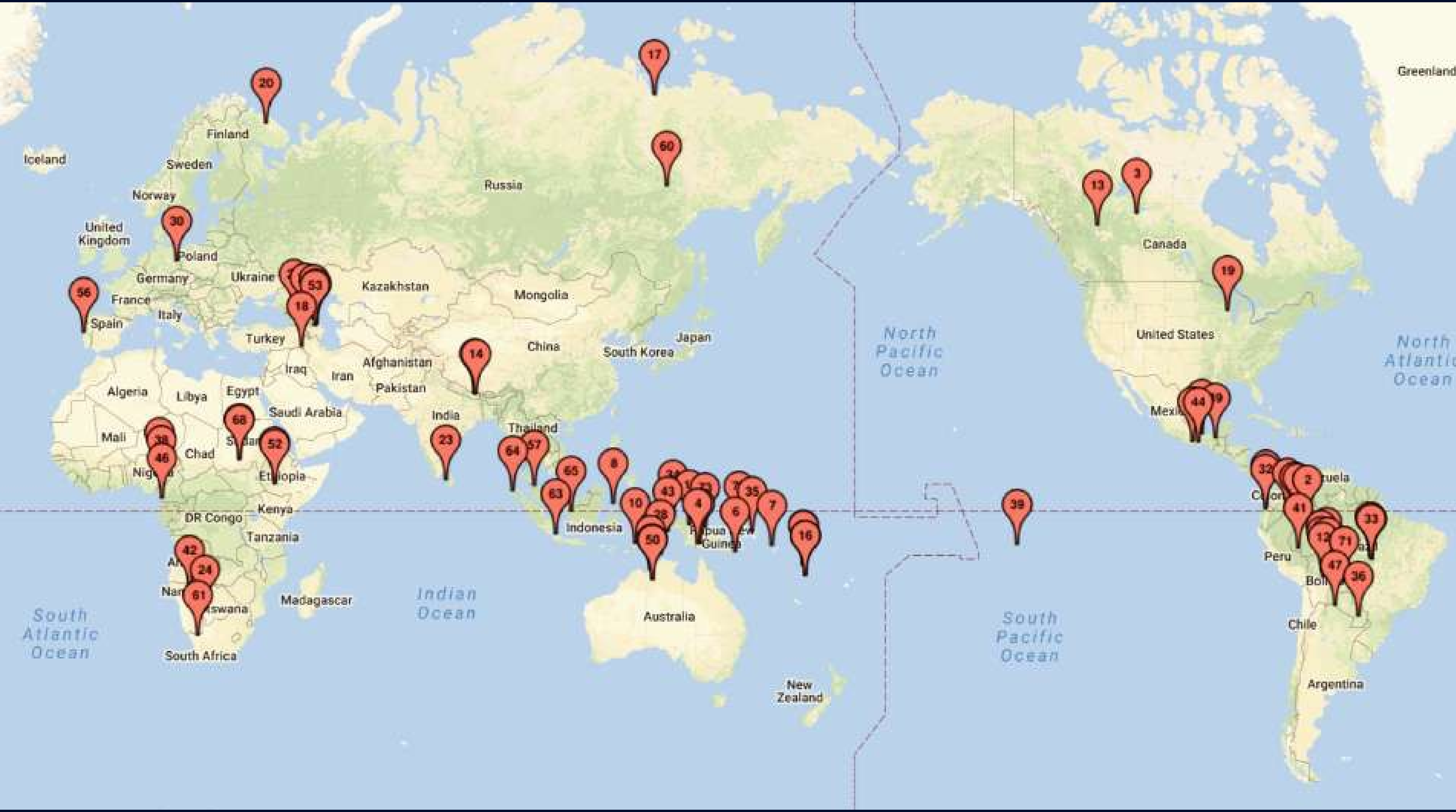
- 50+ languages with audio transcriptions
- 30+ languages with additional morphological analysis (subset)

Bora corpus example (ELAN)



transcription	uuváa			Ilihíyoúvuke			u	ɨ́hdóneri		
morpheme breaks	u	=va	=pe	Ilihíyo	-úvu	-ke	u	ɨhdo	-ne	-ri
morpheme gloss	2SG	=QUOT	=PAST	my_father	-deceased	-ACC	2SG	bite	-INAN	-INST
part-of-speech tag	pronoun	=clitic	-clitic	noun	-suffix	-suffix	pron.	verb	-suffix	-suffix
translation	After you bit my father									

DoReCo candidate languages

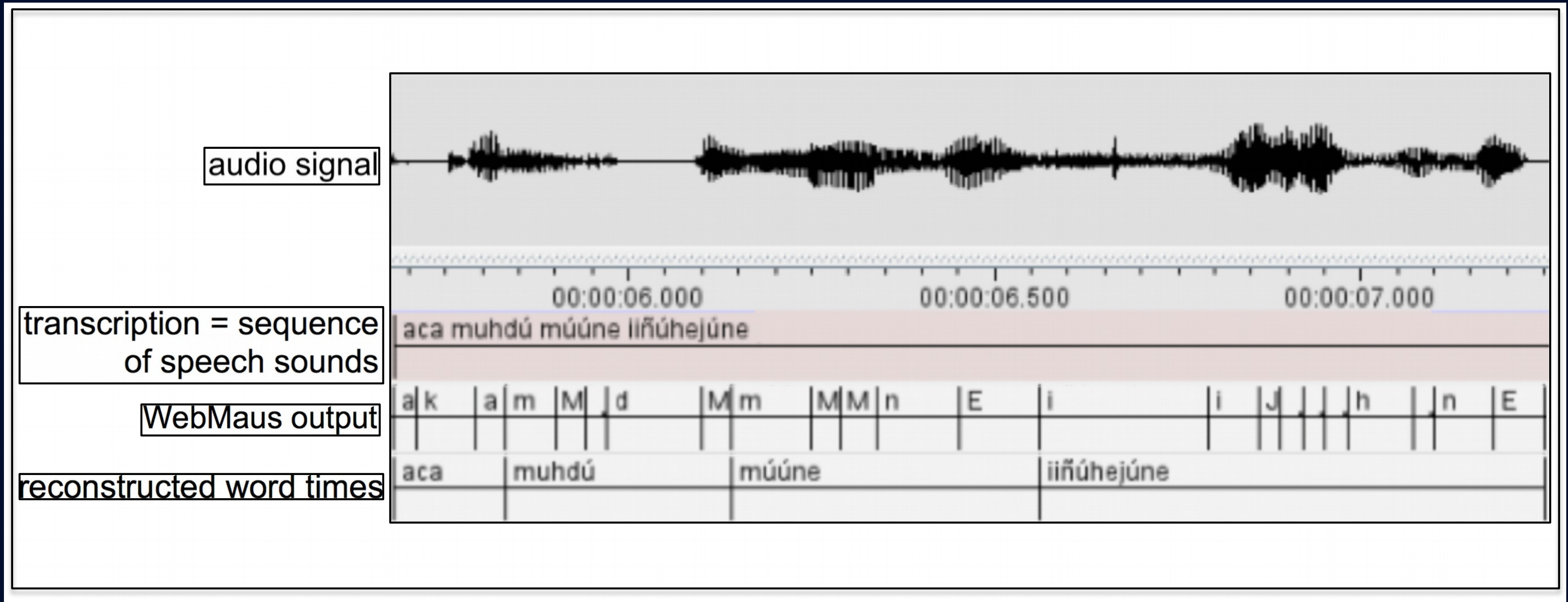


1. Corpus creation

- Corpora are shared by language documenters (with permission from language communities)
- Variety of file formats
  - ELAN, Flex, Toolbox, EXMARaLDA, Pangloss, …
  - Developing file conversion tools
- From a variety of archives
  - TLA, ELAR, PARADISEC, Pangloss, …

1a. Phoneme alignment

- MAUS (Munich Automatic Segmentation System)
- Universal language model
- Evaluating performance on 50 languages (LREC, Marseille 2020)



2. Analyses of corpus

- Testing universal claims about language production on a diverse language sample
- Two focus areas
  - Articulatory phonetics
  - Information rate / packaging

2. Analyses of corpus

- Phonetic research questions
  - Final lengthening before prosodic boundaries
  - Relative compressibility of different phonological segments

2. Analyses of corpus

- Information rate
  - Optimized, universal “attractor state” for information rate
  - Tendency to package comparable amounts of information within inter-pausal units

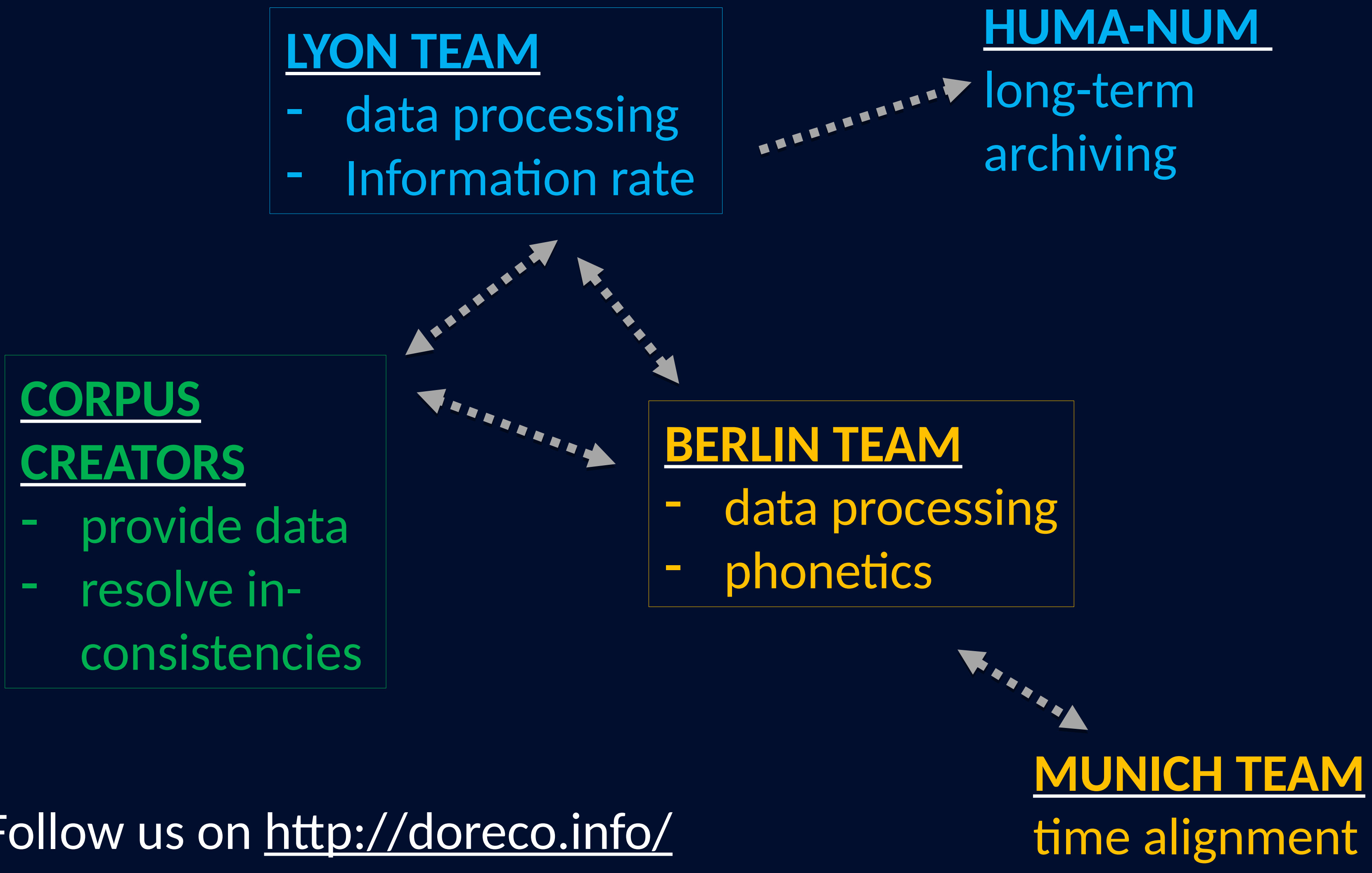
3. Connections with scientific community

- Corpus will be archived in Huma-Num
  - Annotations only, not audio
    - links to audio files in existing archives
  - Interoperability with other linguistic databases
    - WALS, Glottolog, CLLD
  - Publicly accessible in third year of project

3. Connections with scientific community

- Further resources to be shared publicly
  - Conversion tools between different file formats
    - ELAN, FLEX, Toolbox, EXMARaLDA, Pangloss, others
  - TEI encoding for long-term archival
  - ELAN MAUS input
- Actively seeking users and collaborators
  - Other time-alignment systems?
  - Further phonetic and other research questions?

DoReCo workflow and summary



LYON TEAM

- data processing
- Information rate

HUMA-NUM

long-term archiving

CORPUS CREATORS

- provide data
- resolve inconsistencies

BERLIN TEAM

- data processing
- phonetics

MUNICH TEAM

time alignment

Follow us on http://doreco.info/