Global Open Resources and Information for Language and Linguistic Analysis (GORILLA)

Damir Cavar & Malgorzata Cavar

Indiana University

Background

- 2012 AARDVARC NSF grant, multiple workshops, goal new technologies for archives (“language graveyard” motivation)
- 2013-2014 initial projects: cooperation with Tanja Schulz (ASR), Monica Macaulay, Arienne Dwyer to work on initial ASRs
- 2014 Relocation to Indiana University
- Cooperation with the Archive of Traditional Music (ATM) at IU
- Large Digitization project at IU for the archive content ($ 15 mil, now $ 50 mil)
- Suggested cooperation with Latin American countries

Background

since 2015

- Cooperation with the AHEYM project at IU, Dov-Ber Kerler
- Cooperation with Hilaria Cruz from UT Austin to create a first speech corpus on Chatino, Forced Aligner, NLP tools
- Launch of GORILLA, legal arrangements for CC-content
- Multiple Language Resources, technology projects
- grant proposals to DEL, DARPA
- Collaboration with The Language Conservancy, Lakhota.org
- LLOD cooperation, CLARIN groups

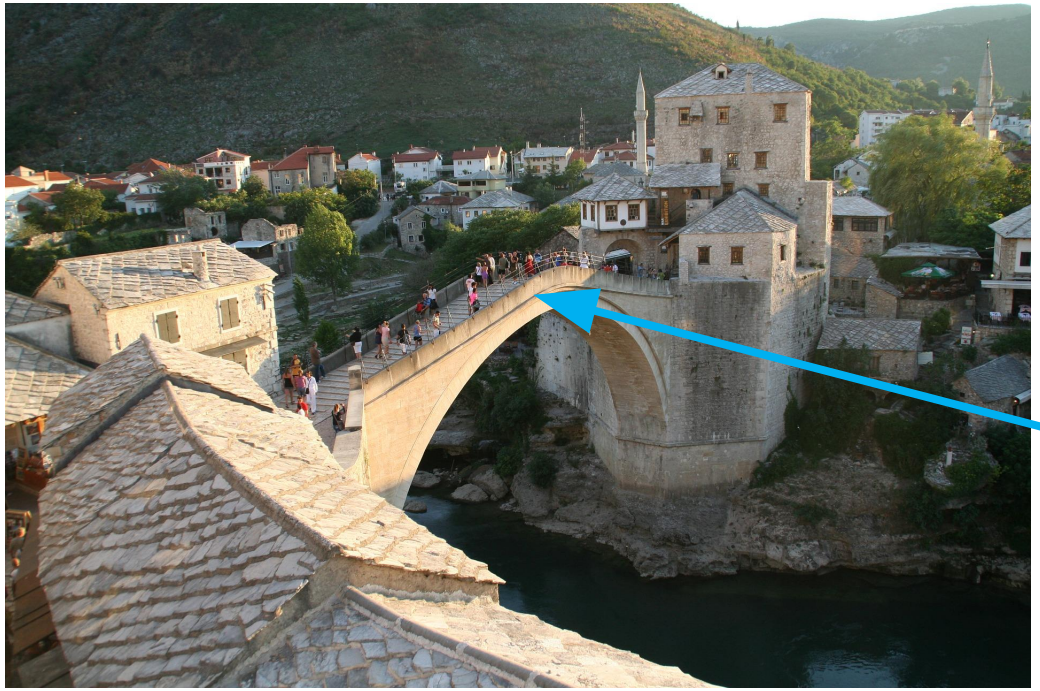
Background

since 2015

- Toyota Technological Institute at U Chicago (Sadaoki Furui): proposed speech corpora for all languages, budgeted $ 50 mil
- LFG Annual conference with focus on endangered languages
- Midwest Speech and Language Days: William Lewis (cooperation with Lakhota.org)
- Endangered Language Alliance cooperation with Chatino corpus
- Cooperation with UT Austin, Hans Boas and Texas German archive
- Exchange with Australia (Initial Summit, now IU-ANU cooperation) related to ATM

About goals

Language Archive ”Language Graveyards”



Language data producers/users: Endangered and under-resourced languages

Top layer of service for digital language data

About goals

Language Communities

Language Documenters: Transcription bottleneck

NLP:
Language resource bottleneck
Technological mono-culture

Disparate needs and techniques

Philosophy: Global

- No focus on a particular language family or geographical area
- Focus on endangered and under-resourced languages
- Promoting standards in terms of data formats and data use

Philosophy: Open

- Free of charge
- Accessible without limits imposed by the communities, researchers, or the archive that stores the data
- Free to be used under Creative-Commons Attribution-ShareAlike (CC BY-SA) license or freer for any use, including derivative and commercial use
- Using only common, open formats (XML-based annotation standards, Praat TextGrid, TEI, etc.)

Content: Resources

- Audio and video corpora, including parallel corpora
- Tree banks
- Lexicons and dictionaries
- Language models
- Other tools for NLP, for corpora analysis and processing

Example: SJQ Chatino

- Earlier work
- Primarily a spoken language
- Recently developed practical orthography (Cruz & Woodbury, 2014)
- AILLA: 107 hours of audio and video recordings (including restricted files), out of that perhaps 10 hours transcribed (out of that mainly grammatical elicitation),

Chatino

- Starting with a text in practical orthography used as a “transcription,” the initial speech corpus has been created
- Initial speech corpus has been manually annotated/time-aligned (cut-and-paste method)
- Initial annotation contains also POS tags and translation.
- A pronunciation dictionary has been created.
- A simplified pronunciation dictionary has been created omitting tone information (improving type-token ratio)
- Tools like ELAN2Split
- Automatic Time Aligner trained on the initial annotations
- Automatic Time Aligner to facilitate creation of more annotated material to ultimately train an initial speech recognition module.

Other languages

Speech Corpora:

- Croatian (Dalmatia), Yiddish, Korean, Baharlu (Turkish, Iran), Egyptian Arabic, German, Burmese, Polish, …
- Transcriptions and annotations of existing Texas German recordings (in cooperation with Hans Boas from UT Austin)

Language Resources

Data and Software Resources:

- ELAN2Split, scripts and environments, for creation of speech corpora for ASRs, Forced Alignment training
- Free Linguistic Environment (FLE)
  - Grammar Engineering Platform for morphologies, syntax, semantics using LFG and related frameworks
- Morphologies (Finite State Transducers) for various languages
  - including extensions of previous projects by other colleagues
- Video and Audio tagging (multi-modal approach)

Technologies

Transcription Bottleneck

- Forced Alignment
- Speech Recognition

Issues

- Limited scope of technologies
- Cost-factor

Technologies

Language Resources Bottleneck

- Lack of corpora, data sets
- Lack of technologies (speech, language)
- Creation of new resources is impeded and expensive due to a lack of bootstrapping technologies

Our Focus

- NLP technologies and processes for:
- Transcription Acceleration Project (TAP) (coined by Nick Evans, ANU, in a joint meeting)
- Grammar Engineering (hybrid Deep Linguistics and Deep Learning) platforms
- New approaches to speech processing: Deep Learning, KALDI, …
- DOIs, ISLRN for data and language models
- Semantic Web infrastructure (RDF, OWL, grammar models)
- Make resources available on sites like Kaggle, organizing Datathon and Hackathon on language data and data science approaches

Next steps

Workshops with a Ken Hale link

- Documentation in the 21st century
- Technology for language data

In cooperation with ANU, Jane Simpson and Nick Evans, and others.