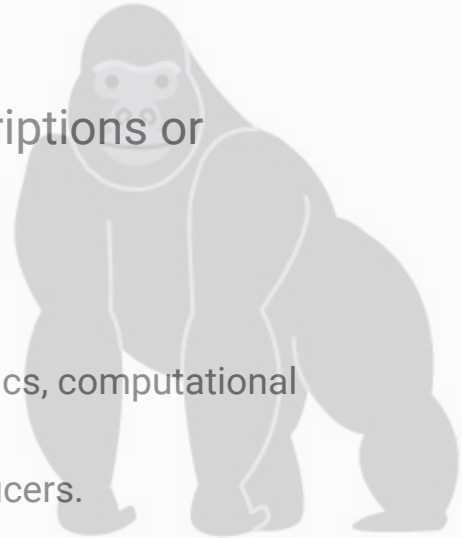
Concept and work behind GORILLA

Damir Cavar
Indiana University

Global Open Resources and Information for Language and Linguistic Analysis

- Emerged from AARDVARC (NSF #1244713)
- Audio and Video language resources and missing transcriptions or annotations.
- Obvious need:
  - Tools and technologies for transcription, annotation, processing.
  - Bridging the worlds of language documentation, theoretical linguistics, computational linguistics and natural language processing.
  - Layer between archives and linguistic community or resource producers.

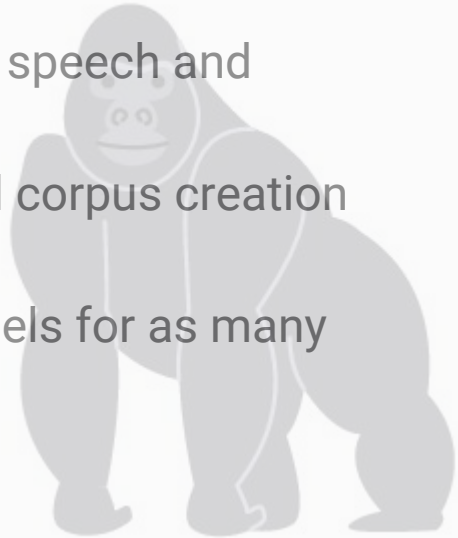


GORILLA

- Cooperation with the Archive of Traditional Music (ATM) at Indiana University
  - Indiana University's Media Digitization Preservation Initiative
    - https://youtu.be/8r4e0xWxrO8
    - Window of opportunity 10 to 15 years
    - Included for example: collection by Charles Voegelin (ATM founded shortly after his hire as the first professor of Anthropology at IU)
  - Using archive (e.g. Hydra, Fedora) and IU IT infrastructure

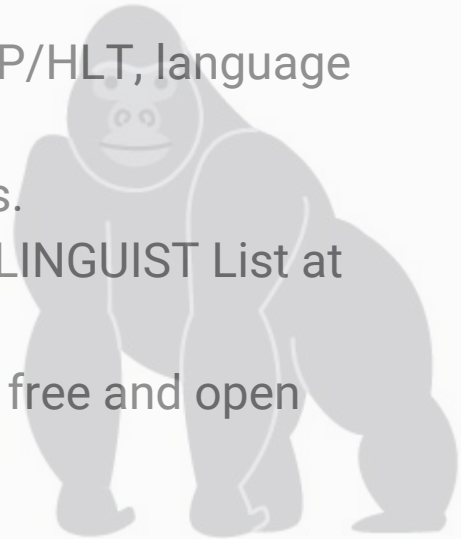
GORILLA

- Link between digital language archives and resources for speech and language technologies (e.g. NLP, HLT).
- Facilitate transcription, linguistic analysis, annotation and corpus creation for documentary linguistic data.
- Aggregate and disseminate language resources and models for as many low-resourced languages as possible.



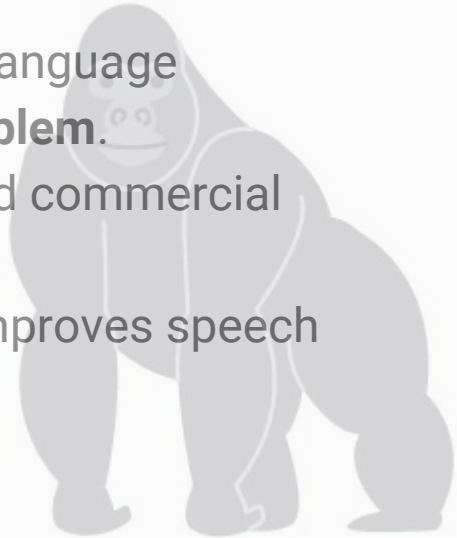
GORILLA

- Facilitate synergies between documentary linguistics, NLP/HLT, language archives, and speaker communities.
- Establish Linked Open Data resources and infrastructures.
- A cooperation between Archive of Traditional Music and LINGUIST List at Indiana University.
- Providing an archival service and infrastructure based on free and open standards.



GORILLA

- Work on a solution for documentary efforts: speech and language technologies to address the **transcription bottleneck problem**.
- Transcribed data can be used for research, education, and commercial goals.
- Development of language data impacts and potentially improves speech and language technologies.



Infrastructure experiments

- UIMA-based linguistic component pipelines, e.g.
  - TEI XML processing and linguistic annotation
  - Phonetic transcription
  - FST-based morphological analyses (e.g. XFST, Foma)
- Web-based integrated standalone applications
  - JavaScript-based complex linguistic apps
  - Data viewers

Access and Licenses

- Observations:
  - For the development of language resources (as opposed to data) CL and NLP need full access to the data.
  - Removing obstacles by:
    - Access to portions of data
    - Creating own data for target languages
    - Sharing all resources (e.g. language resources, models, technologies) freely (not just openly)
  - Licenses for GORILLA (not ATM):
    - CC BY-SA and Apache 2.0

Access and Licenses

- Licensing:
  - Individual licenses not feasible, uniform CC BY-SA/Apache 2.0
  - Donation agreements (e.g. voice, content)
- Full access to all data components of a resource:
  - Audio
  - Transcription
  - Linguistic annotation

Infrastructure

- Independent of ATM/IU IT
  - Virtual server instances (Linux-based platform, Django-based web applications)
  - Existing OLAC connection via LINGUIST List
  - DOI-infrastructure (initially own handle service, DOI-service provided by Purdue in cooperation with IU)

Infrastructure

- CLARIN connectivity
  - CMDI
  - OAI-PMH
  - DOI (Handle, and also ISLRN, see ELRA, LDC initiative)
  - Shibboleth single-sign on
  - WebLicht
- Linked Linguistic Open Data (http://linguistic-lod.org/)
  - Linked metadata
  - RDF of language data and models

Infrastructure

- Current architecture:
  - Python 3 and Django using HTML 5 and various kinds of JavaScript
  - PostgreSQL
  - URL resolution:
    - http://gorilla.../deu/

Resources

- Corpora
  - Audio, transcription, PoS-tagging, translation
  - Chatino, Croatian, Khorasani Turkic, Yiddish, etc.
  - Partially from free conversation, most recorded speech
  - Types:
    - Time-aligned speech
    - PoS-tagged corpora
    - Parallel text
  - Goals:
    - Treebanks, NLP-components

Operation

- LINGUIST List environment
- Internship program
- Individual research projects

- Service open for cooperation