Demo of the Linguistic Field Data Management and Analysis System - LiFE

Siddharth Singh, Ritesh Kumar, Shyam Ratan, Sonal Sinha
Department of Linguistics, Dr. Bhimrao Ambedkar University, Agra
sidd435@gmail.com   ritesh78_llh@jnu.ac.in
shyamratan2907@gmail.com   sonalsinha2612@gmail.com

Abstract

In the proposed demo, we will present a new software - Linguistic Field Data Management and Analysis System - LiFE - an open-source, web-based linguistic data management and analysis application that allows for systematic storage, management, sharing and usage of linguistic data collected from the field. The application allows users to store lexical items, sentences, paragraphs, audio-visual content including photographs, video clips, speech recordings, etc, along with rich glossing / annotation; generate interactive and print dictionaries; and also train and use natural language processing tools and models for various purposes using this data. Since its a web-based application, it also allows for seamless collaboration among multiple persons and sharing the data, models, etc with each other.

The system uses the Python-based Flask framework and MongoDB (as database) in the backend and HTML, CSS and Javascript at the frontend. The interface allows creation of multiple projects that could be shared with the other users. At the backend, the application stores the data in RDF format so as to allow its release as Linked Data over the web using semantic web technologies - as of now it makes use of the OntoLex-Lemon for storing the lexical data and Ligt for storing the interlinear glossed text and then internally linking it to the other linked lexicons and databases such as DBpedia and WordNet. Furthermore it provides support for training the NLP systems using scikit-learn and HuggingFace Transformers libraries as well as make use of any model trained using these libraries - while the user interface itself provides limited options for tuning the system, an externally-trained model could be easily incorporated within the application; similarly the dataset itself could be easily exported into a standard machine-readable format like JSON or CSV that could be consumed by other programs and pipelines. The system is built as an online platform; however since we are making the source code available, it could be installed by users on their internal / personal servers as well.

1 Introduction

Linguistic data management and analysis tools have always been a requirement of field linguists. A huge amount of data is collected and analysed by field linguists for a large number of languages including relatively lesser-known, minoritised and endangered languages of the world and these need to be properly stored, analysed and made accessible to the larger community. On the other hand, there are a huge number of languages across the globe (including the kinds mentioned above), whose data is not available for building any kind of language technology tools and applications. In order to tackle this multi-faceted problem of storing, processing, retrieving and analysing the primary linguistic data, an integrated system with an easily-accessible and user-friendly interface aimed at linguists needs to be made available. "LiFE" is developed with the intent of providing a practical intervention in the field by making available an organised framework for management, analysis, sharing (as linked data) and processing of primary linguistic field data including development of digital and print lexicons, sketch grammars and fundamental language processing tools such as part-of-speech tagger and morphological analysers. The software provides an easy-to-use, intuitive interface for performing all the tasks and there is an emphasis on automating the tasks as far as possible. For example, given some initial input, the system incrementally trains automated methods for inter-linear glossing of the dataset (which improves as more data is stored in the system) and subsequent generation of sketch grammar as well as NLP tools for the language. Similarly, the system automatically in-

fers and links the entries in the lexicon and interlinear glossed data using Lemon (more specifically OntoLex-Lemon) (McCrae et al., 2017) and Ligt (Chiarcos and Ionov, 2019).

2 Motivation and Features

Linguistic field data storage, management, sharing and linked data generation has largely developed independent of each other. As such while there are quite a few tools and applications aimed at field linguists (or community members interested in fieldwork for their own language) for collection and management of data as well as generating lexicon, such as FieldWorks Language Explorer (FLEx)[1] (Butler and Volkinburg, 2007) (Manson, 2020); Toolbox[2] (Robinson et al., 2007); Lexique-Pro[3] (Guérin and Lacrampe, 2007); WeSay[4] (Perlin, 2012) (Albright and Hatton, 2008) and a few other platforms for archiving and providing access to the data, the prominent ones being Endangered Languages Archive (ELAR)[5] (Nathan, 2010); The Language Archive (TLA)[6] (Cho, 2012); SIL Language and Culture Archive[7], etc. The Open Language Archives Community (OLAC)[8], which is a consortium of over 60 participating linguistic archives of various kinds (including the ones mentioned above and others for storage and access of linguistic data, especially of endangered languages) has also recently joined the Linguistic Linked Data Open Cloud which paves the way for providing a large amount of such data as linked data (Simons and Bird, 2003). However none of the tools and platforms directly provide an interface for storing or (largely) automatically generating the primary linguistic data as linked data or provide a seamless two-way between the NLP tools and libraries and linguistic data management softwares.

On the other hand, the linked data community has developed tools for supporting generation of linked data, especially linked data lexicons. One of the best-known tools for this is **VocBench (VB)**, which is a fully-fledged open-source web-based thesaurus management platform with the feature of collaborative development of multilingual datasets

compatible with semantic Web standards. It provides the facilities of generating lexicons, thesauri, and linked data ontologies to the large organisations, companies, and user communities (Stellato et al., 2020). However tools like these focus on generating Linked Data which is generally not very user-friendly for field linguists nor do they provide options for automating the tasks or linking to the NLP ecosystem.

The primary motivation for building this platform is to provide a tool that acts as a bridge between field linguists (who are primarily engaged in data collection from low-resource and endangered languages, building lexicons, writing grammatical descriptions and also producing educational and other kinds of materials for the communities that they work with), linked data community (who are primarily engaged in meaningfully connecting data from different languages and resources using the semantic web techniques) and the NLP community (who primarily makes use of the linguistic data from multiple languages; could potentially provide support in automating the tasks carried out by field linguists; and also provide tools and technologies for the marginalised and under-privileged linguistic communities). As such in its current state the app provides the following functionalities -

- It provides a user-friendly interface for storing, sharing and making publicly available the linguistic field data including interlinear glossed text, lexicon and associated multimedia content.

- It provides reasonable automation for tasks such as generating lexicon, sketch grammar, etc by providing interfaces for training as well as using pre-trained NLP models needed for automating various tasks. The tool currently supports training various algorithms of the scikit-learn and HuggingFace Transformers library as well as using the models trained using these libraries.

- It provides interface for exporting the data in structured formats such as RDF, JSON and CSV that could be directly used for NLP experiments and modelling.

During the demo we will present these features and the interface of the tool in detail and also briefly train the participants in using it.

---

[1]https://software.sil.org/fieldworks/
[2]https://software.sil.org/shoebox/,
https://software.sil.org/toolbox/
[3]https://software.sil.org/lexiquepro/
[4]https://software.sil.org/wesay/
[5]https://www.elararchive.org/
[6]https://archive.mpi.nl/tla/
[7]https://www.sil.org/resources/language-culture-archives
[8]http://www.language-archives.org/archives

3 Presenters

The demo will be given by the developers of this application which include the following -

1. **Ritesh Kumar** is Assistant Professor of Linguistics and coordinator of the masters program in computational linguistics at Dr. Bhimrao Ambedkar University, Agra. he is working in the field of computational linguistics and language documentation and description for over last 10 years. He has conceptualised, mentored and co-developed this app.

2. **Siddharth Singh** is a software engineer and is currently pursuing his MSc in Computational Linguistics from Dr. Bhimrao Ambedkar University. He is the principal developer of the app,

3. **Shyam Ratan** is pursuing his Mphil in Computational Linguistics and is a co-developer of the app.

4. **Sonal Sinha** is pursuing her Mphil in Computational Linguistics and is a co-developer of the app.

References

Eric Albright and John Hatton. 2008. Wesay, a tool for collaborating on dictionaries with non-linguists. *Documenting and revitalizing Austronesian languages*, 6:189 – 201.

Lynnika Butler and Heather Volkinburg. 2007. Review of fieldworks language explorer (flex). *Language Documentation Conservation*, 1.

Christian Chiarcos and Maxim Ionov. 2019. Ligt: An llod-native vocabulary for representing interlinear glossed text as RDF. In *2nd Conference on Language, Data and Knowledge, LDK 2019, May 20-23, 2019, Leipzig, Germany*, volume 70 of *OASICS*, pages 3:1–3:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Julia Cho. 2012. *The Language Archive*. Dramatists Play Service.

Valérie Guérin and Sébastien Lacrampe. 2007. Lexique pro. *Language Documentation Conservation*, 1(2):293 – 300.

Ken Manson. 2020. Fieldworks linguistic explorer (flex) training 2020 (ver 1.1 august 2020).

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: Development and applications. Brno. Lexical Computing CZ s.r.o.

David Nathan. 2010. Archives 2.0 for endangered languages: From disk space to myspace. *International Journal of Humanities and Arts Computing*, 4:111–124.

Ross Perlin. 2012. Wesay, a tool for collaborating on dictionaries with non-linguists. *Language Documentation & Conservation*, 6:181 – 186.

Stuart Robinson, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with toolbox and the natural language toolkit. *Language Documentation Conservation*, 1.

Gary Simons and Steven Bird. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Computing Research Repository - CORR*, 18:117–128.

Armando Stellato, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, Anikó Gerencsér, Anne Waniart, Eugeniu Costetchi, and Johannes Keizer. 2020. Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11:1–27.