Proceedings of the EURALI Workshop @LREC2020, pages 90–94
Marseille, 20 June 2022

Towards a Unified Tool for the Management of Data and Technologies in Field Linguistics and Computational Linguistics - LiFE

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

Abstract

The paper presents a new software - Linguistic Field Data Management and Analysis System - LiFE for endangered and low-resourced languages - an open-source, web-based linguistic data analysis and management application allowing systematic storage, management, usage and sharing of linguistic data collected from the field. The application enables users to store lexical items, sentences, paragraphs, audio-visual content including photographs, video clips, speech recordings, etc, with rich glossing and annotation. For field linguists, it provides facilities to generate interactive and print dictionaries; for NLP practitioners, it provides the data storage and representation in standard formats such as RDF, JSON and CSV. The tool provides a one-click interface to train NLP models for various tasks using the data stored in the system and then use it for assistance in further storage of the data (especially for the field linguists). At the same time, the tool also provides the facility of using the models trained outside of the tool for data storage, transcription, annotation and other tasks. The web-based application, allows for seamless collaboration among multiple persons and sharing the data, models, etc with each other.

Keywords: LiFE, Web-based, Linguistic Data Management, Linked Data, NLP Interface

1. Introduction

Linguistic data analysis and management tools are always being required by field linguists. A large amount of data is collected, and needs to be properly stored, analysed and made accessible to the larger community by field linguists for a large number of languages including relatively lesser-known, minoritized and endangered languages of the world. On the other hand, hardly any dataset is publicly available for building any kind of language technology tools and applications for a huge number of languages across the globe.

An integrated system with an easily-accessible and user-friendly interface aimed at linguists needs to be made available, to tackle this multi-faceted problem of storing, processing, analysing and retrieving the primary linguistic data. "LiFE"[1] intends to provide a practical intervention in the field through an organised framework for management, analysis, sharing (as linked data) and processing of primary linguistic field data including digital and print lexicons, sketch grammars and fundamental language processing tool development, such as part-of-speech tagger and morphological analysers. The software aims to provide an easy-to-use, intuitive interface for performing all the tasks and emphasise on automating the tasks as far as possible. For example, 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, by providing initial input. Likewise, the system automatically links and infers the entries in the lexicon and inter-linear glossed data using Lemon (more specifically OntoLex-Lemon) (McCrae et al., 2017) and Ligt (Chiarcos and Ionov, 2019). We have also integrated the recent transformer-based unsupervised and transfer learning-based ASR models (such as wav2vec 2.0 (Baevski et al., 2020) and wav2vec-U (Baevski et al., 2021)) which provides the whole automated pipeline from transcription to inter-linear glossing and free translation for the field linguists. At the same time the data itself could be used for improving the models for ASR, part-of-speech tagging, morphological analysers and machine translation. In addition to these, the system also enables storage and semi-automatic linking of the dataset to some of the largest linked data sources such as Wikipedia and DBpedia.

2. Motivation

The development of linguistic field data storage, sharing, management and linked data generation has been largely done independent of each other. There are some applications and tools aimed at field linguists (or community members interested in fieldwork for their own language) for management and collection of data as well as generating lexicon. One of the most popular tools in the field is FieldWorks Language Explorer (*FLEx*)[2] which is used for the collection, management, analysis and sharing of linguistic and cultural data (Butler and Volkinburg, 2007), (Manson, 2020). *Toolbox*, earlier called Shoebox[3] is a precursor to the FLEx and one of the the oldest softwares produced by

[1] https://github.com/kmi-linguistics/life

[2] https://software.sil.org/fieldworks/

[3] https://software.sil.org/shoebox/ ,https://software.sil.org/toolbox/

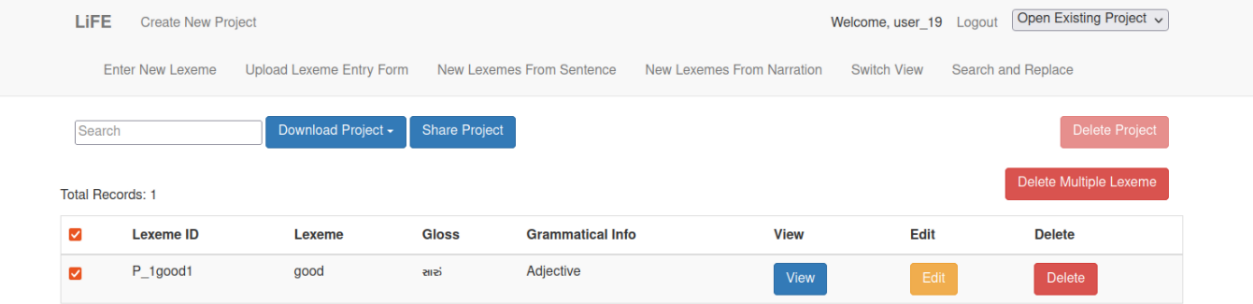

Figure 1: Dictionary View Interface of LiFE

SIL (The Summer Institute of Linguistics) that was essentially meant for anthropologists and field linguists to put their text data in the tool and build dictionaries (Robinson et al., 2007). *LexiquePro*[4] is a software for creating / formatting lexicon databases and easy sharing with others (Guérin and Lacrampe, 2007). *WeSay*[5] is created by SIL for providing support to the non-linguists/native speakers in building dictionaries for their own languages (Perlin, 2012), (Albright and Hatton, 2008).

There have been some efforts at development of tools focussed on data collection as well. (Vries et al., 2014) discusses the development of a tool named *Woefzela*[6], which is a smartphone-based (Android Operating System) data collection tool for speech data collection. It can function without Internet connectivity and allows multiple sessions for data collection. It works well for the quality control of collected data. This tool is demonstrated in the South African data collection project, where almost 800 hours of speech data were collected from remote and rural areas.

There are few other platforms for archiving and providing access to the data, the prominent ones being Endangered Languages Archive (*ELAR*)[7], which is a digital repository for preserving and circulating documentation of endangered languages (Nathan, 2010). The Language Archive (*TLA*)[8] is a hub for language resources that holds language corpus in audio, video format, along with preserving and documenting the dying languages (Drude et al., 2012). *SIL Language and Culture Archive*[9] contains works collected, compiled and created by SIL. The *Open Language Archives Community (OLAC)*[10], which is an association of more than 60 participating linguistic archives of different kinds (including the ones mentioned above and others for access and storage of linguistic data, specifically of endangered languages) has also newly joined the Linguistic Linked Data Open Cloud which covers the way for providing a large amount of such data as linked data (Simons and Bird, 2003).

However, none of the platforms and tools directly provide an interface for storing or (largely) automatically generating the primary linguistic data as linked data or provide a flawless two-way integration between the linguistic data management softwares and NLP libraries and tools. Most of these tools aimed at field linguists do not provide interfaces for generating linked data or even utilising the modern NLP models for automating the tasks

On the other hand, for supporting generation of linked data, the linked data community has developed tools for generating linked data lexicons. One of the renowned tools for this is *VocBench* (VB)[11]. It is a full-fledged open-source web-based thesaurus management platform, which provides feature of collaborative development of multilingual datasets compatible with semantic web standards (Stellato et al., 2020). In addition to these there have been quite a few attempts at transporting the non-linked datasets to the linked data repositories. These efforts are largely carried out manually and end up in producing high-quality linked data. For example, (Samarin, 1967) talked about the lexical data migration from textual e-dictionaries to lexical databases. Earlier Serbian Morphological Dictionaries (SMD) were developed in *LeXimir*, an application for the development and management of lexical resources. Now, a new lexical database model for the SMD is based on the lemon model with a thesaurus. This database improves the existing resources.

While work like these are tremendous efforts, these may not be scalable for a large number of cases. Moreover, given the fact that these efforts are extremely resource-intensive, it may not be at all feasible for endangered and low-resource languages. Hence it is a better option to create the new resources itself as linked data instead of later converting those to linked data. On the other hand, a tool like VocBench which focuses on

[4] https://software.sil.org/lexiquepro/
[5] https://software.sil.org/wesay/
[6] https://sites.google.com/site/woefzela/
[7] https://www.elararchive.org/
[8] https://archive.mpi.nl/tla/
[9] https://www.sil.org/resources/language-culture-archives
[10] http://www.language-archives.org/archives
[11] http://vocbench.uniroma2.it/

creating news resources as linked data, is not very user-friendly for field linguists nor do they provide options for automating the tasks or linking to the NLP ecosystem.
Given this general unavailability of common interfaces and tools that could act as a bridge between the three group of researchers working with linguistic data - field and documentary linguists, linguistic linked data community and NLP practitioners - and the linguistic community itself, a communication among these groups is almost non-existent. Our aims are, thus, to provide the following -

- Provide an interface for Field and Documentary Linguists such that it not only gives a user-friendly interface for putting their data in a structured format but also provide access to the state-of-the-art NLP for use without the need to navigate through complex instructions and workflow of most NLP tools.
- Access to the data from endangered and low-resource languages (if the community and researchers choose to make it available) in a structured format for NLP practitioners. Also an interface for training and testing the model on this data via the interface.
- A (semi-)automated method of linking the data to some of the largest linked data databases.

3. Features of System

As mentioned above, the central motive of building this platform is to provide a tool that acts as a bridge between field linguists (who are chiefly engaged in data collection from poor-resource and endangered languages, writing grammatical descriptions, building lexicons and also producing educational and other kinds of stuff for the communities that they work with), linked data community (who are chiefly engaged in resources using the semantic web techniques and meaningfully connecting data from different languages.) and the NLP community (who chiefly makes use of the linguistic data from many languages; could certainly contribute 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 present state the app provides the following operations -

- It provides a user-friendly interface for storing, making and sharing publicly available the linguistic field data including lexicon, interlinear glossed text and associated multimedia content.
- It provides a pipeline for automatic extraction of text and its POS tags using the unsupervised (using wav2vec-U) and transfer learning methods (using wav2vec2.0). It provides interfaces for training as well as using pre-trained NLP models needed for automating these tasks of ASR and POS tagging. The tool presently supports training various algorithms of the HuggingFace Transformers library and scikit-learn as well as using the models trained using these libraries.
- It provides an interface for exporting the data in structured formats such as RDF, JSON, HTML, XML, LATEX and CSV that could be directly used for NLP experiments and modelling (Singh et al., 2022).
- It generates linked data for dictionaries and interlinear glossed text using vocabs like LiGT and OntoLex-LEMON and then internally linked to the other linked lexicons and databases such as DBpedia and WordNet - this will help in the other tasks as well.

4. System Mechanism

This section contains information related to the architecture of the tool. The tool uses the Python-based Flask[12] framework and MongoDB (as database) in the backend and HTML, CSS and Javascript at the frontend.
There are six collections in the database of the tool:

- projects : containing a list of all the projects in the system,
- userprojects: containing projects created by and shared with each user and current active project of the user,
- projectsform : contains lexeme details form created for the project,
- lexemes : collection storing the details about each lexeme, This is stored as linked data entries using the relevant vocabularies.
- fs.files stores the file's metadata and fs.chunks stores the binary chunks of files (image, video, audio, etc).

There is a login interface where a registered user can login to the application and new user can register. Then there is a navigation bar leading to the interface to create new project; alternatively the user may select an existing project to work from a dropdown list as displayed in Figure 1.
The option to create new project will lead to the form for creating one's the fields required for the current project as shown in Figure 2. This makes the interface of the tool completely customisable which could be designed as per the need of the given project.

[12] https://flask.palletsprojects.com/en/2.0.x/

Once the fields are created then the user could use the relevant buttons to enter 'New lexeme' or 'New Sentence'. One could fill up the form with the required details to complete an entry as shown in Figure 3. The entry made for a specific lexeme through this customised form will be visible in the dictionary view as shown in Figure 1. Dictionary view contains a view button to view details of a particular lexeme and an edit button to edit the details. It has also two delete buttons to delete single or multiple lexemes.

The share button on the user's dashboard as well as in the lexicon view interface provides a multi-level option to share the project with other users - the users will have full control over which parts of the project could be shared and what kinds of access rights the sharee have. These access rights include such fine-grained classification as viewing rights for specific entries or all entries, editing rights for the entries, deleting rights for the entries, right to add new entries, share it with other users (with equivalent or lower rights), etc.

Finally, one can download the complete project in JSON format along with the files uploaded for the project and can share that with others. The lexicon could be downloaded in various formats such as RDF, CSV, HTML, PDF, DOCX, XLSX, etc.

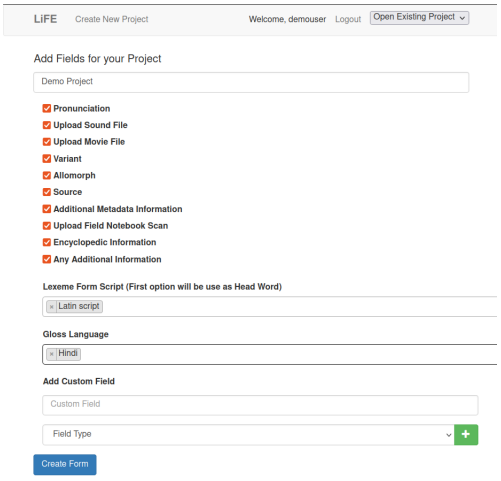

Figure 2: Create New Project Form

Figure 4 shows the model diagram of "LiFE", showing all the functions available to a user .

5. The Way Ahead

The platform is currently under active development and some of the features in the pipeline include the following

- Allow searching across multiple languages and generating concordances / parallel entries from multiple languages - options to search by language families, regions, and other available information.
- Allow for automatically generating glosses, example sentences, etc from different languages (especially those belonging to same language family / closely related), when working on a new dictionary - this will make the dictionary-making quicker. Also linked data could be used for doing this.

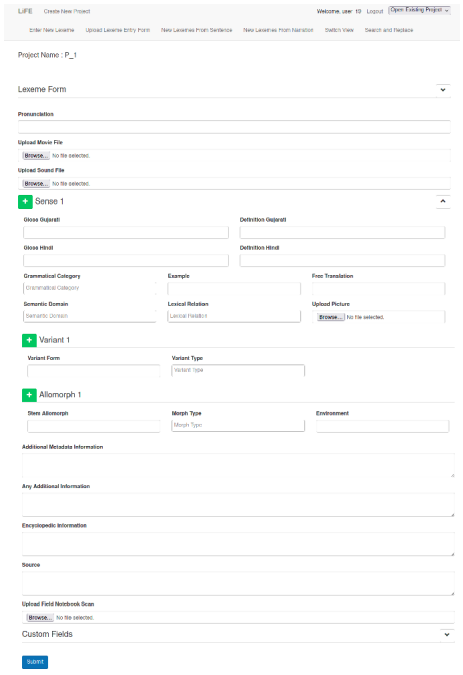

Figure 3: Lexeme Entry Form

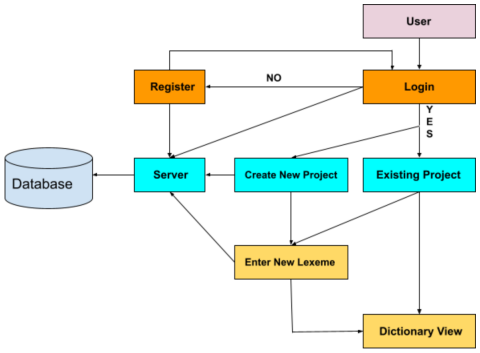

Figure 4: Model Diagram of LiFE

- Interface for training and automating morph analysers, parsers and machine translation system. This will make the whole pipeline after uploading of the speech data automated.
- Automatic sketch grammar generation.
- Template for Android app, which could generate mobile apps for dictionaries automatically (given the database).

6. Bibliographical References

Albright, E. and Hatton, J. (2008). Wesay, a tool for collaborating on dictionaries with non-linguists.

Documenting and revitalizing Austronesian languages, 6:189 – 201, 12.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Baevski, A., Hsu, W., Conneau, A., and Auli, M. (2021). Unsupervised speech recognition. *CoRR*, abs/2105.11084.

Butler, L. and Volkinburg, H. (2007). Review of fieldworks language explorer (flex). *Language Documentation and Conservation*, 1, 06.

Chiarcos, C. and Ionov, M. (2019). Ligt: An llod-native vocabulary for representing interlinear glossed text as RDF. In Maria Eskevich, et al., editors, *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.

Drude, S., Broeder, D., Trilsbeek, P., and Wittenburg, P. (2012). The language archive — a new hub for language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3264–3267, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Guérin, V. and Lacrampe, S. (2007). Lexique pro. *Language Documentation and Conservation*, 1(2):293 – 300, 12.

Manson, K. (2020). Fieldworks linguistic explorer (flex) training 2020 (ver 1.1 august 2020). 08.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: Development and applications. Brno. Lexical Computing CZ s.r.o.

Nathan, D. (2010). Archives 2.0 for endangered languages: From disk space to myspace. *International Journal of Humanities and Arts Computing*, 4:111–124, 10.

Perlin, R. (2012). Wesay, a tool for collaborating on dictionaries with non-linguists. *Language Documentation & Conservation*, 6:181 – 186, 12.

Robinson, S., Aumann, G., and Bird, S. (2007). Managing fieldwork data with toolbox and the natural language toolkit. *Language Documentation and Conservation*, 1, 06.

Samarin, W. (1967). *Field Linguistics. A guide to Linguistic Field Work*. Holt, Rinehart and Winston., New York, NY.

Simons, G. and Bird, S. (2003). The open language archives community: An infrastructure for distributed archiving of language resources. *Computing Research Repository - CORR*, 18:117–128, 06.

Singh, S., Kumar, R., Ratan, S., and Sinha, S. (2022). Demo of the linguistic field data management and analysis system – life.

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., and Keizer, J. (2020). Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11:1–27, 05.

Vries, N., Davel, M., Badenhorst, J., Basson, W., Barnard, E., and de Waal, A. (2014). A smartphone-based asr data collection tool for under-resourced languages. *Speech Communication*, 56:119–131, 01.