A data repository for the management of dynamic linguistic datasets

Thomas Gaillat, Leonardo Contreras Roa, Juvénal Attoumbre

A data repository for the management of dynamic linguistic datasets

Thomas Gaillat
LIDILE - EA 3874
University of Rennes, France
thomas.gaillat@univ-rennes2.fr

Leonardo Contreras Roa
LIDILE - EA 3874
University of Rennes, France
lcontrerasroa@gmail.com

Juvénal Attoumbre
LIDILE - EA 3874
University of Rennes, France
juvenalak@gmail.com

Abstract

This paper addresses the issue of using Nakala, a dynamic database technology, for the management of language corpora. We present our ongoing attempt at storing and classifying multimedia documents of a corpus of language learner oral and written productions with universal resource identifiers. The architecture supports query APIs compatible with R packages and other tools which will facilitate the generation of linguistically enriched datasets for a more effective corpus-based study of language acquisition.

1 Introduction

For several decades, many corpora have been made public. The open-science initiatives which insist on the importance of making data Findable, Accessible, Interoperable and Reusable (*FAIR* (Wilkinson et al., 2016)) further reinforce this trend. However, even if it is important to make corpora public, it should not be the sole objective of the move towards more versatile and flexible data. It is equally important to think about the way corpora are made accessible. Current trends show that most shared corpora are static (Sérasset et al., 2009) in the sense that once downloaded, the data are not updated and run the risk of obsolescence.

When research experiments are conducted on external data, the first stage usually consists in data download prior to pre-processing (curation, part-of-speech (POS) tagging, etc.) Subsequent stages depend on the initial data and all pre-processing and analytical tasks rely on the same data as initially downloaded. However, if, in the meantime, corpus authors have modified their data by updating or correcting observations, experiments automatically rely on out-of-date data. The problem is to find a method to include the most up-to-date corpora as part of the pre-processing stage of experiments.

One way to approach the problem could consist in designing a data architecture ensuring a seamless workflow from the data ingestion to the querying stage. In this paper we present a use case based on the exploitation of a learner corpus stored in a database.

2 Current corpus sharing practices

Over the last few decades many corpora have been developed and a gradual shift towards public access has been observed. In the last decade, different platforms have been set up to make corpora available to the research community (ELRC, META-SHARE, among others). CLARIN illustrates the case as it includes access to data repositories. These repositories provide resources such as corpora which rely on persistent URLs. For instance, Huma-Num Ortolang (Pierrel, 2014), the French platform for language resources, includes a number of resources which are downloadable thanks to such URLs. As far as we know, most corpora are available under single URLs. Other infrastructures also exist such as ELRA's[1] catalogue. As of 2013, the repository provides open access to two sets of non-commercial language resources [2] classified according to a number of criteria that form metadata. As accessible as they are,

[1] The European Language and Resources Association

[2] http://portal.elda.org/en/catalogues/free-resources/

these resources are provided at corpus level rather than observed subject level. It is not possible to update or retrieve specific items from the datasets.

In parallel, the European Language Equality consortium (Rehm et al., 2020) has launched a project to establish the European Language Grid in which language resources of many types will be interconnected. Data sets, tools, models will be interoperable to allow the construction of natural language processing (NLP) pipelines dedicated to specific language tasks. There is a clear move towards an integrated set of resources and services. To support this infrastructure, one essential point is to make corpus items accessible in their up-to-date version. Persistent data sources will support this process by making corpus items queriable and retrievable.

However, the current type of corpus distribution is done mostly via platforms that provide static recordings of data in the sense that, even if the data is updated in the back end, the downloadable version remains in its initial state until a new static version is ready. We propose a dynamic solution to store and query a corpus whereby the corpus architecture gives controlled access to corpus items via APIs, while a robust online repository – in our case the Nakala database (Huma-Num, 2021)– ensures long-term storage and allows constant updates and maintenance.

The data workflow implemented by Nakala is illustrated in Figure 1 below. This workflow is being tested for the upload of our own multimedia corpus of the written and spoken productions of learners of English and French as foreign languages.

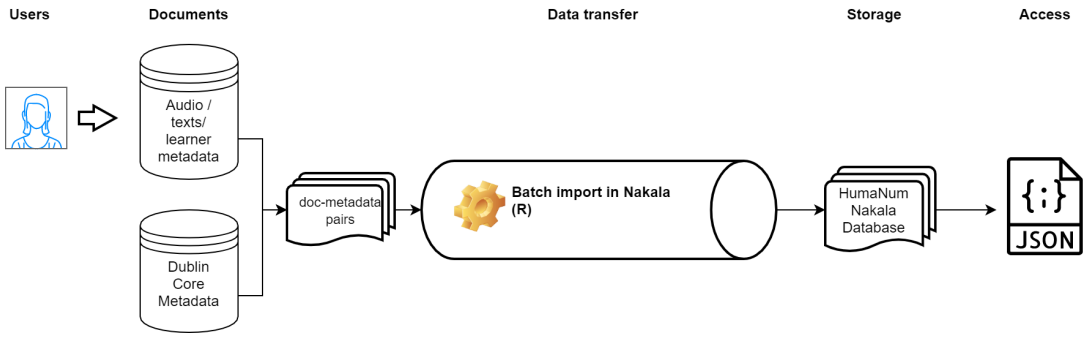

Figure 1: Nakala data workflow

3 Use case: A data repository for a learner corpus

3.1 Storing a learner corpus in a data repository

Making up-to-date corpus data constantly available implies the use of database technologies. This is because, as opposed to static file collections, database management systems allow queries both for augmenting and retrieving data.

We are currently developing a dynamic database management system for the CIL corpus (*Corpus InterLangue*), a collection of recordings and written productions of learners of English and French as a foreign language (L2). The corpus has been collected for the past fifteen years by Master students of the Linguistics and Didactics program of the University of Rennes 2 in France , following roughly the same data collection protocol (Arbach, 2015, p. 239). The outcome of this protocol is a complex set of learner data, composed of the following files for each of the surveyed learners: 1. A recording of the learner being interviewed by a native speaker of the L2 (.WAV) 2. An orthographic transcription of the interview (.CHA/.EAF[3]) 3. A recording of the learner reading a text in the L2 (.WAV) 4. A page-long handwritten production made by the learner in the L2 (.PDF) 5. A digitised transcription of the handwritten production (.TXT) 6. A sociolinguistic questionnaire filled in by the learner (.PDF/.CSV) 7. A consent form signed by the learner (.PDF).

[3]CHA: A proprietary plain-text format for single-tier text-to-audio alignment generated by the software CLAN.
EAF: An XML-based format for multiple-tier, time-aligned transcriptions generated by the software ELAN.

All of these files are stocked in one sub-directory of the corpus, each sub-directory corresponding to a single speaker. As of April 2021, a total of 115 speakers compose the CIL corpus. Each year, a new generation of recordings is added to the corpus, which demands for a system of dynamically updated storage.

The data can be accessed in two different ways. Nakala provides a web interface through which users can browse the data and its directories. Figure 2 illustrates how the user can play an audio file, access the transcription (.CHA file) or download the written production (.PDF file) of a learner whose ID is *fre_al_tr_99_m_20*. The persistent link specific to the audio file is displayed underneath the audio player and the persistent link for the entirety of the data of that specific learner is available on the top-left corner of the browser. The same data can also be accessed through APIs and queried/filtered. A metadata file (.CSV) is included in each sub-directory to keep track of the number and type of files in the corpus, and to serve as search-word tag information for future queries. Once uploaded, Nakala assigns a persistent URL and a DOI to each file or each collection of corpus items, allowing full corpus querying. However, to date, Nakala does not support global-state versioning. This structure allows authors to feed the corpus and users to access it from the same back-end.

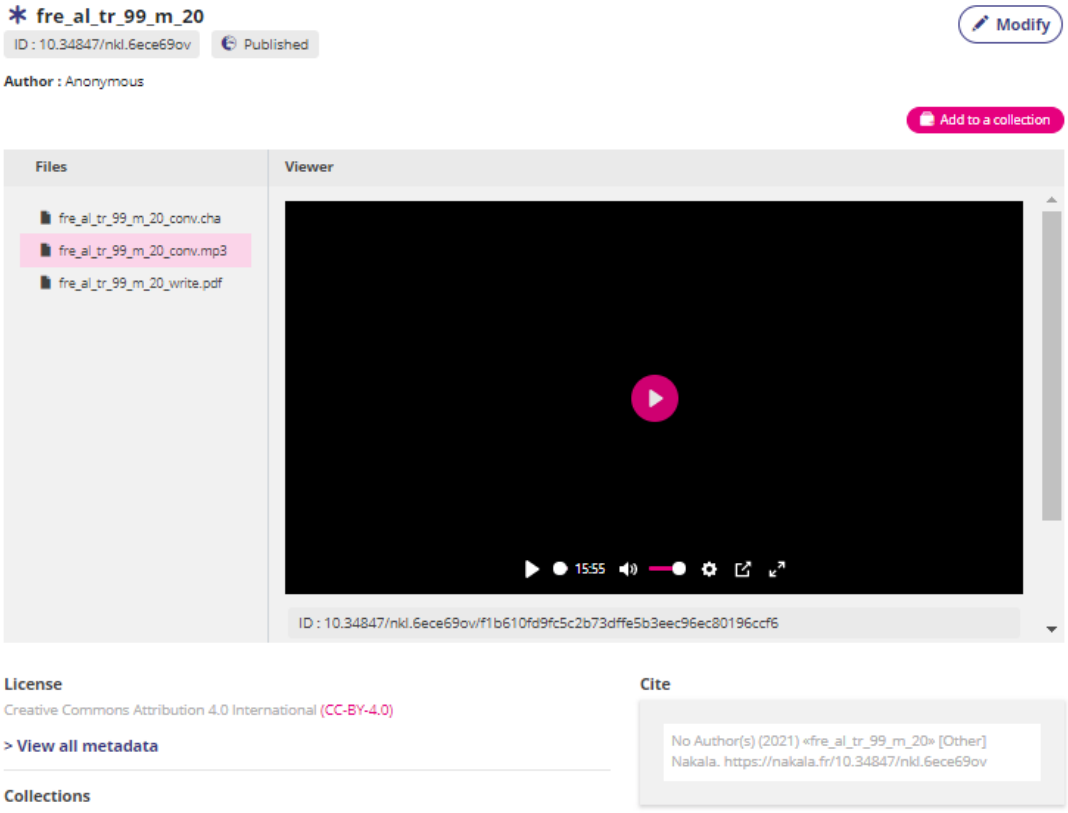

Figure 2: Example of data display on the web layout of the Nakala database

3.2 Generating Linguistic datasets

As well as providing up-to-date data, our purpose is to let researchers select which type of linguistic annotation they want to include in their datasets. Language resource users may not be interested in certain linguistic features while others might be of paramount importance. Here again, flexibility can provide an attractive answer.

We provide modular R programs[4] that make it possible to merge the data ingestion, curation and querying phases with that of linguistic annotation. These programs can be run in an IDE such as R Studio and automate on-demand creation of datasets containing linguistic data as well as learner metadata, i.e.

[4] https://github.com/LIDILE/CIL_query

age of the learner, their level of education, their age of first contact with the L2. Furthermore, in order to analyse aspects of syntax, POS-Tagging adapted to learners can be performed via UDpipe learner model (Wijffels, 2020). Syntactic parsing can also be conducted by UDPipe. In addition, a number of semantic features such as aspect and gender can also be of interest and obtained with this package. By running these scripts, users query Nakala repository via their IDE and generate datasets locally. They can then turn to modelling tasks.

An analogue process could be applied to process oral data from a phonetic/phonological point of view. Automatic text-to-phoneme alignment can be achieved, for example, by attaching the Nakala query pipeline to that of BAS WebMaus (Kisler et al., 2017). Vowel-formant extraction can also be envisaged via the emuR package for speech database management (Winkelmann et al., 2021). Even though the annotation and extraction provided by the two aforementioned methods usually require manual correction, especially if applied to learner speech, they can provide a rich amount of readily available data to automatically obtain a quick first glance at a learner's pronunciation.

4 Discussion and perspectives

This project is a first step in the direction of making a language learner corpus dynamically available. The Human-Num Nakala architecture provides persistent identifiers for single items of the corpus. Researchers can initiate queries via APIs in order to retrieve relevant data and metadata from the corpus.

Further work involves the development of customised R scripts and datasets according to the specific needs of researchers: semantic, morphosyntactic, phonological and phonetic annotations can be performed by branching our query pipeline with specific R scripts and packages aimed at enriching linguistic data. Various state-of-the-art annotation tools may be exploited depending on the research questions. For instance, in the case of morphosyntactic POS tagger comparisons, it could be necessary to use a dataset including tokens which have been POS-tagged with several tools and different tagsets. The resulting enriched linguistic data can be exploited to carry out research on the interlanguage of learners of English and French, which can in turn help develop didactic tools adapted to learners' specific needs. In our use case, we seek to extract the transcripts of interviews conducted with learners of French in order to model their proficiency according to the levels proposed by the Common European Framework of Reference (Gaillat et al., 2021).

Also, it would be beneficial for Nakala to provide a global state versioning functionality, akin to the IDs that Git generates for each user commit. This would allow researchers not only to create datasets relying on the latest version of the corpus collections, but also to query previous versions of the same corpus collections. Such a functionality would support longitudinal comparisons within the same corpora.

Once achieved, the workflow will also be made available in order to allow other systems to extract and enrich linguistic data via Nakala's APIs. With this approach we seek to contribute to a general shift towards corpus interoperability, towards the dynamisation of automatic linguistic annotation of spoken corpora, and ultimately, towards a better understanding of the process of language acquisition.

References

Najib Arbach. 2015. *Constitution d'un corpus oral de FLE : enjeux théoriques et méthodologiques*. Phd thesis, Université Rennes 2.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2021. Predicting CEFR levels in learners of English: the use of microsystem criterial features in a machine learning approach. *ReCALL*.

TGIR Huma-Num, 2021. *Documentation NAKALA*. Documentation des services de la TGIR Huma-Num.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326 – 347.

Jean-Marie Pierrel. 2014. Ortolang. Une infrastructure de mutualisation de ressources linguistiques écrites et orales. *Recherches en didactique des langues et des cultures. Les cahiers de l'Acedle*, 11(11-1). Number: 1 Publisher: Acedle (Association des Chercheurs et Enseignants Didacticiens des Langues Étrangères).

Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajič, Stelios Piperidis, and Andrejs Vasiļjevs, editors. 2020. *Proceedings of the 1st International Workshop on Language Technology Platforms*. European Language Resources Association, Marseille, France.

Gilles Sérasset, Andreas Witt, Ulrich Heid, and Felix Sasaki. 2009. Multilingual language resources and interoperability. *Language Resources and Evaluation*, 43(1):1–14.

Jan Wijffels, 2020. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. R package version 0.8.5.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9. Number: 1 Publisher: Nature Publishing Group.

Raphael Winkelmann, Klaus Jaensch, Steve Cassidy, and Jonathan Harrington, 2021. *emuR: Main Package of the EMU Speech Database Management System*. R package version 2.2.0.