LREC 2012

http://www.lrec-conf.org/lrec2012

LREC 2012 Language Library

- Final Submission -

Hélène Mazo

Published : Wednesday 28 September 2011
Modified : Monday 20 February 2012
Created : Wednesday 27 September 2017

Help for the Language Library

The Language Library is currently available at www.languagelibrary.eu.

Motivation

The Language Library is the new feature of LREC 2012. The rationale behind this initiative is that accumulation of massive amounts of multi-dimensional data about language is the key to foster advancement in our knowledge about language and its mechanisms. The objective is to gather and share part of the linguistic knowledge the field is able to produce, starting a movement aimed at collecting all possible annotations/encodings at all possible levels.

As a first experiment of a community-built repository that allows sharing of multidimensional and multi-level processed/annotated resources, **it needs a small effort from each of you** to put into place new ways of collaboration within the language resources and technology community.

Download the data to be processed

You can download the processable data from the START submission page and you are invited to process the data using the tools you are working on/have available.

For the Written modality, raw data has been chosen from small Wikipedia entries and the Universal Declaration of Human Rights in several languages (providing both comparable and parallel data).

For the Speech modality, data have been provided by ELRA . They consist of brief audio samples of broadcast news, telephone speech etc. They are available for a limited number of languages (here the list of languages).

From the LREC2012 Language Library section of the START Submission page, you are invited to proceed as follows:

1. Confirm your interest in contributing to the Language Library by checking I wish to contribute (it s important that you do this as soon as you know that you are willing to process some data).
2. Select the modality and language(s) you would like to process; alternatively, you can choose to download the full raw data set using the *Download All* button.
3. Download the data: you can download the raw data and upload the processed data until the final paper submission deadline.

Contributing new processable data

Not all languages, neither all modalities, are covered by the LREC 2012 raw data. Please send an email to [E-mail] if you cannot find the language(s) you work on, or if you want to contribute to the Library with other raw (or raw and processed) data to be made available to all.

Important: Data provided must not be copyrighted.

Process and upload the data

Processing and uploading the data can be done for at least 1 month after paper submission deadline:

1. Process/annotate the data using the tools and type of processing/annotation you are working on.
2. For **Written** data: Please **process/annotate plain text files without changing/deleting any part of it**. This is important in order to preserve the integrity and the comparability of the processed/annotated data. Most specifically, if you want to provide stand-off annotation the offsets should refer to the plain texts. Notice that for each Wikipedia entry we provide, in addition to the plain text file which you are asked to process as it is, also the source HTML as reference only. For Speech data: please don t change/split the input audio samples.
3. The only requirement we ask is to keep track of which plain text files were used to produce each output file you are going to upload. Should your tool rename input files, please keep the mapping between old and new names because you will need it when you submit processed data
4. Provide some basic metadata information upon upload (we will recommend some basic metadata).
5. Upload the processed data (of any kind and in any format). **You will receive instructions by email on where and how to submit your processed data** (a mail will be sent to those declaring their willingness to contribute and download some raw data).

Availability to all

All the processed data will be available to everyone before LREC in a special LREC Repository.

Licensing

Wikipedia entries are available under the Creative Commons Attribution-ShareAlike License.

Send us comments

Given the experimental nature of the Library we are very interested in receiving your comments and suggestions to improve it! Please send them to [E-mail].