The Language Library: supporting community effort for collective resource production

Riccardo Del Gratta, Francesca Frontini, Francesco Rubino, Irene Russo, Nicoletta Calzolari

ILC-CNR
Consiglio Nazionale delle Ricerche
Via Moruzzi, 1 - Pisa, Italy
name.surname@ilc.cnr.it

Abstract

Relations among phenomena at different linguistic levels are at the essence of language properties but today we focus mostly on one specific linguistic layer at a time, without (having the possibility of) paying attention to the relations among the different layers. At the same time our efforts are too much scattered without much possibility of exploiting other people's achievements. To address the complexities hidden in multilayer interrelations even small amounts of processed data can be useful, improving the performance of complex systems. Exploiting the current trend towards sharing we want to initiate a collective movement that works towards creating synergies and harmonisation among different annotation efforts that are now dispersed. In this paper we present the general architecture of the Language Library, an initiative which is conceived as a facility for gathering and making available through simple functionalities the linguistic knowledge the field is able to produce, putting in place new ways of collaboration within the LRT community. In order to reach this goal, a first population round of the Language Library has started around a core of parallel/comparable texts that have been annotated by several contributors submitting a paper for LREC2012. The Language Library has also an ancillary aim related to language documentation and archiving and it is conceived as a theory-neutral space which allows for several language processing philosophies to coexist.

Keywords: annotation, metadata, scientific crowdsourcing

1. Introduction

The existence of complex relations among phenomena and properties at different linguistic levels is one of the main characteristics that emerges from the analysis of natural language; yet current trends in the production of language resources tend to over-simplify annotation tasks, focusing mostly on one specific linguistic layer at a time, without (having the possibility of) paying attention to the relations among the different layers. At the same time our efforts are too much scattered without much possibility of exploiting other people's achievements. Even small amounts of processed data can contribute to improve the performance of complex systems. This evidence has led to the creation of many algorithms, methodologies, tools and annotation schemes that encode our knowledge of syntactic, semantic and pragmatic features of every language. However these efforts have been scattered and produced annotated/processed data that often lack interoperability. Today we potentially have enough capability and resources to address the complexities hidden in multilayer interrelations. Moreover, we can exploit the current trend towards sharing and initiate a collective movement that works towards creating synergies and harmonisation among different annotation efforts that are now dispersed. In this paper we present the general architecture of the Language Library, an initiative which is conceived as a facility for gathering and making available through simple functionalities the linguistic knowledge the field is able to produce, putting in place new ways of collaboration within the Language Resource Technology (LRT) community.

2. General Description of the Language Library

The Language Library (LL) wants to make a better use of the sharing trend, promoting a real paradigm shift towards a collaborative, open and accessible repository. We believe that Language Resource (LR) building can be conceived as a collaborative "common shared task".

The rationale behind the LL initiative is that accumulation of (high-quality) multi-dimensional data about language is the key to foster advancement in our knowledge about language and its mechanisms, in particular for finding previously unnoticed interrelations between different linguistic levels and among different languages. It differs with respect to Language Commons (Abney and Bird, 2010) which only focuses on a minimal annotation level that is functional for Machine Translation (MT). The Language Library is the first step towards a community-built space where the entire LRT community shares data about language resources and annotated/encoded language data. The Library is going to be:

(i) open, in that its content is accessible to the community without restrictions[1];

(ii) multilingual and -ideally- multi-domain;

(iii) multi-user and community oriented;

(iv) multidimensional, containing multiple layers of annotation of the same text, possibly by multiple contributors;

[1] As explained in section 5.1., the content of the LL is available according to different CC-like licenses.

(v) collaborative, in the sense of collaboration among experts, and also academics and Natural Language Processing (NLP) companies;

(vi) reuse-oriented, promoting the reuse of annotated resources and annotation schemes;

(vii) maintainable, endorsing the use of annotation standards;

(viii) expandable, starting with a first working prototype with a limited number of data and then progressively adding new features and data.

The more the library grows, the more new contributors will be encouraged to participate, building on existing layers of processing to develop their own, which will be in turn added to the resource and become available to the NLP community. Encouraging analysis of linguistic interrelations is one of the aims of this initiative. Notice that the Library can be seen as a place where the theoretical and the applied linguistics communities could meet, in that the provided annotation can be both manually and automatically produced.

In its mature stages the Language Library will consolidate by focusing on the enhancement of interoperability, encouraging the use of common standards and schemes of annotation. The interoperability effort should not be seen as a superimposition of standards but rather as the promotion of a series of best practices that might help other contributors to better access and easily reuse the annotation layers provided. In fact the Language Library is conceived as a theory-neutral space which allows for several language processing philosophies to coexist.

The Language Library has also an ancillary aim related to language documentation and archiving. Even if it's not focused on the recording of linguistic practices it collects textual materials in many languages, analyzed at different levels, consequently promoting awareness about those specific languages. Even though it cannot be described as a pure digital archive, like AILLA[2] neither as a metadata container like OLAC[3] (Simons and Bird, 2003) or the LREMap (Calzolari et al., 2010), the LL can complement existing digital archives suggesting and promoting tools to easily and effectively manage/annotate linguistic data, fostering resource sharing and facilitating networking between people working on the same language but belonging to different communities. The Library itself uses a set of metadata for describing its data providing an access to described resources.

3. Language Library Philosophy

The idea of the Language Library was conceived to put as few limitations as possible to the kind of contribution the contributors can make. Therefore no requirements on compatibility or formats have been made so far. Ideally some sort of agreement on standards or best practices should emerge from the users themselves as soon as more people are trying to reuse annotations provided by others.

[2]http://www.ailla.utexas.org

[3]http://www.language-archives.org

This means that the architecture must be as open as possible, allowing for multiple uses of the same resources by several users while preserving the integrity and retrievability of each file. There are just a few requisites that the architecture should make sure of:

- each user of the library must be identified by a user id; no anonymous contributions can be accepted;

- each processed file that is uploaded has some link to another file (either a source or another processed file). This means that the Library is based upon a sort of RDF triple: $P_f \overset{\tilde{R}}{\sim} S_f$, the processed file P_f has a relation $\tilde{R}$ with source file S_f. The Library should not be a mere repository of corpora or lexicons but a web of interconnected files, with a great focus on reuse and improvement;

- in principle no file can be modified or deleted from the library (even by the providers themselves). In other words if provider A uploads a POS-tagged file of an English text and then improves its tagger, A can upload a new version of the tagged file, that is going to be recorded in the library as a different file:

$$P_{A_1} \text{ is\_pos-tag\_of } S_0$$

$$P_{A_2} \text{ is\_pos-tag\_of } S_0$$

In this scenario, if a provider B has meanwhile used version P_{A_1} to produce a parsed file can refer to the exact version (s)he used when the contribution is uploaded:

$$P_{B_1} \text{ has\_parsed } P_{A_1}$$

- the Library is *interactive*. Any kind of linguistic data can be a "source" for processing in the Library: text, audio and video samples or even multimodal data. However, not all languages, neither all modalities, are covered by the Library data, so users can contribute to the Library with other processable or processed data to be made available to all;

- each file in the Library is described by a minimal set of metadata cf. section 5.3. that allows providers to describe their files and interested people to search data from the Library.

In order for the integrity of the source/processed files to be preserved these five requirements must be combined with a best practice which users have to adhere to: that is the requirement not to change the source in any way by adding or removing parts of it before processing. In this way when a source is processed by different providers, the processed files can be compared/ combined more easily. Since each user can potentially upload files having the same name as others already in the library, and we don't want to ask users to follow a naming convention, the uploaded files are renamed with a combination of original name, provider name, timestamp and original extensions. By doing this two subsequent versions by the same users as well two simultaneous uploads of homonym files by two different users will not create a conflict.

4. Language Library Architecture

In this section we present the architecture of the Language Library. Basically, the LL is a client-server application, where the server side of the application contains the repository of the annotated/processed texts, while the client side part is rendered by the browsers of the contributors.
The web application provides the following modules:

Search The *search* module presents the users with different views of the data (both raw and processed) contained in the Library;

Download The *download* module allows contributors to download raw and/or processed data from the Library;

Upload The *upload* module allows contributors to upload processed files into the Library. Contributors can also upload raw data and share them among the community;

Login This module manages the login of the contributors of the Language Library. To be logged is mandatory for uploading and/or downloading data from/to the Library. The figure 1 describes the architecture and presents the connections between the modules.

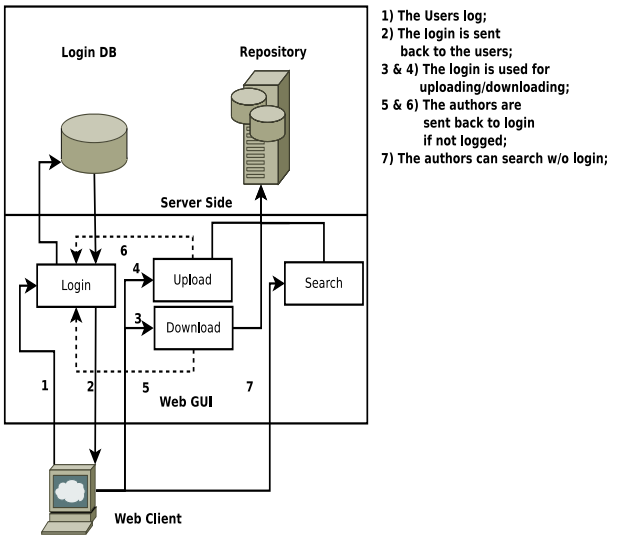

Figure 1: General architecture of the Language Library and connection between different modules

The LL initiative started similarly to the LREMap[4]: strictly connected to the submission phase of the LREC2012 conference but designed to be enriched by more and more capabilities. For this reason the Language Library will go through (at least) 3 different stages:

stage 0 This is the first stage of the Language Library. Authors who submitted papers to the LREC2012 conference have been invited to download a set of raw files from the conference host according to modalities and/or languages they can process. Once processed, these files can be asynchronously uploaded using a web application and then stored in an external repository along with authors' information and files' metadata, cf. section 5.;

stage 1 In this stage the Language Library is disconnected from the conference's submission phase. The LL is enriched with a login system and provides the capabilities for downloading raw files according to modality and/or language directly from the Library repository (as in **stage 0**) and for uploading processed data as well as new raw data to be shared among the community. This stage of the Library provides the possibility of downloading already processed files to add new linguistic annotations. This aspect of the Library defines a sort of "provenance" among processed files: source file S_0 is processed to generate the processed file P_1 that can be processed again to generate P_2 and so on;

stage 2 In addition to the capabilities included in **stage 1**, the main feature offered by the Library at this stage will be a GUI capable of displaying the same text annotated with different schemes, at different levels and by different contributors.

According to these stages, the Language Library will grow both vertically by adding annotation layers and horizontally, by adding new files with different languages and modalities. Figure 2 presents these two dimensions of the Library and adds a third dimension which shows how the same file can be processed by a different contributor.

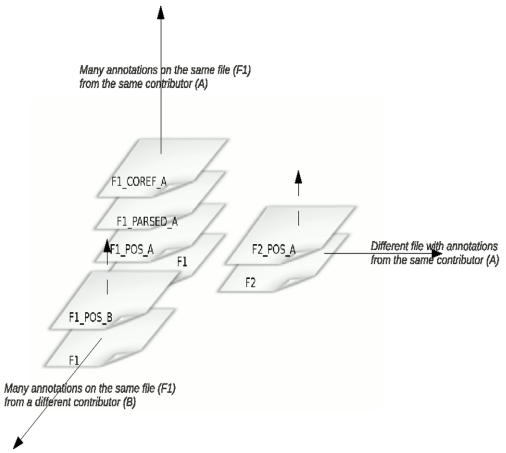

Figure 2: The 3 dimensions of the Language Library

Figure 3 reports a concrete example extracted from the data contained in the Language Library.

[4]http://www.resourcebook.eu

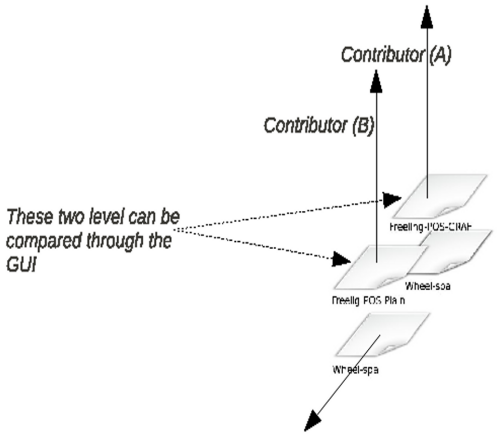

Figure 3: How the Language Library manages the files

In this example, the same source file *Wheel_interwiki-Written-spa.txt* has been processed by two different Spanish providers to produce the same level of annotation, *Part Of Speech*, using the same tool, *FreeLing*. The resulting files are in different format, being one encoded in GraF, the other a plain text.

5. LREC initiative

A first population round of the Language Library has started around a core of parallel/comparable texts that are meant to be annotated over and over again by several contributors submitting a paper for LREC2012.
In this first experiment we implemented the architecture described in **stage 0**, therefore we prepared a repository (a file system) hosting a number of raw data for written and speech modalities in many languages. When filling in the paper submission form, authors were invited to download and process selected texts in the appropriate language(s) and in one or more of the possible dimensions that their submission addresses (e.g. PoS-tag the data, extract/annotate named entities, annotate temporal information, disambiguate word senses, transcribe audio, etc.) and put the processed data back into the newly created Language Library, cf. figure 4.

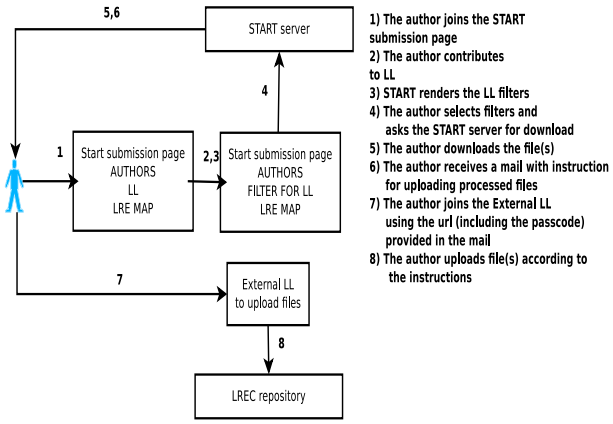

Figure 4: LL architecture **stage 0**: the workflow inplemented for LREC2012

5.1. Data Preparation

In the first downloadable LL data sets - available during the submission process - we provided 20 Wikipedia entries for 64 languages[5] and the written and spoken[6] (mp3) versions of The Human Rights Declaration (in 68 and 58 languages respectively).
Text files have been preprocessed in order to provide *only* the plain text.
The first data sets are distributed according to the following licenses:

(i) Wikipedia entries are distributed under the Creative Commons Attribution-ShareAlike License;

(ii) Human Rights entries both spoken and written are distributed under a Public domain License.

5.2. Upload Processed Files

Authors that accept to contribute to the Language Library receive an email from the START tool with instructions for uploading processed files. They are redirect to the Language Library web interface and automatically logged on the system. By uploading processed files they automatically accept the Language Library licensing system, that is to say that data provided must be either Public Domain (for written and spoken Human Rights Declarations) or Creative Commons Attribution-ShareAlike License (for Wikipedia entries). This aspect is essential for the LL since assures that when the processed files are downloaded to be used by other NLP tools, they are provided according to the same set of licenses.

5.3. End User Interface and Metadata for the processed files

The end user interface determines the usability of a system composed of a repository and a search engine. Even if much effort is spent in the design of the repository and in the organization of the data that will be retrieved, it is the end user interface that determines whether the archived

[5]For the Wikipedia entries we provided 2560 files: 1280 plain texts and 1280 HTML complete files.

[6]From http://librivox.org/the-universal-declaration-of-human-rights-by-the-united-nations/

data can be easily retrieved from a search engine. The current trend is the creation of rich metadata set, rather than "services" that allow users to describe and locate the data (Hughes and Kamat, 2005).
We went into the opposite direction: we defined a minimal set of "high-level" metadata that allows the users to describe the data they are uploading. The end user interface contains four different sections along with their set of metadata: we ask for the specification of values related to provider, contact person and source file(s)[7] used because the infrastructure is based on a clear and retrievable mapping between all source and processed files. But the core set is obviously the specification of the file(s) that have been processed: we ask for modality, type and mode of processing, language. Information about encoding, standard or best practice, tools and documentation is not mandatory[8].

6. First analysis of results

In this section we analyze the data received during the LREC2012 Initiative. These data has been sliced according to the set of metadata described above. In the following discussion, however, not all possible combinations of metadata is analyzed.
We received 686 processed files. Table 1 shows the most frequent languages that are represented in the library, with the number of processed files for each language. As you can see English is the most submitted language but it is not predominant. Less resourced languages such as Burmese are also represented.

English	189
Spanish	110
Catalan	80
Russian	79
Arabic	54
...	...
Bulgarian	22
Japanese	21
Burmese	18
Serbian	11
Uyghur	7
...	...

Table 1: Processed languages

Not surprisingly the predominant processing mode is annotation, covering 548 processed files, cf.figure 5.

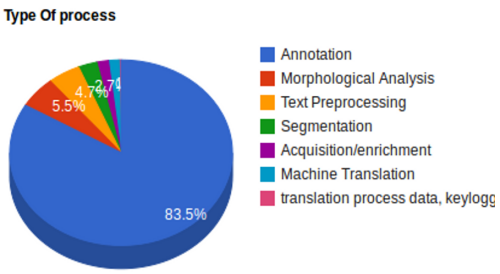

Figure 5: Type of Process distribution

Figure 6 shows the most frequent annotation levels. Even in this early phase, processed data are spread over a great variety of annotation levels, going from *Temporal Expression* to *Semantic Classification*, thus covering both syntax and semantic. Apart from the standard levels of annotation, such as segmentation, tokenization and PoS tagging, which are necessary for almost every kind of processing, information extraction-related annotation seems to prevail. Most particularly Named Entity and Temporal Expression Recognition and and annotation are a pre-requisite in tasks such as question answering, which are among the hottest topics in current NLP research.

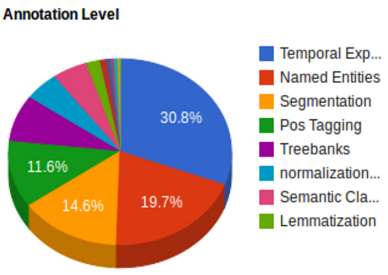

Figure 6: Different Annotation Levels

An interesting aspect of the data we collected is related to the type of annotation, i.e. *inline* vs. *standoff*. Figure 7 clearly shows that for the Library data the *inline* annotation is the most used with a percentage of 73.2%.

[7] By source file we mean the file downloaded from the Library and processed.

[8] A detailed description of metadata proposed is available at http://www.languagelibrary.eu/help/help_upload.html

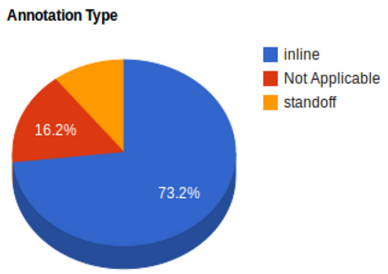

Figure 7: Different Annotation Types

Table 2 shows the levels of annotation available for each language that is currently represented in the Library.

English	Segmentation Lemmatization Pos Tagging Treebanks Semantic Classes Semantic Relations Semantic Roles Named Entities Temporal Expressions
Myanmar (Burmese)	Segmentation Lemmatization Pos Tagging Semantic Classes Named Entities Alignment
Spanish	Pos Tagging Lemmatization
Japanese	Segmentation Pos Tagging
Serbian	Named Entities Events
German	Temporal Expressions
Dutch	Temporal Expressions
Russian	Sound to Text Alignment

Table 2: Annotation levels by language

Obviously English is the language with more annotation levels; moreover these levels come from more than one provider. Myanmar, on the contrary, has several levels of annotation, but they are the outcome of a single annotation process that was provided by a single user.
As reported in section 5.3., standard and/or tools used are non mandatory fields, so that a number of uploaded data do not have such parameters as descriptors. Tables 3 and 4 present the distribution of standard and tools used when declared.

GrAF format	80
Timex3	66
Weblicht	21
XCES	10
TEI P5	8
Hybrid LMF extended with ULex-XML MARKUPS	5
UTF-8 plain text	1
IPA character set in UTF-8 encoding	1
CoNLL 2009	1

Table 3: Standard or best practices when declared

Freeling	186
HeidelTime	62
Athena	22
Sense Substituter based on Resource described in submission	21
BulTreeBank Bulgarian Language Pipeline	21
Humor	21
http://cogcomp.cs.illinois.edu/page/software_view/4	20
Buckwalter, Aragen	18
ETAP-3 parser, StrEd annotation tool	14
Unitex corpus processing tool	11
Sentence alignment (Hunalign)	10
ULex mobile online ...	7
The Sketch Engine	2
GRAMPAL tagger	2

Table 4: Tools used when declared

Although many files declare xml based formats, the output of specific tools is often tab separated and represents a sort of "de facto" standard as well (such as the output of *Freeling*, but also *CONLL 2009*, which is the output format of several parsers). Furthermore some levels of annotation, often Named Entity Recognition (NER) and Temporal Expressions (TE) are often inserted inline as xml tags in the text (cf. figure 7), something that contradicts current trends in standardization, which seem to favor stand-off annotation (GrAF).
Currently, users are not yet allowed to download and further process files provided by others; later on, when this becomes possible, so that users will hopefully become more aware of the importance of maximizing the re-usability of their data by providing all necessary information on standards and used tools[9].
The last dimension we have analyzed is the mode of processing. This dimension shows that the current data are mostly machine processed (more than the 90%), although some files contain manual processing, figure 8.

[9]This can be enhanced by making these metadata recommended (they are currently optional), cf. section 5.3.

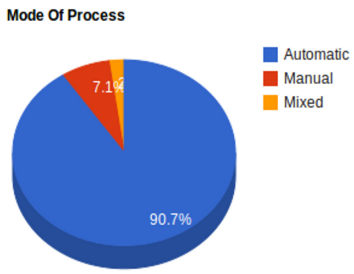

Figure 8: Different Modes Of Process

7. Current implementation

The above described first collection of data at LREC concludes **stage 0** of the Language Library project. The processed data so far collected will be made available to LREC participants during the conference. Presently a first protype of the **stage 1** implementation is already available at **www.languagelibrary.eu**, where all users who want to contribute can login (independently from LREC2012) download and upload of source/processed files. After LREC the new architecture will also allow users to search for, download and re-process data that have already been processed by others.

At a later stage, processed data will also be available through META-SHARE as a special META-SHARE LREC repository[10]. Notice that the metadata of the Language Library are compatible with the minimal set of the META-SHARE schema (Gavrilidou et al., 2011); as a consequence portability is guaranteed. Contextually to this, the Language Library will also be linked to the LRE Map (Calzolari et al., 2010) through the description of resources and tools used.

8. Conclusions and future work

In this paper we have described the first working prototype of the Language Library, an infrastructure that is meant to support the community building of linguistic resources. We also described how this first prototype can evolve into a framework for experimenting interoperability in a multilingual perspective.

With the Language Library we start a large international initiative that connects annotation efforts, offering the possibility to work collaboratively on many common texts for many languages with all possible types of processing, annotation layers, and tools. We think that it can be considered an experiment of "scientific crowdsourcing", i.e. an online collaborative paradigm of interaction and collaboration that is widely used by enterprises but also in the academics to collect data processed by multiple users to gain improvements on a particular task. Through crowdsourcing, the overall quality of scientific data can be improved, constituting a network working on similar tasks. For academic purposes crowdsourcing is successful if it is designed for enlarging the pool of collaborators, and if it leads to invention and innovation maintaining diversity. Beyond extrinsic motivations such as monetary incentives, it gains consensus if it improves social connection, encouraging learning and self-achievement. For this reason we tried to motivate LREC2012 authors to contribute remembering that we offer good opportunity to promote new tools/annotation guidelines. Similarly, we solicited European projects contributors belonging to FLaReNet[11] as national contact points to make visible the results of their tools/resources.

Once all LREC2012 related contributions have been collected, the Language Library will open to the public for search and download, and a call for analysis of the available processed data will be made, open to anyone who wants to test the potentialities of the Library.

[10] www.meta-share.eu

9. Acknowledgments

We thank the META-NET project (FP7-ICT-4 249119: T4ME-NET) for supporting this work. The Language Library started as an initiative within FLaReNet - Fostering Language Resources Network.

10. References

Steven Abney and Steven Bird. 2010. The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 88–97, Uppsala, Sweden, July. Association for Computational Linguistics.

Nicoletta Calzolari, Claudia Soria, Riccardo Del Gratta, Sara Goggi, Valeria Quochi, Irene Russo, Khalid Choukri, Joseph Mariani, and Stelios Piperidis. 2010. The lrec map of language resources and technologies. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Maria Gavrilidou, Penny Labropoulou, Stelios Piperidis, Monica Monachini, Francesca Frontini, Gil Francopoulo, Victoria Arranz, and Valérie Mapelli. 2011. A metadata schema for the description of language resources (lrs). In *Proceedings of the Workshop on Language Resources, Technology and Services in the Sharing Paradigm*, pages 84–92, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Baden Hughes and Amol Kamat. 2005. A metadata search engine for digital language archives, February.

Gary Simons and Steven Bird. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *CoRR*, cs.CL/0306040.

[11] http://www.flarenet.eu