The Multilingual Language Library

@ LREC 2012

Let's build it together!

Nicoletta Calzolari

with Riccardo Del Gratta, Francesca Frontini,

Francesco Rubino, Irene Russo

Istituto di Linguistica Computazionale - CNR - Pisa

glottolo@ilc.cnr.it

The trend

- Make a better use of the **sharing trend**

- In Europe we are building the **META-SHARE** platform, to share LRs and tools
- It is a big step ...

BUT

- We **need a real Paradigm shift**, towards

Collaborative iResources

- **LR building as a collaborative "common shared task"**
  - **New methodology of work**
  - **Interoperability** acquires even more value

Context & Vision

The context

- **NLP is data intensive**
  - Every paper in our conferences speaks about **"data"**
  - Annotation is at the core of training, acquiring, testing, ...
  - But our efforts are still very scattered, with not enough possibility of exploitation

Vision

A Multilingual "Language Library"
As a Large International Initiative

- **MANY** (parallel?) **texts** for **MANY languages**
- With **ALL possible types of processing, annotation layers, ...**

- Similar to **more mature sciences**, e.g. physics, or the Genome project, …
- with *thousands of people working together* on the same big experiment

A Language Library

Rationale

- Accumulation of massive amounts of multi-dimensional data is the key to foster advancement in our knowledge about language & its mechanisms

Strategy

- Create an **infrastructure** for a **large Language repository**
  - Where we **accumulate all the knowledge we have about language**
  - Encourage **analysis of linguistic interrelations**

As a Collaborative Resource: in the sharing paradigm

The **major challenges**:

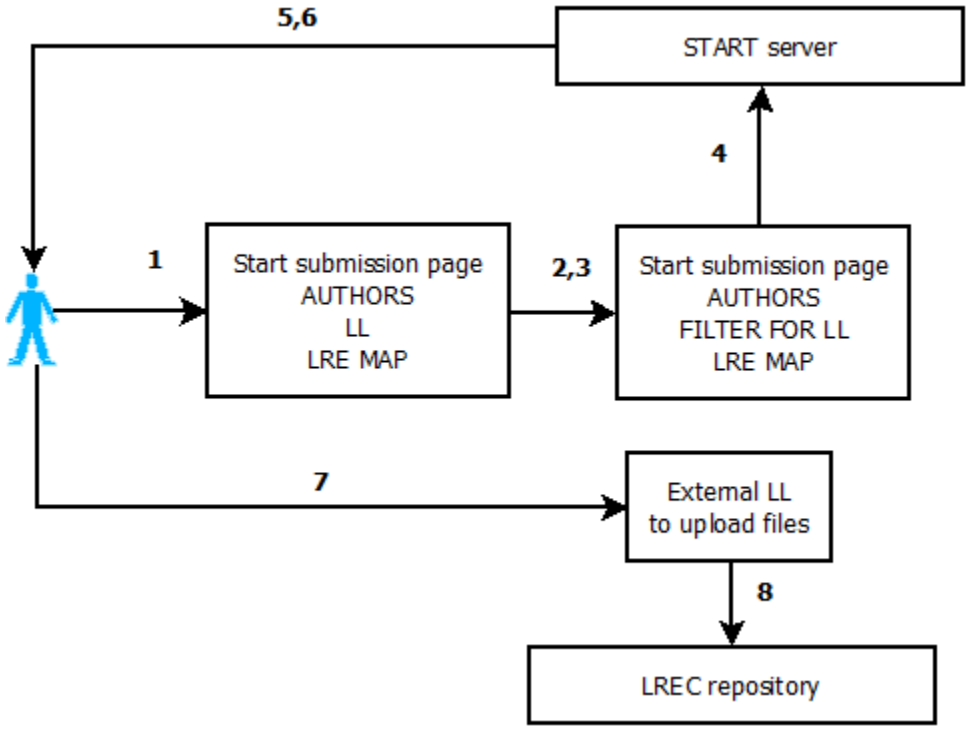
- At the **organisational/design** level?
- At the **community involvement** level?

The first step a new feature @ LREC

- **We: An LREC Repository**
  - Hosting a number of (comparable/parallel) resource
  - In as many languages as possible
  - On all modalities (speech, text, images, etc.)
  - Also as a contribution to META-SHARE

- **Authors:** are invited to **process data**
  - In the language(s) they can process
  - In one or more of the possible dimensions they can address (e.g. POS-tag the data, extract/annotate named entities, annotate temporal information, disambiguate word senses, transcribe audio, translate, etc.)
  - Upload the **processed data back in the LREC Repository**
  - Can also contribute with own raw or processed data, sending to languagelibrary@lrec-conf.org

Flow



1) The author joins the START submission page
2) The author contributes to LL
3) START renders the LL filters
4) The author selects filters and asks the START server for download
5) The author downloads the file(s)
6) The author receives a mail with instruction for uploading processed files
7) The author joins the External LL using the url (including the passcode) provided in the mail
8) The author uploads file(s) according to the instructions

Some data: Languages

Processed files

179	**English**
111	Spanish
80	Catalan
64	Russian
54	Arabic
54	Burmese
40	Japanese
27	Burmese, English
22	Bulgarian
22	Serbian
21	German
20	Dutch
7	Uyghur
3	English, Italian, …

We offer data in 64 languages

Some data: Annotation type

61	Temporal Expressions *(for English, German, Dutch)*
48	Named Entities
41	Pos Tagging
38	Segmentation
20	Lexical substitution
13	Lemmatization
10	Normalization of named entities
10	Semantic Classes
9	Alignment
2	Sound to Text Alignment
1	Events
1	Semantic Relations
1	Semantic Roles
1	Treebanks

Some data: Tools used

187	FreeLing
61	HeidelTime
28	Athena
22	Unitex corpus processing tool
21	BulTreeBank Bulgarian Language Pipeline
21	Sense Substituter based on Resource described in Submission
20	Illinois Named Entity Tagger
18	Buckwalter, Aragen
7	ULex mobile online corpus enrichment tool for language documentation and local language speech technology
4	GRAMPAL tagger
3	Sentence alignment (Hunalign)
2	The Sketch Engine
312	[no tool declared]

Some data: Standards

80	GrAF format
69	Timex3
21	Weblicht
7	CoNLL 2009
3	XCES
5	Hybrid LMF with ULex-XML extension
1	IPA character set in UTF-8 encoding
431	[no standard declared]

Availability

- The processed data will be made **available to all the LREC participants** before the conference, to be compared and analysed

Processed data will be **visible through META-SHARE**

as a special META-SHARE **LREC repository**

- This first experiment on
  - **annotation/transcription/extraction/…**
  - **over the same data** and
  - **on a large number of processing dimensions**
- **May set the ground for a large Language Library**
- **Where everyone can deposit/create processed data of any sort – all our "knowledge" about language**

Collaborative & Interoperability

- Means a change of mentality: going **beyond "my approach"**
- To some "compromise" allowing to go for big amounts, building on each other …

AND … **Interoperability issues**

- Could be a **framework for experimenting interoperability**
- Also **multilingually**

Please contribute here:
http://languagelibrary.eu/