clarin:el

an infrastructure for documenting, sharing and processing language data

Stelios Piperidis, Penny Labropoulou, Maria Gavrilidou
(Athena RC / ILSP)
the problem
use case
requirements

• what does he need?
  • language data
    • corpora (texts authored by the politicians but also general language corpora)
    • dictionaries / domain vocabularies / semantic lexica …
  • tools
    • for processing and annotating
    • for analyzing and visualizing results

• where will he find them?
  • original providers
  • archives and collections
  • other universities & research centers
  • conference proceedings
  • his colleagues’ drawers

• at what state?
  • raw text – ideally, in digital & processable form
  • annotated

• further requirements
  • technical know-how
  • computational power
  • legal problems
the solution: CLARIN
aims

• CLARIN (www.clarin.eu) aims to construct an integrated interoperable research infrastructure bringing together Language Resources and Technologies
• fighting against the current fragmentation
• offering a stable, robust, user-friendly and extensible virtual workplace for accessing language data
• at the service of all researchers (focusing on SSH)
A researcher from her office at Corfu will be able to:

- log in with her academic account and be authenticated only with that (Single Sign-On)
- search, find and get the ok to use texts from Oxford, Bergen and Leiden,
- select the texts she wishes to work with and save her choice as a new collection
- run on this collection semantic taggers from Athens and statistical tools from Budapest
- using the computational power of another computational centre, when and where required
- save the process and results of this analysis and share them with her collaborators in Paris, Vienna and Helsinki
that is…

• CLARIN aims to integrate
  • Language Datasets: digital content of any media type (text, sound, image, video) raw and annotated, lexica, ontologies, grammars etc.
  • Language Technology tools: voice recognisers, lemmatisers, taggers, term extractors etc.

• in a federation of trusted repositories which will be available to all researchers

• through national networks of organisations within each country, focusing on and promoting resources of/about/for the languages of each country (today: more than 200 members from 33 countries)
the Greek infrastructure clarin:el
clarin in Greece

- at the construction phase (2012 – 2015)
- member of CLARIN ERIC since February 2015
- clarin:el implemented as two closely interrelated subsystems for:
  - documenting, depositing, sharing + searching, retrieving and downloading language resources (resources infrastructure)
  - processing language data by web services of language processing and producing new data (processing infrastructure)
- as a first step, aspires to offer the resources a researcher needs
  - in 5 steps
  - with an additional 6th step, when this is possible...
1. keyword search, e.g. Greek corpus in the legal domain
2. browsing of the results
2a. filtered search
3. viewing a selected resource
4. licensing

OROSSIMO Corpus - Law

Licence Agreement – CC-BY

Creative Commons Corporation (“Creative Commons”) is not a law firm and does not provide legal services or legal advice. Distribution of Creative Commons public licenses does not create a lawyer-client or other relationship. Creative Commons makes its licenses and related information available on an “as-is” basis. Creative Commons gives no warranties regarding its licenses, any material licensed under their terms and conditions, or any related information. Creative Commons disclaims all liability for damages resulting from their use to the fullest extent possible.

Creative Commons Attribution 4.0 International Public License

By exercising the Licensed Rights (defined below), You accept and agree to be bound by the terms and conditions of this Creative Commons Attribution 4.0 International Public License (“Public License”). To the extent this Public License may be interpreted as a contract, You accept and agree to be bound by it.

Section 1 – Definitions.

I agree to these licence terms and would like to download the resource. 

download

contact resource maintainer

< resource view
5. downloading
6. language processing

- services for processing monolingual corpora
  - tokenisation and sentence splitting
  - part of speech tagging
  - lemmatisation
  - syntactic parsing
  - term recognition and extraction
  - named entity recognition

- services for processing multilingual corpora
  - for the Greek-English pair
    - all of the above
  - for the Greek-X (=EU language) pairs
    - sentence alignment
behind all these,
a documentation model for describing
language resources and services
ontology – description units
resource typology (1)

• classification on the basis of two main criteria: resource type & media type

• resource type
  • **corpus** (written / spoken / multimedia / multimodal corpora)
  • **lexical / conceptual resource** (e.g. dictionary, terminological resource, word list, ontology etc.)
  • **language description** (e.g. language model, computational grammar etc.)
  • **tool / service** (e.g. lemmatiser, annotator, machine translation tool etc.)
resource typology (2)

- media type: text, sound, image, video

- written corpora
- spoken corpora
- pictures
- videos
resource typology (3)

search for "text"

written corpora

spoken corpora

pictures

videos
metadata schema

- **ResourceInfo**
  - mandatory
  - recommended

- hasComponent
  - IdentificationInfo
  - VersionInfo
  - ContactPerson
  - DistributionInfo
  - UsageInfo
  - ValidationInfo
  - MetadataInfo
  - relationInfo
  - ResourceCreationInfo
  - ResourceDocumentationInfo
architecture of the infrastructure and network
clarin:el architecture

Portal
registration – authentication - authorisation

- search / browse
- licence
- download
- statistics
- mappings
- reporting
- recommenders
- billing/payment

central aggregator

metadata harvesting

user support services

resource provision services

clarin:el inventory

external repos

inventory LR repo

inventory LR repo

inventory LR repo

inventory LR repo
clarin:el users

- behind the infrastructure, the content!
- users: both consumers and providers
- why should one share their resources;
- sharing, enrichment, exploitation, proliferation
clarin:el network

- only LR providers
- network members can be
  - institutions (with or without their own institutional repos)
  - collaborating researchers (at the Hosting Repository)
- current members – Construction Phase (until 31/12/2015)

- Pilot Operation Phase (from 1/1/2016)
  - All academic and research institutions
more information

• [www.clarin.gr](http://www.clarin.gr)
• [http://inventory.clarin.gr](http://inventory.clarin.gr)

info@clarin.gr

clarin.gr

@CLARIN_el

https://www.linkedin.com/grp/home?gid=8309819
Thank you!