Setting up the national infrastructure clarin:el

Stelios Piperidis
Institute for Language and
Speech Processing,
Athena Research Centre,
Greece
spip@ilsp.gr

Maria Gavriilidou
Institute for Language and
Speech Processing,
Athena Research Centre,
Greece
maria@ilsp.gr

Abstract

This paper presents the Greek national infrastructure for language resources, clarin:el, member of CLARIN ERIC since 2015. It describes the design principles for the creation of the network and lists the current members; it describes the infrastructure and its architecture and briefly elaborates on the resources and services offered.

1 Introduction

The Greek network clarin:el is the network of organizations that co-operate for the construction, operation and maintenance of the national research infrastructure (www.clarin.gr). Greece joined CLARIN ERIC in 2015 and the national infrastructure was built as part of the National Research Infrastructures Roadmap in the framework of a project co-funded by Greece and the EU.

The design and development of clarin:el extensively drew upon previous experience gained through similar infrastructural initiatives, like Clarin (Varadi *et al*, 2008), (Wittenburg *et al*, 2008), META-SHARE (Piperidis, 2012), Flarenet (Soria *et al*, 2012) and QTLaunchPad (Piperidis *et al*, 2016).

2 The Greek national network

At the core of the network lies the project consortium; that is, the organizations that implemented the infrastructure, which are the founding members of the network; namely, Athena RC with two institutes, ILSP (www.ilsp.gr) as coordinator, and IMIS (www.imis.gr), NCSR DEMOKRITOS (https://www.iit.demokritos.gr/skel/) and the Greek Research and Technology Network (www.grnet.gr).

Besides the founding members, five more members were selected according to both geographical and scientific principles. Geographic criteria ensured broad coverage of the country's regions, while scientific criteria guaranteed a broad coverage of fields related to language research, linguistics and language technology. Thus, next to the three founding members, the following institutions were selected: the University of Athens, Dept. of Linguistics (http://en.phil.uoa.gr/the-faculty/department-of-linguistics.html) based in Central Greece, two from Northern Greece, namely the Aristotle University of Thessaloniki, Dept. of French Language and Literature (http://www.frl.auth.gr), with expertise on Linguistics, Translation and Language Resources, and the Centre for the Greek Language (www.greeklanguage.gr), an institution of the Ministry of Education responsible for language policy and language teaching; the Ionian University (www.ionio.gr) at Corfu, Western Greece, specializing on Translation studies and Information science; the University of the Aegean at Rhodes (http://dms.aegean.gr/en) with expertise in language resources collection and documentation.

These organisations constitute the network's nodes; they have set up their own repositories (Institutional Repositories), in which they document, store and manage the resources they share through the infrastructure. An alternative type of repository is the Hosted Resources Repository, where organisations not wishing to set up a repository can store, document and manage their resources. Besides organizations, individuals can also share resources, technologies and services using the Hosted Resources Repository for this purpose. Providers (legal entities or individuals) are responsible for the curation of their language resources, technologies and services and for the clearance of the related intellectual property rights.

3 The infrastructure

3.1 The architecture

The architectural design of the clarin:el network and infrastructure has been inspired and, to a large extent, based on the design principles of the META-SHARE infrastructure. As such, clarin:el is a network of distributed repositories, each being set up for and maintained by each member organization of the clarin:el network. Unlike META-SHARE, clarin:el:

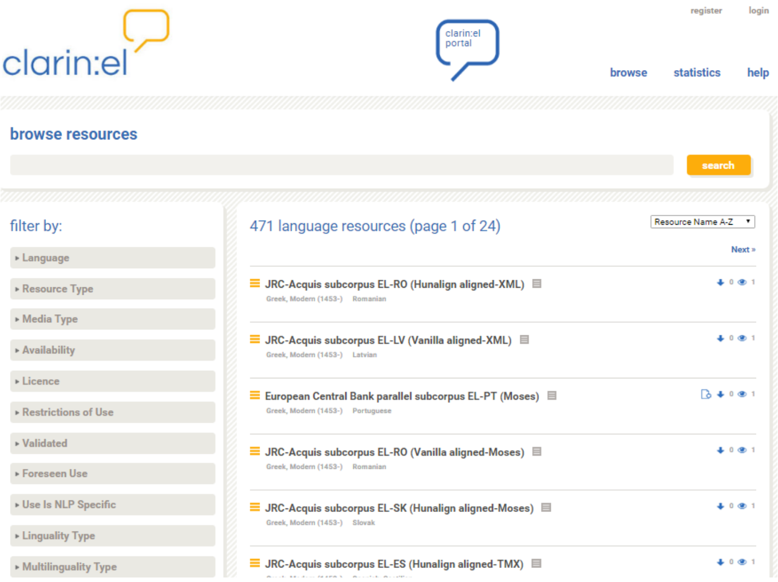
a. features a single central repository that harvests metadata and hosts a central inventory for the whole network, and

b. adopts a multi-tenant approach implementing a single database server for all repositories.

It is the central inventory that will assume all functionalities and responsibilities for connecting clarin:el to CLARIN ERIC (e.g. providing OAI-PMH endpoints for metadata harvesting, ensuring secure access to the CLARIN ERIC federation).

These architectural choices centralize and facilitate the creation, management and administration of repositories, thus taking the burden of local hosting from member-organizations wishing to set up an institutional LRs repository. Repositories are automatically set up on ~okeanos (https://okeanos.grnet.gr/home/), the research cloud infrastructure of the Greek Research and Technology Network and actual datasets are stored in its distributed permanent storage pithos+ (https://okeanos.grnet.gr/services/pithos/). In this way, maintenance operations are simplified to a large extent and the respective costs are substantially reduced.

clarin:el has adopted the META-SHARE metadata schema (Gavriilidou *et al*, 2011), which has been mapped to CMDI 1.0 (Broeder *et al*, 2008), (Broeder *et al*, 2012). Resources are persistently identified using handles (www.handle.net) issued upon being published in the repositories. Metadata support simple and faceted search, as shown in Figure 1.

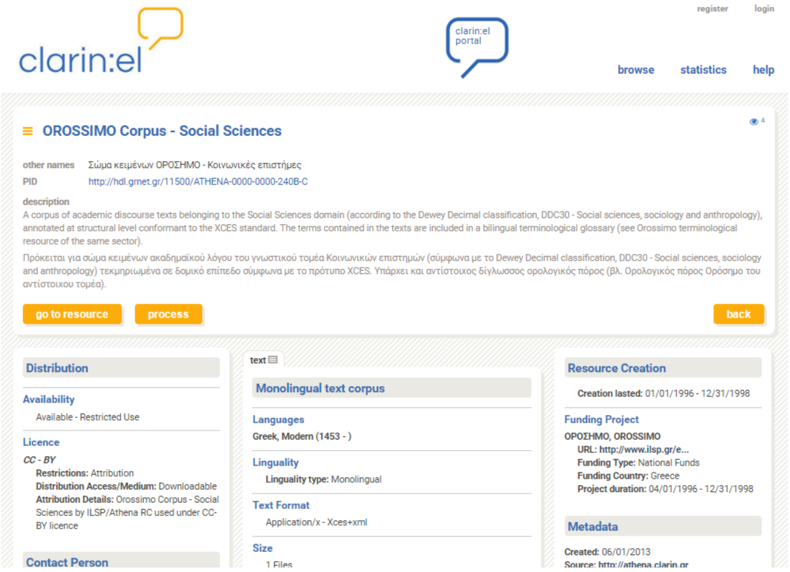


471 language resources (page 1 of 24)

- JRC-Acquis subcorpus EL-RO (Hunalign aligned-XML) — Greek, Modern (1453-) Romanian — 0 1
- JRC-Acquis subcorpus EL-LV (Vanilla aligned-XML) — Greek, Modern (1453-) Latvian — 0 1
- European Central Bank parallel subcorpus EL-PT (Moses) — Greek, Modern (1453-) Portuguese — 0 1
- JRC-Acquis subcorpus EL-RO (Vanilla aligned-Moses) — Greek, Modern (1453-) Romanian — 0 1
- JRC-Acquis subcorpus EL-SK (Hunalign aligned-Moses) — Greek, Modern (1453-) Slovak — 0 1
- JRC-Acquis subcorpus EL-ES (Hunalign aligned-TMX) — 0 1

Figure 1: Browsing and searching the clarin:el inventory

Figure 2 shows the record of the chosen resource with the full metadata description.



≡ OROSSIMO Corpus - Social Sciences

other names Σώμα κειμένων ΟΡΟΣΗΜΟ - Κοινωνικές επιστήμες
PID http://hdl.grnet.gr/11500/ATHENA-0000-0000-240B-C

description
A corpus of academic discourse texts belonging to the Social Sciences domain (according to the Dewey Decimal classification, DDC30 - Social sciences, sociology and anthropology), annotated at structural level conformant to the XCES standard. The terms contained in the texts are included in a bilingual terminological glossary (see Orossimo terminological resource of the same sector).

Πρόκειται για σώμα κειμένων ακαδημαϊκού λόγου του γνωστικού τομέα Κοινωνικών επιστημών (σύμφωνα με το Dewey Decimal classification, DDC30 - Social sciences, sociology and anthropology) τεκμηριωμένα σε δομικό επίπεδο σύμφωνα με το πρότυπο XCES. Υπάρχει και αντίστοιχος δίγλωσσος ορολογικός πόρος (βλ. Ορολογικός πόρος Ορόσημο του αντίστοιχου τομέα).

Distribution

Availability
Available - Restricted Use

Licence
CC - BY
Restrictions: Attribution
Distribution Access/Medium: Downloadable
Attribution Details: Orossimo Corpus - Social Sciences by ILSP/Athena RC used under CC-BY licence

Contact Person

Monolingual text corpus

Languages
Greek, Modern (1453 -)

Linguality
Linguality type: Monolingual

Text Format
Application/x - Xces+xml

Size
1 Files

Resource Creation
Creation lasted: 01/01/1996 - 12/31/1998

Funding Project
ΟΡΟΣΗΜΟ, OROSSIMO
URL: http://www.ilsp.gr/e...
Funding Type: National Funds
Funding Country: Greece
Project duration: 04/01/1996 - 12/31/1998

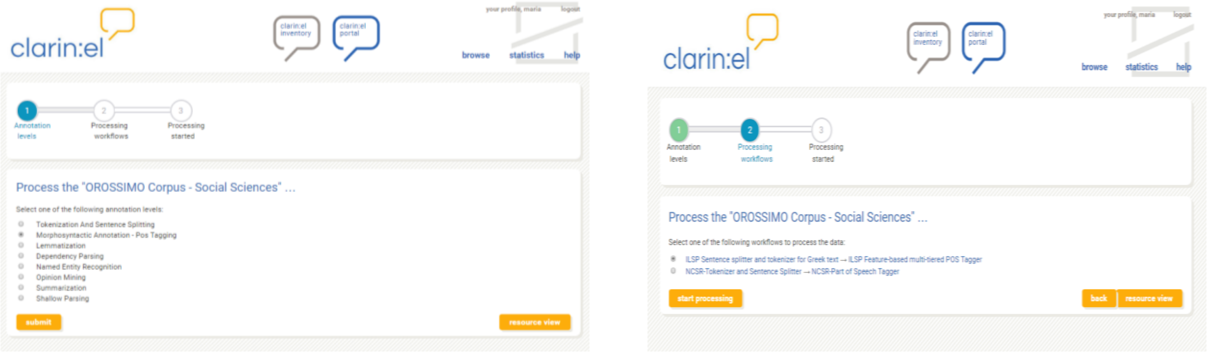
Metadata
Created: 06/01/2013
Source: http://athena.clarin.gr

Figure 2: Full metadata record

The legal component draws on the META-SHARE toolkit and fosters the use of Creative Commons 4.0 licences, the recently upgraded META-SHARE NoRedistribution 2.0 licenses and existing standard open source software (FOSS) licenses for tools and services. The rights and restrictions of use of the resource are under the control of the resource owners, while metadata elements documenting the rights of use are obligatory (no resource can be deposited in the infrastructure without license metadata).

3.2 Language processing services

The primary repository functionalities have been enhanced by an additional language processing mechanism for the processing of language datasets with appropriate services. Inspired by previous initiatives (Piperidis *et al*, 2016), language processing services are documented with the appropriate metadata in the repository and are provided as web services through the language processing layer. In the typical usage scenario, when a clarin:el user selects to process a dataset, a list of all available relevant annotation/processing levels and respective services is automatically generated based on the properties of the dataset (Figure 3), i.e. its resource type, language(s), domain, format, annotation level, and licensing conditions. Currently, services for processing Greek and English datasets are offered at the following levels: tokenization & sentence splitting, POS tagging, lemmatization, dependency parsing, named entity recognition (all for Greek and English), opinion mining and summarization (for Greek only).



Process the "OROSSIMO Corpus - Social Sciences" ...

Select one of the following annotation levels:

○ Tokenization And Sentence Splitting
● Morphosyntactic Annotation - Pos Tagging
○ Lemmatization
○ Dependency Parsing
○ Named Entity Recognition
○ Opinion Mining
○ Summarization
○ Shallow Parsing

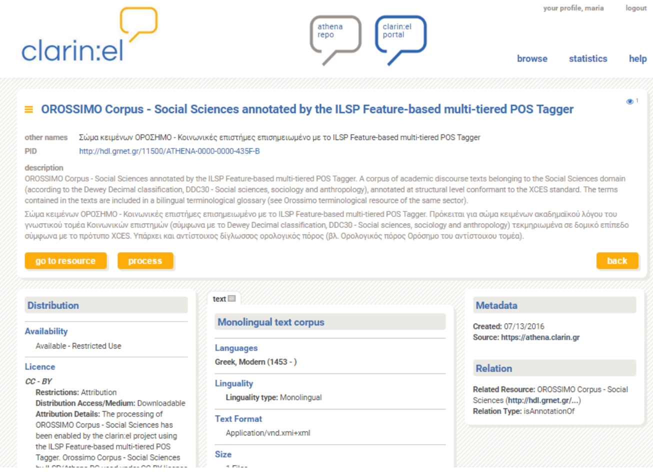
Process the "OROSSIMO Corpus - Social Sciences" ...

Select one of the following workflows to process the data:

● ILSP Sentence splitter and tokenizer for Greek text → ILSP Feature-based multi-tiered POS Tagger
○ NCSR-Tokenizer and Sentence Splitter → NCSR-Part of Speech Tagger

Figure 3: Presenting the relevant annotation levels for a resource and selecting the desired service

In view of integrating language processing services, the metadata model was extended to facilitate the documentation of language processing services as well as processing-related properties of datasets. As soon as the user selects a service, the server dispatches the dataset to the specific web service(s) for processing. When the processing has been completed, the new (annotated) dataset is automatically documented according to the metadata schema, stored and indexed in the native repository (Figure 4). Automatic creation of metadata records for newly-created datasets combines the metadata descriptions of the original dataset and the language processing tool/service used; it also creates relations between the original dataset and the annotated one.



OROSSIMO Corpus - Social Sciences annotated by the ILSP Feature-based multi-tiered POS Tagger

other names Σώμα κειμένων ΟΡΟΣΗΜΟ - Κοινωνικές επιστήμες επισημειωμένο με το ILSP Feature-based multi-tiered POS Tagger

PID http://hdl.grnet.gr/11500/ATHENA-0000-0000-435F-B

description

OROSSIMO Corpus - Social Sciences annotated by the ILSP Feature-based multi-tiered POS Tagger. A corpus of academic discourse texts belonging to the Social Sciences domain (according to the Dewey Decimal classification, DDC30 - Social sciences, sociology and anthropology), annotated at structural level conformant to the XCES standard. The terms contained in the texts are included in a bilingual terminological glossary (see Orossimo terminological resource of the same sector).

Σώμα κειμένων ΟΡΟΣΗΜΟ - Κοινωνικές επιστήμες επισημειωμένο με το ILSP Feature-based multi-tiered POS Tagger. Πρόκειται για σώμα κειμένων ακαδημαϊκού λόγου του γνωστικού τομέα Κοινωνικών επιστημών (σύμφωνα με το Dewey Decimal classification, DDC30 - Social sciences, sociology and anthropology) τεκμηριωμένα σε δομικό επίπεδο σύμφωνα με το πρότυπο XCES. Υπάρχει και αντίστοιχος δίγλωσσος ορολογικός πόρος (βλ. Ορολογικός πόρος Ορόσημο του αντίστοιχου τομέα).

Distribution

Availability

Available - Restricted Use

Licence

CC - BY

Restrictions: Attribution
Distribution Access/Medium: Downloadable
Attribution Details: The processing of OROSSIMO Corpus - Social Sciences has been enabled by the clarin:el project using the ILSP Feature-based multi-tiered POS Tagger. Orossimo Corpus - Social Sciences

text

Monolingual text corpus

Languages

Greek, Modern (1453 -)

Linguality

Linguality type: Monolingual

Text Format

Application/vnd.xmi+xml

Size

Metadata

Created: 07/13/2016
Source: https://athena.clarin.gr

Relation

Related Resource: OROSSIMO Corpus - Social Sciences (http://hdl.grnet.gr/...)
Relation Type: isAnnotationOf

Figure 4: Automatic creation of metadata record for a derivative resource

4 Conclusion

The national clarin:el infrastructure has been set up involving 7 institutional repositories and 1 cloud computing and storage infrastructure provider. It has been in operation since November 2015 offering language data and processing services to all academic users in Greece. Users can login with their academic credentials, if they have such credentials and their academic organisation has joined the Greek Authentication Authorisation Infrastructure (AAI), managed by GRNET, or register and create an account with clarin:el, if they are not affiliated with a Greek academic institution.

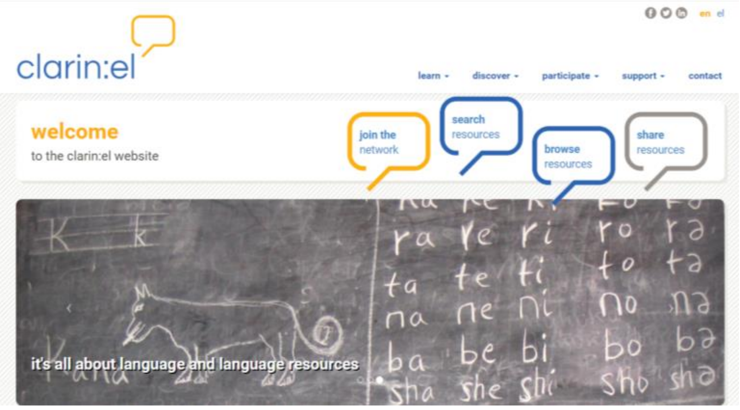


Figure 5: The clarin:el portal

clarin:el is accessible through its portal (www.clarin.gr) which provides connection of the national to the European infrastructure CLARIN ERIC, informative material and a range of support facilities to all its users (Figure 5).

5 References

Broeder, D., T. Declerck, E. Hinrichs, S. Piperidis, L. Romary, N. Calzolari and P. Wittenburg (2008). Foundation of a Component-based Flexible Registry for Language Resources and Technology. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias(Eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Broeder, D., Van Uytvanck, D., Gavrilidou, M., Trippel, T., & Windhouwer, M. (2012). Standardizing a component metadata infrastructure. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Gavrilidou, M., Labropoulou, P., Piperidis, S., Francopoulo, G., Monachini, M., Frontini, F., and Mapelli, V. (2011) A Metadata Schema for the Description of Language Resources (LRs). In *Proceedings of Workshop Language Resources, Technology and Services in the Sharing Paradigm, IJCNLP 2011*, Chiang Mai, Thailand

Piperidis, S., D. Galanis, J. Bakagianni, S. Sofianopoulos (2016) Combining and Extending Data Infrastructures with Linguistic Annotation Services, In Murakami Y. and Donghun L(Eds), *Worldwide Language Service Infrastructure: Second International Workshop, WLSI 2015,* Kyoto, Japan, January 22-23, 2015. Revised Selected Papers" Springer International Publishing, pages 3-17, doi="10.1007/978-3-319-31468-6_1"

Piperidis, S. (2012). The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Soria, C., Bel, N., Choukri, K., Mariani, J., Monachini, M., Odijk, JEJM, Piperidis, S., Quochi, V., and Calzolari, N. (2012) The FLaReNet Strategic Language Resource Agenda. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Eds), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Váradi, T., Krauwer, S., Wittenburg, P., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common Language Resources and Technology Infrastructure. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias(Eds) *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Wittenburg, P., N. Bel, L. Borin, G. Budin, N. Calzolari, E. Hajicova, K. Koskenniemi, L. Lemnitzer, B. Maegaard, M. Piasecki, J.M. Pierrel, S. Piperidis, I. Skadina, D. Tufis, R.v. Veenendaal, T. Váradi, M. Wynne (2010) Resource and Service Centres as the Backbone for a Sustainable Service Infrastructure. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias(Eds) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).