# Preparing the Field for an Open and Distributed Resource Infrastructure: the Role of the FLaReNet Network

**Nicoletta Calzolari, Claudia Soria**
CNR-Istituto di Linguistica Computazionale "A. Zampolli"
Via Moruzzi 1, 56124 Pisa, Italy
E-mail: nicoletta.calzolari@ilc.cnr.it, claudia.soria@ilc.cnr.it

**Abstract**

In order to overcome the fragmentation that affects the field of Language Resources and Technologies, an Open and Distributed Resource Infrastructure is the necessary step for building on each other achievements, integrating resources and technologies and avoiding dispersed or conflicting efforts. Since this endeavor represents a true cultural turning point in the LRs field, it needs a careful preparation, both in terms of acceptance by the community and thoughtful investigation of the various technical, organisational and practical aspects implied. To achieve this, we need to act as a community able to join forces on a set of shared priorities and we need to act at a worldwide level. FLaReNet – Fostering Language Resources Network – is a Thematic Network funded under the EU eContent program that aims at developing the needed common vision and fostering a European and International strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide. In this paper we present the activities undertaken by FLaReNet in order to prepare and support the establishment of such an Infrastructure, which is becoming now a reality within the new META-NET initiative.

## 1. Introduction

Language Technologies (LT), together with their backbone, Language Resources (LR), provide an essential support to the challenge of Multilingualism and ICT of the future. The main task of language technologies is to bridge language barriers and to help creating a new environment where information flows smoothly across frontiers and languages, no matter the country, and the language, of origin.

To achieve this, we need to act as a community able to join forces on a set of shared priorities.

Currently, however, the field of LR&LTs suffers from an excess of individuality and fragmentation: there is no substantial sharing of what are the priorities for the field, where to move, not to mention a common timeframe.

This lack of coherent directions is partially also reflected by the difficulty with which fundamental information about LR&LTs is reachable: basically, it is very difficult, if not impossible, to get a clear picture of the current situation of the field in simple terms such as who are the main actors, what are the available development and deployment methods, what are the "best" language resources, what are the areas for which further development and investment would be most necessary, etc. Substantial information is not easily reachable not only for the producers but also for policy makers and funding agencies.

The field is active, but it needs a coherence that can only be provided by sharing common priorities and endeavours. Under this respect, since some time large groups have been advocating the need of a LR&T infrastructure, which is increasingly recognised as a necessary step for building on each other achievements, integrating resources and technologies and avoiding dispersed or conflicting efforts. A large range of LRs and LTs is there, but the infrastructure that puts LR&Ts together and sustains them is still largely missing; interoperability of resources, tools, and frameworks has recently come to be understood as perhaps the most pressing current need for language processing research. Infrastructure building is thus indicated by many as the most urgent issue and a way to make the field move forward. Time is ripe for going beyond individual research interests and recognise the infrastructural nature of LRs by establishing an Open and Distributed Resource Infrastructure. This will allow easy sharing of data, corpora, language resources and tools that are made interoperable and work seamlessly together, as well as networking of language technology researchers, professionals, users. At the same time, however, this is an endeavour that represents a true cultural turning point in the LRs field and therefore needs a careful preparation, both in terms of acceptance by the community and thoughtful investigation of the various technical, organisational and practical aspects implied. To this end, FLaReNet has promptly embraced the requests coming from the worldwide scientific community and has oriented many of its central initiatives so as to prepare and support the establishment of such an Open Resource Infrastructure.

## 2. The FLaReNet Network

FLaReNet – Fostering Language Resources Network – is a Thematic Network funded under the EU eContent program[1] that aims at developing the needed common vision and fostering a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide. It establishes itself as an international Forum whose essential mission is to act as an observatory to assess current status of the field on Language Resources and Technology and to indicate priorities of action for the future. The FLaReNet goals and activities are of strategic

[1] Grant Agreement no. ECP-2007-LANG-617001, http://www.flarenet.eu

nature, and they can be summarized as follows:

- to gather, consolidate and sustain a **community**: LR&T stakeholders need to be identified and convinced that they are part of a larger body;
- to facilitate **interaction** among LR&T stakeholders, so that exchange of opinions and views is ensured;
- to promote and sustain **international cooperation**;
- to coordinate a community-wide effort to **analyse the sector** of LR&Ts along all the relevant dimensions: technical and scientific, but also organisational, economic, political and legal;
- to identify short, medium, and long-term **strategic objectives** and provide **consensual recommendations** in the form of a plan of action targeted to a broad range of stakeholders, from the industrial and scientific community to funding agencies and policy makers;
- to pave the way to the set up and functioning of an Open Distributed Resource Infrastructure through a number of preparatory initiatives that must be continued and strengthened during the establishment and the life of the infrastructure.

Among the many themes that have emerged from the community consultation, infrastructure building seems to be the main message and the most urgent issue ahead of us. With inductive methods dominating the current paradigms in Language Technology, language resource building, annotation, cataloguing, accessibility, availability and clearance from IPR is what the research community is calling for. The message coming out from the two FLaReNet Forums [2] is only one among the multiple concurring signs now indicating that time is ripe for establishing an Open and Distributed Resource Infrastructure, which allows networking of language technology professionals and their clients, as well as easy sharing of data, corpora, language resources and tools that are made interoperable and work seamlessly together. Such an infrastructure is something that the Language Resources community has been pushing since some time and that is now increasingly recognized as a necessary step for building on each other achievements and avoid scattered or conflicting efforts. As a response to the community needs, the European Commission has granted a Network of Excellence addressing such an infrastructure: the META-SHARE infrastructure will be implemented within the META-NET Network of Excellence[3].

## 3. Preparing the ground for an Open Resource Infrastructure

An important factor for the success of an infrastructure is the acceptance and the active involvement of the community: this has to be carefully prepared. Under this respect, FLaReNet is paving the way to the set up and functioning of such an infrastructure through a number of preparatory initiatives, either already started or planned within FLaReNet, which must be continued and strengthened during the establishment and the life of the infrastructure. As such, these constitute a set of recommended sub-goals that must be pursued by FLaReNet now and together with META-NET in the near future to lay down the basis for the infrastructure:

- Community building and community active involvement
- Analysing the sector
- Enhancing the accessibility of/to information
- Information gathering
- Defining recommendations and priorities
- Promoting cooperation, also at international level.

### 3.1 Community building and sensitizing

It must not be underestimated that a strong and committed community is an essential prerequisite for the success of an infrastructure. Awareness of the need and justification for such an infrastructure must be spread.

FLaReNet has the challenging task of (re)creating a network of experts around the notion of Language Resources and Technologies. To this end, FLaReNet is bringing together leading experts of research institutions, academies, companies, consortia, associations, funding agencies, public and private bodies both at European and international level, users and producers alike, with the specific purpose of creating consensus around short, medium and long-term strategic objectives. It is of foremost importance that the FLaReNet Network be composed of the as widest as possible representation of experiences, practices, research lines, industrial and political strategies. This will allow to derive an overall picture of the field of Language Resources and Technologies that is not limited to the European scene, but can also be globally inspired.

In order to constantly increase the community of people involved in FLaReNet, as well as to ensure their commitment to the objectives of the Network, an active permanent recruiting campaign is running. People wishing to join the Network can do so by filling an appropriate web form available on the FLaReNet web site. The FLaReNet Network is open to participation by public and private, research and industrial organizations. Invitation to join, either personal or by means of mailing lists are used in order to enlarge the community as much as possible.

The Network is currently composed of more than 300 individuals and 81 institutional members from 31 different countries. FLaReNet affiliates belong to academia, research institutes, industries and government, and their number is steadily enlarging through new subscriptions. Such a community needs to grow not only in number, but also with reference to the type of disciplines involved, from the core ones (Natural Language Processing, computational linguistics, Language Engineering) to "neighboring" ones, such as cognitive science, semantic web, etc. Participants are expected and encouraged to

[2] The first FLaReNet Forum was held in Vienna, 12-13 February 2009; the second Forum took place in Barcelona, 11-12 February 2010. See the FLaReNet home page for more information.

[3] http://www.meta-net.eu/

express their views individually as experts but also their organizations views and concerns.
Meetings are the primary means for attracting new members and to reinforce participation of existing ones, but participation is expected and encouraged also by means of online discussions, forum threads, and collaborative documents. The most important community-building events were the FLaReNet Launching Event (Vienna, 12-13 February 2009) and the second FLaReNet Forum (Barcelona, 11-12 February 2010). Both were organised as a collaborative workshop composed of a series of thematic working sessions on specific topics. The intention was to approach each topic trying to identify controversial aspects, risks, missing elements, gaps to be filled, what can/cannot be achieved. The meetings were conceived as a means for raising discussions in an interactive and creative way, thus stimulating open questions, new ideas, and visions for the field towards a multilingual digital Europe. The Vienna and Barcelona Workshops and their thematic sessions have been very successful and demonstrated a great interest for a discussion on the current state and future progress of LR&Ts.

### 3.2 Analyse and survey the LR&T sector at large

In order to define a clear and coherent roadmap that identifies priority areas for public funding in LRs and LT, an accurate map of Language Resources and Technologies is essential. This must cover many different aspects: the methods and models for production, use, validation, evaluation, distribution of LRs and LTs, their sharing and interoperability; different types and modalities of LRs; the applications and products for LR&Ts; the advantages and limitations of standardisation; the different needs and priorities of academy vs. industry and commerce, of data providers vs. users; the traditional and new areas of interest for LRs; the cultural, economic, societal, political issues, etc. One of the major endeavors in this respect is the LREC Map of Language Resources and Tools (Calzolari et al. 2010), which was partially supported by FLaReNet. In addition to this, FLaReNet has performed a number of independent surveys of the LR&T sector. A survey is dedicated to existing language resources and current status of HLT market, mostly from player profile perspective. This survey, which resulted in a FLaReNet deliverable (Choukri et al. 2009), tried to focus on some of the major features that would help understand all issues related to LRs from descriptive metadata to usability in key application, to the composition of various BLARKs for important technologies, to the legal/ethical/privacy issues, etc. Another study is about the identification of the problems occurring in using language resource and language technology standards and to identify emerging needs for future LRT standards (Budin 2009). Here, the approach is based on studying existing documents related to LRT standards, to study existing LRT standards, to evaluate current implementations of these standards, to ask implementers about the problems they have identified in using such standards and to ask all LRT stakeholders about missing standards or other problems they see in this respect. (Parra et al. 2009) contains a survey of automatic production methods for LRs was produced. This comprises a survey of the most demanded resources that are used as the core element of some NLP applications and an overview of the current techniques for automatic construction of LRs. The last academic proposals for automatic acquisition and production of LRs have been also reviewed, in order to confirm the interest that these topics raise in the community of researchers, and as the basic information to start a classification of methods and resources addressed. Finally, the survey by (Odijk and Toral 2010) is centred on an investigation of the available methodologies, campaigns and services for evaluating and validating LRs.

### 3.3 Enhancing the accessibility of information about Language Resources and Technologies

In its first year and a half of life, FLaReNet has already become the “pole of attraction” of the LR&Ts community and has played a central role in spreading awareness of the need, relevance and importance of Language Resources and Technologies.
In coordination with other relevant organisations, associations and initiatives, FLaReNet has prepared the conditions for creating consensus around interoperability issues, thus enhancing accessibility to Language Resources and Tools. In particular, FLaReNet is committed in

- promoting the creation of new language resources and the enhancement of existing ones;
- providing easy and uniform access to the main available catalogues of LR&Ts;
- fostering the knowledge dissemination and availability of language resources and tools that will eventually enter the infrastructure (e.g. through the “LREC Map”);
- the creation of consensus around interoperability issues and the promotion of standardisation activities.

### 3.4 Information pushing and pulling

FLaReNet mobilizes its community by setting up expert groups that will work in synergy on a number of different, yet concurring issues that are crucial for the development of a language resource infrastructure, such as a preliminary assessment of the temporal, technical, organisational, and legal constraints involved in the implementation of an open language infrastructure. To this end, FLaReNet has started very early a campaign for eliciting user needs and requirements on a range of topics, from technical to organisational, related to the functioning of the infrastructure, such as, for instance:

- *temporal*: what is the time frame under which a first implementation can be reasonably expected (considering various steps and building blocks)?
- *technical*: which should be the main principles underlying its architecture, assuming some

evolution over time to account for new developments (web2.0 …)?
- *organisational*: how the involved centres and resources should be regulated? Which are the best organisational and governance models that can ensure efficiency and sustainability?
- *legal*: what are the legal and IPR issues that could be involved by the realization of an open language infrastructure?

A "first" set of basic principles and characteristics for the Infrastructure – as emerging from a set of "concrete scenarios of use of the ORI" anticipated by the FLaReNet Steering Committee – is given in (Calzolari et al. 2009). These principles, still in a conception phase, must be exposed to the entire community, so that many can participate in the discussion and a consensus is gradually formed around them. Some characteristics are accompanied by a first set of recommended actions (see Section 4 below).

### 3.5 Promote and sustain international cooperation

For a Network like FLaReNet, whose aim is the development of strategies and recommendations for the field of Language Resources and Technologies, coordination of actions at a worldwide level is of utmost importance. A number of international cooperation initiatives has been launched with relevant initiatives also outside the EU, in order to establish a kind of global coordination of the LR field. Permanent connections have been established with major European and international players (such as AFNLP, SILT, ELRA, LDC, ISO) and others (e.g. COCOSDA, TEI, Oriental-COCOSDA, ALTA, ETSI) are planned in order to both elicit feedback and disseminate results and recommendations.

Infrastructural issues are widely represented in international cooperations. For instance, FLaReNet entertains permanent connections with its US twin project SILT in order to address interoperability issues in a coherent and concerted way (see Ide et al. 2009).

Most importantly for the goal of a resource infrastructure, FLaReNet tries to boost cooperation between existing infrastructures and initiatives, by favoring occasions where different infrastructural initiatives can compare, define their respective roles, and possibly make synergies emerge. During the LREC 2010 conference, for instance, FLaReNet co-organizes the COCOSDA-WRITE-FLaReNet workshop, where a special session is devoted to a comparison of the different initiatives for sharing language resources.

### 3.6 Define recommendations

In order for the major players in the field of Language Resources and Technologies to consensually work together, a clear direction and priorities for the next years must be indicated. FLaReNet does this under the form of a roadmap for Language Resources and Technologies.

To date, FLaReNet has published two sets of recommendations, the first issued after the FLaReNet Launching Event ("First FLaReNet Forum Highlights"), and the other coming from a consultation of the community. The latter, the "*Blueprint for Actions and Infrastructures*" (Calzolari et al. 2009) gathers the recommendations collected around the many meetings, panels and consultations of the community, as well as the results of the surveying activities carried out under FLaReNet workpackages. The Blueprint encompasses a preliminary Plan for Actions and Infrastructures targeted at HLT players at large, policy-makers and funding agencies.

Among the recommendations developed by FLaReNet some have a direct impact on a resource infrastructure. We list them below, addressing stakeholders and funding agencies, respectively.

## 4. FLaReNet Recommendations for a Language Resource Infrastructure

The recommendations below are those issued by the FLaReNet Network after the first year of consultation of the Language Resources community. They come in two sets, the first addressing language resources stakeholders and the second tailored to funding agencies and policy makers.

### 4.1 Language resources stakeholders

- Implement services and policies to enable sharing of resources
- Adopt a model for tool and resource development based on open advancement and collaborative development, where the community as a whole contributes components, modules, etc. to a common system or framework
- Another strategy proposed and partially implemented is to set up cooperative services by multiple local players to counter one big global player. Since each local player can often offer better quality than the global player for the particular local language or domain, the cooperating local players together can compete against the global player and offer customers the best available quality
- Ways of cooperation and linking among (partially similar) infrastructural initiatives both in Europe and in other continents must be devised and basic principles commonly agreed
- Models for giving credit for data sharing should be devised
- It is necessary to develop models and paths for turning research results into easily shareable results
- Initiatives such as the LREC Map of Language Resources and Technologies should be enforced and widely spread to the entire community, as they can turn into useful measuring tools for monitoring the evolution of LRs over time as well as to gradually lead to a situation where the

notion of citation & publication of LRs becomes accepted and gives academic credit
- Since the lack of documentation and clear information about resources and related technologies is an important issue, harmonization of metadata and, at the same time, enforcing use of a common vocabulary of categories to describe and document resources are important steps towards facilitating surveying and retrieval activities.

### 4.2 Funding Agencies and Policy makers

- Improve access to digital research data for a better exploitation
- Data generated by public sector institutions should be increasingly made available for research and development, following principles similar to the Public Sector Information Directive
- Language resources built in the framework of EU or other funded projects should be made available at fair cost
- Infrastructure building is the most urgent issue. Infrastructures and repositories for tools and language data, but also for information on data (documentation, manuals, metadata, etc.) should be established that are universally and easily accessible by everyone
- In the long term, interoperability will be the cornerstone of a global network of language processing capabilities. The necessary framework and a corresponding infrastructure (i.e. standards and technologies) must be established and made operational. This can only be achieved through a coordinated, community-wide effort that will ensure both comprehensive coverage and widespread acceptance. Not only are data formats to be standardised, but also metadata
- An Open Resource Infrastructure should be established, which allows easy sharing of data, corpora, language resources and tools that are made interoperable and work seamlessly together
- An infrastructure for collecting data is needed. An appeal was made to the EC to support an infrastructure and tools to collect language data for a wide range of applications, as well as for the creation of data, in particular conversational speech data for speech-to-text technology for the whole range of European languages, and make these data available at affordable prices for research purposes and to SMEs
- Different funding agencies should jointly take care of infrastructural priorities
- Enhance current coordination of language resource collection between all involved agencies and ensure efficiency (e.g. through interoperability)
- Repositories of data formats, annotations, and guidelines should be supported as a major help to achieve and promote standardisation
- Since cooperation cannot be limited to a European landscape, alliances and cooperation among (partially similar) infrastructural initiatives both in Europe and in other continents must be organised and favoured.

## 5. Conclusion

The field of Language Resources and Technologies needs a strong and coherent concertation activity to become ready for setting up an open and distributed infrastructure for sharing language data, tools and services. The FLaReNet Thematic Network plays a leading role in preparing the field for this enterprise. First, FLaReNet acts as a facilitator of exchange of information, by promoting discussions among experts about the main issues at stake. Second, it sensitizes the overall community about the importance of an infrastructure, the benefits that can be gained from that, but also the many steps that need to be done in order to reach it. Third, it helps the sector to achieve a broader and deeper self-consciousness with important initiatives, such as the LREC Map, which is the most comprehensive community-built overview of the field to date.

Infrastructures are the future of language resources. FLaReNet is one of the privileged places where they are conceived.

## 6. Acknowledgements

This work has been carried out in the framework of the FLaReNet Thematic Network, EU eContent Work Programme, Grant Agreement no. ECP-2007-LANG-617001

## 7. References

Budin, G. (2009). Identification of problems in the use of LR standards and of standardization needs. FLaReNet Deliverable D4.1. http://www.flarenet.eu/sites/default/files/D4.1.pdf

Calzolari, N., Soria, C., Del Gratta, R., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., Piperidis, S. (2010). The LREC 2010 Resource Map. In *Proceedings of LREC 2010*.

Calzolari, N., Soria, C., Bel, N., Budin, G., Caselli, T., Choukri, K., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., Toral, A. (2009). D8.2° - Blueprint of actions and infrastructures. FLaReNet Deliverable 8.2a. http://www.flarenet.eu/sites/default/files/D8.2a.pdf

Choukri, K., Arranz, V., Mapelli, V., Mostefa, D., Moreau, N. (2009). Up-to-date chart of LR and players and classification along different lines. FLaReNet Deliverable D2.1a. http://www.flarenet.eu/sites/default/files/D2.1a.pdf.

Ide, N., Pustejovsky, J., Calzolari, N., Soria, C. (2009). The SILT and FlaReNet international collaboration for interoperability. In Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, Suntec, Singapore, pp. 178-181.

Odijk, J., Toral, A. 2009. Existing evaluation and

validation of LRs. FLaReNet Deliverable D5.1
http://www.flarenet.eu/sites/default/files/D5.1.pdf

Parra, C., Bel, N., Quochi, V. (2009). Survey and assessment of methods for the automatic construction of LRs. Report on automatic acquisition, repurposing and innovative proposals for collaborative building of LRs. FLaReNet Deliverable D6.1a.
http://www.flarenet.eu/sites/default/files/D6.1a.pdf