Planning the Future of Language Resources: The Role of the FLaReNet Network

Nicoletta Calzolari and Claudia Soria

CNR-ILC, Pisa, Italy
{glottolo,claudia.soria}@ilc.cnr.it

Abstract. In this paper we analyse the role of Language Resources (LR) and Language Technologies (LT) in today Human Language Technology field and try to speculate on some of the priorities for the next years, from the particular perspective of the FLaReNet project, that has been asked 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.

Keywords: Language Resources and Technology, strategic initiatives, priorities.

1 Why are Strategic Initiatives Necessary?

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&Ts 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&Ts 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

A. Gelbukh (Ed.): CICLing 2010, LNCS 6008, pp. 1–11, 2010.

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 Resource Infrastructure (ORI). 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 turnpoint 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.

While there has been considerable progress in the last decade, there remains a significant challenge to overcome current fragmentation and imbalance inside the LR&T community. To this end, it is of utmost importance that *strategic activities* are put into place so as to ensure that the LRs community is made aware of the current status of the field, and at the same time so that new directions of development are indicated in a coherent and clear way.

The entire community behind Language Resources (organizations, institutions, funding agencies, companies, and individuals) needs guidance and assistance in planning for and addressing the needs of the language resources and technologies of the future. Together, and under the umbrella of a shared view of actual priorities, a future can be shaped in which a common market for Language Resources and Technologies is created through coordination of programs, actions and activities.

In order to provide such a strategic view of the future directions, a number of preliminary steps are necessary:

- 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 initiate and carry out a community-wide effort to **analyse the sector** of LR&Ts. The analysis should cover 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.

In this paper we illustrate these steps from the particular standpoint of the FLaReNet[1] project, whose mission is to act as an observatory to assess current status of the field

[1] FLaReNet – Fostering Language Resources Network, www.flarenet.eu– is a Network of Excellence funded under the EU eContent program 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.

on Language Resources and Technology and to indicate priorities of action for the future.

2 An Inventory of Possible Strategic Actions for Language Resources and Technology

2.1 Create and Mobilise a Unified and Committed Community of Language Resources and Technologies players

(Re)creating a network of experts around the notion of Language Resources and Technologies is a challenging task. 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 such a community 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, as well as to ensure their commitment to the objectives of the Network, FLaReNet runs an active permanent **recruiting campaign**. 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 200 individuals and 79 organisations 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 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.

Other general ways for sensitizing and attracting people, as well as for making former members aware of the Network activities, is a massive use of **advertising material, publishing of the Newsletter, and participation in conferences and major events** related to Language Resoruces and Technologies.

Apart from actions for enlarging the FLaReNet community, those aimed at **consolidating** it are especially important. Participants to the community need to feel they belong to a group of people that is actually shaping the field of Language Resources and Technologies by delineating its direction for the next future. The User

Forum, the creation of Thematic Group and “think-tanks” of experts and the launch of closed meetings are the privileged ways for creating close and connected groups of people.

2.2 Define Priorities, Formulate Strategies and Recommendations

Language technologies and language resources are the necessary ingredients for the development of applications that will help bridging language barriers in the global and unified information space, in a variety of means (the Web as well as other communication devices) and for a variety of channels/media (spoken and written language alike but also other associated modalities e.g. gesture). It is of utmost importance, however, to identify a set of priority themes as well as short, medium, and long-term strategic objectives in order to avoid scattered or conflicting efforts. The major players in the field of Language Resources and Technologies need to consensually work together and indicate a clear direction and priorities for the next years, under the form of a roadmap for Language Resources and Technologies. This is the kind of results at which meetings are especially targeted. Actions foreseen to this end are centred around the activity of thematic, general and liaison meetings (see Deliverable 1.4 for further details).

FLaReNet has the challenging goal to act as a “sensor” of current and future trends in Language Resources and Technologies. In order to do this, it must be able to make most pressing issues emerge from its community of players. A number of actions globally converge toward this goal:

- thematic meetings;
- encouragement to propose discussion themes (e.g. through our wiki site);
- requests for topic proposals for Thematic meetings / provoking issues;
- link with major (new) projects & initiatives.

Activities belonging to this category broadly share a common workflow: meetings and events are the privileged places where important issues emerge from the community. These issues are broadly discussed, both at the events themselves and through on-line discussion. Major topics are then distilled and delivered to the community and to the EC under the form of recommendations.

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*” (D8.2a) 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.

2.3 Analyse and Survey the LR&T Sector at Large

The definition of a clear and coherent roadmap that identifies priority areas of LRs and LT that need public funding to develop or improve clearly presupposes the availability of an accurate map of Language Resources and Technologies, under many different respects: 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.

To this end, FLaReNet is involved in surveying the sector of LR&Ts from many different perspectives. A survey was dedicated to existing language resources and current status of HLT market, mostly from player profile perspective. This survey, which resulted in D2.1, 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 was about the identification of the problems occurring in using language resource and language technology standards and to identify emerging needs for future LRT standards (D4.1). Here, the approach chosen 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.

Finally, 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.

2.4 Provide an Assessment of the Current Status of the Field

Work conducted so far in FLaReNet has contributed to draft a first portrait of the current situation in the LR&T sector, in particular for what concerns the types of players and resources (WP2), the various needs for standardisation according to the different communities and the obstructing factors to adoption of standards (WP4), an overview of current practices in evaluation and validation of LR&Ts (WP5), and a review of the innovative methodologies being implemented for the automatic development/processing of LRs (WP6). In addition to the activity of the work packages, input has been collected from a number of events, either organised or co-organised by FLaReNet.

The following is a shortlist of facts that concisely hint at the situation of the LR&T sector as it has emerged from FLaReNet observation.

- Re-use and re-purposing of data is hindered by lack of common data representation
- Documentation of language resources is generally poor
- Clear and easy-to-reach information about resources and related technologies is lacking
- There are too few initiatives around the BLARK concept for European languages
- Little concern is given to the issue of data preservation

- The legal framework is far too complex, and in particular:
  - License models especially suited to LRs are lacking
  - Legal protection modes are different across Europe
  - There are different strata of intellectual property rights
- Sustainability for linguistic tools and language resources needs to be increased
- LRs need to be maintained, in terms of bug reporting, updates and improvements
- More efforts are needed to solve the problem of how to automate the production of the large quantity of resources required, and at the same time how to ensure the necessary quality to get acceptable results in industrial environments
- The evaluation of automatic techniques for LR production is of variable quality. Comparisons among techniques should also be carried out to better assess each of them and their strengths and weaknesses, fostering a greater development in the research on these fields
- Much of the research on automatic acquisition of LRs has focused on small-scale experiments and therefore their usability in applications is largely yet to be demonstrated
- It is very difficult to find information about the characteristics of the language resources that industrial applications use, as well as about the size and granularity of the information contained
- Standardisation is at the core of interoperability. Standardisation issues currently show substantial convergence of opinion and practice, which needs now to be supported to become operational
- LR standards are:
  - too much oriented towards academic/research purposes, not yet mature enough for industrial applications
  - too difficult to understand
  - too abstract, lack concrete examples for implementation, lack of user scenarios or user guides
  - too isolated, existing only on paper but not integratable in digital workflows,
  - too cumbersome to implement, no return on investment in sight for implementers
- Industry-born standards are:
  - too much driven only by specific needs and lack long-term vision
- Given the breadth of current landscape of LR&Ts, a “cultural” change is needed in the sense that there is the need to find ways to monitor how resources are used, to register the resources used or created, to introduce the notion of “publishing” resources and to get academic credit for resources that have been made available.

3 Recommendations for Actions in the HLT Field

The following recommendations are intended both for HLT stakeholders (producers, users and developers of Language Resources and Technologies, both academic and industrial) on the one side and funding agencies and policy makers on the other.

Infrastructure building is the most urgent issue. An Open Resource Infrastructure, which allows easy sharing of language resources and tools that are made interoperable and work seamlessly together, is felt essential. 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.

3.1 Resource Production and Use

- Provide documentation of the produced resources, covering at least the following aspects (metadata): owner/copyright holder, format and encoding issues of the data and the files, languages(s) covered, domains, intended applications, applications in which the data was used, formal parameters that have to be verified, reliability of any annotation that is included
- For documentation, adherence to practices followed by major data centers is advisable
- Ensure quality of language resources (LRs), for instance by performing a basic quality assessment, to ensure that the minimal critical dimensions are documented: availability/reliability of information on technical, formal issues such as media, number of files, file structure(s), file names etc.
- Annotated resources should be provided with a detailed documentation describing the annotation procedures which have been developed in the annotation process of the LR
- Promote the development of new methodologies for assessing the annotation quality of LRs, in particular for semantic annotation
- Information about whether the resources acquired are actually used or, the other way around, of what are the particular characteristics of the actually used resources, needs to be made public.
- Use of best practices or standards in new projects must be enforced, to facilitate data re-use. Projects developing LRs should be requested to adhere to standards for encoding and representation of data and associated annotations
- Policy makers should enforce documentation of resources, including annotation formats
- Priorities in the development of core LRs for languages should be driven by BLARK-like initiatives: support them and encourage countries to develop their own BLARK matrices
- The creation of LRs must be tied to the development of technologies. It is mandatory to produce the basic tools to process the ‘raw’ data
- Support the development of LRs for less-resourced languages
- Invest in the production of parallel corpora in multiple languages
- Support the development of resources and technologies for processing non-verbal, and more generally contextual information encompassed in speech-based interaction

- Actual industrial needs have to be to addressed: information about whether the resources acquired are actually used or, the other way around, of what are the particular characteristics of the actually used resources, needs to be made public. The involvement of industries in the research on automatic methods must be supported
- Public procurement, especially at the EU level, should be used as one of the instruments to boost production and adoption of language technologies.

3.2 Interoperability Issues

- It is important that commonly accepted practices (best practices, de-facto standards or standards, when available) are used for the representation and documentation of data
- Not only are data formats to be standardised, but also metadata
- Standards need tools that support them , to promote and ensure their adoption
- LR standards have to be made *more operational* (both, existing ones and those under preparation), with a specific view on different user communities – most users should not or do not want to know that they are using standards, they should operate in the background and they should be "inherent" to the language technology tools or more generic tools they use
- A crucial step towards consistency and interoperability for a global information exchange is the definition of a work environment for *data category definition and management*
- Aim at new forms and manifestations of standards, as *embedded standards*
- For each standard, *return on investment* and possible *motivations* of users should be elaborated together with potential or real users (early adopters)
- Focus in the *short term planning* on those areas where there is enough consensus so that chances are high that a widely accepted standard can be published in a short period of time
- Increase the *acceptance* of LR standards (and the need for them) in different communities, both research and industry communities, and to directly involve user communities in creating standards
- Analyse the needs and requirements for harmonisation of existing standards
- Develop a *strategy* for LR *standards creation*, taking into account aspects such as: bottom-up vs top-down approaches with an interactive process model needed, and modular component standards rather than a single monolithic standard for all of LR
- Standards maintenance should be a process of *change management*, ideally *in real time*
- Inform more pro-actively on best practices in implementing standards and in successful corporate language standards.

- Try to solve the "standard divide" by which a few languages are very well equipped with language resources and consequently with LR standards needed
- Have an *integrative view* on LR standards: an European Interoperability Framework (EIF) for LR has to be developed (cross-domain, cross-purpose, cross-cultural, etc.)
- Contribute to expand the EIF, e.g. in the context of eGovernment, eHealth, eLearning, etc. where many of the existing LR standards can already contribute effectively to enhance data interoperability
- Bring together research communities and industrial application communities for developing a joint vision on LR standards in general
- Foster cooperation between MT industry and CAT-oriented translation and localization industry, for well-balanced and more integrative LR standards industrially usable yet based on pre-normative research
- Develop a broader vision of LR standards with the inherent inclusion of *multimedia*, *multimodal* and *speech* processing applications
- Create an operational ecology of language resource standards that are easily accessible, re-usable, effective, and that contribute to semantic interoperability
- Aim to a global standardization effort on the well-known line of EAGLES-LIRICS-ISO, a long-term strategy which brings together US-experts with their standards and best practices with the European traditions of EAGLES etc. and with East Asian best practices in the field.

3.3 Licensing, Maintenance and Preservation

- Prevent loss of data along the years, by ensuring appropriate means for data archiving and preservation
- Avoid "home-made" licensing models. When drafting a distribution license, carefully think of making it suitable for subsequent re-use and re-distribution of the resource. Adhere to practices used by distribution agencies whenever possible
- Whenever possible, ensure appropriate means for maintenance of Lrs.
- It is important to ensure that publicly funded resources are made publicly available at very fair conditions. Public agencies should impose that resources produced with their financial support are made available free of charge for academic R&D activities. It is also important to encourage language resource owners to donate them to data centres to be distributed free of charge
- Enforce/sustain initiatives for data archiving and preservation: it should be ensured that the data produced by a certain project/initiative/organisation will survive any possible change of media for distribution

- When funding new LRs, funding agencies should request a plan for their maintenance
- Ensure sustainability of funded resources, e.g. by requesting accessibility and usability of resources for a given time frame
- Sustain initiatives offering legal recommendations/guidelines for the reuse of Language Resources, and investigating appropriate licensing models allowing for re-use and re-distribution.

3.4 Evaluation and Validation

- Work on common and standard evaluation procedures, taking into account normalization for comparison. Techniques should not only be evaluated on scientific grounds, but also by their impact in real scenarios of NLP applications
- Develop tools for automatic validation (fault detection (clipping, noise...), detection of segmentation errors, of weak annotations, confidence measures of speech transcriptions)
- Investigate different solutions for addressing the problem of task- vs. application-oriented, such as:
  - A general evaluation framework, including both kinds of evaluation, such as the ISLE Framework for Evaluation in Machine Translation (FEMTI) approach
  - An integrated evaluation platform
  - In the same framework, remote evaluation distributed over the Internet, which permits to interchange components, allowing comparing various approaches, while also examining the influence of the component on the whole system, and which could be organized as Web services.
- Evaluation of the results of automatic techniques must also foresee complex scenarios where the quality of the final results depends on the quality of the partial results.
- The definition of appropriate evaluation frameworks for automatic acquisition methods is needed. The development of evaluation methods that cover the different automatic techniques is fundamental, in order to allow for a better testing of existing and newly discovered methods. Beyond the evaluation on scientific grounds, it is also recommended that techniques are measured by their impact in real scenarios of NLP applications
- Promote a permanent effort framework to take care of language technology evaluation in Europe.

3.5 Directions for Research and Development

- Invest in the development of resources and technologies for processing non-verbal, and more generally contextual information encompassed in speech-based interaction

- As many of the automatic evaluation measures in the style of BLUE and its descendant are still highly controversial, active research into other types of metrics and other ways of evaluating is desirable
- More efforts are needed to solve the problem of how to automate the production of the large quantity of resources required, and at the same time how to ensure the necessary quality to get acceptable results in industrial environments
- Standards need to *co-evolve* at high speed together with rapid change in science, technology, commerce
- Support the involvement of industries in the research on automatic methods, so as to allow a more precise assessment and evaluation of automatic methods for the development of LRs for real-scale applications
- Support transfer of Human Language Technology to SMEs: instruments should be established to transfer language technologies from projects to the SME language technology community in order to stimulate the availability of new technologies and increase the language coverage.
- New languages that joined recently the Union should be considered as a higher priority in coming EU programs
- Human language resources need to be “de-globalized” and focus on local languages and cultures despite today’s “global” village
- Copyright law should be harmonised at the European and national level in such a way to permit the free use of copyrighted works for academic purposes
- Favour multidisciplinary integration of different communities.

References

1. Calzolari, N.: Approaches towards a “Lexical Web”: the role of Interoperability. In: Ide, N., Fang, A.C. (eds.) ICGL 2008, The First International Conference on Global Interoperability for Language Resources, City University of Hong Kong, pp. 34–42 (2008)
2. Calzolari, N.: Initiatives, Tendencies and Driving Forces for a “Lexical Web” as Part of a “Language Infrastructure”. In: Tokunaga, T., Ortega, A. (eds.) LKR 2008. LNCS (LNAI), vol. 4938, pp. 90–105. Springer, Heidelberg (2008)
3. Calzolari, N., Baroni, P., Bel, N., Budin, G., Choukri, K., Goggi, S., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., Soria, C., Toral, A. (eds.): Proceedings for the FLaReNet Forum, The European Language Resources and Technologies Forum: Shaping the Future of the Multilingual Digital Europe. Istituto di Linguistica Computazionale del CNR, Pisa (2009)
4. Calzolari, N., Soria, C.: The FLaReNet Thematic Network: a Global Forum for Cooperation. In: 7th Workshop on Asian Language Resources in conjunction with ACL-IJCNLP 2009, Singapore (2009)
5. Ide, N., Pustejovsky, J., Calzolari, N., Soria, C.: The SILT and FLaReNet International Collaboration for Interoperability. In: 3rd Linguistic Annotation Workshop in conjunction with ACL-IJCNLP 2009, Singapore (2009)