The European Language Resources and Technologies Forum

Shaping the Future of the Multilingual Digital Europe

Vienna, 12 -13 February 2009

Short Report

N. Calzolari, N. Bel, G. Budin, K. Choukri, J. Mariani, J. Odijk, S. Piperidis, P. Baroni, S. Goggi, M. Monachini, V. Quochi, C. Soria, A. Toral

Istituto di Linguistica Computazionale del CNR - Pisa, ITALY

Introduction by the FLaReNet Coordinator

Nicoletta Calzolari — ILC-CNR

FLaReNet – Fostering Language Resources Network – is an EC eContentPlus Thematic Network (ECP-2007-LANG-617001) whose aim is to create a shared policy and foster a European strategy in the field of Language Resources (LRs) and Language Technologies (LTs). The recent growth of the field should be complemented by a common reflection and by an effort that identifies synergies and overcomes fragmentation. By creating consensus among major players in the field, the mission of FLaReNet is to identify priorities as well as short and long-term strategic objectives, sustain international cooperation and provide consensual recommendations in the form of a plan of action for EC, national organisations and industry. The consolidation of the area is a pre-condition to enhance competitiveness at EU level and worldwide.

Work in FLaReNet is inherently collaborative. A set of Working Groups are clustered in thematic areas and carry out their activities through workshops, meetings, and via a collaborative Wiki platform. The FLaReNet **Thematic Areas** are:

- the Chart for the area of LRs and LT in its different dimensions;
- methods and models for LR building, reuse, interlinking, maintenance, sharing, distribution, …;
- harmonisation of formats and standards;
- definition of evaluation and validation protocols and procedures;
- methods for the automatic construction and processing of LRs.

FLaReNet is bringing together leading experts of many research institutions, companies, consortia, associations, funding agencies, public and private bodies both at European and international level. Anyone can subscribe to the FLaReNet website, joining any of the working groups and participating in their activities. This will offer the advantage of playing a role in the definition of recommendations for future actions, thus shaping the future with respect to the new challenges.

The **FLaReNeT Forum in Vienna** combined the FLaReNet themes with the i2010 objectives to address some of the technological, market and policy challenges to be faced in a multilingual digital Europe. The Forum represented an occasion to identify the grounds for future directions and strategies in the area of LRs and LTs.

The Forum was composed of a series of working sessions where leading experts were invited to present their vision on hot topics in the field of LRs and LTs. A new formula was experimented, whereby the FLaReNet Steering Committee prepared for each session a background document highlighting a set of relevant issues and questions to be addressed by the speakers.

The final session was dedicated to a round-table on International Cooperation, mainly with non-European participants, where future policy and priorities were discussed in a global context. The aim was to initiate a strategic discussion on the utility of promoting international cooperation among various initiatives and communities around the world, within and around the field of Language Resources and Technologies.

The full version of the 1st FLaReNet Forum Highlights (with reports from each session) as well as the full Proceedings are available on the FLaReNet website.

http://www.flarenet.eu
flarenet_coordination@ilc.cnr.it

Preface by European Commission, Unit INFSO-E1 - Language technologies and machine translation

Roberto Cencioni, Kimmo Rossi

FLaReNet plays an important role in the process that will define the actors, the overall direction and the practical forms of collaboration in language technologies and their "raw material", language resources. The main task of language technologies is to bridge language barriers in the global single information space, on the Web and over mobile communication devices, for spoken and written language alike. To achieve this, a community of key people need to work together and show a clear direction and priorities for the next 3-5 years.

One of the concrete tasks ahead of us is to create, for all EU languages, an open language infrastructure which allows networking of language technology professionals and their clients, as well as easy sharing of data, corpora, language resources and tools. Interoperability is a must: the common infrastructure can only succeed if the resources, tools and processes work seamlessly together, now and in the future.

The volume of multilingual information and communication is exploding in the Web. Sharing, collaboration and networking flourish – interactions are more and more instantaneous. This requires more automation: translations are needed on the fly, machine translation systems need to be set up and trained overnight, language resources need to be acquired and annotated automatically, with minimal human intervention.

The new communication and collaboration paradigms create excitement but also confusion. Language technology is a mature field, but the trusted and proven recipes may not work any more. We need new solutions and new partnerships, while securing the basic acquired knowledge base. FLaReNet will have the challenging task to create this network of people, to formulate strategies and to stimulate action in a context that is constantly changing. The demand for cross-lingual technologies is pressing, the expectations are high, and at the same time, the field is suffering from fragmentation, lack of vision and direction. In 2009 citizens will elect a new European Parliament and a new Commission will be nominated. We will deal with decision-makers that do not know us nor our business. This makes it important that we think clearly and express our ideas even more clearly.

In terms of organization, participation and stimulating debate, this two-day forum has been a big success. Now we need to reach out to the public, the policymakers and the business community – not only the academic world. FLaReNet does not have the resources to implement alone the necessary language infrastructure. Reports, meetings, events and contacts are the primary tools to achieve the ambitious goals. The success of FLaReNet relies greatly on simple things such as concise, reader-friendly reports that convey the message at first reading. All FLaReNet partners – but the coordinator in particular – have a crucial role in ensuring that all the communication matches the success of this forum.

An impact assessment has recently been completed on Language & Interaction technology actions funded by the European Commission in 1999-2005. The findings indicate that a lot of work needs to be done especially in three areas: policy, standards and outreach, especially towards the business, markets and end users. FLaReNet is an important instrument in our common effort to address these challenges.

FLaReNet View on Language Resources and Technologies

Language resources are machine-readable (electronic) collections of samples and descriptions of human language: text corpora, speech recordings, grammars, dictionaries/lexicons, grammars, databases of parsed and analysed sentences etc. A wider definition of language resources also includes various automatic language processing tools: spell-checkers, parsers, taggers, editors, annotation tools etc. Language resources (and tools) are the necessary raw material for software and services which can automatically understand, translate and respond to human language. We need a basic set of language resources for every language for which we want to develop automatic language services (e.g. machine translation systems).

The European FLaReNet - Fostering Language Resources Network - was born to enhance European competitiveness in the field of Language Resources and Technologies, especially by consolidating a common vision and a European strategy for the future. FLaReNet is bringing together leading experts of research institutions, academies, companies, funding agencies, with the specific purpose of creating consensus around short, medium and long-term strategic objectives.

In this spirit, FLaReNet gathered more than a hundred players worldwide at the latest Vienna Forum, with the specific purpose of setting up a brainstorming force to make emerge the technological, market and policy challenges to be faced in a multilingual digital Europe.

Over a two-day programme, the participants to the Forum had the opportunity to start assessing the current conditions of the Language Resources and Technologies field and to propose emerging directions of intervention.

Some messages recurred repeatedly across the various sessions, as a sign both of a great convergence around these ideas and also of their relevance in the field. The Forum validated ideas that have been "in the air" for several years and, in some cases, fostered and/or developed by specific groups, as having entered the main stream of thought and practice within the language technology community.

The Challenges

Remedy the lack of resources

Essential language resources for critical Language Technologies applications are still missing. This holds even for the major EU languages and the most demanded applications despite the great advancements in the last decade, and it is even more true for the languages of the States who have recently joined the European Union. Several corrective measures have been identified to this end, among which:

- find reliable methods for assessing the depth and breadth of these gaps, possibly using existing instruments such as the BLARK (Basic LAnguage Resource Kit);
- exploit innovative ways of resource building besides traditional ones. Wikis and social networks can act as cooperative means of Language Resources production that can complement traditional approaches. Automatic procedures for language resource production can also be beneficial to support a faster development of language resources;
- at a political level, simplify legal issues concerning intellectual property, and devise supporting measures that ensure that publicly funded resources are made publicly available at very fair conditions;
- think global, act local: "de-globalize" human language resources and focus on local languages/ cultures despite the today's "global" village, also by devising modalities of cooperation and sponsorship. *Cooperation* and *integration* are the keywords here: public bodies and funding agencies need to ensure cooperation among scattered efforts, so that these are converging.

Attain true resource interoperability

To sustain coordinated actions toward common goals, but also in order to close the gaps in coverage, Language Resources need to be made interoperable. Interoperability of Language Resources means mutual translatability, so that different language resources can be merged, integrated or migrated across formats for being usable by any application or tool.

This involves pushing standardisation forward, building on the achievements that result from years of research. These currently show substantial convergence of opinion and practice, which needs now to be supported. For instance, standards now need tools that support them – this will promote and ensure their adoption. Not only are data formats to be standardised, but also criteria for annotating and producing language resources. The availability of common annotation guidelines and specifications is perceived as a viable solution to current problems in the production of language resources, such as efficiency, quality and interoperability.

Invest in automatic techniques for language resource production

Most language technologies and applications rely on language resources: we must devote more efforts to solve how to automate the production of the large quantity of resources demanded, and of enough quality to get acceptable results in industrial environments.

To guarantee the ability to reach and maintain the necessary quantity of language resources – and their annotation at different levels of complexity – the community must look for techniques to automate the production of resources, to produce large quantities, for all possible domains, for any language, and of the quality necessary to get good results. There is also a need for considering automatic production techniques as components that are usable for industrial applications.

Successful applications, in their turn, will lead to the creation of new and/or access to existing language resources, by verticalisation (adaptation to specific domains) or customization, and extension of language coverage.

Coordinate efforts

Coordination of efforts and initiatives has been repeatedly identified as a key success factor and probably a definitive measure for a substantial leap forward of the field of Language Resources and Technologies.

Coordination is needed at all levels, strategic, political, industrial and academic, and for various aspects related to Language Resources and Technologies, from resource creation to maintenance and evaluation.

The organisation of cooperative/collaborative exercises for building specific large resources, also multilingual ones, was proposed more than once as the *modus operandi* in the future. A compelling case was made for adopting 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.

Cooperation and synergies with other research areas and other economic sectors (content producers) should be also encouraged as a mean to produce Language Resources. It was also stated that market forces will only address languages and areas whose market guarantees a return of investment; which makes a coordinated policy at a European level essential if one wants Human Language Technology to be deployed equally for all languages and countries.

Push evaluation forward

A recurrent issue addressed is that good quality of language technology and applications is essential for making a profitable business, and good quality language technology and applications is (*inter alia*) dependent on the availability of good quality, huge language resources.

Evaluation is a necessary corollary for the advancement of language technology. In the EU, we are still missing a permanent framework to take care of language technology evaluation in a multilingual environment, while it is a recognised difficulty to address the evaluation of all technologies for all languages. The need to establish a permanent public entity for evaluation at EU level was raised.

Overcome current ways of thinking Language Resources and Technologies

We may have reached a point where the traditional notion of language resources needs to be substantially re-thought. New paradigms of Language Resource creation and development are emerging, such as collaborative and social methods.

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 for the whole range of European languages, and make these data available at affordable prices for research purposes and to SMEs. The costs of creating such data, it was claimed, cannot be carried by individual SMEs, and not even by cooperating SMEs, so that government support is called for.

From the point of view of the market, re-thinking the concepts of Language Resources and Technologies means to shift from solutions to *services on demand*. In its turn, this imposes contextual requirements, including an infrastructure, public policies on e-government services, legislation for the adoption of such services, and customer education. Public procurement, especially at the EU level, should be used as one of the instruments to boost production and adoption of language technologies.

Emerging Priorities

A clear *set of priorities* emerged *for fostering the field* of Language Resources and Language Technology. FLaReNet must *see where and how each of these viewpoints informs the roadmap for language technology research and development*, rather than seeing them as alternatives from which we must choose.

Language Resource creation

The effort required to build all needed language resources and common tools should impose on all *players a strong cooperation at the international level* and the community should define how to *enhance current coordination of language resource collection between all involved agencies* and ensure efficiency (e.g. through interoperability).

With data-driven methods dominating the current paradigms, *language resource building, annotation, cataloguing, accessibility, availability is what the research community is calling for*. Major institutional translation services, holding large volumes of useful data, seem to be ready to share their data and FLaReNet could possibly play a facilitating role.

More efforts should be devoted *to solve how to automate the production of the large quantity of resources demanded, and of enough quality to get acceptable results in industrial environments.*

Standards and Interoperability

In the long term, *interoperability will be the cornerstone of a global network of language processing capabilities.* The time and circumstances are ripe to take a broad and forward-looking view in order to establish and implement the standards and technologies necessary to ensure language resource interoperability in the future. This can only be achieved through a *coordinated, community-wide effort that will ensure both comprehensive coverage and widespread acceptance.*

Coordination of Language Technology Evaluation

Looking at the way forward, it clearly appears that *language technology evaluation needs coordination at international level*: in order to ensure the link between technologies and applications, between evaluation campaigns and projects, in order to conduct evaluation campaigns (for ensuring synchrony or for addressing the influence of a component on a system on the same data), in order to produce language resources from language technology evaluation, or to port an already evaluated language technology to other languages (best practices, tools, metrics, protocols…), in order to avoid "reinventing the wheel", while being very cautious that there are language and cultural specificities which have to be taken into account (tone languages, oral languages with no writing system, etc.).

Availability of Resources, Tools and Information

Infrastructure building seems to be one of the main messages for FLaReNet. *For a new worldwide language infrastructure the issue of access to Language Resources and Technologies is a critical one* that should involve – and have impact on – all the community. There is the need to create the means to plug together different Language Resources & Language Technologies, in an *internet-based resource and technology grid*, with the possibility to easily create new workflows. Related to this is *openness and availability of information.* The related issues of *access rights and IPR* also call for cooperation.

Next FLaReNet Actions

Actions for FLaReNet to ensure involvement of a broad – and committed – community are:

- FLaReNet can use its collaborative website to create a think-tank to have a joint reflection, see what can be initiated, how, with whom, and help in creating collaboration possibilities;
- FLaReNet has good practice of standardisation activities and can promote and help in the standardisation-oriented tasks and efforts toward harmonisation, sharing and distribution;
- FLaReNet is going to take the lead in assembling relevant people, institutions, and organisations around the world into a collaborative network to which the institutions and individuals involved are committed (and really, have funding for) whose goal is to collaboratively work toward proper Language Resource coverage, interoperability and Language Technology evaluation, for all corresponding languages;
- FLaReNet will formally promote a new worldwide language infrastructure for easy access to Language Resource and Technologies, in a web-based resource and technology grid. It can even concretely start acting towards this by e.g. exploiting the *LREC Conference* and the *Language Resource and Evaluation* Journal;
- FLaReNet can be the promoter of a communication vector for open source resources and tools: this could be in wiki mode.
- FLaReNet will produce a White Paper summarising ideas for directors of programs of funding agencies, and organise a Forum of directors of funding agencies;
- FLaReNet must establish an International Advisory Board: this group can constitute the nucleus of the Advisory Board and act as the needed International Forum;
- The FLaReNet Advisory Board/International Forum will prepare a Memorandum of Understanding with the main issues discussed and ask members of FLaReNet to sign it when joining the Network.

FLaReNet Steering Committee
Nicoletta Calzolari (ILC-CNR, IT, *Coordinator*)
Núria Bel (Universitat Pompeu Fabra, SP)
Gerhard Budin (Universität Wien, AT)
Khalid Choukri (ELDA, FR)
Joseph Mariani (LIMSI/IMMI-CNRS, FR)
Jan Odijk (Universiteit Utrecht, NL)
Stelios Piperidis (ILSP / "Athena" R. C., GR)

European Commission - DG Information Society & Media - Unit INFSO.E1 - LTs & MT
Roberto Cencioni (*Head of Unit*)
Kimmo Rossi (*FLaReNet Project Officer*)

Speakers
Josep Bonet-Heras (EC - DG Translation, LUX)
Branimir Boguraev (IBM Research, USA)
Nick Campbell (Trinity College Dublin, IRL & NIST, JP)
Key-Sun Choi (KAIST, KR)
Christopher Cieri (University of Pennsylvania - LDC, USA
Thierry Declerck (DFKI, DE)
Marcello Federico (FBK, IT)
Josef van Genabith (Dublin City University - NCLT, IRL)
Edouard Geoffrois (DGA, FR)
Dafydd Gibbon (Universität Bielefeld, DE)
Gregory Grefenstette (Exalead, FR)
Iryna Gurevych (Technische Universität Darmstadt - UKP Lab, DE)
Tony Hartley (University of Leeds, UK)
Henk van den Heuvel (Radboud University Nijmegen, NL)
Harald Höge (SVOX Deutschland GmbH, DE)
Nancy Ide (Vassar College - DCS, USA)
Andrew Joscelyne (TAUS, FR)
Anna Korhonen (University of Cambridge, UK)
Steven Krauwer (Universiteit Utrecht, NL)
Jimmy Kunzmann (European Media Laboratory GmbH, DE)
Gianni Lazzari (PERVOICE S.p.A., IT)
Walther Lichem (Former Ambassador of the Republic of Austria)
Edward Loper (Brandeis University, USA)
Bente Maegaard (University of Copenhagen - CST, DK)
Bernardo Magnini (FBK, IT)
Gudrun Magnusdóttir (ESTeam, SE)
Asunción Moreno (Universitat Politècnica de Catalunya, SP)
Eric Nyberg (Carnegie Mellon University, USA)
Patrick Paroubek (LIMSI-CNRS, FR)
Carol Peters (ISTI-CNR, IT)
Alexandros Poulis (EP - DG Translation - IT Support Unit, LUX)
Gábor Prószéky (MorphoLogic, HU)
James Pustejovsky (Brandeis University - DCS, USA)
Justus Roux (University of Stellenbosch, South Africa)
Marta Sabou (Open University, UK)
Florian Schiel (Ludwig Maximilian Universität München - BAS, DE)
Gary Strong (Johns Hopkins University - HLT Center of Excellence, USA)
Gregor Thurmair (Linguatec, DE)
Jun'ichi Tsujii (University of Manchester - NacTeM, UK)
Dan Ioan Tufiș (RACAI, RO)
Kiyotaka Uchimoto (NICT, JP)
Hans Uszkoreit (DFKI, DE)
Cristina Vertan (Universität Hamburg, DE)
Yorick Wilks (University of Sheffield, UK)

Peter Wittenburg (MPG, NL)
Pierre Zweigenbaum (LIMSI-CNRS, FR)

Discussants
Sophia Ananiadou (University of Manchester - NacTeM, UK)
Luisa Bentivogli (FBK, IT)
Bob Boelhouwer (Instituut voor Nederlandse Lexicologie, NL)
Guy De Pauw (University of Antwerp, BE)
Tomaž Erjavec (Jožef Stefan Institute, SI)
Martine Garnier-Rizet (VECSYS, FR & IMMI-CNRS, FR)
Timo Honkela (Helsinki University of Technology - CIS, FI)
Chu-Ren Huang (Hong Kong Polytechnic University, HK)
Margaretha Mazura (European Multimedia Forum, BE)
Djamel Mostefa (ELDA, FR)
Yohei Murakami (NICT, JP)
Nelleke Oostdijk (Radboud University Nijmegen - DL, NL)
Adam Przepiórkowski (Polish Academy of Sciences - ICS, PL)
Kepa Sarasola Gabiola (University of the Basque Country - IXA Group, SP)
Kiril Simov (LML-IPP-BAS, BG)
Harold Somers (Dublin City University - SC, IRL)
Marko Tadić (University of Zagreb - FHSS - DL, HR)
Frank Van Eynde (Katholieke Universiteit Leuven - CCL, NL)
Folkert de Vriend (Nederlandse Taalunie, NL-BE)

*FLaReNet Coordination Group (*ILC-CNR, IT)
Paola Baroni
Sara Goggi
Monica Monachini
Valeria Quochi
Claudia Soria
Antonio Toral