The European Language Resources and Technologies Forum

Shaping the Future of the Multilingual Digital Europe

Vienna, 12 -13 February 2009

Extended 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 Web site, 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 Proceedings of the FLaReNet Forum as well as a short version of this report are available on the FLaReNet Web site.

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.

Thursday 12th February 2009

Opening Session – 10:00-11:00
Chair: Nicoletta Calzolari

Roberto Cencioni (EC - DG Information Society & Media - Unit INFSO.E1 - LTs & MT, LUX / *Head of Unit*)
Walther Lichem (*Former Ambassador of the Republic of Austria*)
Nicoletta Calzolari (ILC-CNR, IT / *FLaReNet Coordinator*)
Gerhard Budin (Universität Wien, A / *FLaReNet Local Host*)

S1. Broadening the Coverage, Addressing the Gaps – 11:30-13:30
Chair: Joseph Mariani - *Rapporteur*: Khalid Choukri

Introduction by the Chair
Talks:
Steven Krauwer (Universiteit Utrecht, NL) & Khalid Choukri (ELDA, FR), *"Coverage & BLARKS"*
Christopher Cieri (University of Pennsylvania - LDC, USA), *"Practical Considerations in Resource Creation Tied to Human Language Technology Development"*
Justus Roux (University of Stellenbosch, South Africa), *"An African Perspective on Language Resources and Technologies"*
Dafydd Gibbon (Universität Bielefeld, DE), *"Coverage of What? – Gaps in What? On De-globalising Human Language Resources"*
Asunción Moreno (Universitat Politècnica de Catalunya, SP), *"Shared Language Resources Production"*
Pierre Zweigenbaum (LIMSI-CNRS, FR), *,"A Dynamic View of Comparable and Specialized Corpora"*
Nick Campbell (Trinity College Dublin, IRL & NIST, JP), *"Technology for Processing Non-verbal Information in Speech"*
Discussants
Adam Przepiórkowski (Polish Academy of Sciences - ICS, PL)
Marko Tadić (University of Zagreb - FHSS - DL, HR)
Kepa Sarasola Gabiola (University of the Basque Country - IXA Group, SP)
Folkert de Vriend (Nederlandse Taalunie, NL-BE)

S2. Automatic and Innovative Means of Acquisition, Annotation, Indexing – 14:30-16:30
Chair: Stelios Piperidis - *Rapporteur*: Núria Bel

Introduction by the Chair/Rapporteur
Talks:
Jun'ichi Tsujii (University of Manchester - NacTeM, UK), *"Richly Annotated Corpora and Their Inter-operability"*
Yorick Wilks (University of Sheffield, UK), *"Dialogue corpora remain a problem."*
Gary Strong (Johns Hopkins University - HLT Center of Excellence, USA), *"Trends in Language Resources and New Work in ASR Data Labeling"*
Dan Ioan Tufiș (RACAI, RO), *"Going for a Hunt? Don't Forget the Bullets!"*
Anna Korhonen (University of Cambridge, UK), *"Automatic Lexical Acquisition - Bridging Research and Practice"*
Gregory Grefenstette (Exalead, FR), *"The Democratisation of Language Resources"*
Marta Sabou (Open University, UK), *"Web3.0 and Language Resources"*
Iryna Gurevych (Technische Universität Darmstadt - UKP Lab, DE), *"Exploiting Crowdsourced Language Resources for Natural Language Processing: 'Wikabularies' and the Like"*
Discussants
Kiril Simov (LML-IPP-BAS, BG)
Sophia Ananiadou (University of Manchester - NacTeM, UK)
Guy De Pauw (University of Antwerp, BE)

S3. Evaluation and Validation – 16:45-18:30
Chair: Jan Odijk - *Rapporteur:* Joseph Mariani

Introduction by the Chair/Rapporteur
Talks:
Henk van den Heuvel (Radboud University Nijmegen, NL), *"The 'Standard Deviation' of LR Quality"*
Carol Peters (ISTI-CNR, IT), *"Evaluation of Technology for Multilingual Information Access: the Next Step"*
Bente Maegaard (University of Copenhagen - CST, DK), *"Can Evaluation Be Application-Independent?"*
Edouard Geoffrois (DGA, FR), *"Language Technology Evaluation: which Funding Strategy?"*
Bernardo Magnini (FBK, IT), *"Toward an Integrated Evaluation Framework"*
Patrick Paroubek (LIMSI-CNRS, FR), *"Evaluation: a Paradigm that Produces High Quality Language Resources"*
Harald Höge (SVOX Deutschland GmbH, DE), *"A Proposal to Launch a Support Centre for 'Remote' Evaluation and Development of Language Technologies"*
Cristina Vertan (Universität Hamburg, DE), *"Evaluation of HLT-Tools for Less Spoken Languages"*
Discussants
Djamel Mostefa (ELDA, FR)

Nelleke Oostdijk (Radboud University Nijmegen - DL, NL)
Luisa Bentivogli (FBK, IT)

Friday 13th February 2009

S4. Interoperability and Standards – 9:00-10:45

Chair: James Pustejovsky - *Rapporteur:* Nancy Ide

Introduction by the Chair/Rapporteur

Talks:

James Pustejovsky (Brandeis University - DCS, USA) & Nancy Ide (Vassar College - DCS, USA), *"SILT: Towards Sustainable Interoperability for Language Technology"*
Eric Nyberg (Carnegie Mellon University, USA), *"Interoperability, Standards and Open Advancement"*
Peter Wittenburg (MPG, NL), *"Is the LRT Field Mature Enough for Standards?"*
Edward Loper (Brandeis University, USA), *"Interoperability via Transforms"*
Key-Sun Choi (KAIST, KR), *"Ontology of Language Resource and Tools for Goal-oriented Functional Interoperability"*
Thierry Declerck (DFKI, DE), *"Towards Interoperability of Language Resources and Technologies (LRT) with Other Resources and Technologies"*

Discussants

Tomaž Erjavec (Jožef Stefan Institute, SI)
Chu-Ren Huang (Hong Kong Polytechnic University, HK)
Timo Honkela (Helsinki University of Technology - CIS, FI)
Yohei Murakami (NICT, JP)

S5. Translation, Localisation, Multilingualism – 11:00-12:45

Chair: Gerhard Budin - *Rapporteur:* Stelios Piperidis

Introduction by the Chair/Rapporteur

Talks:

Hans Uszkoreit (DFKI, DE), "*Language Resources and Tools for Machine Translation: Trends, Demands, Predictions*"
Marcello Federico (FBK, IT), *"Outlook for Spoken Language Translation"*
Josef van Genabith (Dublin City University - NCLT, IRL), *"Three Challenges for Localisation"*
Tony Hartley (University of Leeds, UK), *"Assessing User Satisfaction with Embedded MT"*
Josep Bonet-Heras (EC - DG Translation, LUX), *"Institutional Translators and LRT"*
Alexandros Poulis (EP - DG Translation - IT Support Unit, LUX), *"Language Technology in the European Parliament's Directorate General for Translation: Facts, Problems and Visions"*
Andrew Joscelyne (TAUS, FR), *" 'Cloud Sourcing' for the Translation Industry"*

Discussants

Frank Van Eynde (Katholieke Universiteit Leuven - CCL, NL)
Harold Somers (Dublin City University - SC, IRL)

S6. Enhancing Market Places/Models for Lrs: New Challenges, New Services – 13:45-15:15

Chair: Khalid Choukri - *Rapporteur:* Jan Odijk

Introduction by the *Chair/Rapporteur*

Talks:

Gregor Thurmair (Linguatec, DE), *"No Resources Without Applications"*
Gianni Lazzari (PERVOICE S.p.A., IT), *"Buy a License or Pay for Service?"*
Gudrun Magnusdóttir (ESTeam, SE), *"Cheap or Expensive - What Works?"*
Gábor Prószéky (MorphoLogic, HU), *"Enhancing HLT Market with Cooperative Services"*
Jimmy Kunzmann (European Media Laboratory GmbH, DE), *"Speech-to-Text Solutions for the European Market: a SME View to Language Scalability"*

Discussants

Bob Boelhouwer (Instituut voor Nederlandse Lexicologie, NL)
Martine Garnier-Rizet (VECSYS, FR & IMMI-CNRS, FR)
Margaretha Mazura (European Multimedia Forum, BE)

Closing Session – 15:15-16:30

Chair: Nicoletta Calzolari

FLaReNet Sessions Rapporteurs

S1. Khalid Choukri (ELDA, FR)

S2: Núria Bel (Universitat Pompeu Fabra, SP)
S3. Joseph Mariani (LIMSI/IMMI-CNRS, FR)
S4. Nancy Ide (Vassar College - DCS, USA)
S5. Stelios Piperidis (ILSP / "Athena" R. C., GR)
S6. Jan Odijk (Universiteit Utrecht, NL)
Nicoletta Calzolari (ILC-CNR, IT)
Kimmo Rossi (EC - DG Information Society & Media - Unit INFSO.E1 - LTs & MT, LUX / *FLaReNet Project Officer*)
Roberto Cencioni (EC - DG Information Society & Media - Unit INFSO.E1 - LTs & MT, LUX / *Head of Unit*)

International Cooperation Round Table – 16:30-18:30

Chair: Nicoletta Calzolari

Participants

Nancy Ide (Vassar College - DCS, USA)
James Pustejovsky (Brandeis University - DCS, USA)
Gary Strong (Johns Hopkins University - HLT Center of Excellence, USA)
Jun'ichi Tsujii (University of Manchester - NacTeM, UK)
Christopher Cieri (University of Pennsylvania - LDC, USA)
Branimir Boguraev (IBM Research, USA)
Key-Sun Choi (KAIST, KR)
Nick Campbell (Trinity College Dublin, IRL & NIST, JP)
Eric Nyberg (Carnegie Mellon University, USA)
Kiyotaka Uchimoto (NICT, JP)
Chu-Ren Huang (Hong Kong Polytechnic University, HK)
Margaretha Mazura (European Multimedia Forum, BE)
Justus Roux (University of Stellenbosch, S. AFRICA)
Hans Uszkoreit (DFKI, DE)
Yohei Murakami (NICT, JP)
European Commission - DG Information Society & Media - Unit INFSO.E1 - LTs & MT:
Roberto Cencioni (*Head of Unit*)
Kimmo Rossi (*FLaReNet Project Officer*)
FLaReNet Steering Committee:
Nicoletta Calzolari (ILC-CNR, IT)
Khalid Choukri (ELDA, FR)
Stelios Piperidis (ILSP / "Athena" R. C., GR)
Gerhard Budin (Universität Wien, AT)
Jan Odijk (Universiteit Utrecht, NL)
Núria Bel (Universitat Pompeu Fabra, SP)
Joseph Mariani (LIMSI/IMMI-CNRS, FR)

Executive Summary

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 LR 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 Web site 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 LR 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.

S1 – Broadening the Coverage, Addressing the Gaps

Introduction: overview, rationale

In a Multilingual Digital Europe, large coverage of languages and of the major economic/social/cultural sectors should be ensured as the first priority. This can only be achieved through the supply of numerous applications and technologies, which should be fed with the necessary language resources (multilingual, multimedia, multimodal). If language resources are missing for a language, then the development of language applications are seriously hampered. To this end it is important that all languages are equipped with a minimum set of resources that are considered indispensable for application development. This is the concept behind the BLARK (Basic LAnguage Resource Kit): the BLARK and its companion ELARK (Extended LAnguage Resource Kit) are matrices depicting, for every language, the minimum set of language resources (in terms of corpora, lexicons, basic tools to manipulate them, skills required, etc.) required to do any pre-competitive research for that language

Under BLARK terms, there is considerable variation across languages as for the quantity and type of language resources available. The first session focused on Language Resource coverage and the gaps that need to be filled. It was structured around two dimensions: **Language Coverage** and **Topic Coverage**. The primary objective was to draw a picture of the current landscape, to assess the current coverage for the various languages (national, regional, etc.) and for the various modalities (text, speech, gesture), and to evaluate the efforts to fulfil the major users' requirements.

Highlights, specific problems/issues

From the views of the major data centres (ELRA, LDC) and infrastructures (CLARIN) about existing resources, it is clear that **essential resources are still missing even for the major EU languages and the major topics**[1].

The use of BLARK makes it easy to spot the gaps to be filled with regards to language resources and technologies needed for specific applications and for as many languages as possible. Not only it helps identifying available language resources, or enabling customers or providers of language resources to fill gaps, but in particular it may have a "political" use as an instrument for coverage assessment, roadmapping and language policy planning: to promote the production of new language resources and to present to funding agencies the gaps to be filled for a language to be considered technologically mature.

A crucial question is **the cost of a BLARK for a given language**. Examples from an US project on the Less Commonly Taught Language show that this is close to 1M$/Language (including the set up of a complex infrastructure to harvest data and process it). If one considers the European scene and the difference in the manpower costs between countries, differences in the status of regional languages, linguistic properties and language similarities, and availability of 'raw' resources, one could derive the BLARK costs for the needed resources. A **huge effort would be required for all the EU languages**, as needed in the multilingual Europe.

It is also crucial to **tie resource creation to the development of technologies.** It is mandatory to **produce the basic tools to process the 'raw' data**. Again a standard performance "baseline" is likely to result in very different costs per language. It is also important to stress that having common tools and clear specifications would help share the work between partners joining efforts while relying on some found resources.

The item on Language coverage was exemplified by the African case, where recent years have seen **a renewed linguistic and cultural awareness on indigenous languages.** This has to be channelled towards the development of resources useful to language technologies. But **also in Europe the language resources and technology situation is very unbalanced** and new languages that joined recently the Union should be considered as a higher priority in coming EU programs.

Referring to the globalization process, it was clearly stressed that we need to **"de-globalize" human language resources** and focus on local languages/cultures despite today's "global" village. Economically wealthy countries bear a responsibility towards our social and scientific environment and should carefully

[1] For instance, it was mentioned that French has, publicly available, about 100h of transcribed broadcast news speech produced thanks to a national effort, to be compared with Arabic (1200 h) or Mandarin (1300 h), produced thanks to a US-funded effort. This exemplifies the rationale of public support devoted to the production of language resources, should it be for cultural and societal reasons, or for geopolitical ones.

consider the language sensitive issues and the relationship to underprivileged colleagues (cooperation, sponsorship, but also develop and share tools, help them meet the international level).

Finally, also **the need to process non-verbal, and more generally contextual information encompassed in speech-based interaction** was presented. This requires exploiting non-verbal features, depicting emotion, affect, interest, etc. and contextual features, such as goal, trust, belief, etc.: these are key elements of such resources, very much neglected[2]. This is essential for R&D activities in the long run, and would boost technology performance, for all applications where human-machine interaction is considered in a real-life "wider" context.

Suggested solutions

It is crucial that such resource needs get prioritized, accounting for the requirements of novel research areas and innovative applications. It is therefore important to have an authoritative and broadly accepted definition of the BLARK and established mechanisms (and funding) for the creation or completion of BLARKs for each language. The Web site, established by ELRA at www.blark.org, could serve as a starting point. The BLARK/ELARK concepts should be regularly redefined and updated as these notions evolve with the technology landscape (technology offer for R&D/Innovation needs), bearing in mind that current technology "baselines" must be established per language, per application, etc. with a clear picture of important barriers and threats.

A few ideas were mentioned for supporting resource production: from getting support from the international/regional organizations (UN, UNESCO, EU), national agencies, private sector players, etc. to find means to set up an International Investment Fund for language resources. It is also important to try to involve private foundations (such as B&M Gates, Google, Qatar IT) in this production process (also for minority languages). **Cooperation and synergies with other research areas and other economic sectors (content producers)** should be also encouraged as a means to produce language resources.

Given the effort and the available investments, **coordination (on methodologies, best practices, standards, interoperability…) should be seriously encouraged.**

In order to highlight some of the production processes that were widely used within the EU, description of the **"shared language resource production model"** was introduced. The basic idea is to form consortia with players that have similar interests and that would produce one database/language each. If the database complies with the pre-defined quality then the database owner will exchange it with the databases coming from all the other members of the consortium that have to produce a database of comparable quality. Consortia allow producers to share costs and join efforts for the production process. This model could be extended for a more general scheme, involving international bodies, individual countries, regional governments, based on their interest to support a given language or set of languages.

Other production approaches were discussed and some felt that we should **envisage new language resource production paradigms, using web2.0, Wiki, social networks,** enough to fill all possible BLARKs[3].

Given the shortage or unavailability of "parallel corpora" in multiple languages used for addressing cross-lingual issues, it was addressed the need to establish a measure of the similarity of texts prior to their alignment. This similarity measure is actually needed in many language applications (machine translation evaluation, source detection, etc.).

The "*dynamic*" aspect of language resources was also stressed, together with the need to build constantly updated text collections, and on-demand selection of subsets, carefully specified to suit applications and users' needs.

The **legal and IPR issues** are other crucial topics. In addition to the current good practices advocated for by the major data centres, it is also suggested to work towards a simplification of the current European legal framework through the introduction of a **"Research Fair Act"** as is the case in the USA. This requires a legislative procedure within the European Union parliament and a lot of lobbying efforts.

[2] The cost and time-effort for a Japanese resource of this type is about 5M$ and over 5 years of recordings for the first data.

[3] In this spirit, it was proposed for example to devote a collaborative effort to produce a basic language resource for all languages consisting in word forms and the corresponding lemmas, part-of-speech tags, frequencies, etc. and could also include translations. This would be Open Source, no license, wiki editable, simple format, widely available, and at low production cost (10 cents/word ≈ 10 K€ for one language).

Lessons for FLaReNet, next steps

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**). It is therefore important to ensure coordination in monitoring the various international programs but also ensure that each is conducting sound **evaluations** to assess the progress being made in filling the identified gaps.

Another important consideration is to identify and **promote applications/technologies of "greatest exposure"** that incorporate **multilingual** aspects, boost sectors of activities than can be "early/today" adopters and use them as window-dressing for language technology to convince decision makers (both the politicians and the financers). **Public procurement, especially at the EU level, should be used as one of the instruments to boost production and adoption of language technologies.**

Many production projects benefit from some public funding and 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.

S2 – Automatic and Innovative Means of Acquisition, Annotation and Indexing of Language Resources

Introduction: overview, rationale

The use of language technologies can be looked at as a source of competitive advantage, especially if they are considered as general purpose technologies that can add value to most ICT products dealing with language in whatever manifestation. But multilingual technologies are located on the production-side of the economic equation: they are intermediate products used to produce final goods and services, and therefore they are valued for what they actually do. And what language technology based applications currently do is hampered by the fact that eventually they fail when they need to cover a new word, or a new domain, or a new language.

Language resources are the necessary ingredients for any language application, but their production is long and expensive; when they are ready, they are already "old".

An additional challenge to the robustness, coverage and performance of the tools and applications is presented by the current language use on the various Web communities, social networks, blogs and the like, which show some differences from the "standard" written language, and therefore pose problems to current automatic language processing applications. Moreover, language on the Web and on other information and communication platforms (radio, TV, etc.), converging today through advances in telecommunications engineering, is tightly interlinked to other media, notably images, video and sounds. The unavailability of the appropriate resources is a hindering factor for systems and application development and full deployment.

As also shown/pointed out in Session 1, most languages still need basic language resources.

It is therefore of uttermost importance to develop and deploy methods for an automatic construction, linking and repurposing of new and existing language resources that can satisfy such demand. Automatic methods of language resource production are likely to offer a solution to the production bottleneck.

The main goal of this session was to identify the key points about current and future production of language resources. Language data is now a core component of Human Language Technologies and the supply of data in the quantities required depends on the ability of the community of finding new methods and models for building, validating and maintaining these data. Those applications based on techniques that heavily rely on the availability of enough language resources will not be able to cope with new needs if we cannot guarantee their "sustainability": i.e. the capacity to reach and maintain the necessary quantity and quality of language resources for LT applications to properly and equally work for any language, any domain and any genre. Questions related to the requisites that the huge volume of data and its actual coverage pose on the production methods, whether automatic methods are ready to be used, and to what extend the existence of standards could affect positively the reduction of the production costs are some critical issues.

Highlights, specific problems/issues

The issue "quantity vs. quality" was the more controversial and recurring topic. It comes out that is still difficult is to have large quantities of data of high quality due to the costs of production. However, the prioritization of different action lines according to the actual requirements of the different applications divides the community, because acquisition and production of both types of resources have specific problems. Those applications that require large quantities of textual data face the coverage problem: of languages, of genres, etc. Even more, certain types of resources will hardly be available on the Web – an important source of language data today – dialogues being a notable example.

Although the Web cannot be the source of a full range of LRs, it was proposed that, with proper design, it might be used by organized communities to cooperatively build large specific language resources. On the other hand, the quality of current community built resources, such as Wiktionary, is still questioned. And since real industrial applications require materials of sufficient good quality, the question was also raised about how quality can be guaranteed in such a cooperative scenario if non experts are involved, especially when LRs involving semantic encoding are needed. The need for producing good, highly accurate resources must be also taken seriously as different experiments show that good annotated data can deliver better results than just more raw data, but tools for improving the performance of these high quality annotation tools must be developed. This seems to be an area where strong and focussed research needs to be fostered and coordinated.

Suggested solutions

There were different suggestions **for ensuring the production of good quality language resources**. Here they are given in order of priority.

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. **The recommendations for future actions** are to go for a further effort in the **standardisation not only of formats but also of criteria for annotating and producing language resources**, taking into account that some of the necessary annotations are not exclusively linguistic.

From the current necessities, it follows that, in order to guarantee sustainability, 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. In that respect, there is a need for building automatic production techniques as components that are usable for industrial applications. The results of such production components should be evaluated against their use in specific tasks. Besides, more research should be devoted to the particular problem of handling an evolving and changing object such as natural language, because normally automatic acquisition techniques take as evaluation material old compiled source materials that do not reflect current use.

The future actions suggested were to look for sources other than the internet. However, new knowledge components and structured repositories of knowledge are being developed in the construction of the future internet (Web 3.0), which can be useful for the production of new language resources.

Lessons for FLaReNet, next steps

The main conclusion is that there are many technologies and applications that rely on language resources right now and that others count on them for future techniques and applications. These current and future necessities reinforce the idea that **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**.

S3 – Evaluation & Validation

Introduction: overview, rationale

The session dealt with two topics: (i) **Validation of Language Resources**, i.e. evaluation and validation of the quality and quantity of language resources produced for a given objective (conduct research investigations, develop a product, etc), and (ii) **Evaluation of Language Technologies**, i.e. evaluation of language technologies and production/distribution of the language resources necessary for developing and testing the corresponding language technologies.

Both topics were addressed with the objective to share good and bad findings, based on experience gained in language resource validation and language technology evaluation and look into the future: Are there new needs, new trends?

Validation of Language Resources concerns the evaluation and validation of the quality and quantity of LRs which are produced for a given objective (conduct research investigations, develop a system, a product, etc). The size of the LRs should be large enough in order to provide a good representation of all the research issues which are to be studied. Or to offer results closest to the optimal achievable target, if it used to train systems. Similarly, the content of the LR should be in agreement with its description, and specifications.

This is especially important for LRs which are used for developing or testing a system. Automatic language processing systems based on statistical methods are trained on a set of two LRs, as input and output (for example the speech signal and its written transcription, or a parallel text in the source and target languages). The system will automatic adjust the parameters of its underlying model in order to reflect the way to transform the input data in the output data, for the training data and later on for the data which will have to be processed in operational conditions, and which is hoped to be well represented by the training data. If there is a mismatch between the input and output data, the parameters will be wrong and the system performances may be poor. In the test phase, the systems are fed with input data, and their performance is compared with a reference corpus corresponding to the input data. If there are errors in the reference, the performance of the systems will be wrongly measured. However, the designers of those systems usually carefully check if the reference corpus is correct during the adjudication phase, if it exists.

LR validation is usually conducted by validation centres which should be different than the centre that produces the LR. The availability of a dedicated infrastructure could facilitate the process in a distributed way.

Evaluation of Language Technologies concerns the evaluation of LTs, including the production and distribution of the LRs which are necessary for developing and testing the corresponding LT (see above). Comparative evaluation of language processing systems has been proposed in the US by Darpa with the support of NIST for organization and LDC for LR production in 1987, after a large previous US program didn't allow for assessing and comparing the quality of systems developed in its framework, as they addressed different tasks.

Comparative evaluation allows for measuring the performance of a system regarding a specific task, comparing the performances of the different systems tested on the same data, therefore providing some cues on the advantages and drawbacks of the different theoretical backgrounds and approaches adopted in each system, and, by considering the performances of the best system, evaluating the state-of-the-art of technology for a given task. It also allows for measuring the progress which has been achieved by a research team or by a research community in the period of time between two consecutive evaluation exercises. The initial use of comparative evaluation was in the area of speech recognition. It was then adopted in many areas of spoken and written language processing, and more generally of intelligent artificial systems. Conducting an evaluation campaign necessitates the choice of a metrics for measuring the performance of a system (which in some cases (such as Machine Translation) is a research topic per se), the determination of the evaluation protocol, the choice, production and distribution of training and testing data, the report on results and the share of results and corresponding methods among the participants. After the evaluation campaigns, evaluation packages comprising the data, the scoring and the results may be distributed in order to allow other parties to compare themselves with the state-of-the-art.

It clearly appears that **Language Technology Evaluation is now a pre-requisite**. It is nowadays widely accepted, especially in the research community, that paper proposals are most often rejected if they don't contain serious evaluation of the results (which may bring a bias in some cases). One may think that in the

future no contract will be awarded if it doesn't contain a serious evaluation, or even a prior evaluation of the technologies provided to the project by the partners.

Evaluation campaigns are presently mostly organized by a central entity (such as NIST). It could also be distributed in the future, if an adequate infrastructure exists.

There is an excellent 10-year experience of a European initiative supported by the EC on Cross-Language Information Retrieval by the CLEF project, which also includes the preparation and distribution of evaluation packages after the evaluation campaigns. The CLEF effort is already coordinated at the international level with other well-known information retrieval system evaluation initiatives such as TREC in the US, and NTCIR in Japan. However, CLEF focuses specifically on multilingual and cross-language issues, which are of particular interest for Europe.

Highlights, specific problems/issues

Regarding **language resource validation**, it was said that the link between validation and standardisation will become stronger, and that the keys to standardisation are the definition of appropriate metadata sets and automatic content generation. **It was stressed that language resource validation will increasingly take the shape of language technology evaluation.** In connection to that, it is important to develop tools for validation (fault detection (clipping, noise...), detection of segmentation errors, of weak annotations, confidence measures of speech transcriptions...), which also have the nice feature, compared with humans, to bring consistency.

Three main obstacles and problems were identified for language technology evaluation:

1) **In Europe we are missing a permanent effort framework to take care of language technology evaluation**, taking the comparison between CLEF (which is a series of limited duration projects) and NIST or NTCIR (permanent evaluation bodies financed by the federal or national government). Also, it would be needed to fully support (100% funding) the evaluation activity, as it is of infrastructural and non-for profit nature, even though evaluation campaign organizers and participants show a lot of enthusiasm. While the usual project funding scheme is usually based on 50% public funding in Europe.
2) A second problem relates to the **difficulty to address the evaluation of all technologies for all languages**. This is a specific need for the Member States which recently joined the European Union, and which may not profit from the language resources and language technologies produced and made available in previous Framework Programs, and for regional languages, which do not benefit from data produced for the EU "official" languages (as it is the case for Europarl). The **effort appears to be very important and costly** if we consider the number of technologies and of languages. The **"cultural" dimension** adds to this question, and it was stressed that the **localisation** activity doesn't need only cross-lingual technologies, such as machine translation or translation memories, but also many monolingual technologies, such as spelling checkers, grammar checkers, etc.
3) A third problem identified for language technology evaluation is related to the relationship between **Task-oriented evaluation** (also called Technology, Module or Component-oriented) **versus Application-oriented evaluation** (also called Usage or User-oriented):
    - Task-oriented evaluation allows for monitoring research directions but doesn't give the overall system performance and doesn't assess the influence of each component in the system;
    - While Application-oriented evaluation provides the overall system performance but doesn't provide feedback for choosing between theoretical approaches.

Suggested solutions

For addressing these problems, several solutions are proposed:

Regarding the issue of a permanent effort framework, the current EU funding policy should be revised to take into account evaluation activities.

As far as the difficulty to address the evaluation of all technologies for all languages is concerned, solutions may be various and at different level.

They range from more "political" ones, i.e.:

- Share the effort between individual countries for their language(s), and international bodies such as the EC.

to more specific and/or technical suggestions, i.e.:

- Two steps evaluations, which would first evaluate systems for one technology on one language, and then compare results of the same system on other languages.
- Cluster languages by language families.
- Use language technology evaluation campaigns to produce, at low cost, quality language resources, with the ROVER paradigm (which relies on the fact that the system based on the merging of all evaluated systems gets better results than the best of those systems).
- Use bootstrapping approaches such as the one mentioned in the session on Automatic Language Resource production, where after an initial training with carefully human-annotated data, the corresponding language technologies are used to produce language resources (with an example of using an Automatic Speech Recognition system to transcribe raw audio data, the result of which can then be used to build Language Models).

For addressing the problem of task- vs. application-oriented, several solutions were proposed:

- 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.
- It was however mentioned that some large programs already include integrated evaluation, such as the Quaero French program on multimedia & multilingual document processing, where several industrial applications are interfaced with many multimedia technologies, including Language Technology, through an evaluation-based interface. But this innovative, integrative approach has a cost (200 M€ budget over 5 years for the program).

Lessons for FLaReNet, next steps

Looking at the way forward on language technology evaluation, it clearly appears that **it definitely needs coordination**: 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.).

This said, lessons for FLaReNet are therefore that there is definitely here a need for coordination, and that this fits perfectly with the mission of FLaReNet to promote and **bring coordination**.

S4 – Openness, Sharing, and Standards

Introduction: overview, rationale

Interoperability is probably one of the most important ingredients in the glue that allows integration, sharing, interchange and reuse of Language Resources and Technologies.

Interoperability of resources, tools, and frameworks has recently come to be recognised as perhaps the most pressing current need for language processing research. Interoperability is especially critical at this time because of the widely recognized need to create and merge annotations and information at different linguistic levels in order to study interactions and interleave processing at these different levels. Recognition of the urgency for interoperability of language resources and tools is becoming more and more critical because the multilingual scenario in Europe and because new emerging data and tools for strategic languages (Arabic and Chinese) as well as minority languages need to be faced. Interoperability will not only allow saving time and effort, but also constitute major progress towards the ultimate goal of creating sustainable and accessible digital resources. In this scenario, lack of interoperability can have a seriously negative impact, thus reinforcing monopoly and resulting in a high-barrier market.

The session on Openness, Sharing, and Standards was concerned with issues of interoperability of the resources and software that support language technology, which is seen as crucial for both development and deployment. In the near term (at least the next decade), interoperability is critical to enable research and as a result, increasingly rapid development of language processing applications. In the long term, interoperability will be the cornerstone of a global network of language processing capabilities.

Recognition of the urgency for interoperability of language resources and tools is apparent in the recent flurry of activity within the community aimed at achieving this goal, including:

- formation of a sub-committees in the International Standards Organization (ISO TC37 SC4) to develop standard representation formats for various types of linguistic annotation and a general framework;
- global efforts to create linked WordNets and FrameNets;
- development and harmonization of systems and frameworks for linguistic annotation, e.g., the General Architecture for Text Engineering (GATE), MITRE's Callisto, and the Unstructured Information Management Architecture (UIMA);
- recent major meetings within both the linguistics and computational linguistics communities expressly concerned with resource interoperability (E-MELD TILR), and the creation of an international conference devoted to language resource interoperability (ICGL);
- multiple workshops at major conferences addressing issues of standards for both representation formats and linguistic categories;
- establishment of registries and catalogues for linguistic categories (e.g., ISO TC37 SC4 Data Category Registry[4]) and annotation schema (e.g., UIMA Component Registry[5]);
- U.S.-funded efforts to merge and/or harmonize linguistic annotations at different levels;
- a recent EU-funded effort to create a common resource and technology infrastructure for the humanities and social sciences (CLARIN), an EU project to establish a roadmap for achieving interoperability (FLaReNeT), and a parallel U.S.-funded effort aimed at working towards interoperability (INTEROP-SILT);
- formation of a special interest group of the Association for Computational Linguistics (SIGANN) at ACL 2007, one of whose primary aims is to work toward the development of standards for representing and designating linguistic information associated with language data;
- independent work within the Semantic Web community on interoperability of ontologies, another major resource for language processing research.

[4] http://syntax.inist.fr/.
[5] http://uima.lti.cs.cmu.edu:8080/UCR/.

Highlights, specific problems/issues

The **diverse efforts** mentioned above **all aim toward language resource interoperability, but they have tended to develop in isolation**. There is as yet minimal ability to integrate data and components, leading to not only a duplication of effort but also a restricted understanding of better ways of addressing specific technical issues. At the same time, there is enough convergence of opinion and practice among these community activities that the crucial steps required to bring all of the pieces together are beginning to emerge. Perhaps more importantly, advances in technology over the past decade, such as the technologies surrounding the Semantic Web and the emergence of distributed computing and data capabilities, have opened up possibilities for interlinking data, annotations, lexicons, ontologies, etc., that provides new motivation to pursue resource interoperability. The time and circumstances are therefore 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.

While on the one hand much has been and is being done to work toward interoperability, it is also essential at this point to consider the obstacles that still stand in its way. Four major obstacles to achieving interoperability for resources and software were identified:

1) Existing standards are not widely accepted or widely used; it is therefore unclear how or even if standards to support interoperability will be adopted by the community.
2) There is a range of opinions concerning theoretical approaches to linguistic analysis and interpretation, and a lack of agreement on even fundamental linguistic properties such as categories for part of speech.
3) Similarly, there is to date no universally accepted representation format for linguistic data, although there is some convergence in the underlying model of the formats and frameworks that are currently in wide use.
4) Although reliable language resources such as linguistically annotated corpora and analytic software are needed by members of the community involved in language technology research and development, and despite the existence of distribution centers such as LDC and ELRA, access to these resources is not easy. Licensing and copyright restrictions are among the greatest obstacles.

Suggested solutions

A variety of (not necessarily compatible or mutually exclusive) solutions were suggested for each of these problems, as summarized below.

1. *Existing standards are not widely accepted/used*. 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. Interoperability (at some level) is achieved as a necessary by-product. Other suggestions addressed possible changes to the standardization process itself. For example, it was generally agreed that **t**he focus of standardization efforts should be on transformation between representation formats and linguistic annotation categories and schemes, rather than an attempt to establish a single standard for any of these phenomena. This allows researchers and developers to utilize formats and schemes that serve their needs and still have interoperability via transduction to formats suitable for other systems. It was noted that transduction can be fostered by identifying an underlying data model that can be realized superficially in a variety of formats/schemes. The point was also made that standards are often not used because of a lack of tools that support them; providing such tools and ensuring that they are easy to use is essential for widespread adoption.

2. *Disagreement concerning theories/linguistic categories.* The ISO Data Category Registry was pointed to as a major effort to address this problem, by providing a centralized repository for the identification and description of linguistic categories that are used in annotation and analysis. However, the ISO DCR does not at this time seek to establish a standard set of such categories, but only to provide a set of definitions / distinctions that can serve as a reference or a point of departure for defining new or variant categories. It was suggested that some steps toward standardization could be taken immediately by taking a "bottom up" approach and addressing only those areas where there is consensus, focusing on the "lowest common denominator" among phenomena. Other suggestions were to take the approach suggested for standards in general above, by establishing mappings/transductions among different categories (the difficulty of applying

this to categories was acknowledged due to the lack of one-to-one correspondences in many cases). Finally, it was suggested that ontologies of linguistic information will be needed to provide the framework for establishing standard categories for linguistic annotation.

3. *No standard representation format(s)/ frameworks.* There was general consensus that in this area, there is a convergence of practice among several widely used formats, frameworks, and systems, relying on a UIMA-like architecture of configurable pipelines of language processing modules, and representing results using some surface format that serializes an underlying, adequately expressive abstract model for linguistic information. The emergence of generic "pivot" formats (e.g. LAF, LMF) that realize the abstract model, into and out of which various serializations can be mapped (for interchange) is also contributing to convergence.

It was also suggested that rather than focusing on how things are represented at relatively low levels of analysis, we should focus on input/output formats for tools instantiated as Web services.

4. *Lack of Accessibility*. Very few concrete solutions for the problem that resources and tools can be hard to find were suggested. The publicizing of LMF via Wikipedia was cited as a possible solution, and community outreach and education were recommended. Access rights pose another kind of obstacle, and it is clear that there is a growing sentiment in support of open source development, and free and unfettered access to resources and tools. A federation of centres was suggested, which would negotiate access rights to data and software with commercial and other enterprises that hold them.

Lessons for FLaReNet, next steps

The workshop 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:

- It is neither desirable nor possible to impose a single monolithic standard for any aspect of language technology;
- Theory/category standardization, if it can/should be achieved, is a more difficult and fundamentally different problems than format/framework standardization, where there is some convergence of approach;
- We need to develop both "bottom up" and top down;
- We can start now by focusing on standards for areas where there is some consensus, and develop a toolkit implementing these standards that can be used off-the-shelf, so that we can move on to tackling the larger problems.

There was one issue arising from several of the discussions that showed a need for the community to clearly distinguish its near-term and long-term goals. We can begin to visualize the language technology of the future, which may involve distributed Web-based services that process streamed data rather than documents, etc. In these scenarios, many of the problems of interoperable formats and processing frameworks that we are currently addressing will either disappear or be shifted to another level or layer of analysis. It is imperative to keep the longer term vision in mind, but at the same time, we cannot abandon efforts to develop and annotate linguistic resources and tools that are interoperable and reusable for the very substantial research and development efforts that are required *now* in order to move us closer to that vision. The scenario that is off in the future and the current situation (wherein robust, full ranging language processing capabilities are very far from realized) are not at odds—although in the workshop more than one "argument" involved a perceived conflict between the near-term and long-term views.

The work of FLaReNet can be facilitated by seeing 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.

S5 – Translation, Localisation, Multilingualism

Introduction: overview, rationale

Language resources and language technologies have become indispensable tools in order to enable translators, interpreters, technical writers, localisers, and other language professionals to provide high-quality services. Machine translation is increasingly used in industry, public administration, and other areas where millions of pages have to be translated every year. It is now also getting widely used by the grand public, thanks to the tools offered freely over the internet for document or message translation. And it starts being extended to the spoken language, with the perspectives of automatic interpretation of talks, courses and meetings, and the need to understand the huge amount of video now available on the Web. Computer-assisted translation methods such as translation memory systems, terminology management systems, localisation tools, etc. are widely used in SMEs and public services. Multilingual text corpora (aligned corpora, parallel texts, comparable corpora etc.) as well as multilingual lexical corpora, lexicons, term bases, etc. are being prepared and used for diverse application scenarios. Quality management tools such as translation metrics, standards for translation service providers, semi-automated workflow and project management systems are language technology applications that are increasingly used by language professionals.

The purpose of the session was to identify urgent needs, assess current trends, and formulate concrete recommendations for further action in the field of technologies for multilingual communication, in general, in Europe. Among the main issues tackled by the speakers were key problems in the field of translation and localisation and how Language Resources and Technologies can help solve these problems, current users' needs and current trends in language technology developments for the different multilingual technologies, new emerging paradigm shifts.

Highlights, specific problems/issues

While there has been considerable progress in statistical machine translation (SMT) and notably entry barriers to machine translation technology have been substantially lowered, there seem to be two main classes of problems that have to be overcome to enable future progress:

- absence of parallel data for many languages, different domains and text types, for both training and evaluation as well as accurate methods for their discovery and classification;
- inability to handle complex linguistic phenomena as well as treat gaps (lexical and syntactic) in training data.

Similar experiences were voiced by speech-to-speech translation experts.

On the rule-based machine translation (RBMT) front, progress is less visible while, on the contrary, their use is growing especially as far as institutional users are concerned. Progress in RBMT is hindered mainly by inadequate grammar resources for most languages and absence of appropriate lexical resources and methods that would enable correct disambiguation and lexical choice.

Integration of linguistic knowledge, at varying levels, in statistical machine translation, statistical layers through language modelling in rule-based machine translation and **combination of statistical machine translation and rule-based machine translation systems seem to be shaping the current trends** of machine translation technology development.

Evaluation of MT suffers from lack of reliable automatic evaluation methods, in addition to the total **absence of evaluation data** for the majority of language-pairs. Furthermore, metrics reflecting the appropriateness of MT for particular applications are totally missing. As has been proposed in other fora as well, task-oriented evaluation and subsequent error analysis are one of the possible ways forward.

Regarding **localisation**, the main additional problems come from the **sheer volume of content** to be localised (putting extra requirements on technologies to be used), and the need for **personalising** information and content not only to linguistic but also to cultural aspects and changing interface devices.

Furthermore MT **users advocate modular approaches** allowing them to configure technical solutions themselves. Such modularity comes with certain preconditions to be met, notably **compliance of systems and resources with standards**, so that interoperability can be ensured. In addition, modularity also calls for **technology information aggregation services** so that users can have access to what exists and what can be

deployed in specific application scenarios, as well as **comparative evaluation of language technology solutions** (much desired by technology users).

Suggested solutions

Possible ways to overcome the problems and facilitate progress in the emerging paradigms include:

- Discovery and exploitation of hidden parallel data, and also use of monolingual and comparable corpora resources instead of perfectly aligned bilingual corpora.
- Exploitation of knowledge technology methods in MT.
- Application of machine learning techniques for optimising combinations of different paradigms, e.g. SMT and RBMT, and use of machine learning techniques to make MT "smart" and learn from its mistakes.
- Automatic collection of language resources for statistical language technology, given current automatic speech recognition performance.

The **integration of major localisation processes** - authoring, translation, distribution - **under one common platform**, interfacing language technology, information retrieval, adaptive hypermedia, will constitute the so-called "new localisation factory" to respond to future needs.

Users of machine translation technology, especially institutional users, point to the **differing priorities between language technology professionals and technology users**. Three aspects are worthwhile noting:

- Different degrees of automation are required depending on tasks, there is no "one-size-fits-all" solution.
- Consistency in translation should be guaranteed, and guarantees come from traditional technologies, like translation memories, more than contemporary ones (Translation Memories > Rule-Based Machine Translation > Statistical Machine Translation).
- Machine translation technology and Computer-Aided Translation (CAT) tools should be seamlessly integrated in the workflow and their use should be as easy and user-friendly as possible.

Collaboration and intelligent **crowdsourcing** were also proposed as the modus operandi in the future, ensuring the availability of high-quality and trustable translated content, that could be deployed either as retrievable examples from translation memory databases or as training material for language-aware or language-agnostic statistical machine translation systems. Likewise, simple, bottom-up and per domain language resources building is advocated as a solution to fill in existing gaps.

Lessons for FLaReNet, next steps

For this session, as well, **infrastructure building** seems to be the main issue and message for FLaReNet. With inductive methods dominating the current paradigms, **language resource building, annotation, cataloguing, accessibility, availability and clearance from IPR** 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.

Evaluation of MT is a trickier point. Leaving aside hard methodological issues and concentrating on the problem of absence of evaluation data, FLaReNet could possibly act as a facilitator here again.

Last, in the light of integrating linguistic and statistical processing for MT, issues relating to the concepts of **BLARK and ELARK** become absolutely relevant.

Another conclusion for FLaReNet is to **focus on facilitating more interaction and co-operation between different communities** active in this broad field: Machine Translation technology providers, Computer Assisted Translation technology providers, translation and localisation service providers, large institutional language services, Language Resource producers in different domains and different languages, research organisations interested in these topics and processes, and publishing houses.

S6 – Enhancing Market Places/Models for Language Resources: New Challenges, New Services

Introduction: overview, rationale

The setup of distribution agencies, such as LDC in US (1992) and ELRA in Europe (1995), has triggered the establishment of a market place of language resources for language technology players with middlemen playing broker roles. Before that time, bilateral agreements were negotiated between language technology players and data producers e.g. terminology centres, dictionary publishers, corpus producers (mostly newspapers and the like).

The new landscape of the e-business confronts the market with new opportunities and challenges, while increasing the potential number of players.

The market of language resources is dealt with from two points of view: (1) the market of making them accessible and (2) the production of a language resource market.

Though the title of the session refers to market places and models for language resources, most contributions were on market places and models for language technologies, applications and services, thus the scope of the session has been enlarged to comprehend issues related to: market of making language resources accessible (trading language resources), language resource production market, and how are these related to and impacted by the Human Language Technology market, applications/services deployment, user profiles and the niches to tackle.

The major topics are: the nature of the market (small players, global market) and the key issue to address (IPR, licensing, distribution, pricing policy, relationship between the different players (producers, brokers, distributors, users), the corresponding business models and how they match Human Language Technology player's expectations, as well as the current trend towards an e-market place.

Highlights, specific problems/issues

The market for Language Resource is not independent. One of the reasons why the topic of this session was enlarged to technologies instead of focussing on Language Resources is that the market for Language Resource is not independent, but only derivative. Language resources become relevant for commercial organizations only when there is a profitable business opportunity for developing language technology, applications or services that require such language resources. In such a case, companies either buy resources, or make them internally or through mechanisms of cooperation (mainly for R&D purposes), or even buy companies that own them. The needs here are driven by business requirements, and the relation with BLARK does not even come up.

Language Resources are business assets. The language resources developed are an essential ingredient for creating language technology, applications or services and are always considered to be business assets, in some cases offering even unique competitive advantages. Selling or distributing such language resources is not the business focus of Human Language Technology companies, since it does not significantly contribute to revenue or profitability, and may even harm these. Therefore, in many cases, these Language Resources will not be made generally available, even if they can be traded for significant prices.

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 application, 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. 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. In the discussion, however, it was suggested that lack of data may also provide an opportunity for a unique competitive advantage, so that it is better to leave this to the market forces. Others, however, stated that market forces will only address languages and areas that are lucrative enough; which makes a coordinated policy at a European level essential if one want Human Language Technology to be deployed for all languages and countries.

Increased cooperation among companies. An arising trend that clearly showed up was increased cooperation between companies. This is evidenced by the cooperation by a variety of local Human Language Technology providers to counter the threat posed by big players such as Google in the general domain Machine Translation market, as well as by the set-up of a network of SMEs in the area of speech technology and applications. Also the TAUS-initiative, a cooperation on the sharing of translation memories, witnesses to

this trend. Cooperation can be done by companies at all times and for many reasons, and relations between companies are generally quite complex.

High quality of Human Language Technology is essential and requires high quality and large quantity Language Resources. A second 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 and large quantity *language resources*. It is essential for all technologies, but it was mentioned explicitly for speech technology and machine translation. Speech technology requires a strong improvement on robustness of speech and language models. For machine translation, instead, an increase in the quality level is required to make it a profitable business. One possibility to achieve the better quality is going into a *niche*, but for general MT new technological breakthroughs are required. Successful applications, in their turn, will lead to creation of new and/or access to existing Language Resources, by verticalisation (adaptation to specific domains) or customization, and extension of language coverage. The need for human work (annotation, transcriptions, etc.) may lead to new outsourcing/off-shore opportunities into low-labour cost countries.

Shift from solutions to services on demand. Several speakers also pointed out that the market is shifting from *solution*s to *services on demand.* This requires new business models and imposes new technical requirements, e.g. in the area of speech standardization, of acquisition devices and speech coding and makes multilingual coverage more pressing than ever. Apart from technical requirements, it also imposes contextual requirements, including an infrastructure, public policies on e-government services, legislation for the adoption of such services, and customer education. The public procurement is an instrument to boost such technologies and some participants felt that the EU should use such an instrument as is being done in the USA through the DARPA/DOD programs.

Small players versus global players. Concerning the market situation for language technology, applications and services, a key observation is that a few numbers of big global players form a threat to many small local players. This holds particularly in the general machine translation market, where Google is dominant. Two different solutions to counter this threat were proposed.

During the debates some discussants indicated that evaluation is very common nowadays and stated that “if there is evaluation, we are not far from application deployment”.

Suggested solutions

It was proposed to establish a “Public-Private” partnership through a Language Resources Investment Fund for the production of resources.

Niche Market. One strategy proposed is *focusing on small niche markets* where, thanks to the restricted domain and dedicated focus, a quality level can be achieved that makes customers willing to pay high prices, thus making the enterprise profitable. Such niche markets require customising systems to the customer’s specific needs, so these are generally projects rather than off-the-shelf products. Such customisation also requires niche-specific data. In many cases the customer is required to provide such data, but in all cases customers are stimulated to provide these by offering better pricing conditions if they bring in their data and the prospect of improved performance for that specific niche.

The main example of such a niche market discussed was in the area of Machine Translation. Another example is speech dictation for radiologists. But, generally, the number of lucrative niche businesses is very limited, and knowledge about them restricted, so that “we come across such niches very often by accident”.

Cooperative Services. 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.

Lessons for FLaReNet, next steps

Support transfer of Human Language Technology to SMEs. Finally, an appeal was also made to the Commission to establish instruments 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.

FLaReNet International Cooperation Round-table

Introduction: overview, rationale

The field of Language Resources and Technologies needs a strong and coherent international cooperation policy to become more competitive and play a leading role globally. It is crucial to discuss future policies and priorities for the field of Language Resources and Technologies – as in the mission of FLaReNet – not only on the European scene, but also in a worldwide context, from, at least, two perspectives.

- *Multilingualism* is one of the major challenges not only for Europe, since European languages are spoken all over the world. A joint global approach in the development of the technologies for the various languages would facilitate the handling of multilingualism. It is therefore of utmost importance to produce the resources which are necessary for addressing the various European languages, in a coordinated way in order to provide a proper coverage, share best practices and ensure their interoperability. This is one of the challenges for the next few years, for a usable and useful "language" scenario in the global network. Europe must cooperate, in this framework, with countries having a similar interest in LRs and LT, such as the US, Japan, China, but also notably South Africa and India, which consider LRs and LTs and languages as a priority for handling multilingualism at their national level.
- This is even more true when we analyse which *infrastructures* are needed. The growth of the field should be complemented by a common effort that tries to look for synergies and to overcome fragmentation. It is an achievement, and an opportunity for our field, that recently a number of strategic-infrastructural initiatives have started, or are going to start, all over the world. This is also a sign that funding agencies recognise the strategic value of our field and the importance of helping a coherent growth also through a number of coordinated actions.

Cooperation is an issue that needs to be prepared. FLaReNet may be the place where these – and future – initiatives get together to discuss and promote collaboration actions.

The round-table devoted to "International Cooperation" was organised mainly with non-European participants. The aim was to start a discussion on the usefulness and the interest of promoting international cooperation among various initiatives and communities around the world, within and around the field of Language Resources and Technologies. This should lead to future discussion on the modalities of how to organise the cooperation on specific aspects.

Highlights, specific issues, and suggested solutions

Many ideas were put on the table, as proposals of joint activities on which the group can identify itself and possibly join forces. They are clustered here in major areas – even if there are many inter-relations among them.

Standards and Interoperability

Standardisation was mentioned by many as a **natural topic for cooperation**. Standards for interchange formats (for I/O among linguistic processing tools and systems) were agreed to be the appropriate means to pursue interoperability. **Common repositories for tools and language data** should be established that are universally and easily accessible by everyone. Also the **creation of a shared repository with data formats, annotations**, etc. – where to find the most frequently used and preferred schemes – is proposed as a major help to achieve and promote standardisation.

We should try to connect ongoing work done by many groups, also exploiting the collaboration between the NSF project SILT (Sustainable Interoperability for Language Technology) and the EC FLaReNet. A concrete common task could be to instantiate standards (such as the ISO Linguistic Annotation Framework, Lexical Markup Framework, Time-ML, Semeval, etc.) in a bottom-up approach, starting with some small experiment as proof of concept. There is interest in cooperation towards a common type system to combine different toolkits.

It is important to coordinate input to ISO standardisation work also from Asian countries: different "standards" are currently adopted for various resources of the same type, and there is need of harmonisation and distribution initiatives.

Metadata is another topic **to be approached in cooperation**. A metadata catalogue should involve every party, considering metadata as the missing knowledge that connects language resources to actual users.

Cooperation on standardisation is felt as even more important today when standardisation is much more dynamic than in the past; we should push for common work on pre-normative standardisation initiatives, with an impact on real-life applications, also involving companies.

Language Resource Creation

Structured models should be thought, and some initiative could be undertaken to **facilitate Language Resource creation for underprivileged languages**. Cooperation should be encouraged to develop corpora that cannot be easily developed by a single national effort. A **distributed network of data centres could help**. For example, there is a serious lack of reusable Language Resources in Africa, while Africa could contribute to the effort of sharing Language Resources.

Access and Availability of Resources, Tools and Information

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. Many issues are at stake here:

- The need to create the means to plug together different Language Resources & Language Technologies, in a **Web-based resource and technology grid**, with the possibility to easily create new workflows.
- It is important to create conditions to easily share and re-use technologies, to have more **open source** tools to be made available for use also to under-funded groups.
- A platform for cooperation could be thought of around the notions of **BLARK and ELARK**.
- **Interoperability** is obviously a pre-condition. This must be a possibility offered to every researcher, both as technology user and as technology provider (with the need that software is used and tested by others).
- **Evaluation** initiatives can profit of this. Many centres are developing tools that can be used by others, e.g. it was mentioned that an entire speech platform can be distributed for free as open source, or that text-mining tools can be provided to the academy.

We should actually start some work to build some Web services and show that it can work—this is the only way we can really get everyone on board (if it is there, they will use it!).

The related issues of **access rights and IPR also call for cooperation**.

Related to this is **transparency and availability of information**. Proposals are: the availability of a Universal Catalogue, of a Portal for all the Web services, of an information centre (or a network of information centres) for the field also as a deposit of news impacting future technology. Automatic data discovery methods can be used to find information from the many news services.

LREC (the *Language Resources and Evaluation Conference*) could be exploited as a means **to promote sharing Language Resources and Technologies and information**, asking for Language Resources and Technologies, described in submitted papers, to be made available (when possible) and/or to be put in the Universal Catalogue. This would also **promote in a broad community** (more than 900 submissions last time) **the idea of (shared repositories of) open Language Resources and Technologies and of a collaborative way of enriching the catalogue**. We could also have a special dedicated Poster session and a Workshop on this idea/movement at Language Resources European Conference.

Also the *Language Resources and Evaluation* Journal could promote this idea, and could be a forum for discussion and thought on these topics.

Evaluation

The need for a more permanent evaluation structure in the EU and of more communication among international evaluation bodies has been raised. World-wide discussion on evaluation data and on how evaluation resources or toolkits can be made really distributable and usable, e.g. stored in the same place, needs to be coordinated at an international level.

Furthermore, there is interest in task oriented evaluation issues, for the importance of trying to meet the requirements of real users.

Networking and Cooperation

Programs with Latin America, to raise awareness, need to be promoted. There is an e-government project on how to include these countries.

A Forum to accelerate the development of Language Resources and Technologies was recently launched in Japan: cooperation and linking with FLaReNet is welcome. Networking is particularly important in areas where there is still an imbalance and a gap to be filled, e.g. in Asian countries, Africa, and Latin America. Help is needed e.g. to see which are the real issues in Language Technology and to identify potential partners.

A joint reflection on how to use the Web to reflect our ideas and help in creating collaboration possibilities is discussed.

Networking and support actions must be conducted more intensively, with **establishment of international committees** that have formal recognition, and organisation and through participation to common workshops.

In a field that is both fragmented and over-structured, many mentioned the need to have an **International Forum** (a meta-body) to share information, discuss strategies and declare that there are common objectives. Such a Forum can play a role only if it is recognised as influential and authoritative: e.g. a Memorandum of Understanding signed by hundreds of organisations could give authority.

International Cooperation and Funding Agencies

Both the EC and NSF have some programs and calls where money can go also to foreign partners, and this could be exploited. It was remembered that in the past there were jointly organised comparable calls between EC and NSF: this was an interesting experiment that may deserve being pushed again.

Lessons for FLaReNet, next steps

What is really needed is to find a means to do all of this. The work of FLaReNet can be at least in part to find ways to enable these activities, by creating collaborations and identifying specific projects/activities where this kind of thing can be done, even on a small scale at first.

A focus of FLaReNet and the NSF SILT has to be to find a way to really bring everyone on board, and that will involve making sure all the players around the world are involved, that they all have a stake (e.g. by having their own center in a network of centers), and by actually doing some experiments to use current and emerging standards to actually build something that is interoperable and works.

Actions on Infrastructural, Information and Networking Initiatives

Multilingualism is a challenge that implies a huge effort. Necessary conditions to support it are: identify what exists, availability of language resource and technologies, evaluation of language resources and technologies, sharing of best practices.

FLaReNet is voluntary work. Yet some initiatives are well in the possibility of FLaReNet, and these we have to push. We could on one side start from low-level/easier tasks and on the other side act as strongly as possible on all the aspects in need of promotion, mobilisation, collaborative action, etc. which are exactly in the mission and within the possibilities of FLaReNet.

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

- FLaReNet can use its collaborative Web site to create a pool of ideas on which 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 can 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 interoperability;
- FLaReNet can 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 act as and/or promote the need of a communication vector for open source resources and tools: this could be in wiki mode;
- FLaReNet can 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 could 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 Hőge (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