The FLaReNet Strategic Language Resource Agenda

Claudia Soria°, Núria Bel♣, Khalid Choukri•, Joseph Mariani∞, Monica Monachini°, Jan Odijk♦, Stelios Piperidis♥, Valeria Quochi°, Nicoletta Calzolari°

°CNR-ILC, Italy; ♣UPF, Spain; •ELDA, France; ∞IMMI-CNRS, France, ♦Utrecht University, Netherlands, ♥ILSP "Athena" R.C., Greece

E-mail: claudia.soria@ilc.cnr.it, nuria.bel@upf.edu, choukri@elda.org, Joseph.Mariani@limsi.fr, monica.monachini@ilc.cnr.it, J.Odijk@uu.nl, spip@ilsp.gr, valeria.quochi@ilc.cnr.it, nicoletta.calzolari@ilc.cnr.it

Abstract

The FLaReNet Strategic Agenda highlights the most pressing needs for the sector of Language Resources and Technologies and presents a set of recommendations for its development and progress in Europe, as issued from a three-year consultation of the FLaReNet European project. The FLaReNet recommendations are organised around nine dimensions: a) *documentation* b) *interoperability* c) *availability, sharing and distribution* d) *coverage, quality and adequacy* e) *sustainability* f) *recognition* g) *development* h) *infrastructure* and i) *international cooperation.* As such, they cover a broad range of topics and activities, spanning over production and use of language resources, licensing, maintenance and preservation issues, infrastructures for language resources, resource identification and sharing, evaluation and validation, interoperability and policy issues. The intended recipients belong to a large set of players and stakeholders in Language Resources and Technology, ranging from individuals to research and education institutions, to policy-makers, funding agencies, SMEs and large companies, service and media providers. The main goal of these recommendations is to serve as an instrument to support stakeholders in planning for and addressing the urgencies of the Language Resources and Technologies of the future.

Keywords: strategic agenda, language resources planning, recommended priority actions

1. Introduction

The EU has established that cultural and language differences are a unique asset to be preserved, despite the complexity of handling multilingualism. As such, considerable investment has been put in finding means - such as technological ones - to overcome the language barriers to support European citizens and industry in a multilingual globalised world. The large majority of industrial technological applications that handle natural language, i.e. Machine Translation, Crosslingual Information Retrieval, Multilingual Information Extraction, Automatic Document Indexing, Question Answering, Natural Language Interfaces, etc., include Language Resources as critical components. At the same time, it is proved that a critical mass of Language Resources (LR) can make advancement in research and technology development possible and quicker, making Europe the leader of the market related to multilingualism.

Companies such as Google or Microsoft play a dominant role in this framework, as they have access to a huge amount of data in many different languages, devote considerable resources to Language Technologies (LT), have massive computing power and a direct research-to-application pipeline using a new business model based on so-called "free" services. The fact that a US company like Google is delivering some of the most comprehensive LT solutions to support multilingualism should raise concern among EU officials.

1.1 Past Efforts in Large Language Resource Programs

In the US, accumulation of language data was a priority since the early 90's, when the fast pacing of statistical approaches in NLP spread the assumption that "there's no data like more data". Since then, this approach has been extended to other areas of language technology - information retrieval by search engines, Machine Translation, and more generally human-machine communication (including Computer Vision). Statistical methods paved the way for DARPA-style comparative evaluation campaigns led by the National Bureau of Standards (what is now the National Institute of Standards and Technology, NIST [1]) starting back in 1987. The growing need to gather large quantities of data to train systems resulted in the creation of the Linguistic Data Consortium[2] in 1992. Europe has put a similar effort into stimulating the field of LRs, with the launching of the European Language Resource Association (ELRA[3]) in 1995, which later promoted LRs and evaluation through the LREC conferences [4] that began in 1998. The importance of LRs for driving LT was well recognized in Europe by the European Commission, which launched in the '90s a number of initiatives for the development of spoken and written resources as well as of standards for their representation. However, Europe missed the opportunity of creating a permanent evaluation agency comparable to the NIST in the US, although a number of stakeholders including ELRA/ELDA[5] played such a role. Also in light of the pioneering actions of the US and the considerable DARPA funding of this research, the (American) English language has by far the best language data coverage. As a result, much of the scientific community works on English data and reports results about English language phenomena. Technologies and applications grow more and more advanced for English,

[1] http://www.nist.gov/index.html
[2] http://www.ldc.upenn.edu
[3] http://www.elra.info/
[4] http://www.lrec-conf.org/
[5] http://www.elda.org

and in turn produce yet more data and induce the organization of yet new evaluation campaigns, including the study of metrics themselves – now a new research topic in itself.
The data issue varies considerably for other languages. Some are relatively well covered when there are programs that provide the investments needed to produce data and test systems. In the US, this is the case for geopolitically significant languages (in Iraq, Afghanistan or the Balkans) or those involved in human emergencies (such as the Haiti earthquake). Other European countries such as France, Germany and The Netherlands have been funding national programs that accelerate measurable research. But most of the world's languages do not have such support.
In some countries language is seen as a major political issue, either because they want to promote their language (e.g. Baltic countries) or because they have a constitutional obligation to preserve the languages spoken by their citizens (e.g. India and South Africa). These countries prioritize the development of language technologies to preserve their languages and ensure communication in them, even if they have limited financial resources to do this extensively. This sort of political commitment to LT as a support to multilingualism is not yet typical of the European Commission and the 27 Member States of the European Union.

1.2 The FLaReNet Strategic Agenda

Language Resources are key to the development of NLP applications for a multilingual Europe.
However, realizing this multilingual technological vision requires a collaborative and coordinated effort from all stakeholders. While there has been considerable progress in technology development in the last decade, the significant challenge of overcoming current fragmentation and imbalance inside the LT community for all languages still remains an issue.
Thanks to initiatives such as the FLaReNet project, this situation is now starting to be tackled and a new awareness is now spreading about the need and importance of joining forces and building a compact community. If a coordinated and concerted approach is adopted – if all interested stakeholders agree to follow a common plan of actions, chances are that we can improve the current situation for all languages.
The FLaReNet Strategic Agenda highlights the most pressing needs for the sector and presents a set of recommendations for the development and progress of LRTs in Europe. The recommendations are the result of a three-year consultation of the FLaReNet project, which gathered together worldwide representatives from economy (software companies, technology providers, users), government agencies, research organisations, universities, non-governmental organisations and language communities.
The FLaReNet recommendations cover a broad range of topics and activities, spanning over production and use of language resources, licensing, maintenance and preservation issues, infrastructures for LRs, resource identification and sharing, evaluation and validation, interoperability and policy issues.
In principle, the addressees of this Strategic Agenda belong to a large set of players and stakeholders in LTs, ranging from individuals to research and education institutions, to policy-makers, funding agencies, SMEs and large companies, service and media providers. Its main goal is thus to serve as an instrument to support stakeholders in planning for and addressing the urgencies of the LRTs of the future. The recommendations contained in the present document should therefore be taken into account by any player, whether on a European, International, National, local, or private level, wishing to draft a program of activities for his/her own communities[6].
The recommendations are organised around nine different dimensions:
a) *Documentation* b) *Interoperability* c) *Availability, Sharing and Distribution* d) *Coverage, Quality and Adequacy* e) *Sustainability* f) *Recognition* g) *Development* h) *Infrastructure* and i) *International cooperation*.
Some of these dimensions are of a more infrastructural nature, some are more related to research and development, some yet more to political and strategic aspects, but they all must be seriously considered when making up a strategy for the future of the field. All of them eventually have an impact in the development and success of LRs, and represent the areas where actions need to be taken to make the field of LRTs grow.
It is useful to see the various dimensions as a coherent system where each one presupposes the others, so that action at one of the levels requires some other action to be taken at another level. For instance, open availability of data presupposes interoperability (which in turn is boosted by openness); to discover and develop new paradigms, and for data to be usefully exploited, the availability of large quantities of data requires the ability to link the information carried by data. Increased data quantity implies a change in their availability towards openness, and so on.
Taken together, these directions are intended to contribute to the creation of a sustainable LRT ecosystem.

2. Resource Documentation

Accurate and reliable documentation of Language Resources is an undisputable need. Instead, as of today, LRs are still often poorly documented or not documented at all, and, even when available, documentation is often not easy to find. Documentation is also the gateway to LR discovery. Ensuring that Language Resources are discoverable is the first step towards promoting the data economy.
Devise and adopt a widely agreed standard documentation template for each resource type, based

[6] A more detailed version of the recommendations is available in Calzolari et al. 2011.

on identified best practice(s): the variable nature of documentation can hamper the dissemination and replication of LRs and makes it hard for users to read and compare how-to files. Common best practices for writing documentation and guidelines need to be established and enforced. A common, standardized documentation template should be defined, promoted, and enforced for all contracts for publicly funded projects.

Documentation should be as exhaustive as possible, and include information about data format and data content, the production context, and existing possible applications.

When producing a LR, allocate time and manpower to documentation from the start; provide documentation (or links to it) when giving access to a LR: every release of a Language Resource should be accompanied by provision of the corresponding documentation. In every language resource production project, part of the funding should be allocated to documentation and dissemination activities. Policy Makers, both at the National and European level, should support activities for collecting and storing documentation for LRs in appropriate infrastructures.

Ensure that appropriate metadata are consistently adopted for describing LRs: documentation is also the gateway to LR discovery. Ensuring that LRs are discoverable is the first step towards promoting the data economy. Language Resource Providers should always document their resources using standard metadata and unique resource identifiers. Therefore, definition and adoption of standardized metadata must be the first priority and first step for all Language Resource Providers.

Policy makers, on their side, must support metadata creation, also by means of promotional activities. At the European level, for instance, there should be an established set of guidelines and rules for metadata description of available Language Resources.

Set up a global infrastructure of common and uniform and/or interoperable metadata sets: one of the main reasons why it is now difficult to find resources that match specific needs and languages is the lack of compatibility for metadata. Different sub-communities, data distribution centers, archiving institutions and projects, and other providers tend to use their own, non-interoperable metadata sets to describe their data, often at different levels of granularity, depending on who does it. The key priority is therefore to work towards the full interoperability of metadata sets. As there are many differing metadata sets and search engines, harmonization is a central problem for the community.

Develop and support community-wide initiatives such as the LRE Map: useful initiatives in this direction are community-based documentation initiatives, such as the LRE Map[7], by which massive documentation of existing resources is achieved in a limited time frame and with limited effort, with the additional advantage that all resources are documented in a uniform and standard-compliant way.

3. Resource Interoperability

Interoperability of resources and data is an essential prerequisite for successful exploitation of the enormous amount of data that the advent of the Internet has been making available since less than two decades. Interoperability of resources is the extent to which they are compatible, so as to allow, for instance, the merging of data coming from different sources while preserving their semantics. Today the lack of interoperability and compliance with standards costs a fortune. It is estimated that buyers and providers of translation lose 10% to 40% of their budgets or revenues because language resources are not stored in compatible standard formats (van der Meer, 2011).

Ensure syntactic and semantic interoperability of Language Resources: *syntactic interoperability* is the ability of different systems to process (read) exchanged data either directly or via trivial conversion. *Semantic interoperability* is the ability of systems to interpret exchanged linguistic information in meaningful and consistent ways via reference to a common set of reference categories (Ide and Pustejovsky 2011).

Set up an "interoperability challenge" as a collective exercise to evaluate (and possibly measure) interoperability: the design of interoperability tasks will also help to determine which emerging standards are most interoperable. Interoperability tests can also replace aspects of validation.

Create a permanent Standards Observatory or Standards Watch: while, on the one hand, it is increasingly recognised that standards are key to resource sharing, re-usability, maintainability and long-term preservation, LRPs are still largely lacking a clear understanding about why standards should be of any help in representing data, and why there are advantages in adopting standards. One solution would be to work towards the *establishment of a broad-based framework for interoperability* of language resources and language technologies, involving industry in the mix. There should be greater *awareness of the importance of standards for resource producers/managers who want to join the open-access club* and boost the utilization of their resources, so as to increase visibility, and attract more users and funding.

Invest in standardization activities: investment at the supra-national level in standardization activities is of utmost importance. In particular, support is to be given to infrastructural activities for collecting and disseminating information on existing standards and best practices. At the same time, activities for setting up new standards where they do not exist should be funded.

Encourage/enforce use of best practices or standards in LR production projects: the community and funding agencies need to join forces to drive forward the use of existing and emerging standards, at least in the areas where there is some degree of consensus (e.g. external descriptive metadata, meta-models, part-of-speech (POS)

[7] http://www.resourcebook.eu/

and morpho-syntactic information, etc.). The only way to ensure useful feedback to improve and advance is to use these standards on a regular basis. It will be even more important to enforce and promote the use of standards at all stages, from basic standardisation for less-resourced languages (such as orthography normalization, transcription of oral data, etc.) to more complex areas (such as syntax, semantics, etc.). LRPs, on their side, should look for standards and best practices that best fit the LRs to be produced, already at the early stages of design/specifications; *adhere to relevant standards and best practices*; produce LRs that are easily amenable to reuse (e.g. adopt formats that allow easy reuse). The creation of "official" validators to check compliance of LRs with basic linguistic standards is an activity to be pursued and encouraged by funding agencies.

Make standards operational and put them in use: as most users are not very concerned about whether or not they are using standards, there should *be easy-to-use tools that help them apply standards while hiding most of the technicalities*. The goal would be to have standards operating in the background as "intrinsic" properties of the language technology or the more generic tools that people/end-users use. LRPs should encourage the building of tools that enable the use of standards, and step up the availability of sharable/exchangeable data. Funding agencies, on the other hand, should fund the development and/or maintenance of tools that support/enforce/validate standards.

Set up training initiatives to promote and disseminate standards to students and young researchers: educational activities, such as training initiatives to promote and disseminate standards to students and young researchers are also important and effective means to enforce a "standards culture".

Identify new mature areas for standardization and promote joint efforts between R&D and industry: there should be a regular examination of new fields to check whether they are "mature" enough to start a standardisation initiative (for instance about semantic roles and spatial language). To this end a joint effort between academia and industry will again be advantageous and is thus to be promoted also in order to identify new areas that are mature for standardisation activities.

4. Resource Availability: Sharing and Distribution

By *availability* here it is intended the way a given resource is actually made available for use by third parties. This implies decisions about licensing and business models.

Opt for openness of LRs, especially publicly funded ones: the availability of massive quantities of open data could transform the NLP industry, as suggested for translation technologies (van der Meer, 2011). The LR community has started to embrace this view and is inclined to think of open data as digital resources distributed under open source-type licenses allowing them to be used, modified (and redistributed). However, reluctance in fully embracing an open data model is still common. To share resources, both data and tools, has become a viable solution towards encouraging open data, and the community is strongly investing in facilities for the discovery and use of resources by federated members. These facilities, such as the META-SHARE infrastructure[8], could represent an optimal intermediate solution to respond to the need for data variety, ease of retrieval, better data description and community-wide access, while at the same time assisting in clearing the intricate issues associated with IPR[9].

Ensure that publicly funded resources are publicly available either free of charge or at a small distribution cost: the results of a questionnaire carried out by FLaReNet strongly advised that at least those resources that are developed with public funding should be made openly available. For mixed-funded initiatives (private/public), it should be ensured that there is an agreement to make resources available at fair market conditions right from the start. Another suggestion is to ensure openness of resources for most types of uses, making use of standardised licenses where available, and creating LRs in collaborative projects where resources are exchanged among project participants after production.

Clear Intellectual Property Rights at the early stages of production; try to ensure that re-use is permitted: we do not yet have a sufficient grasp of the trans-border legal issues in the EU to support enhanced resource sharing and legally protect LRs against improper reuse, copying, modification etc. The Berne Convention for the Protection of Library and Artistic Works extends copyright protection to creators in countries other than their own, but enforcement is still a national issue and is therefore implemented in different ways. In addition to this, the availability and use of huge quantities of web data as useful resources creates a novel situation that raises further legal problems. On the one hand IPRs (especially authorship) need to be protected; but on the other they tend to restrict accessibility to and usability of language resources. The current trend is towards a culture of free/open use with less protective holders' rights. Creative Commons, for example, is one of the most widely used license models for language resources (see Google, Wikipedia, Whitehouse.gov, Public Library of Science, and Flickr). From a practical point of view, producers of language resources should try to clear IPR at the early stages of production, ensuring that re-use is permitted.

Educate key players with basic legal know how: it is crucial to disseminate a certain amount of legal knowledge/know-how to educate all (major) players in the LRT area. It is also important to inform a number of lawyers about community concerns so they can develop adequate frameworks to address such issues. Moreover, it is important that such legal experts are asked to intervene

[8] http://www.meta-share.eu/

[9] See also DiPersio (2011)

in the initial phases of resource production, to ensure that all legal (and also ethical, privacy and other) aspects are taken into consideration when planning for long-term LR sharing and distribution.

Elaborate specific, simple and harmonized licensing solutions for data resources: the community should avoid one-size-fits-all solutions. There are a large number of licensing schemes already in use today, some are backed by strong players (ELRA, LDC, open source communities such as Creative Commons[10], GNU General Public License, etc.), others have been drafted bilaterally and in some cases by the legal departments of data providers. *It is crucial that such licensing is harmonized and even standardized*. Licensing schemes need to be simplified through broad-based solutions for both R&D and industry. Electronic licensing (e-licenses) should be adopted and current distribution models to new media (web, mobile devices, etc.) should be accepted.

5. Resource Coverage, Quality, Adequacy

With the current data-driven paradigm in force, innovation in LT crucially depends on language resources. Accent is being increasingly put on *high quality and huge size of resources*, and as production (still) takes a lot of effort and is very costly, development of the resources for future technologies and applications must start now in order to positively impact the development of multilingual technologies such as Machine Translation, cross-lingual and Web 3.0 applications.

Despite the vast amount of academic and industrial investment, there are not enough available resources to satisfy the needs of all languages, quantitatively and qualitatively. Language resources should be produced and made available for every language, every register, every domain to guarantee full coverage, high quality and adequacy for the various LT applications. We need *the right amount, the right type and the right quality* of resources.

Increase quantity of resources available to address language and application needs: dependence on data creates new disparities for under-resourced languages and domains. It is estimated that 95% of web pages are in the top 20 languages (Pimienta et al., 2009). Naturally, smaller language communities produce much less data than speakers of the languages dominating the globe. The same problems occur for language data in narrow domains with their own specific terminological and stylistic requirements. To ensure Universal Linguistic Rights and massive deployment of LT applications, language services will need to be provided for everyone in their own mother tongue. Funding must be found to cover all languages (including the world's less-well represented languages) in future multilingual applications by developing language resources for all languages. Thus, provision of high quality resources for all European languages, including minority ones is a priority now, in order to avoid disparity in the future.

Implement BLARKs for all languages, especially less-resourced ones: for the particular advancement of LTs, Basic Language Resource Kits[11] (or BLARKs) should be supported and developed for all languages and, at least, main applications (Machine Translation, Information Retrieval, Question Answering to mention a few). Also, as many of the undocumented languages of our cultural legacy may become extinct in the digital age, minority and fringe languages should be comprehensively represented through spoken and written corpora, and manuscripts should be digitized.

In this direction, first the BLaRK concept needs to be worked out in detail, so that it can be embodied as a standard, and possibly planned revision sessions should be set, as it is intrinsically a dynamic notion that changes in time with the change in technology development in the different countries. Second, regular BLaRK surveys must be conducted to produce a clear picture of technology trends, and establish (and regularly update) a roadmap covering all aspects of LTs. Third, resource production should be funded on the basis of BLaRK-like criteria, i.e. giving priority to the development of "missing" resource types for each language.

Provide high quality resources for all European languages: high quality resources should be regarded as a key driver for effective technology in broad areas (e-content, media, health, automotive, telecoms, etc.). To this end and to reduce the amount of human intervention and revision, automatic techniques should be promoted to guarantee quality through error detection and confidence assessment.

Address formal and content quality of resources by promoting validation and evaluation: the promotion of validation and evaluation can perform a valuable role in fostering the improvement of formal and content quality of resources.

Devise new methods for LR quality check: new tools should be developed and maximal use of existing tools should be ensured for the *automatic or semi-automatic formal and content validation of language resources*. The requirements for language resource quality are to be assessed by a think-tank composed by recognized experts from a broad spectrum of the community, the technologies and the various modalities.

Establish a European evaluation and validation body: evaluation in Europe is currently carried out by individual institutions (such as ELDA and CELCT[12]) and by short-term projects (e.g. the TC-STAR[13] and CHIL campaigns[14]), but there is no sustained European-wide coordination, as there is in the US (NIST) or Japan (NII[15]).

In specific areas, the community has organised itself to

[10] http://creativecommons.org/

[11] http://www.blark.org/

[12] http://www.celct.it/

[13] http://www.tcstar.org/

[14] CHIL (Computers in the Human Interaction Loop): http://chil.server.de

[15] http://www.nii.ac.jp/

carry out regular evaluations (e.g. CLEF 2000-2010[16], and Semeval[17]) but with limited funding and much community good will. It would be of utmost importance to *establish common and standard LT evaluation procedures in Europe*. The establishment of such procedures would boost research around evaluation measures, as already happened in the US.

Create an infrastructure for coordinated LRT evaluation: the creation of a European infrastructure enabling a coordinated evaluation of LRTs is a priority. Setting up an evaluation management and coordination structure would ensure a viable, consistent and coherent programme of activities that can successfully scale up and embrace new communities and technological paradigms. This should be coupled with the establishment of a sustainable technical infrastructure providing data, tools and services to carry out systematic evaluation. This could be a distributed infrastructure involving existing organizations.

Carry out evaluation in real-world scenarios: evaluation should encompass technologies, resources, guidelines and documentation. But like the technologies it addresses, evaluation is constantly evolving, and new, more specific measures using innovative methodologies are needed to evaluate the reliability of semantic annotations, for example. Current evaluation campaigns sometimes create rather artificial settings so they stay 'academically clean', making the tasks they measure somewhat unrealistic. One of the most critical challenges, therefore, is to introduce new types of campaigns, possibly based on task-based evaluation.

Promote evaluation and validation activities of LRs and the dissemination of their outcomes: in order to foster evaluation activities, it would be important that they were highlighted as a major research topic (which includes research on metrics, methodologies, etc.) especially as a PhD subject. Thorough dissemination and information of activities and achievements should be done through LRT evaluation portals (e.g. the ELRA HLT evaluation portal[18]).

Assess maturity of technologies for which resources should be developed and draw a list of top twenty technologies for which language resources should be developed: it is important to assess the availability of existing resources with respect to their adequacy to applications and technology requirements. This involves assessing the maturity of the technologies for which new resources should be developed. We recommend, to this end, to closely monitor research developments through publications and patent filing, and to draw a list of the top 20 technologies in order to ensure that crucial resources are developed in at least ten of these, in a publicly and fully funded framework. Regular evaluation campaigns to assess the progress made by such technologies with respect to the state-of-the art is also desirable, especially if conducted inside an evaluation framework along the lines depicted above.

[16] http://www.clef-initiative.eu/

[17] http://www.cs.york.ac.uk/semeval-2012/

[18] http://www.HLT-evaluation.org/

6. Resource Sustainability

Sustainability covers preservation, accessibility, and operability (among other things) that all have mutual influences. Currently, most resource (data and software) building and distribution is based on short-term projects, which often leads to the loss of resources when the projects end.

LRs must be accessible over the long term. This means a) archiving and preserving the data by the production unit, and also archiving them off-site (e.g. in very-long term archiving/data centres); b) maintaining LRs in an appropriate way and c) making sure that linguistic tools and resources are sustainable, e.g. by requesting resource accessibility and usability for a given time frame.

Ensure that all LRs to be produced undergo a sustainability analysis as part of the specification phase: a sustainability analysis must thus be part of a resource specification phase, and it is important that funding agencies impose a sustainability plan mandatory for those projects that are concerned with production of language resources.

Foster use of a sustainability model: the FLaReNet project has developed an analytic model of sustainability in which extrinsic and intrinsic factors are taken into account[19]. Use of this or similar models should be fostered by the entire community.

7. Resource Recognition

LRs (both data and software) are time-consuming, costly and increasingly require a considerable share of research budgets. *The entire ecosystem around LRs needs substantial support and recognition*. Small labs and individual researchers are not keen on depositing or sharing their resources because there has been little incentive to do so. There are almost no rewards for researchers and institutions to share, preserve and maintain resources, and this now poses a number of serious problems.

Give greater recognition to successful LRs: or instance by means of prizes, seals of recognition and the like.

Develop a standard protocol for citing LRs: LRs deserve credit and should be cited in a similar way to sources in scientific publications. A model for citing LRs would therefore be highly desirable such as a standard citation framework that would allow for citing LRs in a uniform way (this would also enforce the use of minimal metadata descriptions) and for which LRP will be responsible and credited for.

Along the lines followed in other fields, especially in Biology, a "Language Resources Impact Factor (LRIF)" should be defined in order to enforce the practice of citation of resources on the model of scientific paper authoring and to calculate actual research impact of resources.

[19] For a detailed account of this model, see Calzolari et al. 2011b, Chapter 2.

Support training in production and use of LRs: there should be more training in production and use of LRs, and LRs should also be used more widely in education. Training in the production and use of LRs should become part of curricula especially in Computational Linguistics and Language Technology.

8. Resource Development

Development of language resources refers to the entire production cycle of a resource.

The proper management of the "life cycle" of language resource creation has attracted less attention and has been largely overlooked in our community.

Ensure strong public and community support to definition and dissemination of resource production best practices: a reference model for creating LRs will help address the current shortage of resources in terms of breadth (languages and applications) and depth (data quality and volume). Such reference model should also include an accurate estimate of the production costs.

The creation of new resources from scratch should be discouraged wherever resources can be found for a given language and/or application. We should *encourage re-use and re-purposing via a "recycling" culture to ensure the reuse* of development methods, existing tools, and translation/transliteration tools, etc.

Work towards the full automation of LR data production: with production costs constantly increasing, there is a need to invest in innovative production methods that massively involve automatic procedures, so as to reduce human intervention to a minimum. We must improve existing tools and introduce new automation techniques, especially for higher-level semantic, content-related and multilingual tasks. We must also foster the evaluation of real-life applications so that research can gradually approach industry needs in terms of information volume and granularity. Support must be given to academic and industrial involvement in research on automatic methods for production and validation of LRs, to allow a more accurate assessment of the automatic methods for building LRs for real-life applications.

Invest in Web 2.0/3.0 methods for collaborative creation and extension of high-quality resources, also as a means to achieve better coverage: given the high cost of language resource production, and that in many cases it is impossible to avoid the manual construction of resources (e.g. if accurate models are requested or if there is to be reliable evaluation) it is worth considering the power of social/collaborative media to build resources, especially for those languages where there are no language resources built by experts yet.

Production and annotation of LRs can be carried out as collaborative projects. Existing LRs should be "opened up" for collaborative annotation and reuse of the annotated results. At the same time, new tools are to be developed and existing tools are to be adapted to the needs of collaborative work. However, the use of crowd-sourcing raises ethical, sociological and practical issues for the community. It is not yet clearly understood for example whether all types of LRs can be obtained collaboratively by using naïve annotators; more research is therefore needed on both the technical (e.g. accurately comparing the quality and content of resources built collaboratively and those built by experts) and ethical aspects of crowd-sourcing.

Start an open community initiative for a large Language Knowledge Repository: there are currently insufficient resources and sources to solve the problem of creating free, large-scale resources for the world languages, even for those with a reasonable web presence. The collaborative accumulation and creation of data appears to be the best and most practicable way to achieve better and faster language coverage and in purely economic terms could well deliver a higher return on investment than expected.

9. An Infrastructure of Language Resources

The need for an infrastructure for Language Resources was the first recommendation since the beginning of FLaReNet and derives historically from the recognition of the infrastructural role of LRs as essential "building blocks" for language technologies. Such an infrastructure will ease recovery and use of LRs through appropriate facilities that allow their availability, visibility, and easy accessibility.

Build a sustainable facility for discovering, accessing and sharing data and tools: this infrastructure will help, *in primis*, to make language resources available, visible and easily accessible. Second, the infrastructure will facilitate sharing and exchange of language resources. An initiative of this kind needs continuous support by Policy Makers to ensure steady development; also, promotional activities on the LRPs' side are needed to secure visibility and participation. The basic principles of an infrastructure for language resources and technologies require a community approach that brings together and builds on current experiences and endeavours. It is necessary to define and agree on the basic criteria and dimensions for an appropriate governance, and define the basic data and software resources that should populate this infrastructure. Multilingual coverage, the capacity to attract providers of useful and usable resources, improvements in sharing mechanisms, and collaborative working practices between R&D and commercial users are key aspects. There must also be a business-friendly framework to stimulate the commercial use of these resources, based on a sound licensing facility, ease of access, ease of conversion into uniform formats.

Establish an international hub of resources and technologies for speech and language services, by creating a mechanism for accumulating speech and language resources together with industries and communities: the content of the infrastructure should not be limited to data, though. Instead, it has to be seen as an international hub of resources and technologies for speech and language services from industries and communities. The development and proposal of (free) tools and more

generally Web services (comparable to the Language Grid platform [20]), including evaluation protocols and collaborative workbenches is deemed essential in such LR infrastructure. The accumulation and sharing of resources and tools in a single infrastructure would lower the cost of R&D for new applications in new language resource domains.

10. International Cooperation

Cooperation among countries and programs is essential to drive the field forward in a coordinated way and avoid duplication of efforts and fragmentation.

It is crucial to discuss future policies and priorities for the field of LRTs not only on the European scene, but in a worldwide context. This is true both when we try to highlight future directions of research, and – even more – when we analyse which infrastructural actions are needed. The growth of the field must be complemented by a common effort that looks for synergies and overcomes fragmentation.

Maintain a public survey on the LT and LR situation worldwide, based on FLaReNet and META-NET: the availability of up-to-date surveys on the situation of language resources and language technologies worldwide is of foremost importance. Both the FLaReNet and META-NET projects have produced such surveys, and it is recommended that they are maintained and further expanded. Similarly, community-driven initiatives such as the LRE Map, META-NET Language Matrixes, and FLaReNet Network of International Contact Points are valuable assets that would deserve continuous maintenance with public funding.

Share the effort of production of LRs between international bodies and individual countries: a coordinated effort at the international level, with shared effort of supra-national and national bodies, would help by providing less resourced languages with examples and best practices, such as defining a commonly agreed on set of basic LRs that have already proven necessary for producing LTs efficiently for better represented languages. This kind of international effort should also try to identify the gaps and draw up an appropriate roadmap to fill them.

Establish an International Forum to share information, discuss strategies and declare/define common objectives: networking and support actions must be conducted more intensively, with establishment of international committees that have formal recognition. 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/define 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.

11. Conclusion

The FLaReNet Strategic Language Resource Agenda gathers, in a coherent organization, the major high-level recommendations collected around the many meetings, panels and consultations of the community, as well as the results of the surveying and analysis activities carried out in the framework of the EU FLaReNet project. It is therefore the result of a permanent and cyclical consultation that FLaReNet has conducted inside the community it represents – with more than 300 members – and outside it, through connections with neighboring projects, associations, initiatives, funding agencies and government institutions. In this paper we have synthesized for Language Technology and Natural Language Processing players at large, policy-makers and funding agencies, a preliminary plan for actions and infrastructures that could become the basis for future initiatives in the field.

[20] http://langrid.org/en/index.html

12. Acknowledgements

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

13. References

Calzolari, N. Núria Bel, Choukri, K., Francopoulo, G., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., Soria, C. (2011). *Final FLaReNet Deliverable. Language Resources for the Future – the Future of Language Resources. The Strategic Language Resource Agenda*. FLaReNet 2011.

DiPersio, D. (2011). Is our relationship with open data sustainable?. In *Proceedings of the Third FLaReNet Forum*. Venice, Italy, 26-27 May 2011.

Ide, N., Pustejovsky, J. (2011). An Interoperability Challenge for the NLP Community. In *Proceedings of the Third FLaReNet Forum*.Venice, Italy, 26-27 May 2011.

van der Meer, J. (2011). Imagine we have 100 Billion Translated Words at our Disposal. In *Proceedings of the Third FLaReNet Forum*. Venice, Italy, 26-27 May 2011.