The Strategic Language Resource Agenda

Nicoletta Calzolari, Valeria Quochi, Claudia Soria

CNR - Istituto di Linguistica Computazionale "A. Zampolli", Italy

with the contribution of

Núria Bel, University Pompeu Fabra, Spain

Gerhard Budin, Universität Wien, Austria

Khalid Choukri, ELDA, France

Joseph Mariani, LIMSI/IMMI-CNRS, France

Monica Monachini, CNR-ILC, Italy

Jan Odijk, Universiteit Utrecht, Netherlands

Stelios Piperidis, ILSP/"Athena" R.C., Greece

Introduction

Despite the complexity of handling its languages, the European Union has established that cultural and language differences are a unique asset to be preserved. Europe needs to find means - such as technological ones - to overcome the language barriers to support 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. Although Language Technologies may consist of language independent engines, they depend on the availability of language-dependent knowledge under the form of Language Resources for their real-life implementation. At the same time, it is proved that a critical mass of Language Resources 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, 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 Language Technology solutions to support multilingualism should raise concern among EU officials.

What has been done already?

In the US, accumulation of language data was a priority since the early 90's, when the fast pacing of statistical approaches in Natural Language Processing 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 Language Resources, with the launching of the European Language Resource Association (ELRA[3]) in 1995, which later promoted Language Resources and evaluation through the LREC conferences[4] that began in 1998. The importance of Language Resources for driving Language Technologies was well recognized in Europe by the European Commission, which launched in the '90 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 and reports results on English language phenomena. Technologies and applications grow more and more advanced for English, 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 Language Technology as a support to multilingualism is not yet typical of the European Commission and the 27 Member States of the European Union.

Language Resources are key to the development of NLP applications for a multilingual Europe.

Realizing the appropriate multilingual Language Resource infrastructure needs a collaborative and coordinated effort from all stakeholders. While there has been considerable progress in technology developments in the last decade, the significant challenge of overcoming current fragmentation and imbalance inside the Language Technologies community for all languages still remains an issue.

Together, and under the umbrella of a shared view of today's priorities, a future can be shaped in which full deployment of Language Resources and Technologies is consolidated through coordination of programs, actions and activities. 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 do cheaper and better for new languages. In this Strategic Agenda we tell you how.

How to use the FLaReNet Strategic Agenda

This Strategic Agenda highlights the **most pressing needs** for the sector and presents **a set of recommendations** for the development and progress of Language Resources and Technologies in Europe. The recommendations are the result of a three-year consultation of the FLaReNet project (www.flarenet.eu), 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 Language Technologies, 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 Language Resources and Technologies 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 nine FLaReNet dimensions

We can think of nine dimensions that are relevant for the field of Language Resources: 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*.

---

[6] A longer version of the recommendations is available in Calzolari et al. 2011b.
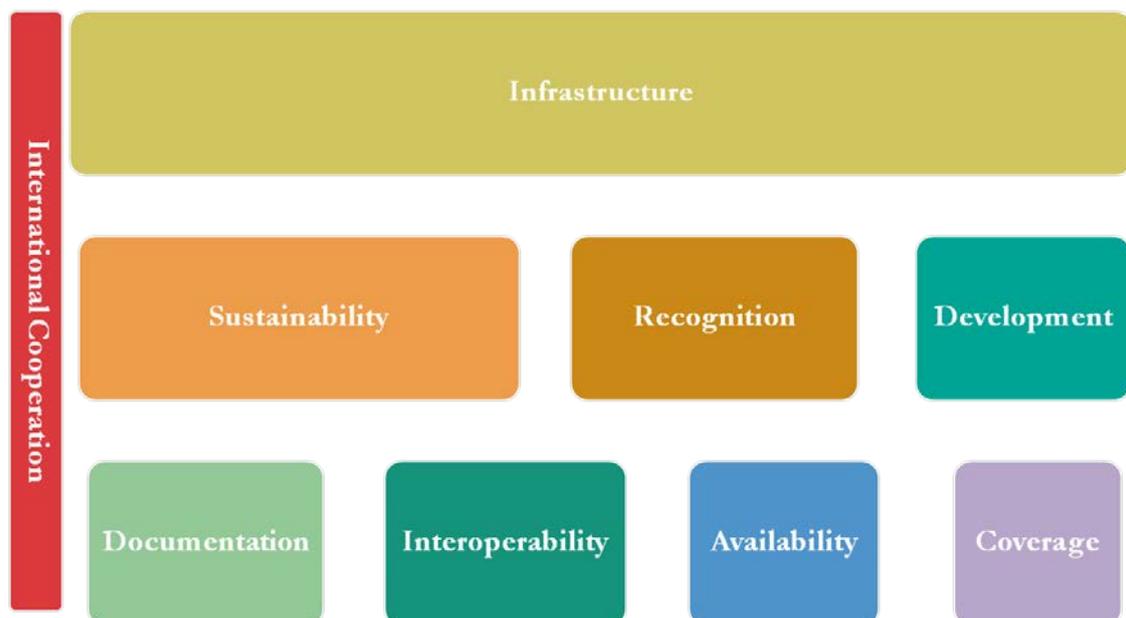
Some of these dimensions are of a more infrastructural nature, some are more related to research and development, some yet more to political and strategic aspects, but they all must be seriously considered when making up a strategy for the future of the field. All of them eventually have an impact in the development and success of LRs, and represent the areas where actions need to be taken to make the field of Language Resources and Technologies 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 Language Resources and Technologies ecosystem**.

Recommendations are mainly targeted to two wide groups of players: Language Resource Producers (LRP, broadly encompassing academic and industrial developers) and Policy Makers (PM, i.e. National or supra-national funding agencies, politicians, etc.)[7].

1.1 Resource Documentation

"Ensure that Language Resources are accurately and reliably documented"

Accurate and reliable documentation of Language Resources is an undisputable need. Instead, as of today, Language Resources are still often poorly documented or not documented at all, and, even when available, documentation is often not easy to find.

LRP: **Devise and adopt a widely agreed standard documentation template for each resource type, based on identified best practice(s)**

Documentation allows Language Resources to be used by people different from those who designed and developed them. "Researchers and practitioners will be able to find, access and process the data they need. They will be confident in their ability to use and understand data and they can evaluate the degree to which data can be trusted" (Giaretta 2011). 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

---

[7] These abbreviations will be used throughout the document, in the colour boxes summarising the main recommendations.

enforced. This entails developing standard specifications for LR documentation.

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

Human users need information that helps them:
   a) Find a resource and assess its usefulness for a given application
   b) Understand the production process, the use of best practices, and intended exploitation
   c) Assess the quality of a resource
   d) Replicate processes and results
   e) Handle idiosyncrasies or documented errors.

Machines need (machine-understandable) information to:
   a) Discover and compare resources
   b) Validate formats and annotations
   c) Process annotations appropriately
   d) Retrieve relevant parts of a resource for a given use
   e) Enable other as yet unexplored new uses.

Uniform documentation is vital. A common documentation template should be defined, promoted, and enforced for all contracts for publicly funded projects.

Documentation should include:
   a) A high-level description giving the non-expert but interested reader a good idea of what is in the resource, including general information such as owner/copyright holder, format and encoding of the data and the files, languages(s), domains, intended applications, applications in which the data has been used, and details about basic quality assessment (in particular for availability/reliability of the encoded information).
   b) Information on the theoretical framework, background, and/or the "philosophy" of the resource.
   c) Specifications of the methodology used to create the resource so that others could replicate the process.
   d) Annotation specifications (with data categories and their semantics) and guidelines, e.g. guidelines used by annotators.
   e) Information on the use of standards (at all levels: production, annotation, validation, etc.).
   f) Specification of the methodology or guidelines used to assess the quality of the resource (if validation is conducted) and the report on such validation.
   g) Estimates of the effort required to create the resource (in any reproducible unit, e.g. person/month).

Language Resource Providers should always allocate time and manpower to documentation from the very start of production of Language Resources. 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.

LRP: **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**

An effort must be made to collect all existing Language Resource documentation and make it easily available. To this end, the design and construction of a (virtual) repository of specifications, guidelines, and documentation of Language Resources, starting with reference resource

models or widely known and used resources (e.g. WordNet, Penn TreeBank, …) is a priority task[8].

Documentation is also the gateway to Language Resource discovery. Ensuring that Language Resources are discoverable is the first step towards promoting the data economy. In order to make Language Resources discoverable, Language Resource Providers should always document their resources, **using standard metadata** and **unique resource identifiers**. Therefore, definition and adoption of standardised metadata must be the first priority and first step for all Language Resource Providers.

LRP and PM: **Ensure that appropriate metadata are consistently adopted for describing LRs**

A useful best practice in this line is to document one's resource in an open catalogue, such as the ELRA Universal Catalogue[9], the META-SHARE[10] or CLARIN[11] catalogues. By providing standard APIs to query the catalogues, it would then be possible to harvest individual catalogues and compile unified catalogues by metadata aggregation. *Language Resource Providers must provide appropriate metadata description for all language resources distributed*, preferably in one of the widely used metadata schemes. *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.

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 centres, archiving institutions and projects, and other providers tend to use their own, non-interoperable metadata sets to describe their data. Resources are also described at different levels of granularity depending on who does it. Another drawback is that it is often impossible to combine data from multiple sources to create new metadata sets for specific uses.

LRP: **Set up a global infrastructure of common and uniform and/or interoperable metadata sets**

The key priority is therefore to work towards the **full interoperability of metadata sets**. As there are many differing metadata sets and search engines, harmonisation is a central problem for the community.

Useful initiatives in this direction are community-based documentation initiatives, such as the *LRE Map*[12], 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.

LRP and PM: **Develop and support community-wide initiatives such as the LRE Map**

It is important to shift the metadata mind-set towards the creation of **machine-understandable metadata**, i.e. pieces of information about (digital) resources that can be processed automatically (e.g. they must have a formal syntax and a declared semantics). This will make metadata browsable/accessible from various tools for various purposes. **Formalized metadata elements** are to be used as much as possible by Language Resource Providers, and they should contain also all aspects of the documentation of the Language Resource that can be formalized. It has to be ensured that all data categories used in the metadata and in the data are registered in a **data category registry** to ensure semantic interoperability. The development of techniques for automating the process of metadata creation would also help

---

[8] An activity along these lines has been started within FLaReNet, see
http://www.flarenet.eu/?q=FLaReNet_Repository_of_Standards_and_Guidelines

[9] http:// universal.elra.info/

[10] http:// www.meta-net.eu/meta-share/meta-share

[11] http:// www.clarin.eu/vlo/

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

spread the adoption of machine-understandable metadata.

1.2   Resource Interoperability

"Design and set up an interoperability framework for Language Resources and Technology"

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.

We can distinguish between *syntactic* and *semantic* interoperability. **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).

LPR: **Ensure syntactic and semantic interoperability of Language Resources**

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).

Interoperability of resources and data is also 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. Data access and links within and across this data is as important as the actual quantity, and data interoperability is essential to it.

The design of interoperability tasks will also help to determine which emerging standards are most interoperable. Interoperability tests can also replace aspects of validation.

LRP and PM: **Set up an "interoperability challenge" as a collective exercise to evaluate (and possibly measure) interoperability**

While, on the one hand, it is increasingly recognised that standards are key to resource sharing, re-usability, maintainability and long-term preservation, Language Resource Providers 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. As a result, many types of resources and many levels of information and annotations are not standardised.

Most existing resources use unique representation formats and conventions, so other people have to first understand the format and then build ad hoc conversions in order to use the resource data for their own activities. This makes it especially difficult to draw on different sources to build on-demand resources needed by emerging web technologies. The lack of standardisation also makes it difficult to evaluate the quality and value of resources for a given application. A basic level of standardisation is particularly vital for so-called "Less-Resourced" Languages.

PM: **Create a permanent Standards Observatory or Standards Watch**

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. Industry involvement in standardisation initiatives will grant that there is broad-based adoption of standards. An initiative towards these goals has been established by FLaReNet with a document (Calzolari et al. 2011a) proposing an overview of the current scene towards an Interoperability

Framework. This document acts as a reference point for the current standards encouraged by the community for adoption[13].

Investment at the supra-national level in standardisation 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.

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) 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.). Language Resource Producers, 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 Language Resources that are easily amenable to reuse (e.g. adopt formats that allow easy reuse). The creation of "official" validators to check compliance of Language Resources with basic linguistic standards is an activity to be pursued and encouraged by funding agencies.

However, enforcing standards cannot be a purely top-down process. It must be backed by information about contributions from different user communities. 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.

Language Resource Providers 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.

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".

At the same time, 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.

1.3   Resource Availability: Sharing and Distribution

> *"Make resources easily and readily available, within an adequate IPR and legal framework"*

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

PM: **Invest in standardisation activities**

PM: **Encourage/enforce use of best practices or standards in LR production projects**

LRP: **Make standards operational and put them in use**

LRP and PM: **Set up training initiatives to promote and disseminate standards to students and young researchers**

LRP and PM: **Identify new mature areas for standardisation and promote joint efforts between R&D and industry**

---

[13] This initiative is in close synchronization with other relevant ones such as CLARIN, ELRA, ISO and TEI and META-SHARE.

business models.

There is strong impulse towards ***open data*** nowadays[14], in the sense of data that are easily obtainable and can be used with few, if any, restrictions. "Producers of data will benefit from opening it to broad access and will prefer to deposit their data with confidence in reliable repositories (Giaretta 2011)". According to the Open Knowledge Foundation[15], data is open if "you are free to use, reuse, and distribute it – subject only, at most, to the requirement to attribute and share-alike". The language resource 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). While the majority of Language Resources experts advocate for data openly available and reusable, it is a fact that 55% of the resources documented by the LRE Map are freely available. However, reluctance in fully embracing an open data model is still common. As pointed out during the third FLaReNet Forum, a clear understanding is needed of the pros and cons of closing or opening up data: "What do we gain by closing data? Are business models based on closed and heavily guarded Language Resources actually successful? Are we losing opportunities for growth by not systematically exploiting common sharing and synergies? What do we lose by opening up data? Can the direct income lost be compensated by direct or indirect financial gains?" (Sellis 2011). 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[16], 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[17].

LRP and PM: **Opt for openness of LRs, especially publicly funded ones**

A study by KPMG Inc. on the Canadian Spatial Data Infrastructure (Sears 2001), concluded that closed, restricted data has major economic harm: "the consequences [of cost recovery] for businesses are higher marginal costs, lower research and development investments and threatened marginal products. The results for consumers are negative: higher prices and reduced products and services. The overall economic consequences... are fewer jobs, reduced economic output by almost $2.6 billion and a lower gross domestic product." One branch of Language Technology, Question Answering, has benefited greatly from the availability of open data resources, especially the research datasets created for the yearly TREC[18] question answering track. Thinking along these lines, ***the availability of massive quantities of open data could transform the NLP industry***, as suggested for translation technologies (van der Meer 2011). The success of the sharing approach is well represented by TAUS[19], where many translation leaders have started sharing their translation memories in the Taus Data Association repository.

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

LRP and PM: **Ensure that publicly funded resources are publicly**

---

[14] See, for instance, the Open Data Strategy for Europe, recently launched by the European Commission.

[15] http://okfn.org/

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

[17] See also DiPersio (2011)

[18] http://trec.nist.gov/

[19] http:// www.translationautomation.com/

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 Language Resources in collaborative projects where resources are exchanged among project participants after production.

Of course, certain types of data are and will probably remain not shareable, either because containing personal data, or for confidentiality or competitive reasons. Non shared data can still be exploited, for instance through **shared services**.

In the meantime, it is important to define appropriate criteria for different levels of openness and to define "best practices" for making resources available that address different constraints that may be faced.

However, achieving large-scale, open datasets is only one of the many general requirements for rapid, open advancement of language technologies. As pointed out during the last FLaReNet Forum, it only makes sense to talk about data openness after clearing the discovery and reusability stumbling blocks. Before data can be openly usable, they need to be a) easily retrievable and b) easily reusable. Point a) is addressed by the Documentation issue in 1.1, while reusability has to do with resource Interoperability (see 1.2). Whether making data open or not is a matter of choice at the licensing level, it crucially depends on choices made at the very early stage of resource planning and production, i.e. by ensuring content and formal interoperability and proper documentation.

On the legal side, it must be borne in mind that Intellectual Property Rights (IPR) issues are crucial to facilitating growth in our sector, yet they pose real 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.

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.

LRP: **Clear IPR 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 Language Resources 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. ***Legislation is lagging behind the technology***. 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).

The Language Resource community is also facing major questions about how to use blogs, newsgroups, web video, SMS and social network sites as data, as there are virtually no laws, regulations or court decisions governing these media. This means that most resources are kept in-house for research or other internal use, otherwise individuals and organisations risk law suits due to some form of infringement.

LRP and PM: **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 Language Resource and Technology 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 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 Language Resource sharing and distribution.

The challenge for both the Language Resource and Technology community and policy makers is to **push for the development of a common legal framework that would facilitate resource sharing efforts** that do not break the law. It is of foremost importance that legislation regarding Language Resource use be harmonised for all types of Language Resources, and that free use of Language Resources be allowed, at least for research or non-profit purposes.

The community should also 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[20], open source communities such as Creative Commons[21], 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 harmonised and even standardised**. 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.

LRP and PM: **Elaborate specific, simple and harmonised licensing solutions for data resources**

1.4   Resource Coverage, Quality, Adequacy

"Address appropriate coverage in terms of quantity, quality and adequacy to technological purposes"

With the current data-driven paradigm in force, innovation in Language Technology 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.

"The development of killer applications crucially depends on the availability of large quantities of data. Cross-lingual knowledge extraction, for instance, is a challenging high impact task for the near and mid future. Today, the tasks seems to be achievable because critical mass of technology is collected" (Mladenic and Grobelnik 2011).

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 Language Technology applications.

We need *the right amount, the right type and the right quality of resources*.

Resource Quantity

One thing that must be borne in mind is that 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

LRP and PM: **Increase quantity of resources available to address language and application needs**

---

[20] http://www.ldc.upenn.edu/

[21] http://creativecommons.org/

narrow domains with their own specific terminological and stylistic requirements. 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.

To ensure ***Universal Linguistic Rights*** and massive deployment of Language Technology applications, language services will need to be provided for everyone in their own mother tongue. This priority is also evidenced by the number of localisation projects for most existing applications, be they proprietary or open-source. As the quality of volunteer-based only localisation is not high, 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. New methods of resource development can be exploited to achieve better coverage, for instance shared or distributed ones.

Specifically for the advancement of Language Technologies, ***Basic Language Resource Kits*** (or ***BLARKs***[22]) 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.

<div style="float:right; font-style:italic;">

LRP and PM: Implement BLARKs for all languages, especially less-resourced ones

</div>

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 Language Technology. 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.

Allocating funding to cover all languages (in particular less represented ones) and all basic needs of language technology remains thus a high priority for ensuring multilingual applications in the future.

Resource Quality

High quality resources should be regarded as a key driver for effective technology in broad areas (e-content, media, health, automotive, telecoms, etc.).

LRP and PM: **Provide high quality resources for all European languages**

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.

PM: **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.

New methods for the quality check of language resources should be provided. 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.

LRP: **Devise new methods for LR quality check**

The community and policy makers should ensure that a quick quality check

---

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

can be carried out, at least for existing resources. Projects/players should be pushed to specify a validation procedure before Language Resource production starts. A "***Quality Seal of Approval***", on the model of the "Data Seal of Approval[23]" should be defined and endorsed by the community.

Evaluation in Europe is currently carried out by individual institutions (such as ELDA and CELCT[24]) and by short-term projects (e.g. the TC-STAR[25] and CHIL[26] campaigns), but there is no sustained European-wide coordination, as there is in the US (NIST) or Japan (NII[27]).

In specific areas, the community has organised itself to carry out *regular evaluations* (e.g. CLEF 2000-2010[28], and Semeval[29]) but with limited funding and much community good will. It would be of utmost importance to ***establish common and standard Language Technology evaluation procedures in Europe***. The establishment of such procedures would boost research around evaluation measures, as already happened in the US.

In the US, NIST plays a very important role in coordinating technology evaluation as it enables Language Resources to be controlled and streamlines the research and development of applications with genuine commercial promise.

The ***creation of a European infrastructure*** enabling a coordinated evaluation of Language Resources and Technologies 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.

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. For practical purposes, it would be helpful to have guidelines and do's/don'ts for this.

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 Language Resource and Technology evaluation portals (e.g. the ELRA HLT evaluation portal[30]).

Resource Adequacy

Not only do we need more data, but the *typology of data* needed has also

PM: **Establish a European evaluation and validation body**

LRP and PM: **Establish common and standard Language Technology evaluation procedures**

LRP and PM: **Create an infrastructure for coordinated LRTs evaluation**

LRP: **Carry out evaluation in real-world scenarios**

PM: **Promote evaluation and validation activities of LRs and the dissemination of their outcomes**

---

[23] http://www.datasealofapproval.org/

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

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

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

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

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

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

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

increased. "Nowadays data can be emails, Facebook walls, and exchanges on Twitter. Today, data is gathered not only from the Internet but also from supermarket receipts, mobile phones, cars, planes and soon even refrigerators, ovens and any type of electronic device we use will provide data. Much of the data that previously simply disappeared after having been used for a specific purpose, is now stored, distributed and even resold for analysis, interpretation or other purposes of which the best if not most frequent case is innovation" (Segond 2011).

First, 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. For instance, it appears that speech technologies require a more natural approach to voice input that can only be achieved by producing appropriate spoken language resources. 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.

> *LRP and PM*: **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**

1.5   Resource Sustainability

"Ensure sustainability of language resources"

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.

Collecting and preserving knowledge in the form of existing Language Resources instead should be a key priority.

Language Resources must be accessible over the long term. This means:

– Archiving and preserving the data by the production unit, and also archiving them off-site (e.g. in very-long term archiving/data centres).

– Maintaining Language Resources in an appropriate way.

– Making sure that linguistic tools and resources are sustainable, e.g. by requesting resource accessibility and usability for a given time frame.

> *LRP and PM:* **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.

The FLaReNet project has developed an ***analytic model of sustainability*** *in which extrinsic and intrinsic factors are taken into account*[31]. Use of this or similar models should be fostered by the entire community.

> *LRP and PM:* **Foster use of a sustainability model**

---

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

1.6 Resource Recognition

"Promote the LR ecosystem"

Language Resources (both data and software) are time-consuming, costly and increasingly require a considerable share of research budgets. ***The entire ecosystem around Language Resources 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.

Greater recognition to successful language resources (and their producers) should be given, for instance by means of prizes, seals of recognition and the like.

> *LRP and PM:* **Give greater recognition to successful LRs**

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

> *LRP and PM:* **Develop a standard protocol for citing LRs**

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.

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

> *LRP and PM:* **Support training in production and use of LRs**

1.7 Resource Development

"Define a reference model for future Language 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. A reference model for creating Language Resources instead 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.

> *PM*: **Ensure strong public and community support to definition and dissemination of resource production best practices**

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.

> *LRP and PM*: **Go Green: enforce recycling, reusing and repurposing**

The experience gained for one language can be used to process others. It is encouraging to see high-level applications for Less-Resourced Languages (instead of just the usual "taggers") such as Automatic Speech Recognition for Amharic, as these can pave the way for designing baseline systems for these languages.

Similarly, most language technologists use existing language resources as input and create content as by-products that could form useful language resources for others. Yet so far very few of these resources are made commonly available at the end of the production cycle. 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.

LRP: **Work towards the full automation of LR data production**

The coverage problem is so enormous that we must promote strategies that approach or **ensure full automation for (high-quality) Language Resource data production**. "Innovation needs data, but also the collection of data needs innovation" (Vasiljevs 2011).

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 Language Resources, to allow a more accurate assessment of the automatic methods for building Language Resources for real-life applications.

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.

LRP and PM: **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**

There are several experiments in crowd-sourcing data collection and Natural Language Processing tasks (Dolan 2011), and most of them look promising. Crowd-sourcing (such as Amazon's Mechanical Turk) makes it possible to mobilize large armies of human talent around the world with just the right language skills so that it is feasible to collect what we need when we need it, even during a crisis such as the 2010 earthquake in Haiti or the flood in Pakistan (Church 2011). For instance, it has been estimated that Mechanical Turk translation is 10 to 60 times less expensive than professional translation (Callison-Bruch and Dredze 2010).

Production and annotation of Language Resources can be carried out as collaborative projects. Existing Language Resources 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 Language Resources 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[32].

A particularly sensitive case is that of **less-resourced languages**, where language technology should be developed rapidly to help minority-language speakers access education and the Information Society. Basic language resources for all the world's languages could be created building a Web 2.0 site (using the same community computing power that generates millions of blogs) starting with the 446 languages currently present on the web. **Collaborative**

LRP: **Start an open community initiative for a large Language Knowledge Repository**

---

[32] See for instance (Zaidan and Callison-Burch 2011) about mechanisms for increasing quality of crowd-sourced data.

and Web 2.0 methods for data collection and annotation seem particularly very well-suited for collecting the data needed for the development of Language Technology applications.

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.

1.8   An Infrastructure of Language Resources

"Our vision is a scientific e-Infrastructure that supports seamless access, use, re-use and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance (High Level Expert Group on Scientific Data 2010)"

As a matter of fact, researchers and developers still have to spend quite some time consulting multiple catalogues or searching the web to find relevant Language Resources, and often they fail. Available resources are very often difficult to access for various reasons. Some are available from distribution centres (notably ELRA and LDC), others from portals of projects or associations, or directly from the web pages of the laboratories or researchers who developed the resource, or even from the owner itself. In many cases, unless the potential user already knows something about the resource (s)he might want to use (e.g. name, owner, or project), (s)he would find it difficult to discover new or as yet unknown resources. The identification and discovery of "new/unknown" resources is therefore a key priority right now and should be accompanied by the spread of a new culture of sharing and/or collaborative resource creation.

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. A small number of infrastructure initiatives (cf. CLARIN, META-SHARE) have emerged in order to solve these problems. However, without proper coordination between them, there is a risk of further fragmentation. An additional problem is that some of the current "infrastructure plans" are projects with limited time duration. It is therefore time to act more decisively to build synergy between all stakeholders in this field.

Thus first, ***a sustainable facility for sharing resource data and tools is to be built***. Broad participation is essential in building the infrastructure, so that acceptance by the public and contributors is ensured.

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 Language Resource Providers' 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

LRP and PM: **Build a sustainable facility for discovering, accessing and sharing data and tools**

LRP: **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**

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.

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[33]), including evaluation protocols and collaborative workbenches is deemed essential in such Language Resource 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.

communities

LRP: Develop and propose (free) tools and more generally Web services, including evaluation protocols and collaborative workbenches in the LR infrastructure

1.9   International Cooperation

"Promote synergies among initiatives at international level"

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 Language Resources and Technologies 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.

A coordinated effort at the international level 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.

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.

LRP and PM: Maintain a public survey on the LT and LR situation worldwide, based on FLaReNet and META-NET

For what concerns international cooperation for the development of resources, FLaReNet recommends to share efforts for the production of language resources between international bodies and individual countries/regions.

PM: Share the effort for production of LRs between international bodies and individual countries

International cooperation between infrastructure initiatives is also important for avoiding the duplication of effort, ensuring that standards are

---

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

truly international, and encouraging the free exchange of ideas.

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.

PM: **Establish an International Forum to share information, discuss strategies and declare/define common objectives**

References

Chris Callison-Burch, Mark Dredze. 2010. "Creating Speech and Language Data With Amazon's Mechanical Turk". In *Proceedings NAACL-2010 Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk*.

Nicoletta Calzolari, Monica Monachini, Valeria Quochi, Núria Bel, Gerhard Budin, Tommaso Caselli, Khalid Choukri, Gil Francopoulo, Erhard Hinrichs, Steven Krauwer, Lothar Lemnitzer, Joseph Mariani, Jan Odijk, Stelios Piperidis, Adam Przepiórkowski, Laurent Romary, Helmut Schmidt, Hans Uszkoreit, Peter Wittenburg. 2011. *The Standards' Landscape Towards an Interoperability Framework. The FLaReNet proposal*. FLaReNet 2011.

Nicoletta Calzolari, Núria Bel, Khalid Choukri, Gil Francopoulo, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, Claudia Soria. 2011. *Final FLaReNet Deliverable. Language Resources for the Future – the Future of Language Resources. The Strategic Language Resource Agenda*. FLaReNet 2011.

Kenneth Church. 2011. "Plan B". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

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

Bill Dolan. 2011. "Parallel Multilingual Data from Monolingual Speakers". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

David Giaretta. 2011. "Preparing to share the effort of preservation using a new EU preservation e-Infrastructure". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

High Level Expert Group on Scientific data. 2010. "Riding the Wave: How Europe can gain from the rising tide of scientific data". *Final report of the High level Expert Group on Scientific Data*. October 2010.

Nancy Ide, James Pustejovsky. 2011. "An Interoperability Challenge for the NLP Community". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

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

Dunja Mladenic, Marko Grobelnik. 2011. "Cross-lingual knowledge extraction". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

Daniel Pimienta, Daniel Prado and Álvaro Blanco. 2009. *Twelve years of measuring linguistic diversity in the Internet: balance and perspectives*. UNESCO, Paris.

Frederique Segond. 2011. "Turning water into wine : transforming data sources to satisfy the thirst of the knowledge era". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

Garry Sears. 2001. "KPMG Consulting Inc. for GeoConnections Policy Advisory Node", *Canadian Geospatial Data Policy Study*, March 2001.

Timos Sellis. 2011. "Open Data and Language Resources". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

Andrejs Vasiljevs. 2011. "How to get more data for under-resourced languages and domains?". In *Proceedings of the Third FLaReNet Forum*, Venice, Italy, 26-27 May 2011.

Omar Zaidan, Chris Callison-Burch. 2011. "Crowdsourcing Translation: Professional Quality from Non-Professionals". In *Proceedings of ACL-2011*.

Dimension	Challenge	Recommended actions	Target
Documentation	Ensure that Language Resources are accurately and reliably documented	• Devise and adopt a widely agreed standard documentation template for each resource type, based on identified best practice(s)	LRP
		• 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	LRP
		• Ensure that appropriate metadata are consistently adopted for describing LRs	LRP/PM
		• Set up a global infrastructure of common and uniform and/or interoperable metadata sets	LRP
		• Develop and support community-wide initiatives such as the LRE Map	LRP/PM
Interoperability	Design and set up an interoperability framework for LRs and LT	• Ensure syntactic and semantic interoperability of Language Resources	LRP
		• Set up an "interoperability challenge" as a collective exercise to evaluate (and possibly measure) interoperability	LRP/PM
		• Create a permanent Standards Observatory or Standards Watch	PM
		• Invest in standardisation activities	PM
		• Encourage/enforce use of best practices or standards in LR production projects	PM
		• Make standards operational and put them in use	LRP
		• Set up training initiatives to promote and disseminate standards to students and young researchers	LRP/PM
		• Identify new mature areas for standardisation and promote joint efforts between R&D and industry	LRP/PM

Availability: Sharing and Distribution	**Make resources easily and readily available, within an adequate IPR and legal framework**	• Opt for openness of LRs, especially publicly funded ones	**LRP/PM**
		• Ensure that publicly funded resources are publicly available either free of charge or at a small distribution cost	LRP/PM
		• Clear IPR at the early stages of production; try to ensure that re-use is permitted	LRP
		• Educate key players with basic legal know how	LRP/PM
		• Elaborate specific, simple and harmonised licensing solutions for data resources	LRP/PM
Coverage, Quality, Adequacy	**Address appropriate coverage in terms of *quantity*, *quality* and *adequacy* to technological purposes**	• Increase quantity of resources available to address language and application needs	LRP/PM
		• Implement BLaRKs for all languages, especially less-resourced languages	LRP/PM
		• Provide high quality resources for all European languages	LRP/PM
		• Address formal and content quality of resources by promoting evaluation and validation	PM
		• Devise new methods for LR quality check	LRP
		• Establish a European evaluation and validation body	PM
		• Establish common and standard Language Technology evaluation procedures	LRP/PM
		• Create an infrastructure for coordinated LRT evaluation	LRP/PM
		• Carry out evaluation in real-world scenarios	LRP
		• Promote evaluation and validation activities of LRs and the dissemination of their outcomes	PM
		• 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	LRP/PM
Sustainability	**Ensure sustainability of language resources**	• Ensure that all LRs to be produced undergo a sustainability analysis as part of the specification phase	LRP/PM
		• Foster use of a sustainability model	LRP/PM

Recognition	**Promote the LR ecosystem**	• Give greater recognition to successful LRs and their producers	LRP/PM
		• Develop a standard protocol for citing language resources	LRP/PM
		• Support training in production and use of LRs	LRP/PM
Development	**Define a reference model for future LR development**	• Ensure strong public and community support to definition and dissemination of resource production best practices	PM
		• Go Green: enforce recycling, reusing and repurposing	LRP/PM
		• Work towards the full automation of LR data production	LRP
		• 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	LRP/PM
		• Start an open community initiative for a large Language Knowledge Repository	LRP
An Infrastructure of Language Resources	**Build and sustain the proper Language Resource Infrastructure**	• Build a sustainable facility for sharing resource data and tools	LRP/PM
		• Establish 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	LRP
		• Develop and propose (free) tools and more generally Web services, including evaluation protocols and collaborative workbenches in the LR infrastructure	LRP
International Cooperation	**Promote synergies among initiatives at international level**	• Maintain a public survey on the LT and LR situation worldwide, based on FLaReNet and META-NET	LRP/PM
		• Share the effort for production of LRs between international bodies and individual countries	PM
		• Establish an International Forum to share information, discuss strategies and declare/define common objectives	PM