



**ECP-2007-LANG-617001**

**FLAReNet**

**Final FLAReNet Deliverable**  
**Language Resources for the Future – The**  
**Future of Language Resources**

**The Strategic Language Resource Agenda**

<b>Deliverable number/name</b>	<i>D2.2b, D4.2, D4.3, D5.2, D6.2, D7.2, D7.3, D8.2c</i>
<b>Dissemination level</b>	<i>Confidential</i>
<b>Delivery date</b>	<i>30 September 2011</i>
<b>Status</b>	<i>Final</i>
<b>Author(s)</b>	<i>Nicoletta Calzolari, Nuria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, Claudia Soria</i>



***eContentplus***

This project is funded under the *eContentplus* programme<sup>1</sup>, a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

---

<sup>1</sup> OJ L 79, 24.3.2005, p. 1.



## Table of Contents

<b>TABLE OF CONTENTS.....</b>	<b>2</b>
<b>INTRODUCTION .....</b>	<b>4</b>
THE CONTEXT.....	4
WHAT ARE LANGUAGE RESOURCES .....	4
THE FLARENET MISSION, POLICY AND STRATEGY .....	5
THE FLARENET COMMUNITY .....	5
THE FLARENET DATA.....	6
THE FLARENET RECOMMENDATIONS: A COLLECTIVE ENTERPRISE.....	6
ORGANISATION OF THE BOOK.....	6
<b>ACKNOWLEDGEMENTS .....</b>	<b>7</b>
<b>CHAPTER 1. THE FLARENET BLUEPRINT OF ACTIONS AND INFRASTRUCTURES.....</b>	<b>8</b>
HOW TO USE THE BLUEPRINT .....	9
1.1 RECOMMENDATIONS AT A GLANCE .....	9
1.2    INFRASTRUCTURE OF LANGUAGE RESOURCES.....	13
1.3    RESOURCE DOCUMENTATION.....	14
1.4    RESOURCE DEVELOPMENT.....	17
1.5    RESOURCE INTEROPERABILITY .....	20
1.6    RESOURCE COVERAGE, QUALITY, ADEQUACY.....	22
1.6.1    RESOURCE QUANTITY .....	22
1.6.2    RESOURCE QUALITY .....	23
1.6.3    RESOURCE ADEQUACY .....	25
1.7    RESOURCE AVAILABILITY: SHARING AND DISTRIBUTION.....	26
1.8    RESOURCE SUSTAINABILITY .....	28
1.9    RESOURCE RECOGNITION.....	29
1.10    INTERNATIONAL COOPERATION .....	30
REFERENCES .....	31
<b>CHAPTER 2 - IDENTIFYING MATURE AND SUSTAINABLE LANGUAGE RESOURCES .....</b>	<b>32</b>
2.1    KEY SUSTAINABILITY FACTORS.....	34
2.2    KEY AREAS OF RESEARCH INTEREST.....	36
2.3    LR MATURITY AND SUSTAINABILITY IN MAJOR RESEARCH AREAS.....	37
2.4    EVALUATION PACKAGES AND RESOURCES .....	39
2.5    MULTIPLE-APPLICATION RESOURCES AS A "MATURITY/SUSTAINABILITY" FACTOR.....	41
2.6    MATURITY AS REFLECTED BY THE LRE MAP (2010) DATA .....	41
2.7    MATURITY IN TERMS OF OBJECTIVE EVALUATIONS.....	43
2.8    FINAL RECOMMENDATIONS.....	45
REFERENCES .....	46
<b>CHAPTER 3 - STRATEGIES FOR LANGUAGE RESOURCE MIXING, SHARING, REUSING AND RECYCLING .....</b>	<b>47</b>
3.1    REUSING RESOURCES.....	47
3.2    INTERLINKING RESOURCES.....	48
3.3    REPURPOSING RESOURCES .....	48
3.4    GENERAL CONSIDERATIONS ON REUSE.....	49
REFERENCES .....	54
<b>CHAPTER 4 - A STRATEGIC ACTION PLAN FOR AN INTEROPERABILITY FRAMEWORK .....</b>	<b>56</b>
4.1    THE CURRENT STANDARD FRAMEWORK .....	56
4.2    BARRIERS AND MAJOR PROBLEMS.....	58
4.3    SCENARIOS FOR USING STANDARDS.....	59
4.4    RECOMMENDATIONS AND NEXT STEPS .....	61
4.5    TOWARDS AN INTEROPERABILITY FRAMEWORK.....	65
REFERENCES .....	67



<b>CHAPTER 5 - THE EVALUATION AND VALIDATION OF RESOURCES.....</b>	<b>70</b>
5.1    RECOMMENDATIONS FOR VALIDATION.....	70
5.2    RECOMMENDATIONS FOR EVALUATION.....	74
<b>CHAPTER 6 - STRATEGIC DIRECTIONS FOR AUTOMATING LR DEVELOPMENT .....</b>	<b>78</b>
6.1    SURVEYING THE STATE OF THE ART.....	79
6.2    STRATEGIC DIRECTIONS AND RECOMMENDATIONS .....	81
REFERENCES .....	84
<b>CHAPTER 7 - TOWARDS AN LR ROADMAP .....</b>	<b>85</b>
7.1    URGENT LR REQUIREMENTS .....	86
7.2    FINDINGS.....	89
7.3    INTERNATIONAL FRAMEWORK.....	93
7.4    STRATEGY .....	95
7.5    CONCLUSIONS AND PERSPECTIVES.....	96
REFERENCES .....	96
<b>GLOSSARY OF ACRONYMS.....</b>	<b>97</b>



## Introduction

*Nicoletta Calzolari, Claudia Soria*

### ***The context***

Language Technologies (LT), together with their backbone, Language Resources (LR), provide an essential support to the challenge of Multilingualism and ICT of the future. The main task of language technologies is to bridge language barriers and to help creating a new environment where information flows smoothly across frontiers and languages, no matter the country, and the language, of origin.

To achieve this goal, all players involved need to act as a community able to join forces on a set of shared priorities. However, until now the field of Language Resources and Technology has long suffered from an excess of individuality and fragmentation, with a lack of coherence concerning the priorities for the field, the direction to move, not to mention a common timeframe.

This lack of coherent directions is partially also reflected by the difficulty with which fundamental information about LR&Ts can be reached: basically, it is very difficult, if not impossible, to get a clear picture of the current situation of the field in simple terms such as who are the main actors, what are the available development and deployment methods, what are the “best” language resources, what are the areas for which further development and investment would be most necessary, etc. Substantial information is not easily reachable not only for the producers but also for policy makers and funding agencies.

Under this respect, since some time large groups have been advocating the need of a LR&T infrastructure, which is increasingly recognised as a necessary step for building on each other achievements, integrating resources and technologies and avoiding dispersed or conflicting efforts. A large range of LRs and LTs is there, but the infrastructure that puts LR&Ts together and sustains them is still largely missing; interoperability of resources, tools, and frameworks has recently come to be understood as perhaps the most pressing current need for language processing research. Infrastructure building is thus indicated by many as the most urgent issue and a way to make the field move forward, together with real sharing of resources, and an effort to make resources available for all languages.

The context encountered by the FLaReNet project was thus represented by an active field needing a coherence that can only be given by sharing common priorities and endeavours. FLaReNet has contributed to the creation of this coherence by gathering a wide community of experts and making them participate in the definition of an exhaustive set of recommendations.

### ***What are Language Resources***

The term “Language Resources” traditionally refers to usually large sets of language data and descriptions in machine readable form, to be used in building, improving or evaluating natural language, speech or multimodal algorithms or systems. Typical examples of LRs are written, spoken, multimodal corpora, lexicons, grammars, terminologies, multimodal resources, ontologies, translation memories, but the term is also extended to include basic software tools for their acquisition, preparation, annotation, collection, management and use. The creation and use of these resources span several related but relatively isolated disciplines, including NLP, information retrieval, machine translation, speech, multimodality. FLaReNet endorses this broader definition of LRs, having recognised the need for an “extension” of the term and in the light of recent scientific, methodological-epistemological, technological, social and organisational developments in the application fields of content processing/access/understanding/creation, Human-Machine, Human-Human & Machine-Machine communication, and the corresponding



## Introduction

areas from which the theoretical underpinnings of these application fields emerge (linguistics, cognitive science, AI, robotics).

### ***The FLaReNet mission, policy and strategy***

FLaReNet – Fostering Language Resources Network – is an international Forum, composed by a large community, aiming at developing the needed common vision and fostering a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide. Its goals are:

- to facilitate interaction among LR&T stakeholders and ultimately re-build a community around LR and LTs;
- to promote and sustain international cooperation;
- to coordinate a community-wide effort to analyse the sector of LR&Ts along all the relevant dimensions: technical and scientific, but also organisational, economic, political, cultural and legal;
- to identify short, medium, and long-term strategic objectives and provide consensual recommendations in the form of a plan of action targeted to a broad range of stakeholders, from the industrial and scientific community to funding agencies and policy makers;
- to pave the way to the set up and functioning of an Open Resource Infrastructure through a number of preparatory initiatives.

### ***The FLaReNet community***

FLaReNet brings together leading experts of research institutions, academies, companies, consortia, associations, funding agencies, public and private bodies both at European and international level. This way, we ensure that both LR producers and users are represented and actively involved, in addition to technology developers. And in turn, broad participation of different key-figures and experts ensures that recommendations are formulated through a consensual bottom-up process in which the relevant scientific, technical, organisational, strategic aspects and positions are taken into account.

The FLaReNet Community is composed of a network of 99 Institutional Members and 365 Individual Subscribers, representing 33 different countries. Community members belong to academia, industry, resource distribution agencies, media, publishing houses, and national and European institutional bodies. FLaReNet has also established a Network of National Contact Points, for the purpose of obtaining up-to-date and reliable information about current initiatives worldwide in the area of Language Resources (Data, Tools, Evaluation and Meta-Resources). The Network consists at present of 105 experts from 34 countries or regions in the European Union (26 Member States and 6 regions), 9 non-EU European countries and non-European countries (36 countries). The role of the National Contact Points is to provide information on the initiatives on Language Resources in their country/region, as well as to ensure a permanent link between FLaReNet and their country/region, and to provide further reliable information about the HLT research activities and entities, industries and administrations in their country/regions. A Map showing all the countries for which there is a FLaReNet contact point is available on the FLaReNet web site<sup>1</sup>.

Over the duration of the project, FLaReNet has truly become a community of experts that – by acting as an aggregator and facilitator of information sharing and discussion – is one of the first scientific social networks in the field of Language Resources and Technology. This community represents now a legacy that needs to be sustained to keep the momentum created by FLaReNet.

---

<sup>1</sup> <http://www.flarenet.eu>



### ***The FLaReNet data***

Having recognised the lack of information about existing language resources as one of the major factors hindering the development of the field, FLaReNet has undertaken a number of actions to survey existing resources, inform about them, and enhance their visibility. Community involvement is the underlying leitmotif that is common to all activities.

FLaReNet, together with ELRA, has launched the *LRE Map of Language Resources and Tools*. The Map is an entirely new instrument for discovering, searching and documenting language resources, here intended in a broad sense, as both data and tools. The purpose of the LRE Map is to complement ongoing cataloguing efforts in order to enhance visibility for all resources, for all languages and intended applications.

Other data come from the *Repository of Standards, Best Practices and Documentation*, the *Wiki Survey of the national and transnational initiatives in the area of Language Resources*<sup>2</sup>, the *Language Library*,<sup>3</sup> a *Database* of the various technical and organisational methods, techniques and models used by academic and industrial players in the area of Language Resources for building and maintaining resources, and a *Survey of methods for the automatic construction of LRs*. All these data are collected in the FLaReNet “DataBook”, a companion document to this deliverable.

### ***The FLaReNet recommendations: a collective enterprise***

As a response to its main mission, FLaReNet has identified a number of consensual priorities and strategic objectives for the field of Language Resources and Technology, which have been provided to the community through a series of *Blueprints of Actions and Infrastructures*. The whole community at large has been involved in this task, by means of a continuous direct consultation of key players and stakeholders, in the attempt to reflect the widest possible view. Moreover, FLaReNet has maintained regular cooperation activities with relevant projects and initiatives, both non-European and European, in order to establish a kind of global coordination of the LR field.

The FLaReNet recommendations target a broad spectrum of users:

- HLT stakeholders at large, including producers, users and developers of Language Resources and Technologies, both academic and industrial (for instance, academic or industrial researchers, service and media providers, providers of translation and localization services, etc.)
- Funding agencies and policy-makers, at national and EC level.

### ***Organisation of the Book***

This volume collects and organises the major high-level recommendations collected around the FLaReNet meetings, panels and consultations, as well as the results of the surveying and analysis activities carried out under FLaReNet work packages. The FLaReNet Steering Committee took care of summarising and organising the various recommendations, also as part of the work carried out in the individual Work Packages of the project. As such, this Book is 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 neighbouring projects, associations, initiatives, funding agencies and government institutions.

---

<sup>2</sup> <http://www.flarenet.eu/?q=WG7>

<sup>3</sup> For a description of the goals of the Language Library, see <http://www.lrec-conf.org/lrec2012/>



## Introduction

The different chapters look at the various dimensions that are of relevance in the field, thus providing different viewpoints on the topic of language resources and complementing each other. They start from the current situation to identify areas for future developments.

This is why the reader will find some redundancy in the content of the chapters. This is done on purpose as some issues, such as evaluation, standards, reuse etc. encompass various dimensions and are therefore tackled in the different chapters from different point of views.

The Blueprint, in Chapter 1, provides a comprehensive perspective, highlighting the synergies among the several positions, and summarising the fundamental recommended actions for the development and progress of LRs and LTs, coming from the FLaReNet community and addressing both the community itself and policy makers.

Chapter 2 looks at LR sustainability and maturity on the basis of a descriptive model of sustainability from a LT perspective. It summarizes the description of a sustainability model, elaborated within the project. Chapter 3 focuses on models for reusing and interlinking existing resources and tools as a means to derive, produce and repurpose new LRs (datasets and relevant processing tools). Chapter 4 focuses on the major conditions and recommendations necessary to build the so-called “Interoperability Framework”, as a dynamic environment of language (and other) standards and guidelines, where the former can interrelate coherently with one another and the latter describe how standards work and “speak” to each other. Chapter 5 looks at issues of Language Resource quality in terms of validation and evaluation, and describes and clarifies recommendations for these activities. Chapter 6 establishes some recommendations for strategic actions to be put in force either by the community or by policy makers towards an extensive automation of LR production, which will lead to beneficial improvements for the commercial value of LTs and thus will enforce competitiveness of LT-based enterprises globally. Finally, Chapter 7 presents a strategy to fill in the gaps detected by a survey on the national and cross-border R&D effort in language resources and technologies, and existing data, tools, standards and evaluation methods.

## Acknowledgements

This work could not have been possible without the invaluable contribution of all the people who voluntarily shared their knowledge and understanding of the LRT landscape. We also acknowledge Andrew Joscelyne for professional revision and editing of early drafts of this document.



## Chapter 1. The FLaReNet Blueprint of Actions and Infrastructures

*Nicoletta Calzolari, Valeria Quochi, Claudia Soria*

*with the contribution of Núria Bel, Gerhard Budin, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis*

This chapter presents the final *Blueprint of Actions and Infrastructures* produced by the FLaReNet project and community. The Blueprint is placed as the first chapter of the book as it summarises and prioritises the contents of the other chapters. It is however conceived as an autonomous document that provides the fundamental recommended actions for the development and progress of LRT in Europe, coming from the FLaReNet community and addressing both the community itself and policy makers.

The content of the *Blueprint* is a synthesis of the discussions and activities of the three years of the FLaReNet project, and as such it integrates the previous two Blueprints<sup>1</sup> as well as the outcomes of the stimulating discussions held both within and outside FLaReNet. Thanks to the participation of outstanding experts in the field, the content of this document can be considered as the expression of the Language Resources community setting its own agenda of future actions to be undertaken concerning a broad spectrum of dimensions related to Language Resources<sup>2</sup>.

Recognizing that the development of the sector of LTs is conditioned by various factors, all interested stakeholders need to operate seriously together and forge partnerships to push LTs. FLaReNet tackled this issue by bringing different stake-holders together, and having them discuss about several key topics, such as the lack of infrastructures for the domain, and try to foster joint plans, projects and roadmaps. Some early results are already visible: an EC funded Network of Excellence started in 2010 – META-NET – with one of its main goals being the design and set-up of an infrastructure for sharing language resources and technology at large – i.e. META-SHARE. This was the main recommendation of FLaReNet in its first year.

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. While there has been considerable progress in technology developments in the last decade, the significant challenge of overcoming current fragmentation and imbalance inside the LTs 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 build a compact community.

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 Blueprint belong to a large set of players and stakeholders in Language Technologies (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,

---

<sup>1</sup> <http://www.flarenet.eu/?q=Deliverables>

<sup>2</sup> According to the FLaReNet definition, Language Resources are all language data sets and basic tools, see the Introduction.





## Ch 1. The Blueprint

National, local, or private level, wishing to draft a program of activities for his/her own communities.

### ***How to use the Blueprint***

In this document, the various actions recommended are organised around nine dimensions that are relevant for the field of Language Resources: a) *Infrastructure*, b) *Documentation*, c) *Development*, d) *Interoperability*, e) *Coverage, Quality and Adequacy*, f) *Availability, Sharing and Distribution*, g) *Sustainability*, h) *Recognition* 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 Language Resources and Technologies grow.

It is useful to see the various dimensions as a coherent system where each one presupposes the other, so that action at one of the levels requires some other action to be taken at another one. 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, etc.

Taken together, these directions are intended to contribute to the creation of a sustainable LRT ecosystem.

### ***1.1 Recommendations at a glance***

The first section of this chapter presents an overview of the main recommendations in a synoptic format, organised by dimensions. Each issue is then expanded and described in the following sections. Each dimension is given a prototypical challenge, representing the “vision” or condition to be attained according to the point of view of the FLaReNet community of experts.

The dimension “Infrastructure of Language Resources” is given a prominent and separate role in that it encompasses most of other dimensions.

Recommendations are targeted to Language Resource Producers (LRP, broadly encompassing academic and industrial developers) and/or Policy Makers (PM, i.e. National or supra-national funding agencies, politicians, etc.).

Dimension	Challenge	Recommended actions	Target
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 (comparable to the Language Grid), including evaluation protocols and collaborative workbenches in the LR infrastructure	LRP

Dimension	Challenge	Recommended actions	Target
Documentation	Ensure that Language Resources are accurately and reliably documented	<ul style="list-style-type: none"> <li>• Devise and adopt a widely agreed standard documentation template for each resource type, based on identified best practice(s)</li> <li>• Ensure that appropriate metadata are consistently adopted for describing LRs</li> <li>• Set up a global infrastructure of common and uniform and/or interoperable metadata sets</li> <li>• Develop and support community-wide initiatives such as the LRE Map</li> <li>• Establish links with other communities to gain access to better information on the existence of LRs in their domains, and exchange Best Practices on handling and sharing resources</li> <li>• 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</li> </ul>	LRP LRP LRP LRP LRP LRP
Development	Define a reference model for future LR development	<ul style="list-style-type: none"> <li>• Ensure strong public and community support to definition and dissemination of resource production best practices</li> <li>• Go Green: enforce recycling, reusing and repurposing</li> <li>• Encourage the full automation of LR data production</li> <li>• 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</li> <li>• Start an open community initiative for a large Language Knowledge Repository</li> <li>• Estimate the cost of producing LRs needed to develop an LT for one language</li> </ul>	PM LRP LRP LRP/PM LRP LRP



<b>Interoperability</b>	Design and set up an interoperability framework for LRs and LT	<ul style="list-style-type: none"> <li>• Ensure formal and semantic interoperability of Language Resources</li> <li>• Identify new mature areas for standardisation and promote joint efforts between R&amp;D and industry</li> <li>• Make standards operational and put them in use</li> <li>• Invest in standardisation activities</li> <li>• Encourage/enforce use of best practices or standards in LR production projects</li> <li>• Set up an “interoperability challenge” as a collective exercise to evaluate (and possibly measure) interoperability</li> <li>• Collaboratively build and maintain a repository of standards and best practices, linked to standards-compliant open/freely available data</li> <li>• Create a permanent Standards Observatory or Standards Watch</li> <li>• Set up training initiatives to promote and disseminate standards to students and young researchers</li> </ul>	<p>LRP</p> <p>LRP</p> <p>LRP</p> <p>PM</p> <p>PM</p> <p>LRP</p> <p>LRP/PM</p> <p>PM</p> <p>LRP/PM</p>
<b>Coverage, Quality, Adequacy</b>	Address appropriate coverage in terms of <i>quantity, quality</i> and <i>adequacy</i> to technological purposes	<ul style="list-style-type: none"> <li>• Increase quantity of resources available to address language and application needs</li> <li>• Implement BLaRKs for all languages, especially less-resourced languages</li> <li>• Address formal and content quality of resources by promoting evaluation and validation</li> <li>• Establish a European evaluation and validation body</li> <li>• Define and establish a Quality Certificate or quality score for LRs, to be endorsed by the community</li> <li>• Assess availability of resources with respect to their adequacy to applications and technology requirements</li> </ul>	<p>LRP</p> <p>LRP</p> <p>PM/LRP</p> <p>PM</p> <p>LRP</p> <p>LRP/PM</p>

<p><b>Availability: Sharing and Distribution</b></p>	<p>Make resources easily and readily available, within an adequate IPR and legal framework</p>	<ul style="list-style-type: none"> <li>• Opt for openness of LRs, especially publicly funded ones</li> <li>• Ensure that publicly funded resources are publicly available either free of charge or at a small distribution cost</li> <li>• Create LRs in collaborative projects where resources are exchanged among project participants after production</li> <li>• Make LRs available for most types of use, making use of standardised licenses where reuse is clearly stipulated</li> <li>• Educate key players with basic legal know how</li> <li>• Elaborate specific, simple and harmonised licensing solutions for data resources</li> <li>• Clear IPR at the early stages of production; try to ensure that re-use is permitted</li> <li>• Make sharing/distribution plans mandatory in projects concerning LR production</li> </ul>	<p>LRP/PM PM/LRP LRP LRP LRP/PM LRP LRP PM</p>
<p><b>Sustainability</b></p>	<p>Ensure sustainability of language resources</p>	<ul style="list-style-type: none"> <li>• Ensure that all LRs to be produced undergo a sustainability analysis as part of the specification phase</li> <li>• Assess LR sustainability on the basis of impact factors</li> <li>• Foster use of a sustainability model</li> </ul>	<p>PM/LRP LRP LRP/PM</p>
<p><b>Recognition</b></p>	<p>Promote the LR ecosystem</p>	<ul style="list-style-type: none"> <li>• Develop a standard protocol for citing language resources</li> <li>• Give greater recognition to successful LRs and their producers (Prizes, Seals of Recognition, etc. )</li> <li>• Support training in production and use of LRs</li> </ul>	<p>LRP/PM LRP/PM LRP</p>
<p><b>International Cooperation</b></p>	<p>Promote synergies among initiatives at international level</p>	<ul style="list-style-type: none"> <li>• Share the effort for the production of LRs between international bodies, such as the EC for the EU, and individual countries/regions, such as Member States in Europe</li> <li>• Encourage agencies to open their calls to foreign participants</li> <li>• Establish MoU among existing initiatives, starting with the EC efforts</li> <li>• Ensure the sustainability of the FLaReNet International Contact Points, also through the Survey Wiki</li> <li>• Produce a White paper on LT, including LR, in preparation for FP8. Distribute it within Europe and abroad</li> </ul>	<p>PM PM PM PM LRP</p>



## 1.2 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)”*

	LRP	PM
Build a sustainable facility for sharing resource data and tools	•	•
Make LRs available, visible and easily accessible through an appropriate infrastructure; participate in building the infrastructure by providing feedback on relevant actions in this direction	•	•
Ensure a stable sustainable infrastructure for LR sharing and exchange; support continuous development and promotional activities	•	•
Establish an infrastructure that can help LR producers with legal issues		•
Support the emergence of a tool-sharing infrastructure to lower the cost of R&D for new applications in new language resource domains		•
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	•	
Develop and propose (free) tools and more generally Web services (comparable to the Language Grid), including evaluation protocols and collaborative workbenches in the LR infrastructure	•	

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 should ease recovery and use of LRs through appropriate facilities that allow their availability, visibility, and easy accessibility.

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 most 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.

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.

As a matter fact, researchers and developers still have to spend quite some time consulting multiple catalogues or searching the web to find relevant LRs, 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 (see below, 1.4 and 1.7).

These basic principles for LR infrastructures are related to or incorporated in most of the issues appearing in the following sections, and will therefore not be expanded here.

### 1.3 Resource Documentation

*“Ensure that Language Resources are accurately and reliably documented”*

	LRP	PM
Devise and adopt a widely agreed standard documentation template for each resource type, based on identified best practice(s)	•	
Ensure strong public and community support to production and dissemination of best practices		•
Ensure that appropriate metadata are consistently adopted for describing LRs	•	•
Provide appropriate MD description for all LR distributed, preferably in one of the widely used MD schemes	•	
Support MD creation and promotion activities; set guidelines and rules for MD description of available LRs and support relevant efforts		•
Create machine understandable MD with formal syntax and clear semantics	•	
Use formalized MD elements as much as possible	•	
Put all aspects of the documentation of the LR that can be formalized in formalized metadata elements	•	
Ensure that all data categories used in the metadata and in the data are registered in a data category registry to ensure semantic interoperability	•	
Automate the process of MD creation	•	
Set up a global infrastructure of common and uniform and/or interoperable metadata sets	•	•
Agree on a limited/basic set of MD	•	
Develop structured MD	•	
Develop and support community-wide initiatives such as the LRE Map	•	•
Establish links with other communities to gain access to better information on the existence of LRs in their domains, and exchange Best Practices on handling and sharing resources	•	•
When producing a LR, allocate time and manpower to documentation from the start; collect and provide documentation (or links to it) when giving access to a LR	•	
In a LR production project, part of the funding should be allocated to documentation and dissemination activities; support activities for collecting and storing in appropriate infrastructures documentation for LRs		•

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.





## Ch 1. The Blueprint

Documentation allows language resources to be used by people different from those who designed and developed them. 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 documentation and guidelines writing need to be established and 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.

What users need is 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.

“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”  
(D. Giaretta, 2011)

Guidelines are also useful for replicating and extending resources. They may re-channel language resource efforts in a more coherent way (see the experience in SPEECHDAT<sup>3</sup>) and help develop similar resources for other languages well past the end of a given project. There should therefore be a set of exhaustive and reliable guidelines for every resource type, building on the experience of successful projects.

An effort must be made to collect all existing LR documentation and make it easily available. To this end, the design and construction of a (virtual) repository of specifications, guidelines, and documentation of LRs, starting with reference resource models or widely known and used resources (e.g. WordNet, Penn TreeBank, ...) is a priority task<sup>4</sup>.

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, i.e. guidelines used by annotators.

<sup>3</sup> <http://www.speechdat.org>

<sup>4</sup> An activity along these lines has been started within FLaReNet, see [http://www.flarenet.eu/?q=FLaReNet\\_Repository\\_of\\_Standards\\_and\\_Guidelines](http://www.flarenet.eu/?q=FLaReNet_Repository_of_Standards_and_Guidelines)





- 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).

Documentation is also the gateway to LR discovery. Ensuring that Language Resources are discoverable is the first step towards promoting the data economy. In order to make LRs discoverable, language resource providers (LRPs) should always document their resources, *using standard metadata and unique resource identifiers*. A useful best practice in this line is to document one's resource in an open catalogue, such as the ELRA Universal Catalogue<sup>5</sup>, the META-SHARE<sup>6</sup> or CLARIN<sup>7</sup> 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.

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 data sets for specific uses. We need a solution to the lack of a "one-size-fits-all" resource.

In this context, *standardised metadata* play a crucial role, and their use should be made obligatory in all resource production projects. 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*<sup>8</sup>, 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. Therefore, definition and adoption of standardised metadata must be the first priority and first step.

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. The development of techniques for automating the process of metadata creation would also help spread the adoption of machine-understandable metadata.

---

<sup>5</sup> [http:// universal.elra.info/](http://universal.elra.info/)

<sup>6</sup> [http:// www.meta-net.eu/meta-share/meta-share](http://www.meta-net.eu/meta-share/meta-share)

<sup>7</sup> [http:// www.clarin.eu/vlo/](http://www.clarin.eu/vlo/)

<sup>8</sup> <http://www.resourcebook.eu/>



## 1.4 Resource Development

*“Define a reference model for future LR development”*

	LRP	PM
Ensure strong public and community support to definition and dissemination of resource production best practices		•
Ensure that all "best/good practices" are disseminated to the R&D community through very lightweight methods, instead of the heavy machinery often used		•
<b>Go Green: enforce recycling, reusing and repurposing</b>	•	•
Implement tools that can help crawl “comparable” resources, select a small set that can be aligned and then provide clear legal conditions to ensure they can be shared with the MT community and beyond (assuming these can be annotated for bilingual lexical extraction, named entities, etc.)	•	
Check the availability of existing LRs; try to reuse what is available; check what is available not only for one's own language, but also for languages of the same family; try to see what can be reused in terms of data and tools but also experience; create and exchange tools that help in reuse/repurposing/interlinking	•	
support reuse projects and research activities in the domain; support activities that document reuse, formulate guidelines, etc.; fund promotional and research activities around reuse (e.g. dedicated workshops, publications on successful reuse cases etc.)		•
Promote more efficient uses of available data through a better grasp of repurposing and merging techniques	•	
Analyse the relationship within language families and use joint LR methods to develop LT within those families. Consider the use of pivot languages within those families, based on the existence of parallel corpora	•	
Adapt LT from the language in which it exists to a similar one in the same language family, then improve the quality of the LRs for the specific data in that similar language through bootstrapping	•	
<b>Encourage the full automation of LR data production</b>	•	•
Develop a certain volume of MT-based corpora, post-edit them (semi)automatically to specific quality levels, and then use them iteratively to train new MT systems more effectively	•	
Promote activities that can contribute to the (automatic) production of large scale and high quality resources	•	•
Support academic and industry 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-world applications		•
Investments in research on the automatic production of language resources should be increased to broaden the range of language resources addressed	•	•
Promote research into new methods and techniques for collecting information from unstructured (non-annotated) data	•	•
Invest in new methods such as active learning and semi-supervised learning that	•	



deliver promising results for reducing the need for human intervention		
Use automatic content extraction from the Web, but bear legal issues in mind	•	
Conduct a study and propose a Code of Ethics (regarding privacy and IPR) for automatic information extraction on the internet	•	
Develop tools for anonymizing data	•	
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	•	•
Carry out LR production and annotation as collaborative projects; open up existing LRs for collaborative annotation and the reuse of the annotated results; participate in communities that carry out similar tasks; evaluate and document their results; develop new tools and/or adapt existing tools and platforms to the needs of collaborative work	•	
Support collaborative efforts for large LR production and annotation projects through appropriate funding; support the infrastructure required for collaborative work		•
Promote the collaborative development of semantically/pragmatically/dialogically annotated corpora	•	•
Propose a Code of Ethics for crowdsourcing, in order to guarantee a sufficient salary and compliance with labour rights	•	
Foster the debate and experiments on new outsourcing trends over the web	•	
Start an open community initiative for a large Language Knowledge Repository	•	•
Estimate the cost of producing LRs needed to develop an LT for one language	•	

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.

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. (see Ch. 3).

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 ASR 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 (see also Ch. 6).

“Innovation needs data, but also the collection of data needs innovation”  
(Vasiljevs, 2011).



## Ch 1. The Blueprint

The coverage problem is so enormous that we must promote strategies that approach or ensure full automation for (high-quality) LR data production. 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.

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.

There are several experiments in crowd-sourcing data collection and NLP 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 recent 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).

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<sup>9</sup> (see for instance Zaidan and Callison-Burch 2011 about mechanisms for increasing quality of crowd-sourced data).

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 (see also Ch. 7). Basic language resources for all the worlds' 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 and Web 2.0 methods for data collection and annotation seem particularly very well-suited for collecting the data needed for the development of LT applications.

There are insufficient current 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.

---

<sup>9</sup> See for instance (Zaidan and Callison-Burch 2011) about mechanisms for increasing quality of crowd-sourced data.



## 1.5 Resource Interoperability

*“Design and set up an interoperability framework for Language Resources and Technology”*

	LRP	PM
Ensure formal and semantic interoperability of Language Resources	•	
Define the standard as the lowest common denominator, at the highest level of granularity, a basic principle that was already factored into EAGLES	•	
Standards must generalize over multilingual data and should be tested on multilingual data	•	
Identify new mature areas for standardisation and promote joint efforts between R&D and industry	•	•
Engage in groups that produce (new) standards and provide feedback and comments	•	
Make standards operational and put them in use	•	
Encourage the building of tools that enable the use of standards, and step up the availability of sharable/exchangeable data	•	
Fund the development and/or maintenance of tools that support/enforce/validate standards		•
Invest in standardisation activities		•
Support infrastructural activities for collecting and disseminating information on existing standards and best practices	•	
Fund activities for setting up new standards where they do not exist	•	
Closely monitor the Linked Open Data initiative tightly connected to semantic interoperability, so that we understand the potentialities of this initiative for our field	•	
Encourage/enforce use of best practices or standards in LR production projects		•
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)	•	
Create ‘official’ validators (such as <a href="http://validator.oaipmh.com/">http://validator.oaipmh.com/</a> or the OLAC validator) to check compliance of LRs with basic linguistic standards		•
Set up an “interoperability challenge” as a collective exercise to evaluate (and possibly measure) interoperability	•	•
Create a permanent Standards Observatory or Standards Watch		•
Set up training initiatives to promote and disseminate standards to students and young researchers	•	•

Interoperability of resources is the extent to which they are formally and content 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)



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.

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 Producers are still largely lacking a clear understanding about why standards help represent data, and why there are the advantages to 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.

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. 2011) proposing an overview of the current scene towards an Interoperability Framework and acts as a reference point for the current standards encouraged by the community for adoption<sup>10</sup>.

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 thus 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.).

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. The design of interoperability tasks will also help to determine which emerging standards are most interoperable. Interoperability tests can also replace aspects of validation (see Ch. 5).

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

---

<sup>10</sup> This initiative is in close synchronization with other relevant initiatives such as CLARIN, ELRA, ISO and TEI and META-Share.





advantageous and is thus to be promoted also in order to identify new areas that are mature for standardisation activities.

### 1.6 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 LT crucially depends on language resources nowadays. 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.*

“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)

#### 1.6.1 Resource Quantity

	LRP	PM
Increase quantity of resources available to address language and application needs	•	•
Enforce shared/distributed construction of resources as a means to achieve better coverage		•
Implement BLARKs for all languages, especially less-resourced ones	•	•
Analyse the dialectal variants of languages and include those variants in the production of LRs	•	
Provide sustainable high quality resources for all European languages	•	•
Ensure that resources are developed for all EU languages so that evaluations can be made of language-dependent technologies listed by the major agencies (training and testing data)	•	•

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 narrow domains with their own specific terminological and stylistic





## Ch 1. The Blueprint

requirements. Thus, *provision of high quality resources for all European language, including minority ones is a priority now*, in order to avoid disparity in the future.

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. 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.

Specifically for the advancement of LTs, *Basic Language Resource Kits (or BLARKs<sup>11</sup>)* should be supported and developed for all languages and, at least, main applications (MT, IR, QA to mention some). 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 LT. 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 well represented ones) and all basic needs of language technology remains thus a high priority for ensuring multilingual applications in the future.

### 1.6.2 Resource Quality

	LRP	PM
Address formal and content quality of resources by promoting evaluation and validation		•
Establish a European evaluation and validation body		•
Provide high-quality resources for all European languages	•	•
Promote automatic techniques to guarantee quality through error detection and confidence assessment	•	
Promote technologies that deliver highly accurate, high-confidence results to reduce the amount of subsequent revision (human or automated)	•	•
Establish common and standard Language Technology evaluation procedures	•	•
Organize (a significant part of) research around evaluation tasks, challenges and campaigns	•	
Promote evaluation as a driving force for research in HLT, in particular for the development of techniques for fully automatic evaluation and of fully formalized evaluation metrics		•
Devise new methods for LR quality check	•	
Develop new and/or ensure the maximal use of existing tools for	•	

<sup>11</sup> <http://www.blark.org/>

automatic or semi-automatic formal and content validation of LRs		
Create a think tank with recognized experts from a broad spectrum of the community (academia/industry; technologies; and modalities (written/speech/multimodal), etc.) to assess requirements for LR quality	•	
Use crowd sourcing techniques to carry out validation tasks, but be aware of the ethical and legal issues associated with these techniques	•	
Create an infrastructure for coordinated LRTs evaluation	•	•
Secure an evaluation framework to assess the progress made by technologies with respect to state of the art		•
Set up an evaluation management and coordination structure to ensure a viable, consistent and coherent programme of activities that can successfully scale up and embrace new communities and technological paradigms		•
Set up a sustainable technical infrastructure providing data, tools and services to carry out systematic evaluation. This could be a distributed infrastructure involving existing organizations	•	•
Provide funds for financing the evaluation management and coordination structure and its associated technical infrastructure. This may require new funding strategies and instruments both at the European and at national and regional levels		•
Promote evaluation and validation activities of LRs and the dissemination of their outcomes		•
Carry out evaluation in real-world scenarios	•	
Define and establish a Quality seal of approval, on the model of the “Data Seal of Approval” to be endorsed by the community	•	•

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

Evaluation in Europe is currently carried out by individual institutions (such as ELDA and CELCT) and by short-term projects (e.g. the TC-STAR and CHIL campaigns), but there is no sustained European-wide coordination, as there is in the US (NIST) or Japan (NII). In specific areas, the community may organise itself to carry out regular evaluations (e.g. CLEF 2000-2010, and Semeval) but with limited funding and much community good will.

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

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.



## Ch 1. The Blueprint

Thorough dissemination and information of activities and achievements should be done through LRT evaluation portals (e.g. the ELRA HLT evaluation portal<sup>12</sup>).

Evaluation packages should be distributed among the community or LT players.

New methods for validating and evaluating LRs (and LT in general) should be sought, proposed and widely agreed.

Projects/players should be pushed to specify a validation procedure before LR production starts.

The community and policy makers should ensure that a quick quality check can be carried out, at least for existing resources.

### 1.6.3 Resource Adequacy

	LRP	PM
Assess availability of resources with respect their adequacy to applications and technology requirements	•	•
Assess maturity of technologies for which resources should be developed	•	•
Monitor ongoing research developments through publications and patent filings, and keep track of prize-winning papers and theses		•
Draw a list of top 20 technologies and ensure that crucial resources are produced in at least 10 of these, in a publicly and fully funded framework	•	
Support EU-critical HTL domains one they have been identified		•
Conduct regular surveys of technology transfer		•
Secure an evaluation framework to assess the progress made by such technologies with respect to state of the art		•
Ensure regular evaluation campaigns to assess state of the art		•
Define a small set of crucial topics on which a critical mass of research/researchers should focus through a clear adherence to its principle	•	•
Introduce current industrial needs into the research agenda. Publish data on whether resources are actually used, which aspects of LRs should be made public, and help industry drill down into research on automatic methods		•
Produce appropriate spoken language resources to study and develop a more natural approach to voice input	•	•

Not only do we need more data, but the *typology of data* needed has also 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). However data access and links within and across this data is as important as the actual quantity.

<sup>12</sup> <http://www.HLT-evaluation.org/>



### 1.7 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 in which a given resource is actually made available for use by third parties. This implies decisions about licensing and business models.

	LRP	PM
Opt for openness of LRs, especially publicly funded ones	•	•
Ensure that publicly funded resources are publicly available either free of charge or at a small distribution cost	•	•
Create LRs in collaborative projects where resources are exchanged among project participants after production	•	
Make LRs available for most types of use, making use of standardised licenses where reuse is clearly stipulated	•	
Educate key players with basic legal know how	•	•
Elaborate specific, simple and harmonised licensing solutions for data resources	•	•
Harmonise legislation regarding LR use for all types of LRs, and make provisions for allowing the free use thereof at least for research/non-profit purposes		•
Develop clear agreement templates for sharing components and data, especially to ensure optimal cooperation from commercial parties while safeguarding their interests	•	
Clear IPR at the early stages of production; try to ensure that re-use is permitted	•	
Make sharing/distribution plans mandatory in projects concerning LR production		•

#### Data Openness

*“Producers of data will benefit from opening it to broad access and will prefer to deposit their data with confidence in reliable repositories (Giarretta, 2011)”*

There is strong impulse towards *open data* nowadays, in the sense of data that are easily obtainable and can be used with few, if any, restrictions. According to the Open Knowledge Foundation, 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 that can in turn be used, modified (and redistributed). While the majority of LR 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. Reluctance in fully embracing an open data model is still common. As pointed out by Timos Sellis 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 LR 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, 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 (see also DiPersio, 2011).

A KPMG study 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 question answering track. Thinking along these lines, the availability of massive quantities of open data could transform the NLP industry, as suggested by J. van der Meer for translation (van der Meer 2011). The success of the sharing approach is well represented by TAUS<sup>13</sup>, where many translation leaders have started sharing their translations in the Taus Data Association repository.

A questionnaire carried out by FLaReNet strongly advised that at least those resources that were developed with public funding should be made openly available. Another suggestion is to ensure openness of resources for most types of uses, making use of standardised licenses where available,

Of course, certain types of data are and will probably remain not shareable, either for confidentiality reasons, personal data, or competitive ones. Non shared data can still be exploited, for instance through *shared services* (see also Ch. 4 and 7).

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 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 1.3, while reusability has to do with resource Interoperability (see 1.5). 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.

### **Legal, IPR issues**

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; on the other IPR tends to restrict accessibility to and usability of language resources.

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.

On top of 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).

---

<sup>13</sup> <http://www.translationautomation.com/>



The LR 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 use, otherwise individuals and organisations risk law suits due to some form of infringement.

The challenge for both the LRT 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 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 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.

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, open source communities such as Creative Commons, GPL, 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.

For mixed-funded initiatives (private/public), we should ensure that there is an agreement to make *resources available at fair market conditions* right from the start.

### 1.8 Resource Sustainability

*“Ensure sustainability of language resources”*

	LRP	PM
Ensure that all LRs to be produced undergo a sustainability analysis as part of the specification phase	•	•
Make sustainability plans mandatory in projects concerning LR production	•	•
Assess LR sustainability on the basis of impact factors	•	
Foster use of a sustainability model	•	•

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 LRs instead should be a key priority.

*LRs 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 LRs is 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.





Ch 1. The Blueprint

The challenge here is to establish clear rules for properly archiving and preserving resources without making resource creation and sharing process too complicated.

A top priority is *developing an analytic model of sustainability in which extrinsic and intrinsic factors are taken into account*, together with new methods and practices for sharing and collaboration<sup>14</sup>. The model must be backed by appropriate initiatives to encourage implementation.

**1.9 Resource Recognition**

*“Promote the LR ecosystem”*

	LRP	PM
Develop a standard protocol for citing language resources	•	•
Carry out a rolling survey of conference papers and LRs mentioned in papers.		•
Identify the “weak signals” indicating new trends in LT and LRs internationally		•
Bring together the main LR stakeholders to find a way of attach a Persistent and Unique Identifier (PUId) to LR		•
Give greater recognition to successful LRs and their producers (Prizes, Seals of Recognition, etc. )	•	•
Track the use of LRs in scientific and professional papers	•	
Compute a Language Resource Impact Factor	•	
Support training in production and use of LRs	•	•
Introduce training in the production and use of LR in Computational Linguistics and Language Technology curricula	•	
Continue to organize tutorials on LR production at conferences such as LREC	•	

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 very almost no rewards for researchers and institutions to share, preserve and maintain resources, and this now poses a number of serious problems.

*Language Resources thus 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 LR providers 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.

<sup>14</sup> See the model proposed by FLaReNet in Ch. 2.





### 1.10 International Cooperation

*“Promote synergies among initiatives at international level”*

	LRP	PM
Share the effort for the production of LRs between international bodies, such as the EC for the EU, and individual countries/regions, such as Member States in Europe		•
Encourage agencies to open their calls to foreign participants		•
Establish an International Forum to share information, discuss strategies and declare/define common objectives.	•	•
Establish MoU among existing initiatives, starting with the EC efforts		•
Maintain a public Survey on the LT and LR situation worldwide, based on FLaReNet and META-NET	•	
Ensure the sustainability of the community driven initiatives such as the LRE Map, META-NET Language Matrixes, and FLaReNet Network of International Contact Points and National Initiatives Survey Wiki		•
Produce a White Paper on LT, including LRs, in preparation for FP8. Distribute it within Europe and abroad	•	

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

A coordinated effort at the international level would help by providing less advanced countries/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.

*International cooperation between infrastructure initiatives* is also important for avoiding the duplication of effort, ensuring that standards are truly international, and encouraging the free exchange of ideas.

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.

It is an achievement, and an opportunity for our field, that recently a number of strategic-infrastructure initiatives have started, or are going to start, all over the world. This is also a sign that funding agencies recognise the strategic value of the LR field and the importance of helping a coherent growth also through a number of coordinated actions.

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

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.



## 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.
- 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*.

## Chapter 2 - Identifying Mature and Sustainable Language Resources

*Khalid Choukri, Joseph Mariani*

This chapter looks at LR sustainability and maturity on the basis of a descriptive model of sustainability from a LT perspective. It summarizes the description of a sustainability model, elaborated within the project<sup>1</sup>.

The sustainability as widely understood in the HLT field is the ability of a given Language Resource to survive over time without any explicit and external financial support, which could be referred to as a self-sustained resource. When mentioning such an expression, no reference is made to the availability and use by the wider community (whether R&D or industry), no reference is made to its rights management (e.g. licensing), to the ability to be customized to suit new needs, to its "openness" so new users could reshape it and convert it to an operable resource in their environment, no reference is made to its updates, corrections, improvements, repackaging, etc. These are however crucial parameters and the proposed model intends to list an exhaustive number of factors that impact this sustainability. These factors have also been clustered into Pillars (or categories) to ensure better understanding and ease to use in the prediction of sustainability of LRs.

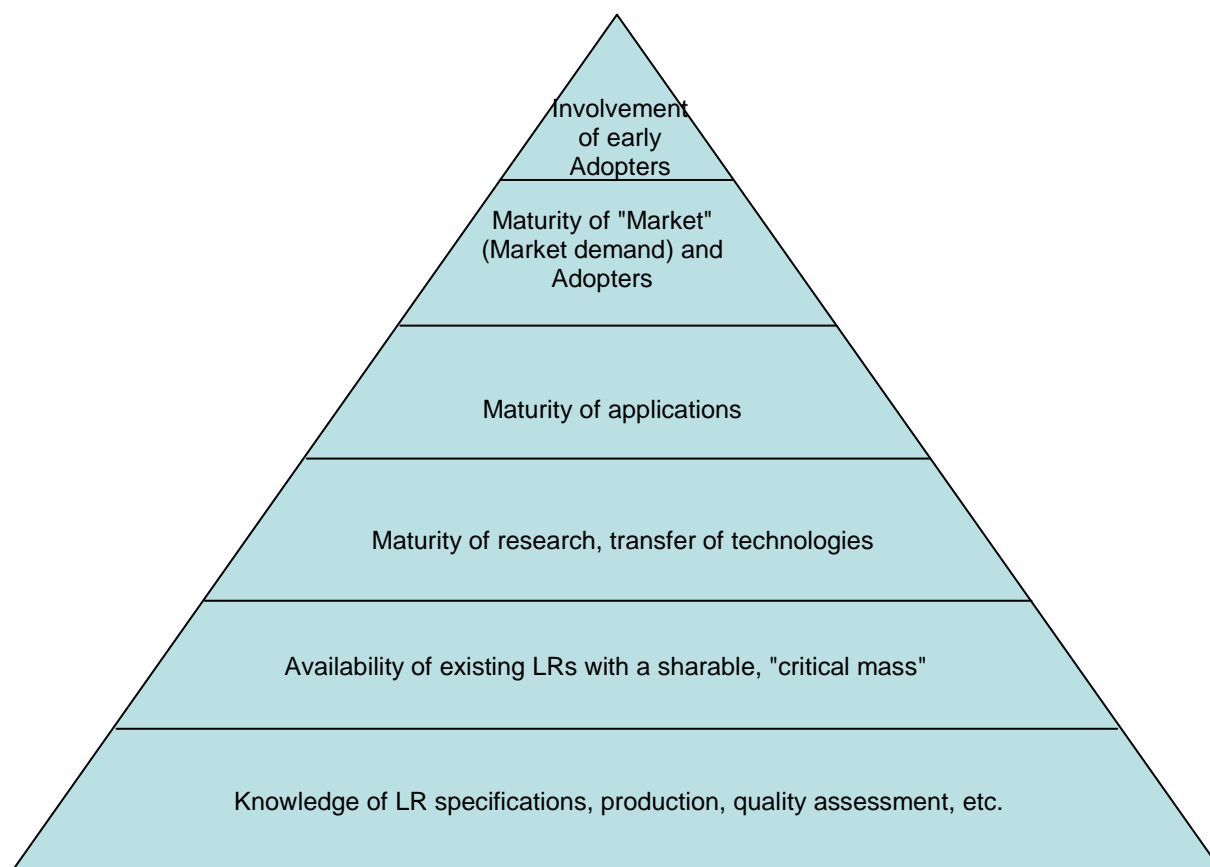
In addition to the "sustainability" term introduced herein, we often encounter other terms such as preservation, perpetuation, etc.

The goal of such descriptive model is to supply LR producers, packagers, maintainers, data-centres/distributors, users as well as funding agencies with appropriate tools that should help them design a rational and cost-effective lifecycle of LRs. Such model will be hinted at and will be implemented as risk-management-model, that could be used as a tool to anticipate on factors that may not comply with the sustainability requirements.

*Maturity* can be defined with respect to the following LT pyramid and the corresponding building blocks as shown in the following diagram:

---

<sup>1</sup> see FLaReNet Deliverable D2.2a for details.



The most important level is the lowest – knowledge of *LR specifications, production, and quality assessment*. It is fairly easy to assess the maturity of an area of research or application involving LRs when there is easily available public knowledge and know-how about the specifications (technical, logistic, linguistic, legal, etc.). Some of this knowledge may even come from established technical standards.

A framework such as SPEECHDAT<sup>2</sup> is an excellent example of resources developed in a well-designed configuration and broadly shared within the speech recognition community. The approaches adopted to develop such resources (adapted to R&D and pre-competitive tasks) have been widely used to produce resources that can be used in real applications. For a large number of technologies, there are well established best-practices (e.g. for SMT and broadcast news speech to text.).

2.1	PM	<i>Ensure strong public and community support to production and dissemination of "best-practices"</i>
-----	----	---

The second most important level in the pyramid is the *availability* of a critical mass of LRs. Mere knowledge about LR production best practices will not boost R&D unless these resources are developed in an adequate framework. This could lead to sharing resources and comparing the outcome of research activities between competing-cooperating labs. This capacity can enhance the maturity of the research and facilitate any subsequent technology transfer. The existence of a critical mass of LRs can trigger two contrasting types of initiatives:

<sup>2</sup> More details on [www.speechdat.org](http://www.speechdat.org)

- Boosting public and/or community support for sharing existing LRs of paramount importance for the EU, or
- Leaving it to the market (or other non EU public bodies).

A third level is the *maturity of research* and of the research community in question for a given topic. It is important to know whether there is a critical mass of research/researchers focused on a particular technology. This requires a careful monitoring of innovative work, scientific publications, patent applications, etc.

If it is unlikely that critical mass will be reached in a given area, policy makers will either deem it a vital/sensitive area (e.g. cultural heritage or security) and launch strong supporting programs, or it will be left to the market, where there is a risk that non-European economies could take the lead (as happened in the Personal Computer industry in the early 1990s).

2.2	LRP	<i>Monitor on-going research developments through scrutinizing publications and patent filings, and keeping track of prize-winning papers and Theses</i>
2.3	PM	<i>Support EU-critical LT domains once they have been identified. MT, Speech to Text (transcriptions), and Human Machine Interactions, are obvious examples</i>

Research maturity can be assessed objectively through evaluation campaigns built on a framework of test sets, metrics, and technology providers (see also chapter 5 on evaluation). It can also be assessed by analysing R&D-to-Market technology transfer.

2.4	LRP PM	<i>Ensure that there are regular evaluation campaigns to assess the state of the art and provide an accurate picture of EU research and technology</i>
2.5	PM	<i>Conduct regular surveys of technology transfer (ideally to stimulate, not simply reflect it)</i>

The fourth level refers to how far a tool, prototype, or component can be used virtually "out-of-the-box" (i.e. with an acceptable level of engineering effort).

Finally market maturity. Market demand and the existence of (early) adopters cannot be guaranteed when there is a full-equipped set of available technology and resources in place. Imperfect technologies can create highly successful applications, and brilliant technologies can fail in the marketplace due to lack of take-up or relevance.

## 2.1 Key sustainability factors

Reviewing the life cycle of several LRs, we identified a large number of extrinsic and intrinsic features that have an impact on the life of these resources. We have tried to be as exhaustive as possible while avoiding overlaps and redundancies. Their impacts on the life of resources have



been given estimated weights based on our experience and educated guesses. The use of these factors and their respective weights led to our sustainability score.

The factors that have been defined<sup>3</sup> as those impacting LR sustainability are listed herein:

- 1) Specifications (including references to best-practices & standards)
- 2) Production and management of the documentation
- 3) Quality assessment and/or quality validation report
- 4) Management of rights, ethics, privacy, consent, and other sensitive legal issues (before and during production)
- 5) Information dissemination including scientific publications
- 6) Formats, encoding, content
- 7) Portability across languages, environments and domains
- 8) Packaging (a compilation of all parts including resource documentation)
- 9) Management of use rights and licenses for sharing and distribution
- 10) Data identification, metadata and discovery
- 11) Versioning and referencing
- 12) Usability assessment and relevance
- 13) Accessibility (in packages, over the web or other)
- 14) Accessibility of LRs in an “open” mode
- 15) Preservation of media to ensure long-term access
- 16) Access charge (for free /for a fee)
- 17) Reference to production and projects, environments in which such resources have been produced/used
- 18) Relevance for other NLP applications and areas
- 19) Maintenance and support over time
- 20) Role and impact of large-scale data centres and archiving facilities.

In order to stress the important “categories” of impacting factors, these have been clustered into 8 groups:

- a) Proper documentation with appropriate documentation management rules
- b) Management of rights, ethics, privacy, consent, and other sensitive legal issues
- c) Information dissemination including scientific publications
- d) Rights and licenses
- e) Data identification, metadata and discovery
- f) Accessibility of packages and media
- g) Accessibility of LRs in “open” mode
- h) Relevance to more advanced NLP applications and areas

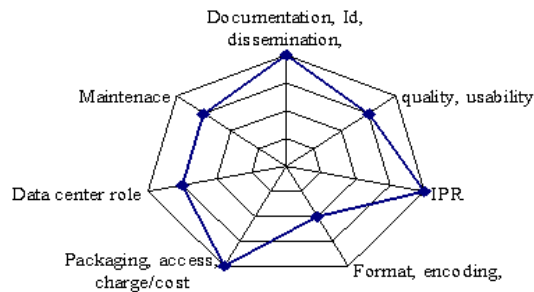
Note that extrinsic factors such as documentation, licensing, and access are among the key criteria for making a resource usable in the wider community. For example, despite theoretical controversies about the nature of the original WordNet, researchers and other users have developed and actively supported a similar resource for their own language (over 70 languages now have WorldNet related projects) largely because the documentation, rights, and access were properly addressed and maintained, making it easier for a critical mass of research and researchers to feel confident about using them.

Example of the scoring diagram could be:

---

<sup>3</sup> See FLaReNet deliverable D2.2a, “Identification of mature self-sustainable LRs versus areas to be sustained and sustainability factors”.





2.6	LRP	<i>Develop a sustainability scoring tool and make it publicly available<sup>4</sup></i>
2.7	LRP PM	<i>Assess LR sustainability on the basis of the sustainability factors and ensure that all LRs to be produced undergo a sustainability analysis as part of the specification phase</i>

## 2.2 Key areas of research interest

Key areas of interest are identified on the basis of the technologies that have been discussed within various forums over the last three years. The domain's own taxonomy has been under discussion for the last decade and there is still no consensus about the best way of cutting up this complex cake. We can look at two obvious dimensions:

In terms of *media* (also called *modality*):

- Text (corpora annotated at various levels: POS, syntax, semantics, co-reference/anaphora; lexica, thesauri, ontologies; translation memories/aligned corpora, etc.)
- Speech, audio, and visual (including multimedia) (transcriptions, signal/phonetic alignments, etc.)
- Video (scene annotations, speaker identification, topic detection, etc.)
- Other (sign languages, etc.)

Or *application type*:

- Web search, retrieval, extraction, indexing, named entity recognition, opinion mining, Q-A answering, multimedia, cross-lingual IR
- Text and discourse analysis & mining, summarisation, plagiarism detection

<sup>4</sup> See a preliminary version at [http://www.elda.org/flarenet\\_sustainability\\_dimensions.html](http://www.elda.org/flarenet_sustainability_dimensions.html)



- Machine translation, including speech to speech translation
- Speech recognition, speaker recognition, identification (voice command and Interactive Response Systems, dictation systems, conversational speech, speech to text from broadcasters, etc.), Speech synthesis (text-to-speech, avatars, etc.)
- Dialogue management (speech and/or text based)
- Language generation (sentence, report, etc.) associated with emotions, speech synthesis, etc.

2.8

PM

*Maturity has to be discussed and elaborated upon according to the two "description" dimensions (modality, application)*

### 2.3 LR maturity and sustainability in major research areas

#### *National Text Corpora*

One key research stream has been the collection and analysis of various features of corpora of written and spoken data. A quick survey of existing corpora shows that most East-European languages have at least one major reference corpus, often referred to as a National Corpus. This is the case for Bulgarian<sup>5</sup>, Czech<sup>6</sup>, Hungarian<sup>7</sup>, Polish<sup>8</sup>, Slovene<sup>9</sup>, Slovak<sup>10</sup>, etc. Similar corpora already exist for UK English (the famous British National Corpus<sup>11</sup>) and now American English (American National Corpus and in particular the freely available sub-set - Open ANC or OANC<sup>12</sup>).

Many of these "National" corpora reflect the current use of the general language and are appropriate for LT purposes, though for a number of them the content is largely skewed to the Social Sciences and Humanities (incorporation of sources for which copyright "period" expired). The major drawback of such resources is they are often copyrighted by third parties and hence are only offered through access tools (via an online service with sophisticated query forms that extract "samples" to better understand the language features of this domain of knowledge). Such approach does not allow exploiting the whole corpus, which is essential for developing LT applications.

Corpora specifically produced for LT tend to gain some legitimacy from the clout of the hosting institute, but also from the wealth of levels of annotation and mark-up (POS, syntax, co-references, anaphora, and other types of annotations).

- ⇒ **Maturity:** Corpora used in LT (e.g. the BNC) have benefited from over a decade of feedback, tuning, and corrections. But many others are used as online services and are of little use for LT (given the access mode that limits their availability).
- ⇒ **Sustainability:** High because of the institute or consortia underwriting the resource.

2.9

LRP

*Ensure that efforts are put in clearing the IPR issues and reformatting the data, so that*

<sup>5</sup> [http://ibl.bas.bg/en/BGNC\\_en.htm](http://ibl.bas.bg/en/BGNC_en.htm)

<sup>6</sup> <http://ucnk.ff.cuni.cz/english/index.php>

<sup>7</sup> [http://mnsz.nytud.hu/index\\_eng.html](http://mnsz.nytud.hu/index_eng.html)

<sup>8</sup> <http://www.nkjp.pl/index.php?page=0&lang=1>

<sup>9</sup> [http://www.fidaplus.net/Info/Info\\_index\\_eng.html](http://www.fidaplus.net/Info/Info_index_eng.html)

<sup>10</sup> [http://korpus.juls.savba.sk/index\\_en.html](http://korpus.juls.savba.sk/index_en.html)

<sup>11</sup> <http://www.natcorp.ox.ac.uk/>

<sup>12</sup> <http://www.americannationalcorpus.org/>



such (national) corpora could be repurposed and repackaged for LT

### **Spoken Corpora**

Speech corpora used in LT have been developed to adapt to the evolution of research topics, mostly influenced by market and security concerns<sup>13</sup>.

The first R&D programs targeted command and control (small vocabularies), followed by dictation, interactive tasks (e.g. ATIS and Sundial<sup>14</sup>), and then to transcription of broadcast news and conversational speech. The main focus today in speech technology is on the transcription of audio speech (broadcast news, meetings, lectures, conversational speech, etc.), information retrieval from audio/transcribed data, speech-to-speech translation, the automatic summarization of spoken data, etc.

There are corpora of transcribed broadcast news for a number of languages (American/British English, Arabic, French, German, Portuguese, Thai, Vietnamese, etc.) but the research community does not always have easy access to them. Major evaluation campaigns (e.g. NIST, ELDA) focus on a few languages (e.g. English, Mandarin Chinese, French and Arabic) so more use is made of English corpora, for example, than any other language.

A few years ago, the telephone-based speech recognition community began a large programme on front-end evaluations called Aurora<sup>15</sup>, led by the Aurora Working Group at the European Telecommunications Standards Institute (ETSI). The purpose was "to determine the robustness of different front ends for use in client/server type telecommunications applications." To support this initiative, a number of the resources developed in this programme have been repackaged, and Aurora databases are available from ELRA.

Other resources for Human-Machine Interactions have been produced but cover artificial settings, as none of the major users of dialogue systems and/or human-human services (such as call-centre operators) are willing to share their data with the R&D community and potential competitors (including reasons of privacy). The "artificial" setting is often based on Wizard of Oz simulations (MEDIA and Port-Media projects at ELDA<sup>16</sup>) or conversations between friends and relatives on pre-defined topics (e.g. Call Friends & Call Home at LDC<sup>17</sup>).

- ⇒ **Maturity** can only be reached when a critical-mass of community<sup>18</sup> players decides to tackle the same issue such as broadcast news, Aurora front-ends, etc. There are still major technical problems with conversational speech and broadcast news collected in low-bandwidth contexts, noisy channels, overlapping speech in intense discussion contexts, etc.
- ⇒ **Sustainability**: Resources from responsible data centres are usually sustainable (i.e. they can be identified and discovered by newcomers), but many resources developed and used internally by individual organizations cannot be accessed and maintained.

### **Corpora for MT**

During the last decade, work on Machine Translation has been fostered by the statistical approach using translation memories and aligned corpora as training and tuning resources. Aligned corpora have stimulated the development of MT systems for pairs of languages for

<sup>13</sup> See performance illustration in section 2.7 as given by NIST.

<sup>14</sup> The ATIS resources are available through LDC while the EC project Sundial resources are lost!!

<sup>15</sup> <http://www.elda.org/rubrique18.html>

<sup>16</sup> <http://www.elda.org/article252.html>

<sup>17</sup> [www ldc.upenn.edu/Catalog/LDC96S49.html](http://www ldc.upenn.edu/Catalog/LDC96S49.html)

<sup>18</sup> Or more often a funding agency!



which traditional rule-based MT systems would have been too expensive to craft. Most of these resources are available either for the major languages used in international bodies (EU and UN), or for those that are used by a large software development community (on the basis of produced technical manuals).

Most EU languages, for example, are well supplied with resources from the Europarl and JRC-Acquis corpora. This is also true for languages used in large international organizations e.g. Arabic, Russian, Chinese, French, English and Spanish at the United Nations, part of whose multilingual holdings have been released. This has enabled baseline MT systems to be built at a low cost by R&D teams, many of which use MOSES open source software (see for instance the baselines developed for English to Arabic within the Medar project<sup>19</sup>).

For other languages, open resources have been packaged, exploiting translated and localised software technical manuals and screen messages e.g. Linux manuals, KDE manuals (KDE Kool Desktop Environment manuals), etc.

One side-effect of this statistically-driven, aligned corpus approach is the development of more research into crawling techniques and tools as well as alignment tools. Exploitation of alignment tools for the extraction of bilingual lexica has been carried out by many labs that have improved such tools.

Emerging hybrid approaches to MT combine statistical with rule-based techniques. Some SMT engines mix together aligned “raw” corpora with aligned POS tagged corpora to make use of more knowledge and leverage existing resources.

The evaluation of MT systems in large campaigns carried out by NIST (annual MT evaluation), ELDA (TC-STAR, CESTA and MEDAR), Euromatrix/EuromatrixPlus projects, have also helped underwrite the SMT approach for a number of language pairs. Existing resources are not enough and more effort should be put into compiling adequate resources in collaborative environments.

- ⇒ **Maturity:** In SMT, (relatively few) fairly small-scale existing resources have been used and reused again, but there are also some larger-scale resources for some EU and UN languages.
- ⇒ **Sustainability:** Many of the language resources involved are supported by strong institutions (the EC Joint-Research Centre in Ispra, ELRA, LDC, NIST, etc.) and are therefore likely to survive.

2.10	LRP	<i>Implement tools that can help crawl “comparable” resources, select a small set that can be aligned and then provide clear legal conditions to ensure they can be shared with the MT community and beyond (assuming these can be annotated for bilingual lexical extraction, named entities, etc.)</i>
2.11	LRP	<i>Develop a certain volume of MT-based corpora, post-edit them (semi)automatically to specific quality levels, and then use them iteratively to train new MT systems more effectively</i>

## 2.4 Evaluation packages and resources

A large number of technology areas are backed by well-designed and well-organised evaluation campaigns, DARPA being an exemplary case in point with sustained work on MT, audio transcription, and speaker identification and other topics that have been evaluated by NIST (and by LDC for the resources). The EC and European national governments have also backed

<sup>19</sup> [http://www.medar.info/MEDAR\\_evaluation\\_package/package\\_medar\\_v5.tar.gz](http://www.medar.info/MEDAR_evaluation_package/package_medar_v5.tar.gz)



evaluation efforts such as Technolangu and Technovision in France (ELDA) and N-best/STEVIN in the Netherlands, while Evalita in Italy has been a voluntary community effort.

This approach highlights mature practices for producing language resources, and assesses the maturity of specific technologies. To ensure comparability of the results, participants have to agree on a common framework for data content as well as formatting and encoding specifications.

- ⇒ **Maturity:** Evaluation campaigns are clearly the best way to assess the maturity of a given type of technology/resource; evaluation centres need to agree on a number of features, and such agreement (consensus) only become feasible if topics have reached a certain level of maturity.
- ⇒ **Sustainability:** Possible when an official body underwrites the availability of the whole evaluation package and resources, but difficult when it comes down to project level.

2.12

LRP

*Define a small set of crucial topics, appropriate for comparable evaluations, on which a critical mass of research/researchers should focus through a clear adherence to principles*

The ELRA HLT evaluation portal<sup>20</sup> compiles information about past and present LT evaluation activities. It has identified a large number of multilingual and monolingual areas where activities have been carried out with specific resources, ensuring that best-practices for developing test data have been agreed upon by those involved.

For instance the ImageCLEF<sup>21</sup> evaluation (in its 9<sup>th</sup> year) has developed methods for releasing test data in areas such as:

- Medical retrieval
- Photo annotation
- Plant identification
- Wikipedia retrieval
- Patent image retrieval and classification in collaboration with CLEF-IP

The same goes for NIST that assesses "progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation" in its video retrieval evaluation (TRECVID<sup>22</sup>).

Areas identified include:

- Machine Translation
- Speech-to-Speech Translation
- Multilingual Text Alignment
- Information Extraction
- Information Retrieval
- Text Summarization
- Parsing
- Speech Recognition
- Speech Synthesis

<sup>20</sup> <http://www.hlt-evaluation.org/>

<sup>21</sup> <http://imageclef.org/2011>

<sup>22</sup> <http://www-nlpir.nist.gov/projects/tv2010/tv2010.html>



## Ch 2. Sustainable LRs

- Multimodal Technologies

These multimodal technologies include technologies combining features extracted from different media (text, audio, image, etc.):

- Audio-visual Speech Recognition
- Audio-visual Person Identification
- Audio-visual Event Detection
- Audio-visual Object or Person Tracking
- Biometric Identification (using face, voice, fingerprints, iris, etc.)
- Head Pose Estimation
- Gesture Recognition
- Multimodal Information Retrieval (e.g. Video Retrieval)
- etc.

Related resources have been developed in projects such as AMIDA, QUAERO, TRECVID, IMAGECLEF, CHIL, VACE, TECHNO-VISION, BIOSECURE, etc.<sup>23</sup>

On the basis of such lists, we can conclude that many of these areas are mature enough in terms of best practices to be able to produce resources for training and/or benchmarking the technologies. Most of these, however, only cover a small set of languages or language pairs. For instance Summarization data can be found in projects such as Must/NTCIR (Japan), TAC/NITS, TIPSTER, TIDES (USA), GERAFF (Canada, French) that in all only cover three or so languages.

2.13

PM

*Ensure that resources are developed for all EU languages so that evaluations can be made of language-dependent technologies listed by the major agencies (training and testing data)*

## ***2.5 Multiple-application resources as a "maturity/sustainability" factor***

Some LRs can be used to develop and test technologies in more than one LT research or technology area. This is important for sustainability as it enables resource production and maintenance costs to be shared across larger communities over longer spans of time.

A good example is the set of large monolingual corpora that are annotated at various analytic levels - morphological, named entity, WSD, and treebank - and are often used by the speech community to build language models, and by the text community to test named entity recognition, syntactic analysis, and so on. The main development problem is not being able to anticipate the appropriate range of coding/encoding features that will ensure that a given resource can be used effectively for new NLP applications as they emerge.

Generally speaking, if a resource is backed by a large community of evaluators (which implicitly indicates adherence of a large number of researchers and developers) , there is better chance that it will be more widely used.

## ***2.6 Maturity as reflected by the LRE MAP (2010) data***

The LRE Map<sup>24</sup> shows the "big" trends, but is not granular and synthetic enough to identify how mature each resource is for each language.

<sup>23</sup> <http://www.hlt-evaluation.org/spip.php?article149>

<sup>24</sup> <http://www.resourcebook.eu>



The Language Matrixes produced in META-NET on the basis of the LRE Map data (Mariani and Francopoulo 2011) provide a synthetic view for the various modalities (Written language, Spoken language, Multimodal/Multimedia) and for different categories of Language Resources (Data, Tools, Evaluation means and Meta-resources (Standards, Metadata, Guidelines), but they still lack granularity and completeness. The Language Matrix on Multimodal/Multimedia data below shows for example that there is a strong prominence of the English language, with 22 of the 82 identified corpora<sup>25</sup>. But it could be interesting to have a more detailed analysis of the different types of corpora, while the situation for the other languages and for the other types of resources is sparser and needs an improved taxonomy for the Types (the initially suggested ones are in bold characters) and more data from new sources.

MM Data (Ranked)	Bulgarian	Czech	Danish	Dutch	English	Estonian	Finnish	French	German	Greek	Hungarian	Irish	Italian	Latvian	Lithuanian	Maltese	Polish	Portuguese	Romanian	Slovak	Slovene	Spanish	Swedish	Other Europe	Regional Europe	Multilingual	L.I.	N.A.	Total
<b>Corpus</b>	0	1	3	4	22	1	2	3	10	2	2	0	4	0	1	0	0	4	0	0	0	7	4	3	1	0	3	5	82
<b>Terminology</b>	0	0	0	0	2	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	7
Database	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
<b>Ontology</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
A complete archive of resources	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Data Collection and Annotation Management System	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
<b>Grammar/Language Model</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
<b>Lexicon</b>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	0	1	3	4	25	1	2	4	12	2	2	0	4	0	1	0	0	4	0	0	0	8	4	3	1	1	7	7	96

E.1.a. Multimodal/Multimedia Data (Ranked order)

The same goes for written language data. Over 950 resources have been identified, the top three covering English with 327 items, French with 80 items, and German with 70. More than 50% of those resources (552) are constituted by corpora, and 25% by lexica (229) but there is a long tail of new Types entered by the authors.

<sup>25</sup> The types of resources that appear in bold characters correspond to the types suggested in the questionnaire. The other ones have been added by the authors.



Written Data (Ranked)	Bulgarian	Czech	Danish	Dutch	English	Estonian	Finnish	French	German	Greek	Hungarian	Irish	Italian	Latvian	Lithuanian	Maltese	Polish	Portuguese	Romanian	Slovak	Slovene	Spanish	Swedish	Other Europe	Regional Europe	Multilingual	L.I.	N.A.	Total	
	<b>Corpus</b>	7	12	6	17	206	3	3	44	43	10	8	1	32	9	4	1	7	19	12	2	5	29	19	19	18	5	9	2	552
<b>Lexicon</b>	6	7	2	8	77	1	2	24	15	3	4	0	16	0	0	0	2	6	7	0	1	19	4	11	8	3	3	0	229	
<b>Ontology</b>	1	2	0	2	18	0	0	3	4	2	0	0	4	0	2	0	1	1	1	0	0	4	0	3	0	1	16	2	67	
<b>Grammar/Language Model</b>	1	1	2	1	11	0	1	4	2	0	1	0	2	0	0	1	2	1	1	1	0	5	1	3	1	0	2	1	45	
<b>Terminology</b>	1	1	0	2	10	1	0	5	3	0	1	0	0	1	1	0	1	0	0	0	0	2	0	2	3	1	1	0	36	
A syntactic judgments database	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	3	
Resource: morphology	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	
Thesaurus	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	
A Knowledge Base with Lexical-Semantic Relations between words	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
A list of categories with examples of language use	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
Controlled Legal Language	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
Corpus-Based Online Dictionary	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
Database	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
Encyclopedic knowledge	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Event Semantics	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Lexicon/corpus	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Online Encyclopedia	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Repository of bilingual lexicons	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Resources integration	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Text Book	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Virtual Game World	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Yahoo!'s local listings in Chicago	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Encyclopedia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Part-of-Speech Tagset	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Psycholinguistic Database	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
<b>Total</b>	16	23	11	30	327	5	6	80	70	15	14	1	56	10	7	2	13	27	21	3	6	61	25	39	30	11	36	5	950	

E.3.a. Written Language Data (Ranked order)

The META-NET Language Whitepapers<sup>26</sup> also contain Language Tables, which provide a subjective and qualitative overview of the situation of Language Technologies and Language Resources for each language.

Again, a detailed analysis of such resources is essential to detect usable resources and the remaining gaps.

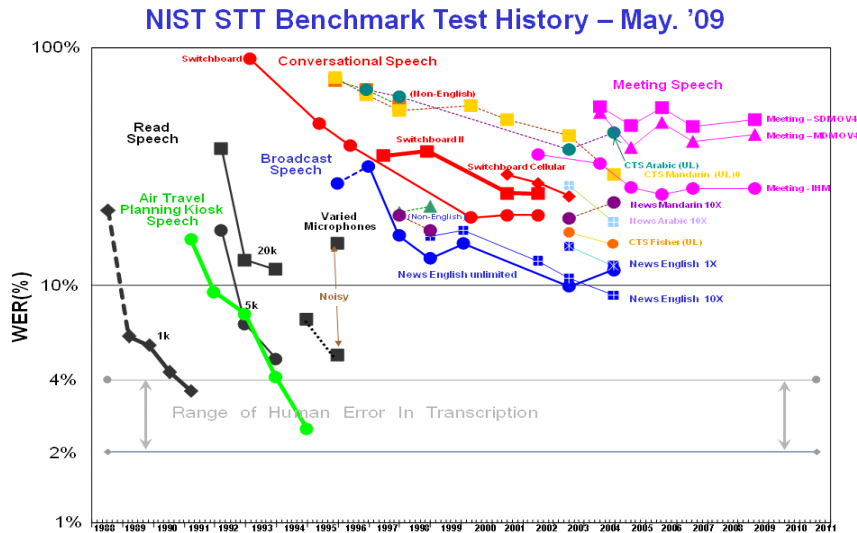
## 2.7 Maturity in terms of objective evaluations

Objective evaluation can show how "mature" a technology is in terms of performance.

As an example, the following diagram charts the progress of ASR systems over the years, on the basis of international evaluation campaigns conducted by NIST (the best performance obtained each year in terms of Word Error Rate (WER)). NIST has conducted multiple evaluations with various tasks, from voice-activated system with a vocabulary of 1,000 words, voice dictation (5 and 20 K words), radio/TV broadcast news transcriptions (English, Arabic and Chinese Mandarin) with "high" quality recordings (e.g. no overlap between speakers), telephone

<sup>26</sup> <http://www.meta-net.eu/meta-share/whitepapers>

conversations (also in English, Arabic and Mandarin), transcripts of meetings, etc., in variable conditions.



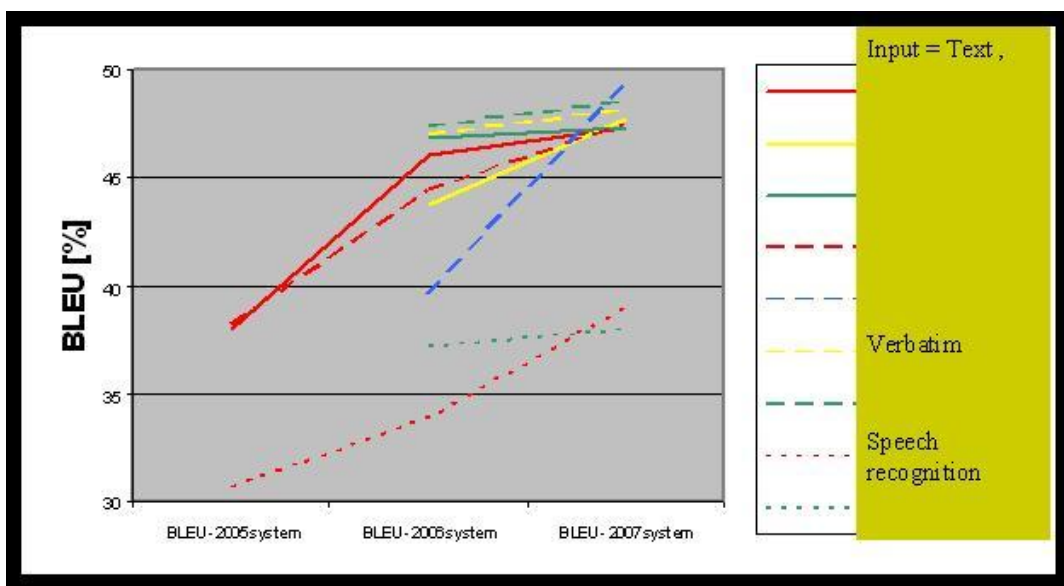
Timeline for Automatic Speech Recognition since 1987 in terms of NIST evaluation campaigns<sup>27</sup>

Today, there are a number of real-world applications in use (transcription of broadcast news, voice dictation, command and IVR)<sup>28</sup> where the technology achieves sufficient performances. As it appears in the chart, those performances are similar to those of humans for some tasks (voice command, voice dictation). But it is also clear that for other more difficult tasks (conversational speech, or meeting transcriptions), progress is very slow. It should also be stressed that some tasks, such as voice search, don't necessitate very high recognition performances. The performances attained by technology should therefore be compared with the performances needed by the application.

In the area of speech to speech translation as explored in the TC-STAR project, the results of evaluation campaigns are as follows: (Joseph comment: The content and the drawings of the figure are not clear. It may also be said that humans achieve a BLEU score of about 80. And the results of the “understanding” measure conducted in TC-Star, showing that the level of understanding by a human of a text translated by a machine is close to the level for the text translated by a human could also be mentioned.)

<sup>27</sup> <http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html>

<sup>28</sup> Many applications are based on the recognition of audio streams but some (broadcast news) are based on the “parrot” approach, that involves a well-trained speaker “repeating” (into a speech recognition system adapted to their voice) the output of a speaker.



MT Evaluation in the TC-STAR project (Input = Output of automatic speech recognizers, verbatim: transcribed speech, text= edited transcriptions)

Again, the underlying technologies are now at work in a large number of real-world applications such as Google-translate.

## 2.8 Final recommendations

2.14	LRP	<i>Draw up a list of the “top 20” technologies – Machine Translation, automatic speech/speaker recognition, multimodal/multimedia information retrieval/extraction, summarisation, etc.</i>
2.15	PM	<i>Ensure that funding agencies and sponsors put efforts so that a critical mass can be realised for at least 10 of these technologies, for which crucial resources should be produced in a framework, publicly and fully funded, that ensure full sharing under the fairest conditions</i>
2.16	PM	<i>Secure an evaluation framework to assess the progress made by such technologies with respect to state of the art, reported elsewhere</i>
2.17	PM	<i>Foster the use of the sustainability model by all resource production players</i>
2.18	LRP PM	<i>Ensure that all “best/good practices” are disseminated to the R&amp;D community through very lightweight methods</i>



### **References**

Mariani, Joseph and Francopoulo Gil, (2011). First Public Version of the META-Matrix. Deliverable D11.1.1. META-NET. [www.meta-net.eu/public\\_documents/t4me/META-NET-D11.1.1-Final.pdf](http://www.meta-net.eu/public_documents/t4me/META-NET-D11.1.1-Final.pdf)



## Chapter 3 - Strategies for Language Resource Mixing, Sharing, Reusing and Recycling

*Stelios Piperidis, Penny Labropoulou*

This chapter focuses on reusing and interlinking existing resources and tools as a means to derive, produce and repurpose new LRs (datasets and relevant processing tools). Together with the automatic production of LRs (dealt with in Chapter 5), these strategies are ways of remedying the lack or scarcity of LRs needed for the advancement of LT research and development.

### 3.1 Reusing resources

*Reuse* in this context involves the use of LRs as a component that can be integrated into a new resource, or extending it in various ways (e.g. linking, translating, etc) to create a new resource; as seen below, it also involves reusing a LR “model”/specs. “Raw” corpora, for example, can be partially or fully re-used to create annotated corpora, a tagger can be used as a component of a corpus management environment, or a morphological lexicon can be combined with other types of lexicons (encoding subcategorisation frames, sense relations etc.) to create a new, enhanced lexicon.

New resources have often been produced by re-using existing content. But this has usually been carried out by the organization that created the original LRs and for their purposes only. Suites of tools (corpus building components, a POS tagger, parser, annotation tools, statistical data extraction components etc.) are usually put together to form an integrated corpus management environment. Similarly, while one part of a large corpus may be morpho-syntactically tagged, a smaller part may be further parsed to form the basis for a treebank, and the same or yet another part may be manually or semi-automatically annotated with semantic information. Re-use has also been extensively applied to the compilation of dictionaries and computational lexicons, extracting information from corpora (e.g. examples) and other dictionaries (e.g. glosses and semantic relations).

The classic case for resource re-use by a broad range of organizations and researchers to create new LRs is WordNet. EuroWordNet was the first project that spurred the creation of new WordNets for other languages, not only using the basic model but also the resource itself as a basis for the development of the new LRs. Other languages with different funding sources have followed.

In addition to its technical merits, the success of WordNet (as mentioned in 2.1) is largely due to the following

- It is easily accessible and freely available for all types of use.
- A large community, organized as an association, has been created and is active all over the world supported by a web site dedicated to the constant exchange of information on wordnets for all languages<sup>1</sup>.
- Different methodologies have been devised to develop new language versions depending on the availability of other resources. The availability of digital monolingual lexical resources with semantic information encoded in them for certain languages led to the selection of the *merge* model, in which the new WordNet is first built independently and then linked to the most equivalent synset in the Princeton WordNet (i.e. by merging the

---

<sup>1</sup> <http://www.globalwordnet.org/>



language-specific wordnets). Then there is the *expand* model, in which the original WordNet is expanded with equivalence links from each synset to synsets in the other languages; this model is favoured for languages in which digital bilingual resources already exist. In certain cases, the resource producers have even translated the original WordNet and exploited the original synset relations to create the semantic network of their own language.

### 3.2 Interlinking resources

Resources can be *interlinked* to create new resources.

For instance, interlinking of LRs across languages can be used to produce annotated corpora. *MultiSemCor*<sup>2</sup>, an Italian semantically annotated corpus, has been created by transferring the annotations of the English original resource (*SemCor*) to the manually translated version, following alignment of the two corpora. In fact, using parallel aligned corpora, one of which is annotated syntactically or semantically and these annotations are then projected on to the other, is a highly promising method for bootstrapping new LRs with less manual effort<sup>3</sup>.

*Interlinking* different resources within the same language can also help produce larger and richer LRs. In the case of corpora, interlinking the different annotation layers of corpora can lead to larger, coherent, multipurpose LRs. The American National Corpus (ANC)<sup>4</sup> has issued an open invitation for contributions of linguistic annotations on all or part of the ANC or the OANC (the freely available version of the ANC) and for such derived resources as word lists for free distribution and use. The OANC has been annotated with various types of linguistic information (WordNet senses, FrameNet annotations, part of speech tags, Penn Treebank syntactic annotations etc.). All the annotations are in GrAF format (the ISO standard for standoff annotation of linguistic data), and can be merged or combined using ANC resources. GrAF annotations can be loaded into annotation tools such as GATE and UIMA and/or transduced to other formats using the ANC2Go tool.

As shown by these examples, successful interlinking is facilitated to a large extent by ensuring interoperability (explored in Ch. 4) which enables the re-usability and mixing of resources and tools.

### 3.3 Repurposing resources

*Repurposing* involves taking a resource originally created for a specific application or domain and tailoring it to the requirements of another application or domain. A good example would be taking linguistic/terminological resources and re-engineering them to create ontologies (see e.g. the NEON project<sup>5</sup>). This kind of transfer of resources to new domains and applications, although it is not easy, seems to be a promising way to reduce production costs and has attracted a lot of attention recently. As scientific methods and techniques get increasingly used across scientific areas (or subareas) and problems (e.g. methods used for automatic speech recognition being carried over to machine translation, but also alignment techniques used in processing both biological data streams as well as linguistic data) repurposing of resources, e.g. use of transcribed speech corpus to derive language models for other purposes, is expected to gain more importance.

---

<sup>2</sup> <http://www.multisemcor.fbk.eu>

<sup>3</sup> See, for instance, Padó and Lapata. (2009), Bentivogli and Pianta (2005), Hwa et al. (2005), Smith and Eisner (2006).

<sup>4</sup> <http://americannnationalcorpus.org>

<sup>5</sup> <http://www.neon-project.org>



### 3.4 General considerations on reuse

The examples presented above show some of the ingredients for successful re-use: ease of access, free availability for research and re-use, clear licensing terms, community building, exchange of ideas and information across related projects, ease of use, interoperability and potential for merging, research ideas and experimentation on re-using existing material in innovative ways.

It is now almost unanimously agreed that the re-use of existing resources can help reduce the cost and time of LR production. Nevertheless, effective re-use comes with its own kind of price-tag.

The re-use of resources is largely dependent on LR availability, identification and easy access. In fact, infrastructural issues – such as interoperability, resource sharing, and easy access to LRT – were recurring messages in all the sessions at the FLaReNet Forums and in other FLaReNet or FLaReNet-related events, as well as in all three META-NET Vision Groups (Translation and Localisation<sup>6</sup>, Media and Information Systems<sup>7</sup>, Interactive Systems<sup>8</sup>)

#### A LRT Infrastructure

Many LR groups, initiatives and individuals have been advocating for some time for the creation of a language resource and technology infrastructure - an open resource infrastructure that enables the easy sharing of data and tools that can operate seamlessly together. This is now increasingly recognized as a necessary step for building on each other's achievements, integrating resources and technologies, interconnecting and coordinating existing structures and avoiding the dispersal or inconsistency of various efforts. The META-SHARE infrastructure is now being implemented by the META-NET (Technologies for the Multilingual European Information Society) consortium to address this requirement.

Operating this infrastructure efficiently and effectively depends on the will of the LR community, who must now endorse it, populate it richly with LRs to meet emerging needs and use it.

3.1	LRP	<i>Make their LRs available, visible and easily accessible through an appropriate infrastructure (e.g. META-SHARE); participate in building the infrastructure by providing feedback on relevant actions in this direction</i>
3.2	PM	<i>Ensure a stable sustainable infrastructure for LR sharing and exchange; support continuous development and promotional activities</i>

#### Legal and licensing issues

Over the last decades we have witnessed a vast growth in digital text and audio-visual content thanks to the advent of the internet, the large-scale digitization of printed and handwritten material, and the production of computer and mobile-mediated texts and multimedia/multimodal data. However, legislation concerning their use is neither clear nor even consistent across countries. US law includes the *fair use act*<sup>9</sup>, while in a number of European countries reference is made to the free use of LRs for educational purposes but not for research purposes, let alone for commercial applications. Moreover, existing legislation, particularly copyright law, is not uniform over all types of LRs, and there are different regulations regarding

<sup>6</sup> <http://www.meta-net.eu/vision/reports/VisionGroup-TranslationLocalisation-draft.pdf>

<sup>7</sup> <http://www.meta-net.eu/vision/reports/VisionGroup-MediaInformationServices-draft.pdf>

<sup>8</sup> <http://www.meta-net.eu/vision/reports/VisionGroup-InteractiveSystems-draft.pdf>

<sup>9</sup> Copyright law of the US: para 107, Title 17.

the treatment of text as compared with the treatment of audio, images or video, and different again for software, technology applications and non-creative material in general. In addition, the rules regarding collective works, databases and works of shared authorship are inconsistent. Copyright limitations and exceptions need to be harmonised across the EU so that:

- a) there are simple and clear rules as to how content that may constitute a LR may be used
- b) the restrictions imposed by copyright owners are kept to a minimum and do not hinder the development of LTs
- c) such restrictions are waived in case LRs are deployed for research purposes.

3.3

PM

*Harmonize legislation regarding LR use for all types of LRs, and make provisions for allowing the free use thereof at least for research/non-profit purposes*

### ***IPR issues***

Clearing the IPR on primary data can create problems for the availability of LRs and should be considered early in the design and specifications stage. This is often neglected and usually undertaken at the very end of the production stage, and even afterwards. When more than one source of primary data is involved, IPR clearance at the final stage can be so time-consuming and manpower intensive that it is simply never done. Collecting spoken data, for instance, is often carried out without the prior, legally formulated permission of the informers; and going through the process of asking informers to sign approvals can at the end of the LR production be more time-consuming than creating a new LR from scratch.

Nowadays, more and more textual data, especially those collected by individuals for their personal work, are being harvested without asking prior permission from the owners. Web crawling is a quick way to obtain large amounts of data for personal research. Inclusion of personal data (e.g. names of speakers in dialogues) can also open up an LR liable to legal problems. Derived resources, such as lexicons extracted from corpora, may be more debatable as to potential copyright law infringement, but they also run the risk of legal prosecution.

LR producers should therefore try to clear the legal rights at kick-off and at the same time make sure that the material they are collecting can be made available for most types of use, including modifications and the creation of derivatives. If they can only obtain this for part of the LR, this should be clearly marked as such and made available to the LR community. When making an LR available, the original material's IPR should be respected and authorised types of use should be specified under an appropriate licensing scheme.

3.4

LRP

*Clear IPR at the early stages of production; try to ensure that re-use is permitted*

3.5

PM

*Establish an infrastructure that can help LR producers with legal issues*

### ***Collaborative efforts***

Due to the growing availability of raw digital data, there is a steady shift of interest towards annotation, especially at the semantic and pragmatic/discourse levels and across modalities. Annotated material is extremely valuable for advances in LT; and there should be a focus on larger volumes and higher quality. As high-level annotation is not yet automated (at least not to an acceptable quality), other means of producing them could be deployed.



One solution is to leverage collaborative effort by field experts and trained annotators. Other methods, such as crowd-sourcing proper via, for instance, the Amazon Mechanical Turk<sup>10</sup>, social tagging and gaming platforms, that rely on eliciting datasets and/or content (mainly annotations) from the general public with little or no training present many advantages; however, there should be more experimentation, metrics for strengths and drawbacks, and suggestions for improvements thereof (see Callison-Burch & Drezde 2010), including elaboration of ethical aspects involved in some of the methods. All collaborative efforts must be co-ordinated to avoid redundancies and multiply gains by building upon each other's achievements. Communities should be created and supported by an infrastructure that allows exchange of ideas, comments, the status of each effort etc., following open invitations for contribution not only of annotated data (such as those issued by the ANC) but also calls for participation of on-going annotation projects.

Platforms that enable the storage and parallel processing of shared data are a key success factor. Standoff annotations or tools that allow existing annotations to be easily stripped off, as well as tools for merging various annotation levels, will prove valuable in this respect.

3.6	LRP	<i>Carry out LR production and annotation as collaborative projects; open up existing LRs for collaborative annotation and the reuse of the annotated results; participate in communities that carry out similar tasks; evaluate and document their results; develop new tools and/or adapt existing tools and platforms to the needs of collaborative work</i>
3.7	PM	<i>Support collaborative efforts for large LR production and annotation projects through appropriate funding; support the infrastructure required for collaborative work</i>

### ***Survey/Cataloguing of existing LRs***

Before embarking on the production of new LRs from scratch, it is important that LR producers take existing resources into account. Obvious cases of reuse are extensions/enrichments of existing LRs, adaptation of tools to new domains and applications and so on. More innovative ways of reusing LRs have been suggested in recent years, as reusing annotations from other languages and translations to produce annotations in one's own language as described above. The roll out of datasets and tools trained on other languages for producing LRs for less-resourced languages should be further investigated. Languages of the same family may show enough similarities to boost mutual LT advancement by using the same set of datasets and tools. More efforts should be dedicated to this line of research.

3.8	LRP	<i>Check the availability of existing LRs; try to reuse what is available; check what is available not only for one's own language but also for languages of the same family; try to see what can be reused in terms of data and tools but also experience; create and exchange tools that help in reuse/repurposing/interlinking</i>
-----	-----	---

<sup>10</sup> The Amazon Mechanical Turk builds on the concept of crowd-sourcing (essentially crowd+outsourcing), where, in general, tasks traditionally performed by an employee or contractor are outsourced via internet to a large group of people willing to contribute for a low fee. Tasks are normally described as human intelligence tasks, and LR production, including various annotation tasks, constitute an example.

3.9

PM

*Support reuse projects and research activities in the domain; support activities that document reuse, formulate guidelines, etc.; fund and promote research activities around reuse (e.g. dedicated workshops, publications on successful reuse cases, etc.)*

### **Interoperability**

Formal and semantic interoperability is a strong desideratum for LR reuse, and standards and best practices are crucial for ensuring interoperability (see also Chapter 4). The survey carried out for the FLaReNet deliverable D3.1<sup>11</sup> has demonstrated both the use/usefulness of existing standards and the adoption of widely used practices as *de facto* standards. This approach should be strongly supported. Easy access to standards and best practices is a must. The FLaReNet site already hosts a repository of standards and best practices<sup>12</sup>, including references to resources that have used them. This initiative should be continued and further enriched. Adoption of existing standards should be encouraged; areas where there are no standards and/or are under way, should be identified and the creation of new standards with the widest possible consent of interested users and field experts should be promoted. To achieve better results, experts from academia and industry should work together.

3.10

LRP

*Look for standards and best practices that best fit the LRs about to be produced, already at the early stages of design/specifications; adhere to relevant standards and best practices; engage in groups that produce standards and provide feedback and comments; produce LRs that are easily amenable to reuse (e.g. adopt formats that allow easy reuse)*

3.11

PM

*Support infrastructural activities for collecting and disseminating information on existing standards and best practices; fund activities for setting up new standards where they do not exist; fund the development and/or maintenance of tools that support/enforce/validate standards*

### **Documentation**

All documentation relating to a LR should be easily accessible by prospective LR users, something that is fairly rare today (an exemplary exception is that of resources equipped with dedicated web sites). ELRA provides documentation for its resources in the ELRA catalogue, FLaReNet for a few individual resources in its repository<sup>13</sup> but this endeavour needs enrichment. Any infrastructure that hosts LRs should also host the relevant documentation; the different types of documentation (specifications, annotation guidelines, user manuals, validation reports, usage reports etc.) should be clearly identifiable. The META-SHARE infrastructure is taking the first steps in this direction, integrating many of these documentation dimensions.

3.12

LRP

*When producing an LR, allocate time and manpower to documentation from the start; collect and provide documentation (or links to it) when giving access to an LR*

<sup>11</sup> See FLaReNet deliverable D3.1 "Report on the scientific, organizational and economic methods and models for building and maintaining LRs"

<sup>12</sup> [http://www.flarenet.eu/?q=FLaReNet\\_Repository\\_of\\_Standards\\_and\\_Guidelines](http://www.flarenet.eu/?q=FLaReNet_Repository_of_Standards_and_Guidelines)

<sup>13</sup> [http://www.flarenet.eu/?q=Documentation\\_about\\_Individual\\_Resources](http://www.flarenet.eu/?q=Documentation_about_Individual_Resources)



3.13

PM

*In a LR production project, part of the funding should be allocated to documentation and dissemination activities; support activities for collecting and storing in appropriate infrastructures documentation for LRs*

### **Metadata**

Metadata descriptions play an important role in making LRs easy to identify. All LRs should therefore be accompanied with appropriate metadata records. Existing and/or new metadata schemes should allow sharing information and reducing the time and effort required to produce these records. LR repositories should be open to metadata harvesting through well-known protocols.

Formal and semantic interoperability of metadata can be guaranteed through the use of a common repository, such as the ISO DCR ([www.isocat.org](http://www.isocat.org)). Metadata schemes should take into consideration the needs of LR users and include information that will help them in identifying, accessing and using LRs. The META-SHARE and CLARIN metadata schemes have been created with these considerations in mind, while serving different communities and realms of science and technology. Appropriate tools for uploading and converting metadata records should be provided by LR repositories.

3.14

LRP

*Provide appropriate metadata descriptions for all LRs distributed, preferably in one of the widely-used metadata schemes*

3.15

PM

*Support metadata creation and promotion activities; set guidelines and rules for metadata description of available LRs and support relevant efforts, including their normalisation*

### **Evaluation and validation**

For LR reuse, it is important to evaluate and validate both content and form to ensure quality. This means that evaluation and validation practices should be encouraged (see Ch. 5). Information on this process and its results should be provided in the LR documentation and LR metadata descriptions.

3.16

PM

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

### **Availability**

When LRs are made available to the LR community, they should ideally be made available for most types of use and reuse, including modification, conversion, production of derivatives etc. Derived resources should also be made available under the same terms as the original resource.

Distribution policies and practices should be revisited in order to fit them better to LT user requirements. For instance, at the technical level, a corpus should be downloadable so for easier reuse, instead of merely being queried in a corpus management environment.

At the legal level, restrictions on the use or reuse of LRs should be reduced to a minimum so that LRs are continuously enriched, and LTs and related services are fully deployed and further developed. There should be standardized licenses clearly indicating what LR users may do with



the LRs for all available LRs. ELRA, CLARIN and the META-SHARE network have prepared a set of licensing templates where reuse is clearly mentioned.

National Funding Agencies should/could support periodic construction of reference resources for their languages assigning funding dedicated to the clearing of IPR issues so that such resources can then be made publicly available, distributable, and exploitable.

3.17	LRP	<i>Make LRs available for most types of use; make use of standardised licenses where reuse is clearly stipulated; make LRs accessible in formats that allow reuse</i>
3.18	LRP PM	<i>Encourage availability of LRs for multiple types (of re-)use; enforce the creation and adoption of standardised licenses where reuse is clearly stipulated</i>

### **Openness**

Ideally all LRs should be open to use, reuse, sharing, improvement and deployment in order to advance research and foster development. LR producers should be encouraged to make their LRs open at least for research purposes. Entirely publicly-funded LRs should be open or shareable, but must carry clear attribution and rights-holders information, be properly documented and be made available with a licence requiring nothing more than attribution. Funding agencies should be aware of this and funded bodies should agree on legal conditions in advance.

3.19	LRP	<i>Opt for making LRs open for most uses</i>
3.20	PM	<i>Encourage openness for all LRs; enforce openness for publicly-funded LRs</i>

META-SHARE has aligned itself with the open data movement and prepared a Charter<sup>14</sup>, a Declaration for Language Resource Sharing, which takes into consideration these recommendations.

### **References**

- Bentivogli, L., E. Pianta (2005) "Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus". *Natural Language Engineering* 11(3): 247-261
- Callison-Burch, Chris and Dredze, Mark (2010) "Creating Speech and Language Data With Amazon's Mechanical Turk". *Workshop on Creating Speech and Language Data With Mechanical Turk at NAACL-HLT, 2010*.
- Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3).
- Padó, S. and Lapata, M. (2009). Cross-lingual annotation projection of semantic roles. *Journal of Artificial Intelligence Research*. 36(1): 307-340.

<sup>14</sup> <http://www.meta-net.eu/meta-share/charter>



### Ch.3 Strategies

Smith, D. A. and Eisner, J. (2006). Quasi-synchronous grammars: Alignment by soft projection of syntactic dependencies. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 23-30, New York



## Chapter 4 - A Strategic Action Plan for an Interoperability Framework

*Nicoletta Calzolari, Monica Monachini, Valeria Quochi*

Today open, collaborative, shared data is the core of any sound language technology strategy. And standards are fundamental to exchanging, preserving, maintaining and integrating data and language resources, and achieving interoperability in any language resource infrastructure.

The older term “reusability” has today evolved into “interoperability” - the ability of information and communication systems to exchange data and enable the sharing of information and knowledge. Interoperability was declared one of LT’s major priorities at the first FLaReNet Forum in Vienna.

An **Interoperability Framework** can be defined as a dynamic environment of language (and other) standards and guidelines, where the former can interrelate coherently with one another and the latter describe how standards work and “speak” to each other. It is also intended to support the provision of language service interoperability.

Language industry companies need this sort of language strategy because if they cannot interoperate with their clients and suppliers, they would go out of business: “The lack of interoperability costs the translation industry a fortune” (report on a recent TAUS survey (2011b)), referring to the high cost of adjusting file formats and the like, esp. when handing off tasks in workflows.

### 4.1 The Current Standard Framework

In the past two decades, there has been increasing awareness of the need to define common practices and formats for linguistic resources, due to the robustness and industrial-scale utilisation of certain types of NLP technology.

In the 1990s, several projects laid the foundations for the standardisation of resource representation and annotation. Among these the Expert Advisory Group on Language Engineering Standards (EAGLES 1996<sup>1</sup>), within which the Corpus Encoding Standard (CES and XCES, see Ide 1998) was developed, the International Standard for Language Engineering (ISLE<sup>2</sup>, see Calzolari et al. 2002) between EC and NSF, and the SAM project (that led 15 years later to the SpeechDat family, and developed SAMPA, phonetic alphabet used in the ASR community instead of IPA). With these projects, Europe was at the forefront in establishing standards for Language Technology.

Today, standardisation is high on the agenda once again. Consensus has begun to emerge, and in certain areas stable standards have already been defined. However, for many other areas, work is still on-going, either because “the emergence of a solid body of web-based standards has dramatically impacted and re-defined many of our ideas about the ways in which resources will be stored and accessed over the past several years” (Ide & Romary 2007), or because there are emerging technologies, such as multimodal systems, that have specific requirements not covered by existing formats and established practices.

We can therefore observe *a continuum of standardisation* initiatives at various stages of consolidation, together with new proposals for standards as various areas of language technology grow sufficiently mature. While some standards are “official”, i.e. designed and

---

<sup>1</sup> <http://www.ilc.cnr.it/EAGLES96/home.html>


<sup>2</sup> [http://www.ilc.cnr.it/EAGLES96/isle/ISLE\\_Home\\_Page.htm](http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm)



promoted by standardisation bodies such as ISO<sup>3</sup>, W3C<sup>4</sup>, (the now defunct) LISA, and ETSI<sup>5</sup>, others are so-called de-facto standards or “best practices”, i.e. widely-used formats and representation frameworks that have gained community consensus (e.g. WordNet (Fellbaum 1998), PennTreeBank (Marcus et al. 1993), CoNLL (Nivre et al. 2007)).

Drawing on a previous report drafted by the CLARIN project (Bel et al. 2009), the original document has been revised and updated (by FLaReNet together with META-SHARE and ELRA) with standards relevant for the broader LT community, also addressing those that are typically used in industry, at different levels of granularity. “The Standards’ Landscape Towards an Interoperability Framework<sup>6</sup>” (Bel et al., to appear) thus lists both current standards and on-going promising standardisation efforts so that the community can monitor and actively contribute to them. This document is conceived like a “live” document to be adopted and updated by the community (e.g. in future projects and networks), so as not to restart similar efforts over and over. It is meant to be a general reference guide for the whole community and particularly useful for LT initiatives such as the META-SHARE infrastructure, as it provides concrete indications about standards and best practices that are important for given tasks or media in LT. These standards are at different stages of development: some are already very well known and widely used, others more LR-specific standards, especially those developed in the framework of the ISO Technical Committee 37 devoted to LR management, are in the process of development or are being revised.

Currently, we can identify a relatively small set of *basic standards* (defined as “foundational standards”) that have gained wide consensus and allow basic interoperability and exchange. These are not necessarily specific to language resources, but provide a minimum basis for interoperability: e.g. Unicode-UTF8 for character encoding, ISO639 for language codes, W3C-XML for textual data, PCM, MP3, ATRAC, for audio, etc.

On top of these come *standards that specifically address language resource management and representation* that should also be considered as foundational: ISO 24610-1:2006  Feature structure representation, Text Encoding Initiative (TEI), and Lexical Markup Framework (LMF) for lexical resources (Francopoulo et al. 2006, 2008). They are increasingly recognized as fundamental for LR interoperability and exchange.

A further set of *standards focusing on specific aspects of linguistic and terminological representation* are also currently in place and officially established: TMF (ISO 2002) for terminology, SynAF (Declerk, 2006) and MAF (Clément and de la Clérgerie, 2005) for morphological and syntactic annotation.

These standards result from years of work and discussion among groups of experts in various areas of language technology and are thought to be comprehensive enough to allow for the representation of most current annotations. Most of them address syntactic interoperability by providing pivot formats (e.g. LAF/GrAF, Ide and Suderman 2007), while today there is a great need for semantic interoperability, which is still an actual challenge. Most of the more linguistically-oriented standards are also periodically revised to make them even more comprehensive as new technologies appear and new languages are brought into the equation. They still need considerable effort to promote them and spread awareness to a wider community.

*Standards for terminology management and translation technologies* are probably the most widespread and consolidated of all LR standards, largely due to real market pressure from the

---

<sup>3</sup>

[http://www.iso.org/iso/standards\\_development/technical\\_committees/other\\_bodies/iso\\_technical\\_committee.htm?commid=48104](http://www.iso.org/iso/standards_development/technical_committees/other_bodies/iso_technical_committee.htm?commid=48104)

<sup>4</sup> <http://www.w3.org/>

<sup>5</sup> <http://www.etsi.org/>

<sup>6</sup> This document also collects input also from the LRE Map, Multilingual Web, the FLaReNet fora, LREC Workshops, ISO and W3C



translation industry. These include: TMF, the overarching foundational standard for all forms of terminology markup; TMX (Translation Memory eXchange), the vendor-neutral open XML standard for the exchange of Translation Memory (TM) data created by Computer Aided Translation (CAT) and localization tools; TBX, the open XML-based standard for exchanging structured terminological data (that has been approved as an international standard by LISA and ISO) and now taken over by ETSI (after LISA ceased to exist); XLIFF, an OASIS standard for the exchange of data for translation. The most recent initiative is the reference architecture OAXAL (Zydron, 2008), a Open Architecture for XML Authoring and Localization Reference Model, made up of a number of core standards from W3C, OASIS and LISA (from 2011 ETSI).

Finally, there is a stream of *on-going standardisation projects and initiatives* focused mainly on recently mature areas of linguistic analysis and emerging technologies such as semantic annotation (e.g. SemAF ISO24617 - 1-6) which includes temporal (ISO 24617-1:2009, TimeML) and space annotation (ISO-Space<sup>7</sup> ISO PWI 24617-4), emotion, i.e. EML (W3C, 2007) and multimodal annotation, i.e. EMMA (W3C, 2009). The community should monitor these initiatives closely and actively participate in them. It is recommended that researchers, groups and companies involved or interested in these areas actively contribute to the definition of such standards.

In addition to the standards mentioned above, specific communities have established practices that can be considered as *de facto* standards, such as WordNet and PennTreeBank. Their availability, accessibility and openly detailed documentation coupled with the possibility to be implemented without restrictions are the key to promote their usage, speed-up the development of modules for their adoption, thus reducing costs and time. A number of tools exist that facilitate their usage. As these need not change in the near future, it is recommended that mappers/converters are developed from these best practices/common formats to support other endorsed/official standards.

To sum up, *there are a number of standards that jointly provide a potentially useful framework ready for adoption.*

## 4.2 **Barriers and major problems**

There are a number of problems that raise barriers to the broad-based usability of the current standards framework. Let's look at some of these problems/barriers, whose analysis calls for a number of recommendations.

- The key issue that hampers broader usage is not so much a lack of certain standards, but, especially with respect to LT-specific standards, a *lack of (open) tools* for using existing standards easily, and a concomitant lack of tool providers.
- The lack of (ideally open-source) reference implementations and documentation that would help others understand clearly what was done and how.
- The lack of a developer/user education and culture for using standards. There is a strong tendency to use idiosyncratic schemes, which cause format incompatibility (even for minor differences) and prevents annotations to merge or being used together. This in turn prevents easy reuse of available data.
- Within ISO, there are two sets of standards: general interest standards (e.g. country codes) and special interest standards. General interest standards are free, but the others are not, and this should be avoided. The problem is that NLP standards are not considered general interest standards, and therefore have to be paid for, acting as a brake on wider adoption. There are now attempts - e.g. in ISO TC 37 - to overcome this

---

<sup>7</sup> <https://sites.google.com/site/wikiisospace/>



situation by allowing direct application of standards free of charge through implemented databases with free access – e.g. the new ISO 12620 ISOCat (Kemps-Snijders et al. 2009).

- In W3C, full documentation of standards is free so it is easy for W3C standards to be broadly accepted and applied. However, participation in the definition and decision-making process for these standards is costly and as a result, there are almost no SMEs involved. Which means, in turn, that in W3C only the big players can set the rules.
- Standards need to be built by consensus which means that standards creation is a slow process.
- As noted in a recent TAUS survey (TAUS/LISA 2011): “the industry lacks an organizing body or umbrella organization capable of leading the effort and monitoring the compliance”.

4.1	LRP	<i>Standards must be open</i>
-----	-----	-------------------------------

4.2	PM	<i>Need of a body organising, monitoring and coordinating standardisation efforts</i>
-----	----	---

### 4.3 Scenarios for using standards

There are various scenarios that critically involve the need for standards, and they provide strong motivation for standards adoption and investing in developing those standards that are still lacking. These include:

- *Using the same tools on different resources; using different tools on the same data*  
 In architectures for knowledge mining, or creating new resources, where the same data have to be used and enriched by (chains of) different tools (Bosma et al. 2009), standards such as common formats become crucial for their success (see KYOTO<sup>8</sup>).  
 The use of different tools on the same data is relevant for testing and comparing tools, and also in collaborative situations to process the same corpus for different purposes.
- *Creation of workflows – Web Service Interoperability*  
 Standards will ensure operability in cases where workflows are needed that chain together tools that were not originally designed to work in a pipeline. Today workflows can mostly be run with tools that were already designed to work together, or can be enabled by using format converters. Experiences such as TS-STAR and PANACEA<sup>9</sup> (Bel 2010, Toral et al. 2011) show that using a common standardised format will facilitate integration.  
 If tools have been built or modified to work directly on common/standard formats, workflows might be easier to design and quicker to run. Although this is not yet possible, new tools could naturally move in this direction once the advantages have been demonstrated.  
 Workflow management should be generalised to cover both web service interfaces and local processing.
- *Integration/Interlinking of resources, components and services*

<sup>8</sup> [www.kyoto-project.eu](http://www.kyoto-project.eu)

<sup>9</sup> <http://www.panacea-lr.eu/>



This has become an important trend for companies providing services. Interoperability among software components is a major issue today.

When integrating (legacy) resources, there is a need not only for standard formats but also methodologies and best practices for resource management and updating. In the example of Propbank and PennTreeBank in SemLink<sup>10</sup>, changes in one resource cause problems in linking the resources together, resulting in a lot of extra manual work<sup>11</sup>. Data lifecycle issues come into play here.

Interoperability is equally critical when integrating new training data sets, and can facilitate the repurposing of existing data.

– *Documentation and metadata*

At a different level, there is critical need for standard templates in resource documentation. This could help users compare available resources, for example.

Consensus on basic sets of Metadata agreed in the community is also of utmost importance for the easy identification and tracking of resources independently of their physical location. This is becoming particularly critical in new emerging infrastructures. There is currently much interest and action in metadata standardisation, not only in Europe and the USA, but also in Australia.

– *Validation of language resources*

To achieve certified quality validation of LRs, conformity to an accepted standard is considered an important criterion (see ELRA Validation Manuals<sup>12</sup>).

– *Evaluation campaigns: shared tasks*

To compare the results of different methods, approaches, or technologies, data must be encoded and annotated in a common format that different groups can process and use. Here standards clearly play a fundamental role. In fact, many de-facto standards start from evaluation campaigns or shared tasks and then spread through the sub-community in question (e.g. CoNLL<sup>13</sup>).

These initiatives play an important role in encouraging/spreading the use of standards or best practices and as a means for awareness-raising in a wider community.

– *Mashups*

Standards are obviously critical for easily integrating data for *mashup* applications combining data and functionalities.

– *Collaborative creation of resources*

Collaborative ways of creating or updating/enhancing language resources represent a recent trend in the LT community, and standards (should) play an important role in these as they facilitate the sharing of data and especially annotation and editing tools.

– *Preservation*

When standards evolve we need to port resources/tools to new emerging standards. Standards should facilitate this updating process and help avoid mismatches. At the same time, standards make data preservation easier.

---

<sup>10</sup> <http://verbs.colorado.edu/semlink/>

<sup>11</sup> Martha Palmer 2011 Oral Communication at the SILT Workshop on Interoperability, Brandeis 13-14/4/2011.

<sup>12</sup> <http://www.elra.info/Validation-Standards.html>

<sup>13</sup> <http://www.cnts.ua.ac.be/conll/>



## 4.4 Recommendations and next steps

### – Semantic/content interoperability

Semantic annotation offers improved service-oriented interoperability.

So far we have at most achieved syntactic interoperability, i.e. the ability of systems to process shared data either directly or using conversion software. Pivot formats such as GrAF (Ide and Suderman 2007), solve the issue of syntactic interoperability, enabling formats to merge and convert easily.

Today we desperately need semantic interoperability, the ability of systems to interpret shared linguistic information in textual, spoken, multimedia content in meaningful, consistent ways (e.g. with reference to a common set of categories). This is naturally far more difficult as different natural language coding come into play, as well as different theoretical linguistic approaches.

4.3 LRP *Semantic interoperability is needed*

A good methodology of work, already defined in EAGLES, was to define the standard as the lowest common denominator.

4.4 LRP *Define the standard as the lowest common denominator, at the highest level of granularity, a basic principle that was already factored into EAGLES*

### – Linked Data and Open Data

4.5 LRP *Closely monitor the Linked Open Data initiative tightly connected to semantic interoperability*

Interoperability via Linked Data could, for example, be able to link our own objects with corresponding objects in other fields, and achieve both in-domain and extra-domain convergence.

### – Tools that make it possible to use standards

4.6 LRP PM *Encourage building tools that enable the use of standards, and step up the availability of sharable/exchangeable data*

### – Web services platforms

Web service platforms offer an optimal test case for interoperability and could be used to showcase critical needs and advantages. So, we should

4.7 LRP *Use the web service model to provide platforms with NLP modules as web services for various applications*



Projects such as Language Grid, U-Compare, and PANACEA<sup>14</sup> can be understood as an abstract model for platforms providing services based on LT.

These platforms need both syntactic and semantic interoperability. They can be evaluated to see if service architectures are the best way to address interoperability issues.

4.8

LRP  
PM

*Projects results could be provided as web-services. Cloud-based service architectures could also be leveraged as enablers for LT development*

– *Collaborative creation of resources and crowd-sourcing*

Using the collaborative paradigm to create language resources could become the fast lane to standardisation, and at the same time share the cost of resource creation.

Crowd-sourcing with respect to shared resources may be linked to interoperability, as it requires commonly accepted specifications.

Such a collaborative development of resources would ultimately create a new culture of joint research.

4.9

LRP  
PM

*Promote collaborative development of resources also as a help to standardisation*

– *A Large-scale Collaborative Multilingual Annotation Plan*

4.10

LRP

*Design a massive multilingual annotation plan, whereby everyone can deposit annotated data at every possible level of annotation (possibly for parallel/comparable resources), as a specific type of collaborative initiative for resource annotation*

This collaborative approach to creating extremely large amounts of annotated data would help maximise the visibility, reuse and use of resources annotated according to common standards and best practices, while at the same time encouraging more exploration of the diversity of linguistic data. A huge multilingual annotation pool, where everyone can deposit data annotated at every possible different linguistic level for the same resources, or for diverse resources, should be defined as a specific type of collaborative initiative for resource annotation (Calzolari 2010). This could create a fruitful (community-driven) loop linking the most widely-used annotation schemes to best practices. This sort of initiative would also be extremely beneficial for META-SHARE.

– *Evaluation campaigns and validation of resources*

---

<sup>14</sup> <http://www.panacea-lr.eu/>



As mentioned in section 4.3, evaluation campaigns help the standards agenda. But the lack of any European evaluation body to coordinate and supervise common format definitions remains a genuine problem (see also Chapter 5).

Shared tasks and shared data largely influence common data formats. This makes shared tasks a key site for interoperability and standards actions for both resources and components.

4.11	PM	<i>Need of a European evaluation and validation body</i>
------	----	--

4.12	LRP	<i>Create 'official' validators (such as <a href="http://validator.oaipmh.com/">http://validator.oaipmh.com/</a> or the OLAC validator) to check compliance of LRs with basic linguistic standards</i>
------	-----	--

This could help provide validation services for resources to be showcased via META-SHARE.

#### – Interoperability Challenge

The idea of an “Interoperability Challenge” was raised by Nancy Ide and James Pustejovsky at a SILT Workshop in April 2011 and could become an international initiative, using the evaluation campaign model, but aimed specifically at organising and supporting shared tasks to speed up the dissemination of standards and drive forward interoperability<sup>15</sup>.

The NLP community as a whole should be involved in this shift to interoperability by building a general challenge involving tasks that explicitly require multiple data formats, annotation schemes, and processing modules, so that participants would be highly motivated to adapt, adopt, and use standards and common formats and reach an understanding of the advantages they offer.

4.13	LRP	<i>Set up an “interoperability challenge” as a collective exercise to evaluate (and possibly measure) interoperability</i>
------	-----	--

#### – Repository of standards and best practices

A small preparatory initiative was started within FLaReNet<sup>16</sup>, but this requires a dedicated effort and must be organised as a collaborative action, so that it acts as a culture builder. A repository of standards could obviously be linked to a repository of standards-compliant open/freely available data to maximise its benefits.

4.14	LRP	<i>Collaboratively build and maintain a repository of standards and best practices, linked to standards-compliant open/freely available data</i>
------	-----	--

#### – Awareness initiatives

<sup>15</sup> see <https://sites.google.com/site/siltforum/files>

<sup>16</sup> [http://www.flarenet.eu/?q=Standards\\_and\\_Best\\_Practices](http://www.flarenet.eu/?q=Standards_and_Best_Practices)



4.15 PM *Launch awareness programs*

Awareness about the existing standards and the motivations behind them is one of the key factors for enabling their adoption. Educational programs should therefore be launched to explain, promote and disseminate standards especially to students and young researchers (e.g. through tutorials at conferences, summer schools, seminars...). Steps could be taken to include standardisation in regular university curricula. Also, effective ways to demonstrate the return of investment (ROI) of interoperability must be sought. Adapting one's tools and resources to standardised common formats in fact requires some investments that players may not be willing to make unless the clearly see advantages.

– *Education and training initiatives*

4.16 LRP  
PM *Set up training initiatives to promote and disseminate standards*

4.17 LRP  
PM *Standardisation as part of university curricula*

– *Standards Watch*

4.18 LRP  
PM *Create a permanent Standards Observatory or Standards Watch*

No mechanism is currently available to watch when a topic/area deserves standardisation. On the European side there is for example a lack of official standardisation initiatives for such relevant topics as space annotation and the representation of Lexicon-Ontology relations. These have an economic potential and the EC is not present enough in standardisation initiatives around these areas.

– *Quality Certificate*

The Data Seal of Approval<sup>17</sup> is a label for resource quality intended to provide a certification for data repositories to keep data visible and accessible and to ensure long term preservation. For example, CLARIN centres are expected to comply with certain standards. This is linked to the concept of LR preservation and sustainability (see Chapter 2).

Infrastructures such as META-SHARE could introduce some mechanism of assigning quality scores to resources and tools and evaluating them for compliance with standards/best practices, and making these evaluations public. "Penalty" systems could also be devised.

---

<sup>17</sup> <http://www.datasealofapproval.org>



4.19	LRP	<i>Define and establish a Quality Certificate or quality score for LRs, to be endorsed by the community</i>
------	-----	---

– *Multilingual web content-related standards*

4.20	LRP	<i>Link-up/collaborate/be in line with ISO, W3C, LISA and other multilingual web content-related standards, which in the case of LT can be seen as more basic levels of representation, to ensure the (potential) integration of LRT into current and future web content products</i>
------	-----	---

Multilinguality should be incorporated into standards, e.g. ISO standards should be instantiated and generalised for as many languages as possible, although this is not always the case today.

A recommendation should be made to standardisation bodies to:

4.21	PM	<i>Allow for testing/applying standards on a multilingual basis</i>
------	----	---

4.22	PM	<i>Standardisation initiatives must be conducted at an international level</i>
------	----	--

## 4.5 Towards an Interoperability Framework

A recurrent request from industrial players at recent meetings such as the META-NET Vision Groups and the META-Council, is: “*give me the standards and give me open data*”.

Standards must not only be usable; they must also be used, otherwise they serve no purpose. An essential step in promoting such use is to make LR standards operational, i.e. coming up with “Operational Definitions of Interoperability” in the form of concrete proposals.

One step towards this is to ensure that standards are ‘open’. The minimum requirements for open standards are availability and accessibility for all, detailed documentation and the potential to be implemented without restrictions. Publicly available standards with public specifications are vital for easy adoption (Perens, 2010; Krechmer, 2006).

4.23	LRP	<i>Ensure that all standards are fully operational</i>
------	-----	--

This is the key recommendation for an interoperability framework.

We need to outline the basic pre-conditions and the essential steps to be taken, some of which have already been mentioned above.

### 4.5.1 Technical conditions

– *Common metadata*

This is a widely recognised pre-condition in all major infrastructure initiatives: ELRA, LDC, CLARIN, META-SHARE, and of course FLaReNet.

– *Explicit semantics of annotation metadata and semantic interoperability*



It is essential to have an explicit semantics of annotation metadata or data categories. One mechanism for this is ISOCat, which is currently the only available instrument, even if there are still many problems with it. Categories can be defined **at** a persistent web location that can be referenced from any scheme that uses them.

High-level metadata is not the only set of values that are recorded in ISOCat. So far, the linguistic categories within ISOCat have been mostly taken from the EAGLES, MULTEXT-East and LIRICS projects (e.g. morpho-syntax, also extended to Semitic, Asian and African languages), and in the case of terminology from LISA and ISO-12620 value sets. Recently, ISOCat has been enriched by the CLARIN project in the field of Social Sciences and Humanities, but these metadata are not precise enough for full-blooded NLP.

This gap between a specific domain and general NLP is currently being addressed by META-SHARE and it will require the efforts of a broad community of resource developers/users. To increase convergence towards common data categories, it would be advisable to create “data category selections” for major standards/best practices. This would be one step towards semantic interoperability.

Another step would be to get funding agencies to encourage the entering of data categories and selections in ISOCat, which will only become useful when it is used on a broad front.

– *Tools that help people to use standards*

It is extremely important to develop (online) tools that mask the complexities of standard formats and allow standards to be used easily and support easy exportation/mapping of data to standards. We recommend the development of mappers/converters from/to the major standards/best practices/common formats to the other endorsed/official standards and/or to major proprietary formats.

This is true in particular for infrastructures like META-SHARE where best practices should also be promoted through tools.

#### 4.5.2 Infrastructural conditions

– *A common (virtual) repository for finding the most appropriate standards easily*

A joint international effort should organize the indexing of different standards and best practices, make it easy to find them and keep track of their status, versions and history. This is critical for an infrastructure such as META-SHARE whose job should be to recommend standards and best practices for the resources they showcase, especially the latest examples (the new ones).

– *Common documentation templates*

Documentation is often inadequate – either too little or too much. This sort of documentation is essential for a common understanding of standards.

A consensus-driven set of templates for resource documentation should be devised and disseminated, with actions to facilitate adoption.

– *Provide a framework that facilitates testing*

We need test scenarios to verify compliance to standards.

– *An interoperability framework for/of web services*

Making standards operational also means grounding them on an interoperability framework for/of web services. This means providing standards-compliant linguistic services, particularly in workflows. In this respect the KYOTO project is both a case study and a success story.

– *“Meta-interoperability” among standards*



Standards must themselves form a coherent framework, and speak with each other within an LR-specific ecology.

4.24

LRP

*Ensure “meta-interoperability” among standards, that must form a coherent framework, in a LR-ecology*

### 4.5.3 Social and cultural conditions

The social and/or cultural conditions on interoperability are just as important as the technical conditions.

#### – *Community involvement*

The community as a whole must be involved in standardisation processes. We recommend that researchers, agencies and companies involved or interested in resource development/annotation actively help in defining LT standards.

There is a need for an effort to change the old-style “community” mind-set into a new “social network” mind-set of collaboration to achieve real interoperability. The wider the participation in such initiatives, the more robust and valid the standards will be.

#### – *Dissemination but not policing*

We need to disseminate standardisation, push standards, and invent incentives for using standards. But people should not feel obliged to conform. Standards must not be seen as another overhead; people should use them because it is in their interest.

#### – *Interoperability as a valid research area*

Interoperability and standardisation issues should become academically acceptable research areas.

#### – *Link to sustainability*

In general, a virtuous circle must be established between standards, use, feedback, and interoperability.

## References

- Bel, N. et al. (2009). *CLARIN Standardisation Action Plan*. CLARIN <http://www.clarin.eu/node/2841>
- Calzolari, N. et al. (2011). *The Standards' Landscape Towards an Interoperability Framework*. FLaReNet Report.
- Bel, N. (2010). “Platform for Automatic Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies: PANACEA”. In *Proceedings of the 26th Annual Congress of the Spanish Society for Natural Language Processing (SEPLN)*, Valencia.
- Bosma, W., et al. (2009). KAF: a generic semantic an-notation format. In: *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*. Pisa.
- Calzolari, N., et al. (2002). “Broadening the scope of the EAGLES/ISLE lexical standardization initiative”. In *Proceedings of the 3rd workshop on Asian language resources and international standardization (COLING '02)*, vol. 12. pages 1-8, Taipei.
- Calzolari N., et al., eds. (2009). *Shaping the Future of the Multilingual Digital Europe*, 1st FLaReNet Forum, Vienna.



- Calzolari, N. (2010). Invited presentation at the COLING 2010 Panel. *23rd International Conference on Computational Linguistics (COLING 2010)*. 23-27 August, Beijing, China.
- Declerck, T. (2006). "SynAF: Towards a Standard for Syntactic Annotation". In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. Pp. 229-232, Genoa.
- EAGLES 1996. <http://www.ilc.cnr.it/EAGLES96/home.html>
- Fellbaum C., ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Francopoulo, G. et al. (2006). Lexical markup frame-work (LMF). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. pages 233-236, Genoa.
- Francopoulo G. et al. (2009). Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*. 43(1): 57-70.
- Ide, N. (1998). "Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora." In *Proceedings of the First International Language Resources and Evaluation Conference (LREC'98)*. pages 463-470, Granada.
- Ide, N and L. Romary (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsén, H., Minker, W. (eds.), *Evaluation of Text and Speech Systems*, Springer, Dordrecht, 263-84.
- Ide, N. and K. Suderman. (2007). "GrAF: A graph-based format for linguistic annotations". In *Proceedings of the Linguistic Annotation Workshop at ACL 2007*, pp. 1-8, Prague.
- International Organization for Standardisation. 2002. ISO:16642-2002. Terminological Markup Frame-work. <http://www.loria.fr/projets/TMF/>
- International Organization for Standardization (2008). ISO DIS 24611 Language Resource Management - Morpho-syntactic Annotation Framework (MAF). ISO/TC 37/SC4/WG 2.
- International Organization for Standardization (2008). ISO DIS 24611- (1,2,3,4,5,6) Language Resource Management - Semantic annotation framework (SemAF). ISO/TC 37/SC4/WG 2.
- Kemps-Snijders M., et al (2009). "ISOcat: Remodeling Metadata for Language Resources". *International Journal of Metadata, Semantics and Ontologies (IJMSO)*, 4(4): 261-276.
- Krechmer K. 2006, Open Standards Requirements. *The International Journal of IT Standards and Standardization Research*, 4(1): 43-61.
- Lionel C. and Éric de la Clergerie (2005). Maf: a morphosyntactic annotation frame work. In *Proceedings of the 2nd Language and Technology Conference (LTC'05)*, pages 90-94, Poznan.
- Marcus, M. P., B. Santorini, M.A. Marcinkiewicz (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19 (2): 313-330.
- Nivre, J. et al. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. pages 915-932, Prague.
- TAUS (2011a) Annual Plan 2011: Translation Innovation Think Tank Interoperability Watchdog. <http://www.translationautomation.com/images/stories/pdf/taus-annual-plan-2011-extended.pdf>
- TAUS (2011b) Report on a TAUS research about translation interoperability. February 25, 2011. <http://www.translationautomation.com>
- Toral A., et al. (2011). "Towards a user-friendly web-service architecture for statistical machine translation in the PANACEA project". In: M. L. Forcada, H. Depraetere, V. Vandeghinste (eds.) *Proceedings of the 15th EAMT 2011*, pages 63-70, Leuven.
- W3C (2007) EML: Emotion Incubator Group, W3C Incubator Group Report, 10 July 2007.
- W3C (2009) EMMA: Extensible MultiModal Annotation markup language, W3C Recommendation, 10 February 2009.



#### Ch.4 Interoperability Framework

Zydron A. (2008). OAXAL. What Is It and Why Should I Care? Globalization Insider.  
<http://www.lisa.org/globalizationinsider/2008>.

## Chapter 5 - The Evaluation and Validation of Resources

Jan Odijk

This chapter looks at issues of Language Resource (LR) quality in terms of *validation* and *evaluation*, and describes and clarifies recommendations for these activities.

**Validation** of an LR means checking whether the LR is compliant with its specification and/or documentation, or, put otherwise, whether "you built the LR right". Validation determines resource quality (see section 1.6.2). Validation can be *formal* ("is the form of the LR compatible with its specification/documentation?") and/or related to *content* ("is the content of the LR compatible with its specification/documentation?"). In principle, validation can be applied to both data and tools/technologies/components. In practice, however, validation has been mainly applied to data, while evaluation has been mainly applied to tools and technologies.

**Evaluation** of an LR (broadly construed as including technologies and applications) is checking whether the LR is suited for a specific task, or, put otherwise, whether "you built the right LR". Evaluation determines resource adequacy (see section 1.6.3 **Errore. L'origine riferimento non è stata trovata.**). Though evaluation can in principle be applied both to data and to technologies and tools, in actual practice evaluation is mostly applied to technologies and applications only.

### 5.1 Recommendations for Validation

Validation of an LR during the production process (*internal validation*) is universally considered to be essential for ensuring the highest quality of the LR.

Validation of LRs by external parties after the production of the LR (*external validation*) is generally (though not universally) considered to be important.

Validation of LRs is possible only if an appropriate specification or documentation of the LR is available. The quality of LR specifications and documentations tend to vary greatly, and even for high quality specifications and documentation, the form can vary greatly.

A good approach to achieve actual use of such requirements and to achieve homogeneity in form is to use templates for LR specifications and documentation. Defining such templates for LRs in general is perhaps possible, but not sufficiently specific. So templates should be defined for specifications and documentation of specific LR types (see REC 1.1 and 1.3)

The experiences in the SPEECHDAT family of LR production projects, in which such templates were defined for certain types of speech resources and in which these templates have actually been used, proves that such an approach is feasible and yields high quality and highly uniform specifications and documentation.

Though almost everybody agrees that validation is important, it requires effort and costs. It will therefore only be done if it is planned and budgeted for in the project plan for the creation of a resource:

5.1

LRP

*Plan and budget internal and external validation in the LR production plan*

It is also generally agreed that the effort and costs to carry out validation of LRs must be minimized.



5.2 LRP *Minimize the effort and cost of LR validation while maintaining high quality*

Even with minimized effort and costs, some believe that external validation is not really affordable. It is even sometimes thought that you don't need to validate a resource at all: if a successful technology can be created on the basis of such a resource, then the resource must be of a sufficiently high quality, it is claimed. But that is not so clear. In actual practice, all technology developers have to make adaptations or improvements to resources, or ensure that errors in resources do not disrupt the technology development, if such resources do not satisfy a minimum of requirements checked during the validation process. Such effort and costs are hidden in the research or development costs and they are repeated over and over in every research or development centre. Validation of resources therefore contributes to minimizing effort and cost in research and development, and might very well be overall more cost-effective than no validation.

### ***Interoperability***

Several methods for minimizing validation effort and costs are imaginable. Some have been proposed and should be promoted: An important method is by ensuring interoperability of the LR with other LRs:

5.3 LRP *Ensure formal and semantic interoperability of LRs (see Ch. 4)*

With interoperable LRs, large parts of what usually fall under validation (cf. e.g. the ELRA Validation manuals<sup>1</sup>) can be replaced by fully automatic interoperability tests:

5.4 LRP *Replace (aspects of) validation by fully automatic interoperability tests.  
Develop test scenarios for optimally using automatic interoperability tests*

If a tool or dataset can be used directly –without any ad-hoc adaptations -- in a pipeline with other tools and data, this mere fact is an indicator for its quality in certain respects. For example, if the output of a newly-developed PoS tagger can be used directly as input to an existing syntactic parser, this is an indicator for the correctness of the PoS-tagger output format and interpretation (incl. the PoS-tags used). By using a test scenario where a given resource is combined with a range of other resources, each with its own functionality, multiple aspects of the quality of the resource are tested. This can be done not only for the resources themselves, but also for their metadata: one can use tools to check compatibility with metadata harvesting schemes<sup>2</sup> but one can also use metadata harvesting itself as a test.

To achieve interoperability, several conditions have to be met (see also Ch. 3 and 4), for which recommendations are given here:

5.5 LRP *Make the LR visible and accessible via a data and tools infrastructure or exchange facility*

<sup>1</sup> <http://www.elra.info/Validation-Standards.html>

<sup>2</sup> <http://validator.oaipmh.com/>



5.6	LRP	<i>Use widely accepted de facto or official standards and best practices for data formats that are supported by the infrastructure and exchange facility in which the LR should function</i>
5.7	LRP	<i>Provide metadata for LRs in accordance with widely accepted standards, e.g. Athens Core<sup>3</sup> in the context of the Component-based Metadata Infrastructure (CMDI)<sup>4</sup></i>
5.8	LRP	<i>Use formalized metadata elements as much as possible (i.e. metadata elements with a restricted set of possible values, specific types, etc.) and use the String type for metadata element values only when it really cannot be avoided</i>
5.9	LRP	<i>Put all aspects of the documentation of the LR that can be formalized in formalized metadata elements</i>
5.10	LRP	<i>Ensure that all data categories used in the metadata and in the data are registered in a data category registry to ensure semantic interoperability</i>

### ***Automating validation***

Several tools, often derived from earlier technology developments, can be used to automate the detection of errors or likely errors in LRs.

5.11	LRP	<i>Develop new and/or ensure the maximal use of existing tools for automatic or semi-automatic validation of LRs</i>
------	-----	--

In the field of spoken LRs, there are tools for fault detection (clipping, noise...), detecting segmentation errors, of weak annotations, providing confidence measures of speech transcriptions, etc.

For written resources there are tools to create simple frequency tables, pivot tables etc. for attributes in the LR and their values, tests of the LRs against lists of permissible values, aggregation functions and clustering functions that are very useful for detecting real or likely errors.

We need to validate LRs both formally but also in terms of their content. Content validation is less easy to automate and therefore it is less easy to minimize effort and cost. This should be a continuous area of research focus.

5.12	PM	<i>Promote the development of methods for maximally automating content validation</i>
------	----	---

Tools such as the ones mentioned in connection with Rec. 5.11 can be used for formal validation and also partially for content validation.

<sup>3</sup> <http://www.isocat.org/rest/group/3733>

<sup>4</sup> <http://www.clarin.eu/cmdl>



These tools are essential for creating large-volume resources that are also of high quality, and therefore require continuous attention.

5.13	PM	<i>Promote the use of tools that can contribute to ensuring quality in the (automatic) production of large scale resources</i>
------	----	--

### ***A Collective Effort***

Work on validation, its methodologies and supporting tools should be done in the whole community, and results should be shared to accelerate progress. It is essential that this happens not only in the academia, but also in the industry:

5.14	PM	<i>Support academic and industry involvement in research on automatic methods for validation of LRs</i>
------	----	---

Many aspects of LR validation and LR quality require constant adaptation to new developments (e.g. support for standards, recommendations on metadata issues, recommendations on data category issues, etc.). To this end, think tanks should be set up.

5.15	LRP	<i>Create a think tank with recognized experts from a broad spectrum of the community (academia/industry; technologies; and modalities (written/speech/multimodal), etc.) to assess requirements for LR quality</i>
------	-----	---

5.16	LRP	<i>Enable users to provide comments and votes, both for resources/tools and standards</i>
------	-----	---

Past projects, especially the SPEECHDAT family of resource-creation projects, have taught us that the cooperative creation of specific LRs (e.g. in a consortium where each partner makes an LR for one language) with clearly specified and agreed requirements and specifications and resource exchange among partners, is an excellent way to obtain large, high quality LRs for a wide range of languages as efficiently as possible.

5.17	LRP	<i>Create LRs in collaborative projects where resources are exchanged among project participants after production</i>
------	-----	---

When set up properly, collaborative projects like this one will very often arrive at de facto standards for data formats that are supported by tools, and are accepted and endorsed by both academia and industry. Setting up such collaborative projects should be easier nowadays than a decade and a half ago due to easier interactivity over the internet.

Certain aspects of validation cannot be automated because they require human knowledge. Some of these 'Human Intelligence Tasks' could be carried out by crowd sourcing, which is fairly easy to organize nowadays. This is used already for resource production, and could equally well be used for resource validation. The task must be made attractive to users, for example by



presenting it as a game (cf. *Google Image Labeler*<sup>5</sup>, or, in a language context, *Phrase Detectives*<sup>6</sup>). Or it could be made part of a task that has to be done in the context of security type interactions such as *ReCaptcha*<sup>7</sup>. A further possibility is to use a service such as the *Amazon Mechanical Turk*<sup>8</sup>. However, there are ethical and legal issues with such services, which require careful consideration before making use of them.

5.18

LRP  
PM

*Use crowd sourcing techniques to carry out validation tasks, but be aware of the ethical and legal issues associated with these techniques*

## 5.2 Recommendations for Evaluation

There is wide consensus that research on LT technologies has largely been driven by systematic evaluation during the last two decades. Such evaluation:

- Enables researchers to objectively compare approaches and reproduce experiments
- Helps researchers make issues explicit, validate new ideas, and identify missing science
- Provides an important tool for judging funding efficiency and determining the maturity of developments for a given application.

It is generally agreed that evaluation should remain an important driving force for this research in the coming years.

5.19

PM

*Promote evaluation as a driving force for research in LT*

Almost every research paper is expected to have an evaluation component, of course, but even so, evaluation is still a fragmented process. To carry out evaluation systematically, research should (in part) be organized around evaluation tasks and challenges.

5.20

PM

*Organize (a significant part of) research around evaluation tasks, challenges and campaigns*

As has become clear in recent years, organizing research in this way has additional benefits:

- It creates focus and mass
- It enables the community to create larger, more representative and better quality evaluation data and tools
- It helps define, promote and use standards and best practices
- Evaluation campaigns have created awareness of the fact that each HLT system has its own strengths and weaknesses, and that by combining multiple HLT systems using a voting system, a new one can be created that outperforms each of the input HLT systems.

<sup>5</sup> <http://images.google.com/imagelabeler/> based on Luis von Ahn's ESP Game ([http://en.wikipedia.org/wiki/ESP\\_Game](http://en.wikipedia.org/wiki/ESP_Game))

<sup>6</sup> <http://www.phrasedetectives.org>

<sup>7</sup> <http://www.google.com/recaptcha>

<sup>8</sup> <https://www.mturk.com/mturk/welcome>



Such a combined system can then be used to create higher quality annotations than before.

But many researchers believe that this is not yet enough, and recommend the following:

5.21

LRP

*Set up an evaluation management and coordination structure to ensure a viable, consistent and coherent programme of activities that can successfully scale up and embrace new communities and technological paradigms*

This structure should, among other things,

- Develop a short term (1-year) and midterm (5-year) action plan to take responsibility for LT evaluation in Europe.
- Propose solutions and actions to include both task- and application-oriented evaluation, e.g.
  - Set-up a general evaluation framework, including both kinds of evaluation, (comparable to the ISLE Framework for Evaluation in Machine Translation (FEMTI) approach)
  - Set-up an integrated evaluation
- Identify key areas for evaluation and relevant evaluation packages.
- Stimulate work on shared, standard evaluation procedures. Evaluation should include an assessment of the practical impact of LRT on real NLP applications
- Dedicate special attention to less-resourced languages, and investigate strategies and funding to include them. In a concerted effort on multiple languages, less-resourced languages can often “piggy back” on better served languages.
- Dedicate special attention to evaluating technology used for resource production and acquisition:
  - Systematic evaluation of automatic techniques for LR production should be promoted to assess their strengths and weaknesses and stimulate more research in these fields
  - Appropriate evaluation frameworks for automatic acquisition methods must be developed to test both current and newly discovered methods.

Since the evaluation of a technology or component under development needs to be carried out frequently, it is essential that this process is automated as far as possible.

5.22

PM

*Promote research into and the development of techniques for fully automatic evaluation*

This very often requires fully formalized evaluation metrics, of which there are many. Some of these, however, are still very controversial (especially in the area of machine translation) or considered to be of limited applicability.

5.23

PM

*Promote research into and the development of fully formalized evaluation metrics*

Many researchers feel that there should be an independent technical infrastructure for evaluation.



5.24

LRP

*Set up a sustainable technical infrastructure providing data, tools and services to carry out systematic evaluation. This could be a distributed infrastructure involving existing organizations*

This infrastructure should provide

- Ubiquitous remote evaluation
- Evaluation for single components, and also of their contribution to a more global technology or application.
- Evaluation packages, including tools (e.g. for computing evaluation metrics) and evaluation data
- Visibility, accessibility and the long term preservation of the components, data and evaluation packages it hosts. Certification of each participating organization in the infrastructure for these via commonly agreed upon guidelines and protocols (e.g. via the *Data Seal of Approval*<sup>9</sup>) is necessary.

For this sort of infrastructure to be successful, the individual components must be interoperable, as must the components and the data. It would therefore be necessary to

- Define clear interfaces between individual components, e.g., by shared communication protocols, and
- Ensure agreement on such interfaces within the research community
- Stimulate the development of conversion tools to overcome differences between interfaces.

This infrastructure could be set up as a web site using web services and work flow systems. It would be a natural move to use the META-SHARE exchange facility currently being developed in the META-NET project as a basis for such an evaluation infrastructure.

This kind of infrastructure could bring additional benefits. The practical need for interfaces between LT technologies, data, and infrastructure components will directly help introduce de facto interoperability outside the domain of evaluation itself. This infrastructure should stimulate open access to shareable components, ideally cost-free based on the mutual sharing and exchange of components.

Sharing components and data has many advantages:

- It avoids duplication,
- It allows research groups to focus on their specialty
- It creates de facto standards and interoperability
- It enables various approaches to be compared easily
- It enables evaluation on both systems and on their component parts.

This will only work if there are clear agreements that will safeguard the interests of commercial parties.

5.25

LRP

*Develop clear agreement templates for sharing components and data, especially to ensure optimal cooperation from commercial parties while safeguarding their interests*

<sup>9</sup> <http://www.datasealofapproval.org/>



The agreement models used by TAUS for Translation Memories<sup>10</sup>, by ECESS in the context of speech synthesis<sup>11</sup>, and by WebForditas<sup>12</sup> for machine translation can provide inspiration.

This kind of management and coordination structure, including a technical infrastructure based on META-SHARE, will naturally require its own funding, as was pointed out by several researchers from the community:

5.26

PM

*Provide funds for financing the evaluation management and coordination structure and its associated technical infrastructure. This may require new funding strategies and instruments both at the European and at national and regional levels*

As mentioned in Chapter 2.8, the level of technology performance needed by an application may vary depending on the application. It is important to make application and technology developers aware of this fact, and also to better define the required level of performance for a given application:

5.27

PM

*Make technology and application developers aware of the fact that the required performance level of a technology depends on the application it is used in  
Bring technology and application developers together to define the performance levels required for common applications*

---

<sup>10</sup> <http://www.tausdata.org/>

<sup>11</sup> <http://www.ecess.eu/>

<sup>12</sup> <http://www.webforditas.hu/?show=textTab&lang=english>





## Chapter 6 - Strategic Directions for Automating LR Development

*Núria Bel, Valeria Quochi*

Automatic LR development means automatically compiling resources, for instance by crawling parallel corpora, achieving grammar induction or the induction of a “polarity” lexicon (listing items that represent contrasting positions in human discourse such as “this is good/this is bad”), or by automatically annotating resources so they can be used for new tasks. Examples would be part of speech tagging for language modelling or annotations for “time” in texts to train a system to track events.

As mentioned in Ch. 5, the automation of LR development can provide solutions to the high cost of manually building and annotating new resources, and therefore to the high investments faced by commercial NLP systems in Europe when localizing their products. At the same time, it can make advancement in research and technology development possible and quicker. The high costs for producing language resource data, which in turn are often critical for the development of good multilingual technology, seems to be the major factor preventing the full deployment of NLP technologies, especially in geographic areas characterized by a rich variety of official languages such as Europe. Furthermore, it is equally expensive to adapt both the content and the systems that manage it to new applications and communication media such as, for instance, SMS and Tweets.

A clear example of the effects of LR shortage is related to applications for Sentiment Analysis/Opinion Mining, which have recently expanded massively due to the availability of content derived from social networks and web reviews in a variety of languages. The aim is to check whether people have a positive or negative view of some issue, product or related item and track trends for a given product or service. The preliminary work goes back to the early 1980s (e.g Jaime Carbonell’s PhD thesis 1979), but real commercial interest surged in the 2000s. Some of the technologies are based on document classification techniques and reach accuracy of 80%. However, it has been proved that employing techniques that use specific language resources (such as polarity lexica tuned for the specific domains of application) can raise this performance to 91.33%. (Liu et al. 2005).

The sector of text-analytics for some time was highly affected by this new application to the point that some companies focused on that area only (e.g., to mention some: Cymfony<sup>1</sup>, Attensity<sup>2</sup>, Motivequest<sup>3</sup>, Sentiment Metrics<sup>4</sup>, Lithium<sup>5</sup>, QuarkRank<sup>6</sup>). Most of the products, however, mainly concentrate on the processing of opinion expressed in English, and sentiment analysis resources are still not available across the language spectrum. Once again, there is on-going research into automatic solutions to create these resources such as information induction and the repurposing of existing resources. But there seems to be little overall interest in leveraging this work over the language spread (in fact, in the time span covered by our survey, only 4 papers describe the application of such techniques and again not for languages other than English). This is just a case study that supports the recommendation of promoting and

---

<sup>1</sup> [www.cymfony.com](http://www.cymfony.com)

<sup>2</sup> [www.attensity.com](http://www.attensity.com)

<sup>3</sup> [www.motivequest.com](http://www.motivequest.com)

<sup>4</sup> [www.sentimentmetrics.com](http://www.sentimentmetrics.com)

<sup>5</sup> [www.lithium.com](http://www.lithium.com)

<sup>6</sup> [www.quarkrank.com](http://www.quarkrank.com)



supporting research and experimentation on the production of resources so that potentially killer applications may approach and satisfy market needs.

A survey was carried out in FLaReNet (see D6.1a and b) to assess the state of the art and the potential for automation of language data resource production, and to express recommendations. Although the automatic compilation of speech and multimodal resources is equally important, they were virtually absent from our survey, except in the form of transcribed speech, which ultimately boils down to written text<sup>7</sup>.

## 6.1 *Surveying the state of the art*

FLaReNet surveyed the demand for language data to identify whether automatic methods were being researched and developed to satisfy this demand, and for which applications or purposes (D6.1a and b).

The first conclusion is that Machine Translation (MT) has been the most demanding area for LRs in recent years. Though, handcrafted grammars, monolingual and bilingual lexicons have of course been in continuous demand for commercial systems for some time; the progressive success of Statistical Machine Translation (SMT) has led to a strong demand firstly for a large quantity of raw text (i.e. large monolingual and parallel corpora) and then for a growing volume of annotated data, such as POS tagged texts, identified named entities, syntactically annotated data, and similar resources.

While some languages are well-provided for with both raw and annotated data (and/or their concomitant tools), making it easy to launch SMT systems (as demonstrated for example in Euromatrix<sup>8</sup>), there is an evident lack of data for the entire span of languages and semantic domains, hence, the research interest and recent efforts in trying to automate resource production as a means to reduce production costs, i.e. tools that can produce such resources without (or with the least possible) human intervention, the most important cost factor.

SMT is also an example of the community response to data shortage problems. The unavailability of parallel corpora (texts in different languages which are a translation of the other for supplying samples to train the SMT system), in particular translation among all of the European languages, has fostered research into automatic comparable corpora compilation. Comparable corpora are a collection of texts in different languages, of similar characteristics, but not translations of one to the other. Current research looks for measures that automatically measure the degree of comparability. Thus, SMT can be considered an advanced case study that can help to predict the future needs of other technologies and applications. As in this specific case of comparable corpora, industry players will probably pay more attention to technologies for the automatic production of other resources when addressing smaller markets (language and domain of use), as accurately hand-crafted LRs are simply too expensive.

In order to paint the current picture of automatic production of language resources, which is still mainly a research issue, FLaReNet systematically surveyed papers on the production/acquisition of Language Resources accepted at various LT and Computational Linguistics conferences worldwide between 2006 and 2010. The conferences surveyed have been selected as representative among those considered most relevant and prestigious in the domain. These sources include the European Chapter of the Association for Computational Linguistics (EACL), North American Chapter of the Association for Computational Linguistics (NAACL), Association for Computational Linguistics (ACL), International Conference on Computational Linguistics (COLING) and Language Resources and Evaluation Conference

<sup>7</sup> To a certain extent this is because topics such as automatic speech or video segmentation are normally dealt with in more physical than symbolic terms.

<sup>8</sup> <http://www.euromatrixplus.net/>

(LREC). A small number of papers in journals<sup>9</sup> were also reviewed, but we only draw on them for qualitative analysis.

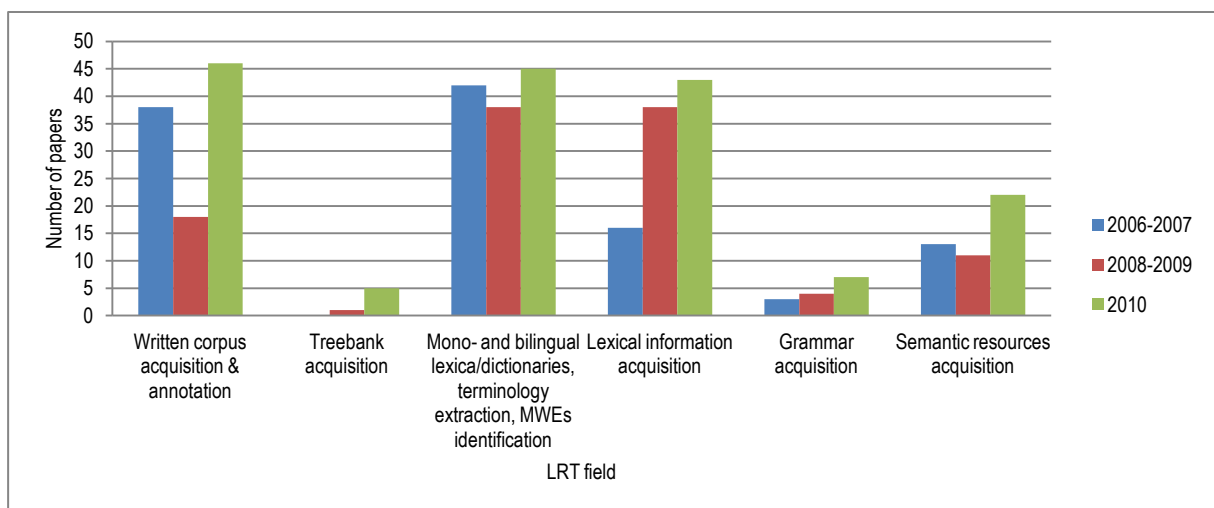
The aim of this survey was to assess whether:

- there is a critical mass of research in automatic LR production that should be supported and/or transferred to industry
- research on automatic LR development meets the demand for on-going developments in applications
- it is possible to pinpoint where the (most promising) results are taking place.

The community has also been consulted, as in the FLaReNet Vienna Forum, where there was a special session on the production of resources. The conclusions of presentations and discussions were also examined. The results of a poll at the FLaReNet Barcelona Forum were also taken into account. Finally, we drew on the presentations and discussions at the *Workshop on Methods for the automatic acquisition of Language Resources and their evaluation methods*, held as an LREC workshop, as well as from direct experience of some of the FLaReNet partners with the PANACEA project<sup>10</sup>.

The survey shows that the most frequent (textual/written) areas of research into the production of resources are the following (see Table 1 and D6.1 for more details):

- Mono- & bilingual lexica/dictionaries, terminology extraction, Multi Word Unit identification, Named Entity Recognition
- Written corpus acquisition & annotation
- Lexical information: e.g. subcategorization frames, selectional preferences, and lexical features like countability, gender, semantic roles...
- Semantic resources (WordNets, ontologies, polarity and sentiment lexica)
- Grammar induction
- Treebank production



**Table 1: Overview papers on automatic acquisition of LRs 2006-2010. Distribution per LRT field.**

<sup>9</sup> *Natural Language Engineering* (Cambridge University), *Computer Speech and Language* (Elsevier Publications), *Journal of Language Resources and Evaluation* (Springer Publications), *Machine Translation* (Springer Publications), *Computational Linguistics* (MIT Press Journals).

<sup>10</sup> [www.panacea-lr.eu](http://www.panacea-lr.eu)



Although there are certainly other areas where resource automation can be beneficial, the survey reports on what the community has selected (through publications at renowned conferences) as being a convincing profile of the field, and which are likely to become promising areas for the development of robust automatic resource creation technology in the near and mid-term future.

## 6.2 *Strategic Directions and Recommendations*

In this section, drawing from the outcomes of the survey and of related FLaReNet activities, we establish some recommendations for strategic actions to be put in force either by the community or by policy makers towards an extensive automation of LR production, which we believe will lead to beneficial improvements for the commercial value of LTs and thus will enforce competitiveness of European LT-based enterprises globally.

6.1

LRP  
PM

*Ensure support and encourage investments in the area of full automation of the full range of language data production*

The list of active and productive research areas in Table 1 cannot be considered a closed list of areas to support. As mentioned above, spoken data resources, such as richly annotated dialogue corpora, are largely under-represented in conference papers, but nonetheless are deemed essential for new research directions in statistical learning approaches to dialogue management, context-sensitive interpretation, and context-sensitive speech recognition. There might well be other areas where at present only manual resources are being produced or where automatic production means are still delivering poor results and thus are not presented/accepted in conferences. The support to automation must be considered a strategic investment for the whole area of NLP applications.

6.2

PM

*Investments in basic and long term research on methods for automatic production of language resources should be increased to achieve results in a broader range of language resources addressed*

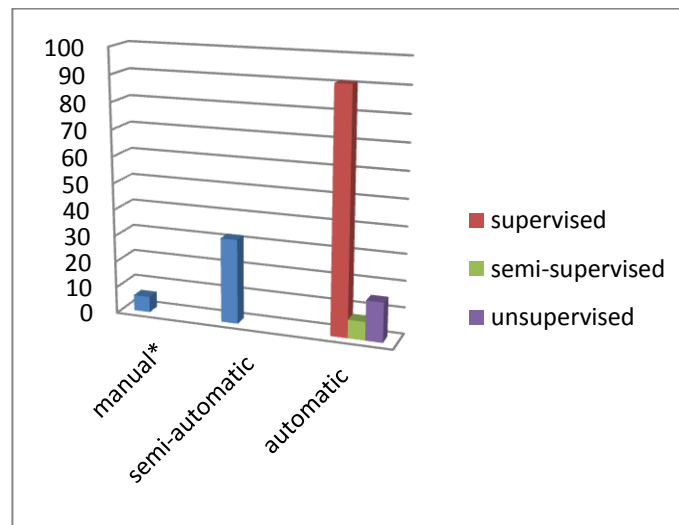
Table 1 shows written corpora and dictionary/lexicons as the most addressed resources. There is some evidence of the interest in the development and annotation of corpora and lexicons for languages other than English, although English still predominates. Thus, it seems that the emphasis is not necessarily on new technology developments, but on the gradual rollout of existing technologies to new languages.

The importance given in terms of the number of papers to domain adaptation is also significant as it could be understood as the community consensus that the development of general purpose resources (developed once and for all applications) has been shown to not lead to optimal results.

In quantitative terms, however, it was the topic of “comparable corpus” acquisition that was driving efforts in 2010, most probably because of the reasons we have previously mentioned.

For the other categories, only lexical acquisition seems to show steady growth, although semantic resources seem to have recovered, especially referring to the creation of polarity or subjective lexica for sentiment analysis, one of the most demanding areas as explained above.

In the case of technologies, data in conference papers (in Table 2) suggest that current research tackling the complete spectrum of approaches to the total automatic production of language resources is the predominant paradigm, although (and because of the difficulties found) semi-automatic methods are starting to show up.



**Table 2: Distribution of conference papers per acquisition methodology (\**manual* here refers only to collaborative development and repurposing of LRs)**

The broad success of supervised machine learning techniques has also influenced automatic resource production. Table 2 shows that most research work uses supervised technologies. Because in order to train supervised systems the availability of previously annotated data is a must, and because systems perform better when large quantities of previously annotated data are used, a vicious circle is created: to produce new data, a sufficient quantity of previously annotated, high quality data is needed. To create this initial seed data is a costly (usually manual) exercise that again is an obstacle for doing the exercise for many different languages or for tuning systems to different domains.

6.3	LRP PM	<i>Promote (shared) gold-standard annotation projects for training supervised methods</i>
6.4	LRP	<i>Promote more efficient uses of available annotated data through a better grasp of repurposing and merging techniques</i>
6.5	LRP PM	<i>Promote research into new methods and techniques for collecting information from unstructured (non-annotated) data also through inter-disciplinary research with other scientific domains working on large volumes of data and monitor current innovative techniques</i>

Because of the cost of manually building gold-standard resources, the automatically produced ones also are used for training purposes. The final results highly depend on the quality of these automatically produced resources. Ensuring such quality, evaluation and correction of these data is also a cost factor and hence it is also noticeable that research has addressed the development of techniques to automate and reduce human intervention in the evaluation of automatically created resources. As a general criterion, methods that maximize precision even at the price of losing coverage should be promoted because of the need to minimize, in all cases, human intervention for evaluating and correcting the results of automatically produced resources. Nevertheless, the problem of error detection also applies to manually produced



resources in general thus, methods for automatically determining the quality, or confidence, of resources and annotations is also a desiderata for language-resource production in general.

6.6 LRP *Research on evaluation frameworks/methodologies is needed to measure the impact of resource/tools dependencies and thus the final quality of the end result (see Ch. 5)*

6.7 LRP *Promote research on automatic techniques for error detection, quality and confidence assessment*

6.8 LRP *Promote technologies that deliver highly accurate, high-confidence results to reduce the amount of subsequent revision (human or automated)*

Because of the cost of getting enough data for the supervised methods previously mentioned, unsupervised technologies have also attracted the efforts of researchers. However, the results of unsupervised methods are still not completely satisfactory. Parallel semi-supervised and active learning<sup>11</sup> techniques are being investigated as a potential solution for building new, high-quality resources with a reduced amount of human intervention. Such methods also appear to have the potential to offset the fact that available data are not large enough for proper training and used for bootstrapping additional data. Although still at an early stage, they will possibly emerge as successful research directions in the near future.

6.9 LRP *Invest in methods such as active learning and semi-supervised learning that deliver promising results while reducing human intervention*

Other proposals to reduce costs of the production of gold-standard data refer to the crowd-sourcing trend of producing resources by using cheaper, non-expert annotators, as already mentioned in Chapter 3. However, issues about how to guarantee the quality of such resources are still open to discussion as severe inconsistencies in data annotation can make the resource useless for practical robust applications.

Several industry players have chosen crowd-sourcing as a breakthrough to the shortage problem. It might look as if the industry is either unaware of research into automatic production of resources, or simply don't want to invest in this line of research.

However it is of utmost importance that the industry finds ways of cutting the cost of language resource production because it consists mostly of SMEs with limited investment capacity<sup>12</sup>. This limited investment capacity is also likely to be the reason of the limited research carried out by these players. The set up and promotion of an "industry-friendly" infrastructure for Language Resource sharing (see Ch. 3) and for research could also be a good instrument for bringing companies and academic research together on the same tasks and development lines.

6.10 LRP *Support the emergence of a LR-sharing infrastructure that can lower the cost of R&D*

<sup>11</sup> Active learning is a form of supervised machine learning in which the learning algorithm can interactively query the user (or some other information source) to obtain the desired outputs at new data points. Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

<sup>12</sup> ELRA in 2010 published data about LR stakes (interest?): 48% in academia and 37% research and technology development in industry. Speech related products are again the main consumer of industrial LR.





*for new applications in new language resource domains*

Finally, in order to bring the research community and industry closer, and thus to boost technological development and commercial competitiveness of European players, it is important that technologies are evaluated in real world scenarios. To this end new evaluation settings must be designed and campaigns organised, with public support (at least initially) to convince the industry of the maturity and utility of certain technologies.

6.11 LRP *Carry out quality evaluation in real-world scenarios*

6.12 LRP *Set up evaluation campaigns to boost automatic evaluation of LR production and promote them across the relevant industries*

### **References**

Liu B., M. Hu, and J. Cheng. (2005). "Opinion observer: analyzing and comparing opinions on the web". In *WWW2005: the 14th international conference on World Wide Web*, pp. 342–351, New York, NY, USA, 2005. ACM Press.



## Chapter 7 - Towards an LR Roadmap

*Joseph Mariani, Khalid Choukri*

LTs are developing at a rate proportional to the difficulty of the research challenges and the complexity of the problems.

NLP research goes back to the very beginning of computer science in the 1950s. A major paradigm emerged in the mid-1980s opened up by research into Automatic Speech Recognition, where approaches based on statistical machine learning outperformed those based on crafting rule sets as expert systems.

Statistical methods paved the way for DARPA-style comparative evaluation campaigns led by the NBS (what is now the NIST) starting back in 1987. The growing need to gather large quantities of data to train systems resulted in the creation of the LDC in 1992 on the general assumption that in NLP “there’s no data like more data.”<sup>1</sup>

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). It has led to substantial advances and successful applications, even though it has by no means solved all the problems of automatic language processing.

Europe has put a similar effort into creating a Language Resource repository, with the launching of the European Language Resource Association (ELRA) in 1995, which later promoted LR and evaluation through the LREC conferences that began in 1998. However, despite several short-term projects, Europe did not create a permanent evaluation agency comparable to the NIST in the US, although a number of players including ELRA/ELDA played such a role, on a limited scale and on short term projects.

To develop language technologies, research and industrial teams need access to a volume of data commensurate with the operational conditions of the application in mind, and equally appropriate Language Technology evaluation mechanisms including metrics and a methodology. Hence the importance of Language Resources and Evaluation (LRE). LT evaluation can measure the performance of the best systems available and compare it to the performances required by the application (which may not necessarily be 100% in every case).

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<sup>2</sup>, 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 programmes 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 countries such as France<sup>3</sup>, Germany and The Netherlands in Europe have also funded national

<sup>1</sup> B. Mercer, Arden House, 1985 as quoted by F. Jelinek in his keynote talk “Some of my best friends are linguists”, LREC 2004, 28 May 2004

<sup>2</sup> See for example the large efforts devoted by IBM to produce the Watson knowledge retrieval system which won the US Jeopardy! Quiz Show in February 2011, while considering that this system works only for English.

<sup>3</sup> Even though US interest has meant that US funded a Mandarin Chinese Broadcast News speech corpus that is 10 times larger than the corresponding corpus in French created under French national programs.



programs that accelerate measurable research. But most of the world's languages do not have such support.

Some countries see language 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). They prioritize the development of language technologies to preserve their languages and ensure communication in them, even if they have limited financial resource 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.

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 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 Language Technology solutions to support multilingualism in Europe should raise concern among EU officials.

ELRA, FLaReNet and then META-NET conducted a survey on the national and cross-border R&D effort in language resources and technologies, and existing data, tools, standards and evaluation methods. This will be available online and continuously updated (the *Language Technology Initiatives Survey*<sup>4</sup>) with an updated set of Language Matrixes which compare the status of resources for various languages based on information provided directly by LR producers or users.

Those language matrices can be used to identify gaps in LT and LR for different languages and different domains, and explain why there are no applications in these domains. The gaps may point to a complete lack of LR and LT, or to the insufficient quality of LR and LT, provided there is a framework for assessing the quality of the LT in question for that language and that application.

Using this analysis, together with an estimate of the investment needed to cover a given language task for a given language, we propose a strategy to fill in these gaps in the most efficient way, at least for EU languages.

## Sources

- The LRE Map<sup>5</sup>
- The META-NET Language Matrixes<sup>6</sup>
- The FLaReNet Wiki Survey of National/Transnational LR initiatives<sup>7</sup>
- The META-NET National/Transnational LT Programs Survey
- The META-NET Summary of detected gaps
- The META-NET Language White Papers<sup>8</sup>.

## 7.1 Urgent LR requirements

There are two dimensions to identifying LR requirements:

---

<sup>4</sup> <http://www.meta-net.eu>

<sup>5</sup> <http://www.resourcebook.eu/>

<sup>6</sup> <http://www.meta-net.eu/>

<sup>7</sup>

[http://www.flarenet.eu/?q=Feedback\\_from\\_Contact\\_Points\\_on\\_National\\_Initiatives\\_in\\_the\\_Area\\_of\\_Language\\_Resources](http://www.flarenet.eu/?q=Feedback_from_Contact_Points_on_National_Initiatives_in_the_Area_of_Language_Resources)

<sup>8</sup> <http://www.meta-net.eu/whitepapers>



1. *Language coverage or diversity:*

This involves identifying the LRs (data, tools and services) needed to develop LT for a given language, and how to address those needs and reduce, if not eliminate, the current ‘two-speed’ language landscape

2. *Topic coverage or topicality:*

This involves identifying the innovative types of language data and tools that constitute the “language resources of the future” for new/strategic applications and sectors, but which are still lacking or are only at the research/experimental stage.

### ***Language coverage or diversity***

To check the availability of LRs for various languages, we explored the data contained in the META-NET Language Matrixes for monolingual LRs, and in Euromatrix/Euromatrix+ for bilingual LRs.

The META-NET Language Matrixes derive from the LRE Map<sup>9</sup>, which comprises information provided by the authors of papers submitted to LT conferences. The initial data obtained from LREC in 2010 is being enriched with material from subsequent conferences- EMNLP’10, COLING’10, ACL-HLT’11, Interspeech’2011, Oriental Cocosda 2011 - journals (*Language Resources and Evaluation* (LRE)) and will be linked to catalogues such as ELRA or LDC. It may initially appear somewhat biased towards NLP, but data from speech communication conferences will be added in due course.

To create the LRE Map, a taxonomy of Language Resources was designed by the LREC 2010 Program Committee. LRs were first divided in 4 major categories: Data, Tools, Evaluation and Meta-Resources (such as guidelines, standards, metadata), and 4 modalities (Written Language, Spoken Language, Multimodal/Multimedia and Sign Language). 24 LR types were put forward (see Mariani and Francopoulo 2011) with an option to add other types. The first analysis was conducted on the 23 official EU languages, on non-EU languages in Europe (such as Norwegian) and on EU regional languages (such as Catalan and Basque).

The Euromatrix/EuromatrixPlus Matrixes are available on the corresponding project Web sites<sup>10</sup> and concern Machine Translation.

Language coverage in these Matrixes emerged as follows from those quantitative measures:

- *Written Language Data* (essentially corpus, lexicon, ontology and grammar/language Models) varies greatly among languages. The most resourced languages are English (30% of the identified written language data are in English), followed by French and German, and then Spanish, Italian, Dutch, Portuguese, Swedish and Romanian. Among the less resourced languages are Estonian, Finnish, Lithuanian, Slovak, Slovene, Maltese and Irish. *Written Language Data* also exist for some non-EU European languages and for several regional European languages.
- Many types of *Written Language Tools* exist, including taggers/parsers, annotation tools, named entity recognizers and tokenizers which form the bulk of these tools. The languages with most coverage are English (25%), French, German, Spanish and Italian. Many languages are poorly resourced, although many of the tools are language-independent.
- *Spoken Language Data* appear less numerous than *Written Language Data*. They also mostly cover corpus, grammar/language models and lexicons. The best-resourced languages are English, French and German, then Spanish and Dutch.
- There are also fewer *Spoken Language Tools* than *Written Language Tools*, with a preponderance of annotation tools. Most of these tools are language-independent.

<sup>9</sup> See Calzolari et al. (2010)

<sup>10</sup> <http://www.euromatrixplus.net/>

- Annotation tools are especially crucial for producing *Multimedia/Multimodal (MM) Resources*. Those tools are mostly language-independent. The most widely covered languages for *MM Data* are English (25%) followed by German.
- *Evaluation Resources* (data, tools, methodologies, packages) exist for some languages, but certainly not all. English is well-served (40%), followed by French, German, Spanish, Italian and Dutch, which all have or have had LT evaluation activities and programmes.
- Finally, there are *Meta-Resources*, especially standards and metadata for certain languages, but again, although most of them are language independent, the most frequently used language is English (30%).
- The situation for bilingual resources (corpora, lexicons, etc.) varies considerably among languages and language pairs, some of them being very well resourced (English (20% of all resources)), others well resourced (French, German, Spanish, Italian, Portuguese, Dutch, Greek, Swedish, Polish (from 10% to 4% of all resources), while others again are considerably under-resourced (Eastern European languages, Baltic languages, Maltese and Irish (with only 2 identified resources)).
- The performance of MT systems as measured by BLEU scores correlates closely to the availability of existing LRs for corresponding language pairs, as well as to the linguistic structure of the languages (some of them being less accurately measured by BLEU scores) and, of course, to the effort devoted to research on those languages.

### **Topic Coverage or Topicality**

To identify the kind of LRs which are needed for new applications, we then analysed the Language Matrixes and checked new areas suggested by the authors. Some of those new areas were already covered under different wording in the initial list that was put forward. Others correspond to the merger of several suggested types. Others were obviously missed when compiling the taxonomy: these will be added in new versions of the questionnaire to be circulated in future conferences. Finally, others correspond to “weak signals” indicating that a new application needs a new technology, and that LRs are being produced to cover that need, at least for one language. We found the following, among others:

- New *Written Language Tools* corresponding to new trends in research and applications of NLP processing at upper levels: *knowledge representation tools, semantic role labellers, semantic network creators, reasoners, sentiment analysis tools*. Some aim at facilitating text analysis such as *frequent item set mining, error miners, key phrase extractors, or rhetorical structure taggers*. Toolkits have been developed for experiments on new algorithmic approaches (*CRF Training Toolkit*) or for carrying out complete tasks: *surface realizer, text simplification tools, proficiency testing tools*, etc.
- There are a few new types of *Written Language Data* also for upper level processes such as a *word aligned corpus with a multilingual ontology or event semantics*.
- *Tools for computer aided language learning (CALL)* are a new important trend in *Spoken Language Tools*.
- Beside *annotation tools* crucial for producing *Multimedia/Multimodal resources*, new kinds of MM tools are emerging such as *MM crosslingual information retrieval*, associated with increasing activity in voice-based video search, and tools in *machine learning* applications. *Talking heads* are also of interest, though they are mostly language-independent.



## 7.2 Findings

Using the previous study of language matrixes, plus a workshop on less-resourced languages organized as a satellite event at LTC'09 in Poznan (November 2009<sup>11</sup>) and a questionnaire circulated among FLaReNet members and discussed at the FLaReNet Forum in Barcelona (February 2010), we have identified several findings which we embody as recommendations to all LT stakeholders - scientific and industrial communities working in this field, educators, the EC, Member States, regional governments and research agencies.

### Language Diversity

In order to properly 'cover' a language:

- Visible cooperation between countries and programs would help set an example and offer Best Practices to less resourced countries. For example, the definition of a consensual basic set of language resources (or BLARK) has helped produce sufficiently high quality technologies for a given language, and also the identification of gaps and roadmaps.

7.1

LRP

*Produce a public Survey on the LT and LR situation worldwide, based on FLaReNet and META-NET deliverables*

- Generally speaking, strong political will (rather than lip-service) about the language dimension and sufficient funds are a *sine qua non*.
- There must also be awareness that LT and LRs are important.
- There should be a critical mass of specialists in the language, requiring the training of young researchers.
- There must be a sufficient background:
  - o A writing system/ transcription code/ agreed orthography
  - o Language Resources (sufficient in quantity and quality)
  - o Tools (especially language independent ones such as those based on statistical training, and ideally open source)
  - o Metadata, annotation schemes, standards
  - o Development platforms
  - o Evaluation facilities (adapted to the language specificities e.g. in the case of the machine translation of morphologically-rich languages).
- The effort should be financed over the long-term, based on a strong foundation, in agreement with the complexity of language.
- Short-term development of a specific product or service for that language (e.g. as a toy), should be avoided. Demonstrating applications that require a strong foundation should be encouraged.

7.2

LRP

*On the basis of the Language Coverage chart, contact officials in national/regional governments to explain the situation regarding language preservation and communication through language, the support that LT can bring and the need of LRs to develop those technologies*

PM

<sup>11</sup> See Mariani et al. (2010).





7.3

PM

*Start with the EU, extend to the regions and to Europe as a whole, and out to the Rest of the World (ROW)*

- Dialectal variants and sociolinguistics should also be taken into account. This does not require very much effort, as there are usually commonalities with the corresponding “main” language.

7.4

LRP

*Analyse the dialectal variants of languages and include those variants in the production of LRs*

- When a majority language also exists, both should be studied together. It would save time and effort to handle a family of languages altogether. Porting a LT from one language to a closely related one can be addressed by focusing at specific levels (phonology, lexicon, syntax) and by producing the corresponding LRs. This is also true for MT using pivot languages.
- Bootstrapping approaches facilitate the coverage of a language.

7.5

LRP

*Analyse the relationship within language families and use joint LR methods to develop LT within those families. Consider the use of pivot languages within those families, based on the existence of parallel corpora*

7.6

LRP

*Adapt LT from the language in which it exists to a similar one in the same language family, then improve the quality of the LRs for the specific data in that similar language through bootstrapping*

- Evaluation must be conducted for all languages, not only for those which attract most of the research community in its bid for scientific recognition.

7.7

LRP

*Create and promote on-line evaluation tools using Best Practices, guidelines, protocols and metrics which already exist for certain languages*

- The related costs could be shared between the corresponding countries or regions, and international bodies (such as the European Commission (EC)) which could also ensure proper coordination (see section 7.3).
- The keywords must be *Interoperability* and *Sustainability*.

### **Topic coverage**

In order to address the question of LRs for innovative applications:

- Information can be extracted from the data entered in the LRE Map

7.8

PM

*Carry out a rolling survey of conference papers and LRs mentioned in papers. Identify the “weak signals” indicating new trends in LT and LRs internationally*

- There is a strong trend towards addressing more semantically focused tasks, such as machine reading, language understanding, knowledge retrieval and human interaction, requiring models of semantics or pragmatics. Research is now shifting from sentences to discourse, documents to dialogue, and artificial to natural interaction. And this means



gathering and annotating the data that would facilitate Machine Learning for the design of working systems.

7.9

LRP *Promote the collaborative development of semantically/pragmatically/dialogically annotated corpora*  
PM

- Web-based applications are particularly interesting due to the availability of online LRs for training and also reflecting the operational conditions of the application. These LRs can then be enriched and the quality of the LT improved as a bootstrapping process.

7.10

LRP *Encourage the development of language based applications on the Web*  
PM

- Applications which used to be specific to written language, such as information retrieval, question & answer systems, or machine translation, are now becoming available for spoken language, with a relatively minor drop in performance when compared to overall performance. They need large amounts of speech data and can be used for (crosslingual) search on video and TV broadcast data, and for automatic indexing and sub-titling/dubbing.

7.11

LRP *Port text data applications to audio data*

- LRs must be created from real situations for processing emotions in affective computing, while also taking intentions into account. There is a need for LRs which include both real data and perceived data.

7.12

LRP *Promote the development of perceptually annotated corpora*  
PM

- Given the importance of the effort needed to produce LRs, explore the automatic extraction of LRs from the Web, which means also checking the quality and liability (rights) of such resources. (see Chapter 6)

7.13

LRP *Use automatic content extraction from the Web, but bear legal issues in mind*

- There may be a need to draw up a Code of Ethics for the automatic extraction of information.

7.14

LRP *Conduct a study and propose a Code of Ethics (regarding privacy and IPR) for automatic information extraction on the internet*  
PM

7.15

LRP *Develop tools for anonymizing data*  
PM

- The question of equal access to information for all, regardless of physical disability, is now a legal matter that will require cross-media technologies, and therefore requires the

production of the corresponding LRs. Similarly, the language barrier for accessing information may appear as a disability, that also requires cross-lingual LRs.

7.16 PM *Develop applications to answer legal accessibility regulations in Europe. Produce the corresponding cross-media resources*

7.17 LRP PM *View the inability to communicate in a foreign language also as a disability, that should have a legal answer*

- Natural interaction other than in telephone applications will ultimately mean removing the close-talk microphone interface. It is necessary to carry out more research into audio-acoustics and design and build especially equipped rooms for applications in meeting transcription, smart homes or robotics.

7.18 LRP *Produce appropriate spoken language resources to study and develop a more natural approach to voice input*

- The NLP community should establish connections with data produced by other communities, such as medical or neuroscience data.

7.19 LRP PM *Establish links with other communities in gaining access to better information on the existence of LRs in their domains, and exchange Best Practices on handling and sharing resources*

### **Infrastructural matters**

- A permanent infrastructure should exist to ensure the interoperability of LR production and facilitate the distribution processes.

7.20 LRP PM *Build the proper LR infrastructure*

- Linguists could help in obtaining high quality LRs, while crowdsourcing and social networks could help produce large quantities of LRs collaboratively. The ethical aspects of schemes such as Amazon's Mechanical Turk should be investigated.

7.21 LRP *Use crowdsourcing for developing LRs in many languages and of many different kinds*

7.22 LRP PM *Propose a Code of Ethics for crowdsourcing, in order to guarantee a sufficient salary and compliance with labour rights (See Chapter 3)*

- The technical resources of the LR infrastructure should be improved with better human computer interfaces for searching and sharing LRs. Web services could help in this process and a common infrastructure would ensure better interoperability.

7.23 LRP *Develop and propose (free) tools and more generally Web services (comparable to the*



		<i>Language Grid), including evaluation protocols and collaborative workbenches in the LR infrastructure</i>
		- The LR ecosystem should be promoted more aggressively, with better recognition for LR producers, on the model of scientific paper authoring.
7.24	LRP PM	<i>Bring together the main LR stakeholders to find a way to attach a Persistent and Unique Identifier (PUId) (a Persistent, Unique and International Standard LR number (ISLRN)) to LRs (this may also lead to attaching such unique identifiers to all researchers, together with an examination of the underlying ethical issues)</i>
7.25	LRP PM	<i>Track the use of LRs in scientific and professional papers</i>
7.26	LRP	<i>Attach information related to the initial designer(s) and further contributor(s) to the LR</i>
7.27	LRP	<i>Compute a "Language Resource Impact Factor" for each LR</i>
7.28	LRP PM	<i>Give greater recognition to successful LRs and their producers (Prizes, Seals of Recognition, etc. )</i>
		- There should be more training in production and use of LRs, and LRs should also be used more widely in education.
7.29	LRP PM	<i>Introduce training in the production and use of LR in Computational Linguistics and Language Technology curricula</i>
		- There is general agreement that experience and knowledge should be shared more effectively to avoid reinventing the wheel in the production of new LRs in different languages or for different tasks.
7.30	LRP	<i>Share all available tools via a LR infrastructure</i>
7.31	LRP	<i>Continue to organize tutorials on LR production at conferences such as LREC</i>

### 7.3 International Framework

The planned production of necessary LRs will be a huge effort if we consider the number of LRs multiplied by the number of languages needed to provide solutions to the problem of multilingualism.

There is unanimous agreement that this effort should be shared by science, industry and government, using national funding, and regional funding where appropriate, and an EC contribution for the European Union share, with the support of the Internet community.

To address multilingualism more effectively, the experience and best practices gained for one language should be leveraged to process others, including the share of information about the needs of LR, sizes and production costs<sup>12</sup>, the use of metadata and standards, the reuse of development methods and of existing tools, such as annotation tools or evaluation protocols and metrics, the use of automatic transcription, translation and transliteration tools which emerged from research, and relationships among languages should be exploited to address language families as a cluster.

The funding mechanisms for a language could be adapted to the size of the effort necessitated by the lack of resources existing for that language.

7.32	LRP PM	<i>Estimate the cost of producing LRs needed to develop an LT for one language</i>
7.33	LRP PM	<i>Establish a joint, coordinated effort using the proper administrative instrument and an appropriate network infrastructure</i>
7.34	PM	<i>Share the effort for the production of LRs between international bodies, such as the EC for the EU, and individual countries/regions, such as Member States in Europe</i>
7.35	PM	<i>Adapt the funding amounts and mechanisms to the lack of existing resources for the corresponding languages</i>

International, National or Regional agencies should open their programs to foreign participants (and plan for the coverage of the corresponding cost even if they don't fund those participants).

7.36	PM	<i>Encourage agencies to open their calls to foreign participants</i>
------	----	---

It is useful to use multilingual data as a knowledge source to conduct multilingual joint-training, and allow for joint knowledge discovery.

7.37	LRP PM	<i>Leverage LRs built with a similar approach in a common framework as a multilingual joint training facility</i>
------	-----------	---

We are now faced with an increasingly complex R&D landscape, due to the sudden upsurge of worldwide initiatives in LR production/distribution and LT evaluation. It will be vital to understand the exact topic coverage or agenda of each of these initiatives and build mutual trust.

7.38	LRP PM	<i>Establish MoU among existing initiatives, starting with the EC efforts</i>
------	-----------	---

It is important to constantly identify the nature and volume of resources in all countries, to meet regularly at dedicated forums such as those organized by FLaReNet, and at conferences such as

<sup>12</sup> For example, it's worth knowing that the annotation of one hour of speech data may need from 15 hours to 50 hours, depending on the complexity of the annotation process (from orthographic transcription to more complex labels) and of the targeted quality.



LREC, and establish a permanent international network of stakeholders (researchers, industry players, laboratories, administrations etc.).

7.39	LRP PM	<i>Ensure the sustainability of the LRE Map, the Language Matrixes and the National Initiatives Survey Wiki</i>
7.40	LRP PM	<i>Ensure the sustainability of the LREC conference and of the LRE Journal</i>
7.41	LRP PM	<i>Ensure the sustainability of the International FLaReNet International Contact Points, also through the Survey Wiki</i>
<p>On this basis, it should be feasible to agree among several countries/regions on a common strategy (contained in a White Paper, or in a series of Memoranda of Understanding (MoU)).</p>		
7.42	LRP PM	<i>Write a White paper on LT, including LRs, in preparation for FP8. Distribute it within Europe and abroad</i>

## 7.4 Strategy

To develop LT effectively for the various applications needed to enhance and support multilingual communication in Europe and worldwide we need to produce and test the quality of LRs in different languages. This is a huge challenge and needs an appropriate strategy to ensure an efficient approach to language coverage.

The LTs available for the languages with most advanced technologies, especially (American) English, necessitated access to available LR of a sufficient size. So the first agenda item is to list these LRs and identify how much effort was involved in producing them, with an estimate of LT performance in relation to LR size. There may be open source tools to produce the LRs, but if not, they should be produced with the aim of being easily portable to many different types of languages. These LTs must also be evaluated. Here again, there may be guidelines, methodologies and software to conduct evaluations, which may be reusable for other languages. If not, they should be produced and shared, with the same aim of being easily portable to many languages with different specificities.

The availability of a network infrastructure such as META-SHARE will help distribute existing data, together with the tools for producing new data and for evaluating LTs. It will also help produce the data itself in a collaborative way and also carry out LT evaluation.

The production of LRs may need the participation of specialists for various processes. For example, a LR developed for speech translation will need people to transcribe what is said, and people to translate the resulting transcription into different languages. Similarly, a semantically annotated speech corpus would possibly need transcription followed by prosodic, syntactic and semantic labelling.

Those tasks could be achieved by a large population through the new trend in crowdsourcing over the web. This would help identify appropriate people capable of addressing a specific language and allow them to work in-country from home. However, the ethical conditions of crowdsourcing should be carefully checked so that contributors receive appropriate wages and legal and the tax regulations are observed in each country.





The LT evaluation campaigns would also help in the production of LRs. If several systems from different laboratories process the same data, their results can be used to annotate the data, by merging them using a mix of automated decision making, and human intervention when results are different among laboratories or when automation cannot be trusted due to low confidence scores. This has already been successfully experimented for morphological or syntactic tagging, and for speech transcription.

The LT evaluation campaigns would also provide insight into the comparative quality of the tools used, and on the adequacy of LT for a given application that requires a minimum performance level (this would contribute to an estimate of its Technological Readiness Level (TRL)). The tools could then be distributed through the same infrastructure, accompanied by information on their quality. This would create a virtuous circle of LR production through LT development and LT evaluation to more LR production and LT improvement.

In the European Union, the cost of this effort should be shared between the European Commission and EU Member States<sup>13</sup>, in agreement with the principle of subsidiarity. The European Commission would have the primary duty of providing the infrastructure necessary to carry out shared research, including the specification of data exchange standards, the collaborative infrastructure, communications facilities for building and maintaining links with all the stakeholders, and an evaluation infrastructure to assess the results and compare different systems. The Member States, in connection with the Regions, would have the primary duty of producing the Language Resources needed for their language(s). The development of the core LT for this can be shared by the various funding agencies.

## 7.5 *Conclusions and perspectives*

Our analysis of LRs and LTs shows that although the English language is relatively well covered, allowing for the availability of the presently operational state-of-the-art applications, many other languages lack the necessary LRs to develop LTs and applications of the right quality. There should therefore be a major coordinated effort to ensure that these LRs become available for all languages.

A Multilingual Europe Technology Alliance offers one model for this, provided that the LRs needed for LT development in 23 languages can be made available, and that there is an infrastructure in place to assess LT performances in those different languages on a regular, systematic basis, and measure their progress and fit to needs. This entire process should be translated into a ten-year program combining EC programs and Member State national programs. The model could later be extended to other language clusters, such as non-EU European languages, European regional languages and any other country's language(s), using a cooperative scheme involving financial contributions from national or regional governments wishing to make their language(s) LT-ready.

## **References**

- Calzolari N., C. Soria, R. Del Gratta, S. Goggi, V. Quochi, I. Russo, K. Choukri, J. Mariani, S. Piperidis, "The LREC 2010 Resource Map", LREC'2010, Malta, 19-21 May 2010.
- Mariani J., K. Choukri, Z. Vetulani, "Report on the Special joint LTC-FLaReNet session « Getting Less-Resourced Languages On-Board ! », LTC'09 Conference, Poznan", FLaReNet, February 2010.

---

<sup>13</sup> Such a scheme is presently being experimented in the ERIC agreements accompanying the ESFRI programs, as in CLARIN on Language Resources for the Human and Social Sciences.



## Glossary of acronyms

ANC	American National Corpus
ASR	Automatic Speech Recognition
BLaRK	Basic Language Resource Kit
BNC	British National Corpus
CALL	Computer Aided Language Learning
CLARIN	Common Language Resources and Technology Infrastructure
CRF	Conditional Random Field
DARPA Defense	Advanced Research Projects Agency (US)
DCR	Data Category Registry
EC	European Commission
ELaRK	Extended Language Resource Kit
ELDA	Evaluations and Language resources Distribution Agency
ELRA	European Language Resources Association
EU	European Union
FLaReNet	Fostering Language Resources Network
ISO	International Standards Organization
LDC	Linguistic Data Consortium (USA)
LR(s)	Language Resource(s)
LRE	Language Resources and Evaluation
LRE Journal	Language Resources and Evaluation Journal
LRE Map	Language Resource and Evaluation Map
LREC	Language Resources and Evaluation Conference
LRPs	Language Resource Producers
LRT(s)	Language Resource(s) and Technology(ies)
LT(s)	Language Technology(ies)
META-NET	Multilingual Europe Technology Alliance
MM	Multimodal/Multimedia
MT	Machine Translation
NBS	National Bureau of Standards (USA) (now NIST)
NIST	National Institute of Standards and Technology (formerly NBS) (USA)
NLP	Natural Language Processing
OANC	freely available ANC
OLAC	Open Language Archives Community
PMS	Policy Makers
POS	Part Of Speech (tags and tagger)
R&D	Research and Development
ROW	Rest Of the World
SLT	Spoken Language Translation
SMT	Statistical Machine Translation
TAUS	Translation Automation Users' Society
TRL	Technology Readiness Level
WER	Word Error Rate
WSD	Word Sense Disambiguation