ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs

Nicoletta Calzolari[1], Khalid Choukri[2], Maria Gavrilidou[3], Bente Maegaard[4], Paola Baroni[5], Hanne Fersøe[4], Alessandro Lenci[5], Valerie Mapelli[2], Monica Monachini[1], Stelios Piperidis[3]

[1]Istituto di Linguistica Computazionale del CNR (ILC-CNR)
Via G. Moruzzi 1 – 56124 Pisa – ITALY
[glottolo, monica.monachini]@ilc.cnr.it
[2]Evaluations and Language resources Distribution Agency (ELDA)
Rue Brillat Savarin 55-57 – 75013 Paris – FRANCE
[choukri, mapelli]@elda.fr
[3]Institute for Language and Speech Processing (ILSP)
Artemidos 6 & Epidavrou – GR-151 25 Maroyssi – GREECE
[maria, spip]@ilsp.gr
[4]Center for Sprogteknologi, University of Copenhagen (CST)
Njalsgade 80 – 2300 Copenhagen S – DENMARK
[bente, hanne]@cst.dk
[5]Università degli Studi di Pisa – Dipartimento di Linguistica (UPI-DL)
Via Santa Maria 36 – 56126 Pisa – ITALY
[paola.baroni, alessandro.lenci]@ilc.cnr.it

Abstract

In this paper we present general strategies concerning Language Resources (LRs) – Written, Spoken and, recently, Multimodal – as developed within the ENABLER Thematic Network. LRs are a central component of the so-called "linguistic infrastructure" (the other key element being Evaluation), necessary for the development of any Human Language Technology (HLT) application. They play a critical role, as horizontal technology, in different emerging areas of FP6, and have been recognized as a priority within a number of national projects around Europe and world-wide. The availability of LRs is also a "sensitive" issue, touching directly the sphere of linguistic and cultural identity, but also with economical, societal and political implications. This is going to be even more true in the new Europe with 25 languages on a par.

Introduction

After considering the strategic and infrastructural role of Language Resources (LRs) within any Human Language Technology (HLT) application (section 1), we focus on the main issues discussed within the ENABLER (European National Activities for Basic Language Resources) Thematic Network (section 2): the survey of LRs, interoperability and multilinguality, open access to LRs, validation methodologies for LRs, industrial and basic requirements. Finally, few recommendations are provided towards the design of general strategies and an overall coordination for the field of LRs (section 3).

1. The Strategic Role of LRs

Language Resources (LRs) – Written, Spoken and, recently, Multimodal – are a central and strategic component of the so-called "linguistic infrastructure" (the other key element being Evaluation), necessary for the development of any Human Language Technology (HLT) application and product. The availability of adequate LRs for as many languages as possible is a pre-requisite for the development of a truly multilingual Information Society. They play a critical role, as horizontal technology, in different areas of the 6th Framework Programme, and have been recognized as a priority within a number of national projects around Europe.

The availability of LRs is also a "sensitive" issue, touching directly the sphere of linguistic and cultural identity, but also with economical, societal and political implications. This is going to be even more true in the new Europe with 25 languages on a par.

The ENABLER Thematic Network of HLT National Projects in European countries – an EC funded IST project, designed and started by Antonio Zampolli, with a clear strategic vision for the field of LRs – is the first broad European initiative which has the mission of explicitly considering together the technical, organizational, strategic and political issues of LRs. In ENABLER these various aspects are put together in a coherent framework, to set up medium- and long-term set of priorities (both technical and strategic) and to promote these at the national and international levels.

Moreover, ENABLER has recognized the importance to promote actions aiming at integrating the different resource types, until now developed independently, and – as a consequence – at promoting the cooperation between the communities of Speech, Text and Multimodality.

In the following we briefly highlight the main issues tackled by ENABLER on the different layers.

2. The Main Issues

2.1. The Survey of LRs

The ENABLER Consortium conducted the *Survey of LRs* to get a global picture of the situation on LRs, in order to be able to compare the various conditions that hold across different languages and – on this basis – to suggest more

sound recommendations. The *Survey* provides an overview of the results of National Projects and activities on LRs of different types (written, spoken, multimodal, lexical resources and related tools). The results of the overview, related to 164 different resources from various countries and languages, concern all the facets needed to properly describe LRs: resource production issues (collection, processing, annotation, validation), legal issues (copyright, availability), the description of the resource (type, language coverage, content) and the distribution policies adopted by the LRs producing organizations. For the design of the set of elements required for describing the LRs, and for the updating procedure, two points of view were taken into account:

1. the one of the LR producers: which elements provide the most accurate description of the resource;
2. the one of the prospective users: which elements constitute the most informative data to facilitate the formulation of queries in order to identify the resource which best suits their needs.

We have developed a set of active web pages taking into account the two points of view. The survey results and the updating mechanism can be found at http://www.ilsp.gr/enabler/. The structure on which the updating mechanism is developed could serve as a basis for the creation of a central information point for the dissemination of information on LRs, which the producers would have the responsibility of updating.

The *Survey* aimed not only at collecting information on LR activities, but also at harmonizing their descriptions and, finally, at leading to a common metadata schema for their description.

2.2. Compatibility and Interoperability of LRs: towards Multilinguality

A way to reach the optimisation of the process of production and sharing of (multilingual) LRs can be found in a common and standardized framework which ensures the encoding of linguistic information in such a way to grant its reusability in different applications and tasks. Standards are, hence, critical to achieve the interoperability needed for effective integration.

The Network promoted the compatibility and interoperability of LRs mainly through cooperative work with: ISLE/EAGLES, for harmonisation of linguistic specifications, in particular for multilingual lexicons (based on MILE, the Multilingual ISLE Lexical Entry) and corpora; ISO TC37 SC4 WG4 Committee, to make European standards truly international ISO Standards; ELRA Validation Committee, for the incorporation of agreed standards in protocols for validation of LRs, both Spoken and Written; INTERA, for the harmonisation of metadata descriptions; Semantic Web communities, to promote synergy between the groups of knowledge management/ontology and of HLT/LRs technology.

One of the recommended steps is to verify whether the MILE (Calzolari et al., 2002) model (with its set of basic notions) can be used as a *common parlance* between different lexicons, by means of an analysis of the mapping conditions between existing lexicons and MILE, and to define which adjustments it needs.

To facilitate the integration of the LRs and tools resulting from all the various LR initiatives of the last decade and, at the same time, to make word-content machine understandable, as it is the aim of the Semantic Web, three critical issues must be addressed:

1. *standards*, which are unavoidable to achieve interoperability and integration;
2. *content*, as the information crucial to be represented is semantic information;
3. *multilinguality*, seen by ENABLER as a critical issue for the immediate future.

Multilinguality is also a strong integrating factor, horizontal with respect to different application areas and LR types (Spoken and Written). It implies not only harmonised technical decisions, but also heavy organisational aspects, which can only be taken into account at the supranational level.

A clear recommendation in formulating a global common strategy for LRs is that research and LR building should be closely interrelated.

2.3. Validation Methodologies for LRs

An edited collection of available validation protocols and procedures for LRs (including descriptions) was produced, representing the current best practice in validation of LRs. It concludes that:

(i) validation must be integrated in the development process;
(ii) validation can frequently provide feedback to correction and to the next development loop;
(iii) validation is primarily focused on the content, i.e. on the correctness of the linguistic encoding;
(iv) a set of commonly accepted validation standards would be useful to future resource producers.

2.4. Industrial Needs of LRs

LRs are needed for the language and translation industry and for any content industry. Large companies produce their own LRs for the languages for which a business can be made, but small companies can not afford this and, often, no resource is built for less spoken languages. Additionally, resources that are built by companies are normally not shareable, as they are seen as a competitive advantage and – consequently – not traded.

As a consequence, an important goal of ENABLER was – besides describing the needs for LRs in general – also to map which type of resources and of LR format and content are required by the HLT industry. The target groups were the language, speech, translation and content industries. The main objectives were to push the exploitation of what already exists and to collect industry's recommendations for future LRs, in particular multilingual, for a better planning of future initiatives. The investigation showed that the market for LRs is large and growing. It also confirmed the need for LRs: not enough LRs are available and the ones that exist do not always meet the requirements stated. This is especially the case for less used languages and Accession country languages. Less used languages are by definition 'uninteresting' from a commercial point of view and, hence, there is a particular need for public support for the development of these resources. This fits very well with previous recommendations made by the EUROMAP project 2003 (Joscelyne, 2003).

3. The Strategic Issues: International Research Infrastructures for LRs

An important goal of ENABLER was to provide recommendations for strategic initiatives to be promoted in the field of LR production and management. Two main lines have been highlighted:

1. *infrastructural initiatives* – ENABLER has promoted the creation of a new international infrastructure for linguistic resources;
2. *coordination initiatives* – these concern both the national dimension and the transnational and transcontinental ones.

These lines of action seek to address the main priorities for LRs and to define a strategy for LRs in the next years. The workshop "International Roadmap for Language Resources" organised by the ENABLER Network in collaboration with the ELSNET Network in Paris on 28th-29th August 2003 has actually laid the basis to build a roadmap for LRs. A first list of main priorities that act as critical issues for the future of LRs was drawn:

- provide basic LRs for a larger set of languages;
- increase multilingual LRs;
- reduce development time of LRs;
- enhance LR content interoperability;
- foster synergies with neighbouring areas (e.g. terminology, Semantic Web);
- develop new methodologies and tools for LR management, quick domain and application adaptation, data-driven tuning etc..

The next sections illustrate for some of these issues how they have been at the centre of the ENABLER achievements, in the context of a wide array of initiatives to promote LRs in the years to come.

3.1. From BLARK (Basic LAnguage Resource Kit) to ELARK (Extended LARK)

ENABLER has adopted and strongly supported the BLARK (*Basic LAnguage Resource Kit*) concept, first launched through ELSNET (Krauwer, 1998) and Nederlandse Taalunie (Binnenpoorte et al., 2002). The promotion of BLARK required to:

(i) specify for every language the minimum set of LRs (in terms of text and spoken corpora, lexicons, basic tools to manipulate them, skills required etc.) to be able to do any pre-competitive research for that language;
(ii) spot the actual gaps to be filled (a matrix highlighting the gaps of LRs for many applications and languages will be accessible and modifiable directly from the ELRA Web site, to enable customers or providers of LRs to fill it, to identify available LRs and to promote the production of new LRs);
(iii) present a summary of the technical, operational and organisational problems to be tackled and provide suggestions for an overall organisation framework for international cooperation.

BLARK must be considered as an evolving notion. A further level was defined as *Extended LAnguage Resource Kit* (ELARK), which will be extensively promoted for its larger adoption.

Among these initiatives, we should not omit the maintenance and updating work on LRs (Macleod, 1998), further to the production work, since the cost of LRs is high enough to take into consideration their reusability on a long-term basis.

3.2. Open and Distributed Framework for LRs

The need of ever growing LRs – testified also by the current US funding strategies – led us to propose and promote a change in the overall model of how to build, maintain and share LRs. In particular, a new paradigm is required and proposed to make the Web usable, i.e. an *open, distributed and collaborative language infrastructure,* based on open content interoperability standards. Semantic Web developers will need repositories of words and terms and knowledge about their relations within language use and ontological classification. The cost of adding this *structured and machine-understandable lexical information* can be one of the factors that delay its full deployment. The effort of making available millions of 'words' for dozens of languages is something that no small group is able to afford. Existing experience in LR development proves that such a challenge can be tackled only by pursuing – on the organisational side – a truly interdisciplinary and cooperative approach and by establishing – on the technical side – a highly advanced environment for the representation and acquisition of lexical information, open to the reuse and interchange of lexical data. We promote the launch of a large initiative, comprising the major LR and HLT groups in Europe and world-wide, for the creation of an open and distributed infrastructure for LRs. The outcome of such an initiative could be the design of a completely "new generation" of LRs.

The Linguistic Infrastructure supported by ENABLER intends to contribute to the structuring and integration of the European Research Area, addressing problems such as the fragmentation of its research base, the under-financing of research and the weakness in converting R&D results into useful economic or society benefits. For this aim it is necessary to pool together and to build on many different and related initiatives.

An important *Declaration on Open Access to LRs* was endorsed by all the participants of the ENABLER/ELSNET Workshop "International Roadmap for Language Resources".

3.3. Contributing to the Design of an Overall Coordination and Strategy in the Field of LR

International cooperation will be certainly the most important factor for the field of LRs in the next years. A report produced by ELDA (Mapelli & Choukri, 2003) presents an analysis of several organisational frameworks, focussing on funding and organisational procedures for providing LRs. These frameworks are classified into five different areas: the European Union framework, the work carried out by data centres such as ELDA and LDC, national programmes (with the examples of the French and Italian programmes) and the Northern American and Asian scenarios.

The pre-requisites to be addressed for the production of interoperable LRs in a cooperative framework belong to different layers: technical (specifications), validation (quality assessment), legal, commercial. For example, clarifying legal issues aims at simplifying the relationship between producers/providers and users of LRs. With the exception of commercial publishing ventures, the core

business of most LR providers is not LR production, collection and/or validation. In practice, most of them develop or acquire resources for their own internal needs. Finally, marketing LRs requires several activities to be dealt with, in particular packaging, distribution and maintenance.

In order to fill the gaps in terms of LRs, cooperation on all combined organisational, funding, technical and commercial issues appears now to be necessary. To strengthen such a cooperation, there is no doubt that an effort in coordinating this cooperation is required. A coordinated operation was already launched in the framework of Speech LRs with the creation of COCOSDA (*International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques*). Major strategic outcomes of ENABLER with respect to international cooperation and to the design of an overall coordination and strategy in the field of LRs are the following ones.

3.3.1. ICCWLRE

A new committee, originally conceived by Antonio Zampolli, has been established in the field of Written LRs, the *International Coordination Committee for Written LRs and Evaluation* (ICCWLRE), which will provide the optimal environment to continue (part of) the ENABLER mission, while, at the same time, enlarging its scope beyond the European boundaries. Possible tasks for this Committee include: information dissemination on LRs, dissemination of standards, promotion, coordination, and enabling activities, copyright and IPR, training and methodology for LRs creation and validation, roadmaps for LRs, political and strategic tasks. The first joint meeting of COCOSDA and ICCWLRE is organised as a satellite event at LREC 2004, with the goal of building a roadmap for LRs, as a joint effort of the communities of Speech and Text, fostering future synergies among them.

3.3.2. LangNet

Last but not least, an initiative – LangNet – is being proposed in the framework of the ERA-Net scheme of the 6th Framework Programme of the European Commission to coordinate national initiatives in HLT all over Europe. LangNet candidates itself to provide the most natural environment to continue the efforts and the momentum gained by the ENABLER Network. Language Technologies seem to be especially well fitted for the ERA-Net scheme, based on the assumption that each country wishes to conduct research activities allowing for the development of systems and applications for their language(s). It therefore seems natural that the individual countries basically take into account all the "(spoken and written) language-dependent" aspects and that the European Commission rather takes into account all the generic, "language-independent" aspects, in agreement with the principle of subsidiarity. But in order to avoid a 2-speed Europe (Maegaard, 2003), linguistically speaking, coordination should be established between the European Commission and the member states and strategies should be drawn in order to ensure a proper balance of language coverage in Europe.

The idea behind these initiatives is to establish some sort of permanent coordination to capitalise on parallel existing (national or international) initiatives on the long run. It goes without saying that such international initiatives will not be able to work properly without a good framework within each organisation involved in this cooperation. Not only a good coordination from the top is required, but also a good response and feedback from the bottom players is needed. We could therefore talk about two levels of coordination actions: a macro-coordination initiative at the upper level and a micro-coordination at the level of each partner.

References

Baroni P., Calzolari N., Lenci A., *Extended Configuration of the Network and Final Report*, ENABLER Deliverable D1.2, Pisa, 2003, Pp. 21[1].

Binnenpoorte D., De Vriend F., Sturm J., Daelemans W., Strik H., Cucchiarini C., A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, in *LREC 2002 Proceedings*, Las Palmas de Gran Canaria, 2002

Calzolari N., Bertagna F., Lenci A., Monachini M. (eds.), *Standards and Best Practice for Multilingual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry)*, ISLE CLWG Deliverables D2.2&D3.2, Pisa, 2003, Pp. 194[2].

Gavrilidou M., Desypri E., *Final Edited Version of the Survey*, ENABLER Deliverable D2.3, Maroyssi, 2003, Pp. 46[3].

Joscelyne A. et al., *Benchmarking HLT progress in Europe*, HOPE, Copenhagen, 2003.

Krauwer S., ELSNET and ELRA: A common past and a common future, in *ELRA Newsletter*, Vol. 3, N. 2, 1998.

Macleod C., A Plea for Consideration of Maintenance of Language Resources, in *LREC 1998 Proceedings*, Granada, 1998.

Maegaard B., Choukri K., Mapelli V., Nikkhou M., Povlsen C., *Language Resources – Industrial Needs*, ENABLER Deliverable D4.2, Copenhagen, 2003, Pp. 19[4].

Mapelli V., Choukri K., *Report on a (Minimal) Set of LRs to Be Made Available for as Many Languages as Possible, and Map of the Actual Gaps*, ENABLER Deliverable D5.1, Paris, 2003, Pp. 22[5].

Mapelli V., Choukri K., *Report Contributing to the Design of an Overall Co-ordination and Strategy in the Field of LRs*, ENABLER Deliverable D5.2, Paris, 2003, Pp. 20[6].

Zampolli A. et al., *ENABLER Technical Annex*, Pisa, 2000, Pp. 39[7].

[1] URL: http://www.enabler-network.org/documents/ENABLER_D1.2.zip.

[2] URL: http://www.ilc.cnr.it/EAGLES96/isle/clwg_doc/ISLE_D2.2-D3.2.zip.

[3] URL: http://www.enabler-network.org/documents/ENABLER_D2.3.zip.

[4] URL: http://www.enabler-network.org/documents/ENABLER_D4.2_FinalVersion.zip.

[5] URL: http://www.enabler-network.org/documents/ENABLER_D5.1_FinalVersion.zip.

[6] URL: http://www.enabler-network.org/documents/ENABLER_D5.2.zip.

[7] URL: http://www.enabler-network.org/documents/ENABLER_TechnicalAnnex.zip.