Community Culture in Language Resources – An International Perspective

Nicoletta Calzolari

Istituto di Linguistica Computazionale – CNR
Via Giuseppe Moruzzi N° 1 – I-56124 Pisa – ITALY
glottolo@ilc.cnr.it

Abstract

I highlight a few issues which I consider of relevance with respect to the infrastructural role of Language Resources. I underline some of the circumstances and attitudes which are specific of the European approach, and sketch how I see the current situation in the LR field and what I think is of highest priority with respect to implementing an open Language Infrastructure. My objective is to show that it is imperative that there is an underlying global strategy behind the set of initiatives which are/can be launched in Europe and world-wide, and that a global vision and cooperation among different communities is necessary to achieve more coherent and useful results.

1. The growth of a Language Resources community culture

1.1. Setting the scene

Since the '80s it has become clear that Language Resources (LR) have progressively acquired a larger role in Human Language Technology (HLT), also in view of developing innovative and robust technologies or to integrate existing ones to achieve more advanced applications. This process achieved a crucial step through the acknowledgment of the infrastructural role of LRs, first recognized by A. Zampolli to whom we also owe the term itself 'Language Resources' [1]. This trend was very influential in the formation of the strategy of the European Commission (EC) in the '90s and in the launching of many European LR related projects and initiatives, the conditions and time being ripe for the speeding up of a major effort in LR development. LRs started to be considered as the necessary common platform on which to base new technologies and applications, a recognition which is nowadays widely accepted for the development and takeoff of our field.

Also the concept of reusability – directly related to the importance of "large scale" LRs within the dominant data-driven approach – has contributed significantly to the structure of many R&D efforts [2]. Many large international projects in this area, on both sides of the Atlantic and in Japan, were motivated by this idea. After the first pioneering EC projects on LRs already in the '80s - ESPRIT BRA ACQUILEX and EUROTRA-7 – there was a flourishing of international projects and activities (see also [3] for an overview) that contributed to substantially advance knowledge and capability of how to represent, create, acquire, access, tune, maintain, standardize, etc. large lexical and textual repositories.

1.1.1. Infrastructural initiatives

The set of these projects of the '90s can be seen as the beginning of a consistent and coherent realization in Europe of a well-thought plan to implement the badly needed infrastructure of LRs [4]. In addition to its "scientific" implications, this large intellectual and economic movement obviously entailed "strategic" considerations, and pushed towards the need to reflect on the situation in the area of LRs in Europe from a very broad perspective. Some of the LR projects, dealing with policy and meta-level issues related to LRs and standards, have been instrumental to define a coherent strategy for the LR field in Europe, and to give Europe a central position in the LR area, leading also to founding independent associations such as ELRA (European Language Resources Association), the European counterpart of the American LDC (Linguistic Data Consortium).

It was perceived as essential to define a general organization and plan for research, development and cooperation in the LR area, to avoid duplication of efforts and provide for a systematic distribution and sharing of knowledge. To ensure reusability, the creation of standards was the first priority. Another tenet was the recognition of the need of a global strategic vision, encompassing different types of (and different methodologies of building) LRs, for an articulated and coherent development of this field.

Even if LRs have a rather short history, they are nowadays recognised as one of the pillars of HLT, and a central and strategic component of the so-called "linguistic infrastructure" (the other key element being Evaluation), necessary for the development of any HLT system, 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 a horizontal technology, in different areas of the EC 6th Framework Programme, and have been recognized as a priority within a few national projects around Europe.

1.1.2. Signs of the wide resonance of LRs

A few signs of the wide resonance LRs have acquired in the last decade can be found, among others, in a number of international initiatives: the LREC Conference (1000 participants in 2004 in Lisbon); bodies such as ELRA and LDC, or COCOSDA (International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques) and WRITE (Written Resources Infrastructure, Technology and Evaluation); the new international journal *Language Resources and Evaluation* [5]; not to mention the vital role of LRs in statistical and empirical methods, in evaluation campaigns, and so on. Moreover, there is a clear and growing industrial interest in the use of LRs and standards, in particular for multilingual applications.

On the one hand, such a solid position of the LR area must be maintained and reinforced, anticipating the needs

of new types of LRs and quickly consolidating (through EAGLES/ISLE-like standardisation initiatives) areas mature enough for recommendation of best practices and standards. A virtuous circle should be established between innovation and consolidation. On the other hand, however, much stronger initiatives are needed to achieve true interoperability (see e.g. the issue of open architectures below), for which I envision the need of a new paradigm – in the sense of Kuhn – for the area of LRs.

New types of initiatives are now underway, such as: a) the EC LIRICS (e-Content) project, aiming to provide ISO ratified standards for LRs & LT, b) the Unified Lexicon project – by ELRA and its Production Committee – linking the LC-Star and PAROLE lexicons to set up a methodology to connect Spoken and Written LRs, and thus establish common standards and new models of LR distribution, or c) the new NEDO Japanese project for developing international standards of LRs for Semantic web applications, specifically geared to Asian languages but with the cooperation of Asian and a European partner.

2. How to shape the future?

We must build on the set of accumulated experience – and data – we have gained so far, but – exactly because of the massive amount of knowledge and data we have been able to gather – we must also reflect if today situation does not require, to make a real step further, a deep change of perspective and a new vision.

2.1. Roadmap for LRs

In recent consultations about LRs, such as the ELSNET/ENABLER Roadmap workshops (Paris, 2003 and Lisbon, LREC2004), a first list of priorities which act as critical issues for the future of LRs was drawn:

- define and provide basic LR coverage for all languages (BLARK/ELARK concepts);
- significantly increase multilingual LRs;
- develop an "Open Source" concept for LRs;
- coordinate the design and creation of LRs (also across languages) with a view to interconnectivity and reusability, to enhance LR content interoperability;
- enhance metadata infrastructure and standards;
- give high priority to methods and tools to quickly develop LRs "on demand" (acquisition, annotation, merging, porting between domains or languages, …), a particularly important issue for industrial exploitation;
- develop LRs for evaluation purposes, and define validation methodologies and protocols for LRs;
- foster synergies between spoken and written areas and with neighbouring areas (e.g. terminology, Semantic Web);
- investigate IPR issues.

2.2. Some LR priorities and challenges

For a better organised field many challenges exist, at various levels of complexity and with various priorities and weights, both at technological and organisational level. I mention some and quickly touch a few:

- Overcome the usual *mismatch between advancement in LRs and in LT*.
- Design *lexicons as dynamic resources* whose content is co-determined by automatically acquired linguistic information from text corpora and from the web. We should push towards innovative types of lexicons: a sort of 'example-based living lexicons' that participate of properties of both lexicons and corpora.
- Eliminate the *lack of communication between the communities of LRs/LT and Semantic Web(SW)/Ontologies*. LT will highly benefit from the SW but the SW needs LT, otherwise there is a clear risk of 're-discovery' of what was done 20 years ago.

  Examples of relations from LRs/LT to SW:
  - *Semantic mark-up*: for the SW task of adding meaning to Web data and make it usable for automatic processing.
  - *LRs as the basis for knowledge representation and sharing*, for interoperability among knowledge based systems.
  - *Ontology learning, ontology design and evaluation of ontologies*: LT is mature enough to be a core technology for the extraction and creation of semantic content.

  Examples of relations from SW to LRs/LT:
  - *LRs/LT as web services, and use of SW representation formalisms*: the SW may crucially determine the shape of the new generation of LRs of the future, consistent with the vision of an open space of sharable knowledge available on the Web for processing.
  - *Open access paradigm, semantic interoperability, information integration*: this is – in my vision – the real target for the next decade for LRs, and implies a complete re-thinking of the current area of LRs.

I'd like also to mention a few types of LRs that should receive attention in the next years.

- New types of *"example-based" context sensitive LRs, Lexicon and Corpus together*, dynamically created.
- The *Web exploited as a multilingual corpus*.
- *Facts and commonsense knowledge*, built in distributed and collaborative way by the community.
- Common sense in *affective classification of text*.

And we cannot forget two issues often neglected:

- *Knowledge transfer across languages*, to take advantage of LRs built for few resource-rich languages and induce knowledge in languages with few LRs.
- *Maintenance of LRs* (updating, tuning, etc.): it is still a big issue that deserves to be organised.

3. LRs in the future HLT

Focusing our view into the future of LRs, a radical modification of perspective is needed, to facilitate integration of linguistic information resulting from all LR initiatives, bridge differences between various standpoints on language structure and linguistic content, put an infrastructure into place for content description and interoperability at European level and beyond, and make LRs usable within the emerging SW scenario [7].

3.1. A new paradigm for LRs

The need of ever growing LRs for effective multilingual content processing requires a change in the paradigm, and the design of a "new generation" of LRs, based on open content interoperability standards. SW developers will need repositories of words and terms, machine-understandable knowledge about their relations within language use and ontological classification. The effort of making available millions of 'annotated words' for dozens of languages is something that no single group

is able to afford. This objective can only be achieved when working in the direction of an integrated *Open and Distributed Linguistic Infrastructure*, where not only the linguistic experts can participate. It is already proved by a number of projects that lexicon building and maintenance can be achieved in a cooperative way. We claim that the field of LRs and LT is mature enough to open itself to the concept of collaborative effort of different sets of communities (e.g. spoken and written, LT and SW, theoretical and application oriented).

3.1.1. Open and distributed architectures for LRs and LT, interoperability, GRID technology

A new paradigm of R&D in LRs and LT is emerging, pushing towards the creation of open and distributed linguistic infrastructures for LRs and LT, based on sharing LRs and tools. It is urgent to create a framework – both technological and organisational – that enables controlled and effective cooperation of many groups on common tasks, adopting the paradigm of accumulation of knowledge so successful in more mature disciplines, such as biology and physics. This implies the ability to build on each other achievements, merge results and have them accessible to various systems and applications. This is the only way to make a clear leap forward. This means emphasizing interoperability among LRs, LT and knowledge bases. Standards are again unavoidable.

This may also mean application of *GRID technology* to tackle the problems of processing extremely large quantities of "facts and their relations", of development of unprecedented large-scale annotated LRs, and of their dynamic linking across many different sources. A difficulty and a challenge is how to coordinate different information sources.

A way to attain the optimisation of the process of production and sharing of (multilingual) LRs relies on a public and standardized framework ensuring that linguistic information is encoded in such a way to grant its reusability in different tasks and applications. The ENABLER [6] project promoted the compatibility and interoperability of LRs endorsing: i) ISLE/EAGLES (http://www.ilc.cnr.it/EAGLES96/isle/), for harmonisation of linguistic specifications, in particular for corpora and multilingual lexicons; ii) ISO TC37 SC4 WG4, to make European standards truly international Standards; iii) ELRA Validation Committee, for integration of standards in protocols for LR validation; iv) INTERA, for harmonisation of metadata descriptions; v) cooperation with Semantic Web communities, to encourage synergy between knowledge management/ontology and HLT/LRs.

3.1.2. Lexicons' integration and interoperability: concrete steps towards a cooperative model

The SW model of open data categories will foster LR integration and interoperability, through links to common standards. With the ISLE approach to lexical standards, and its definition of the MILE (Multilingual ISLE Lexical Entry) [8], new lexical objects can be progressively created and linked to a core set. An increasing number of linguistic data categories and lexical objects stored in open and standardised repositories will be shared and used by different types of users to define their own structures within an open lexical framework.

It will guarantee freedom for the user to add or change objects if that is deemed necessary, but will require an evaluation protocol for the core standard lexical data categories, and verification methods for the integration of new objects. This vision, enabled by MILE, will pave the way to the realisation of a common platform for interoperability between different fields of linguistic activity – such as lexicology, lexicography, terminology – and SW development. The lexicons may be distributed, i.e. different building blocks may reside at different locations on the web and be linked by URLs. This is strictly related to the adoption of SW standards (e.g. RDF metadata to describe lexicon data categories), and enables users to share lexicons and collaborate on parts of them.

In our group we have recently developed LeXFlow, an architectural and practical framework for dynamic semi-automatic integration of lexicons and LRs [9]. LeXFlow is a system – based on XML – that manages lexical workflows where the different agents can reside over distributed places, and thus enables new methods for cooperation among lexicon experts, through collaborative management on various lexicon operations.

4. Technical vs. organisational/strategic issues for a LR infrastructure

The approach to realise a true LR infrastructure requires the coverage not only of a range of scientific and technical aspects, but also organisational, coordination, strategic and political issues play a major – and maybe most critical – role, as was highlighted in the ENABLER project [10]. They in fact acquire a more and more decisive relevance with the growing maturity of the LR field. 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 linguistic information, open to the reuse and interchange of linguistic data.

We should promote together 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.

Linked to this idea, an important *Declaration on Open Access to LRs* was endorsed by all participants of an ENABLER/ELSNET Workshop held in Paris in 2003.

4.1. ELRA role in the field of LRs

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. Coordination should be established between EC and member states, and strategies should be drawn in order to ensure a proper balance of language coverage in Europe. To this end ENABLER and ELRA have adopted and strongly supported the BLARK (Basic LAnguage Resource Kit) concept [11].

A Linguistic Infrastructure intends also to contribute to the structuring and integration of the European Research Area, addressing problems such as the fragmentation of its research base and the weakness in converting R&D results into useful economic or society benefits. To this aim, we claim it is necessary to pool together and build on many

different, but related, initiatives both for Spoken and Written LRs.

International cooperation will be certainly the most important factor for a coherent evolution of the field of LRs – and consequently of HLT – in the next years. A report produced by ELDA [12] presents an analysis of several organisational frameworks, focusing on funding and organisational procedures to provide LRs. ELRA [13], as a promoter of infrastructures for LRs, has in its mission also production and validation of LRs and promotion of standards. The Unified Lexicon project [14] of the Production Committee, defining common standards for spoken and written LRs, aims at overcoming existing barriers among independently built spoken and written LRs. It is the first step to pave the way to innovative methods of tailoring and acquiring LRs starting from available repositories, based on individual requirements. It can be seen as a contribution to solving the current fragmentation of LRs, while capitalising on and reusing results from previous European and national projects and standardisations activities.

4.2. Cooperation among communities

Technologies exist and develop fast, but the infrastructure that puts them together and sustains them is still largely missing. For example, the absence of a specific HLT action line in the European FP6 means not so much a change in the funding scene, but – more dangerous – lack of opportunities to discuss meta-level issues on HLT, difficulty in designing common global long-term strategies, with the risk of being just opportunistic in R&D choices. While there is a pressing need of international research infrastructures for LRs and LT, of bodies where to discuss a broad research agenda, priorities and strategic actions for multilingual and multimedia LRs and LT. To achieve this, cooperation must be enhanced among many communities acting now separately, such as LR and LT developers, terminology, SW and ontology experts, content providers, linguists, humanists. This is one of the challenges for the next years, for a usable and useful "language" scenario in the global network. The implementation of the notion of open distributed infrastructures for LRs and LT could act as a major technological and organisational challenge around which synergies (with other communities) can develop, and can naturally lead to the creation of an International Forum where to discuss about strategies and priorities.

A warning is due: such a language infrastructure may turn into being inherently market driven, since the most widely used language portions may become the best developed and supported. This deserves serious reflection for the political implications.

The idea behind such (past and future) initiatives is to establish some sort of permanent coordination to build on parallel existing (national or international) initiatives. At the end everything is tied together, which makes our overall task so interesting – and difficult. What we must have is the ability to combine the overall view with its decomposition into manageable pieces. No one perspective – the global and the sectorial – is really fruitful if taken in isolation. A strategic and visionary policy for cooperation between various groups has to be debated, designed and adopted for the next few years, if we hope to be successful, but – inside this – a realistic and stepwise approach to solving well-defined and limited aspects must be adopted. To this end, the contribution of the main actors from the various areas involved is of extreme importance. This will be a must for our field to contribute, effectively and globally, to the big challenges of the 'knowledge-based society'. Some of the events of the last years are hopefully moving in this direction.

5. References

[1] Zampolli A., "Towards Reusable Linguistic Resources", *EACL 1991, 5th Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, 1991.

[2] Calzolari, N., "Lexical databases and textual corpora: perspectives of integration for a Lexical Knowledge Base", U. Zernik (ed.), *Lexical Acquisition: Exploiting on-line Resources to build a Lexicon*, Lawrence Erlbaum Associates, Hillsdale, NJ, 191-208, 1991.

[3] Calzolari, N., "An Overview of Written Language Resources in Europe: a few Reflection , Facts, and a Vision", Rubio, A., Gallardo, N., Castro, R., Tejada, A. (eds*.), Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, Granada, Vol. I, 217-224, 1998.

[4] Calzolari, N., Zampolli, A., "Harmonised large-scale syntactic/semantic lexicons: a European multilingual infrastructure", *MT Summit Proceedings*, Singapore, 358-365, 1999.

[5] Ide, N., Calzolari, N., "Introduction to the Special Inaugural Issue", *Language Resources and Evaluation*, Springer, 39(1):1-7, 2005.

[6] Zampolli, A. et al., *ENABLER Technical Annex*, Pisa, 2000.

[7] Calzolari, N, "Computational Lexicons: Towards a New paradigm of an Open Lexical Infrastructure?", G. Willée, B. Schröder, H.C. Schmitz (eds.), C*omputerlinguistik. Was geht, was kommt?. Computational Linguistics. Achievements and Perspectives*, Gardez!, Sankt Augustin, 41-47, 2002.

[8] 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, 194 pp., 2003.

[9] Soria, C., Tesconi, M., Bertagna, F., Calzolari, N., Marchetti, A., and M. Monachini. 2006. "Moving to Dynamic Computational Lexicons with LeXFlow". *Proceedings of LREC2006*, Genova, Italy, 2006.

[10] Calzolari, N., Choukri, K., Gavrilidou, M., Maegaard, B., Baroni, P., Fersøe, H., Lenci, A., Mapelli, V., Monachini, M., Piperidis, S., "ENABLER Thematic Network of National Projects: Technical, Strategic and Political Issues of LRs", *LREC 2004 Proceedings*, Lisbon, 937-940, 2004.

[11] 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.

[12] 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.

[13] Maegaard, B., Choukri, K., Calzolari, N., Odijk, J., "ELRA - European Language Resources Association. Background, Recent Developments and Future Perspectives", *Language Resources and Evaluation*, Springer, 39(1):9-23, 2005.

[14] Monachini, M., Calzolari, N., Choukri, K., Friedrich, J., Maltese, G., Mammini, M., Odijk, J., Ulivieri, M., "Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian". *Proceedings of LREC2006*, Genova, Italy, 2006.