CLARIN: Common Language Resources and Technology Infrastructure

Tamás Váradi, Peter Wittenburg, Steven Krauwer, Martin Wynne, Kimmo Koskenniemi

Hungarian Academy of Sciences (Budapest), MPI for Psycholinguistics (Nijmegen), Utrecht University (Utrecht),
Oxford Text Archive (Oxford), University of Helsinki (Helsinki)
varadi@nytud.hu, peter.wittenburg@mpi.nl, steven.krauwer@let.uu.nl,
martin.wynne@oucs.ox.ac.uk, kimmo.koskenniemi@helsinki.fi

Abstract

This paper gives an overview of the CLARIN project [1], which aims to create a research infrastructure that makes language resources and technology (LRT) available and readily usable to scholars of all disciplines, in particular the humanities and social sciences (HSS).

1. Introduction

The present paper intends to give a general introduction to the CLARIN research infrastructure project, a long-term pan-European effort to build a virtual, distributed research infrastructure to support researchers of all fields dealing with language based material (text, speech, multimodal media), especially Humanities and Social Sciences. Currently working on the preparatory phase, the CLARIN consortium, consisting of 32 partners from 22 countries, intends to build a federation of trusted archive centers that will provide resources and tools through web services with a single sign-on access. The paper will briefly describe what CLARIN is (Section 2.) and what it does not intend to be (Section 3.), what is the rationale (Section 4.) and the current state of the project (Section 5.). The main part of the paper (Section 6.) will focus on the objectives and the planned activities of the preparatory phase. Finally, Section 7. introduces the CLARIN consortium, containing a brief discussion of its organization, governance structure and budget issues.

2. The mission

The CLARIN infrastructure is based on the firm belief that the days of pencil-and-paper research are numbered even in the humanities. Computer aided language processing is already used by a wide variety of sub-disciplines in the humanities and social sciences, addressing one or more of the multiple roles language plays: as a carrier of cultural content and knowledge, as an instrument of communication, as a component building our identity and as an object of study per se.

Current methods and objectives in these disparate fields have a lot in common with each other. However it is evident that to reach the higher levels of analysis of texts that non-linguist scholars are typically interested in, such as their semantic and pragmatic dimensions, requires an effort of a scale that no single scholar could, or indeed, should afford. The cost of collecting, digitising and annotating large text or speech corpora, dictionaries or language descriptions is huge in terms of time and money, and the creation of tools to manipulate these language data is very demanding in terms of skills and expertise, especially if one wants to make them accessible to professionals who are not experts in linguistics or language technology. The benefits of computer enhanced language processing will become available only when a critical mass of coordinated effort is invested in building an enabling infrastructure, which will make the existing tools and resources readily accessible across a wide span of domains and provide the relevant training and advice.

Making resources and tools easily accessible and readily usable to scholars of all disciplines, in particular the humanities and social sciences is the mission of the CLARIN infrastructure initiative.

The purpose of the infrastructure is to offer persistent services that are secure and provide easy access to language processing resources. Our vision is that both the resources for processing language and the data to be processed is made available in usable formats and can be run over a distributed network from the user's desktop.

The CLARIN objective is to make this vision a reality: repositories of data will be equipped with standardized descriptions, language processing tools will be amended to operate on standardized data, legal and access issues will be resolved, and all of this will be available on the Internet using grid architecture. The nature of the project is therefore primarily to turn existing, fragmented technology and resources into accessible and stable services that any user can share or customize for their own applications. This will be a new underpinning for advanced research in the humanities and social sciences, a "research infrastructure".

3. What CLARIN is *not* about

The development of CLARIN is envisaged as a three stage process: preparatory phase, construction phase and exploitation phase. The preparatory phase imposes stringent constraints on the scope of what CLARIN can attempt to do out of EC support. A number of the points listed below, however, are possible to pursue out of national or other funding.

- It is important to remind ourselves that the current phase is not aimed at *building* the infrastructure, it is meant solely to prepare it, assessing the difficulties and removing any known obstacles to the construction of the infrastructure.
- Creating new resources on any significant scale is also out of the scope of the present stage. We intend to use

what exists already and make any adaptations necessary

- CLARIN does not intend to create applications *per se* except for highly specific purposes such as demonstrators.
- It is decidedly not the objective to focus on the major European languages. CLARIN is firmly committed to the principle that all languages are equally important.
- CLARIN is not oriented towards strengthening European industry. Our target audience are social sciences and humanities researchers although we don't necessarily want to exclude anyone.

4. The rationale for a European Infrastructure

The need for a pan-European effort is justified for the following reasons. Even where sufficient language resources and tools do exist, there is too much fragmentation in the field stemming from lack of coordination across languages and countries and the relevant resources or tools are difficult to find. At present the resources and tools have very limited interoperability which either leads to the wasteful reproduction of effort or frustration altogether. The resources and tools suffer not only from lack of visibility but also lack of sustainability. It is typical to find that valuable resources are left to lie about untapped because they do not receive the long-term archiving and curating effort that they would require. A common infrastructure would also mean pooling relevant expertise that may not exist in all countries. Tools can also be shared provided they are language independent and if not, they can still be ported to various languages.
Clearly, the cost, the expertise, the effort and the resources are of such a scale that most countries would not be able to bear them individually hence the need for a shared unified research infrastructure.

5. Where we stand

To avoid any misunderstandings: the CLARIN infrastructure described here does not yet exist. Even if one finds repositories of language data in most European countries, and even if some of them are technologically quite advanced there has never been an attempt in Europe to link the existing repositories across national frontiers and to interconnect them in such a way that the user see it as a single large scale facility (with at least virtually one single entry point) offering access to a broad variety of data and services.
Recently the European Commission has taken initiatives towards a long term roadmap for research infrastructures in Europe, explicitly including the infrastructure needs of the social sciences and humanities. This initiative, called ESFRI, has recently led to a report describing 35 essential research infrastructure proposals for Europe. This report, the ESFRI Roadmap[3] has now been taken up by the EU and by the member states with a view to possible future implementation in a three-stage process: Preparatory Phase, Construction Phase and Exploitation Phase. Last year the EU launched a call for proposals whereby the infrastructure projects selected for the ESFRI Roadmap were invited to submit a proposal for the Preparatory Phase for each of the envisaged infrastructures.
The CLARIN initiative sucessfully submitted a proposal and the project started work at the beginning of 2008. It will have a duration of 36 months, after which – if the project is successful – the Construction Phase will start. In the rest of this paper we will provide more information about our approach, and our activities in the first three years. The CLARIN consortium is led by Utrecht University, and it has 32 partners from 22 EU and associated countries. In addition there is a wider community of CLARIN members including over a hundred institutions with specific expertise in language resources, spread over 33 countries.

6. Objectives of the Preparatory Phase

According to the EC call for proposals the preparatory phase has to aim at bringing the project to the level of legal, organisational and financial maturity required to implement the project. As the ultimate goal is the construction and operation of a shared distributed infrastructure that aims at making language resources and technology available to the humanities and social sciences research communities at large, an approach along various dimensions is required in order to pave the way for implementation. We briefly describe the four main dimensions and the preparatory phase objectives for each of them.
First of all there is the funding and governance dimension. The aim here is to bring together the funding agencies in all participating countries (currently 22) and to work out a ready- to-sign draft agreement between them about governance, financing, construction and operation of the infrastructure.
Secondly there is the technical dimension. A language resources and technology infrastructure is a novel concept. Even if it will be based on existing and emerging technologies (grid, web services) there are no off-the-shelf blueprints for the architecture of such an infrastructure. The technical objective is to provide a detailed specification of the infrastructure, agreement on data and interoperability standards to be adopted, as well as a validated running prototype based on these specifications. The validation should cover the technical, linguistic aspects and user aspects alike (see below). The construction of the prototype will also help to make realistic cost estimations for the construction and exploitation phases.
The third dimension is the language dimension. For the validation of the specifications of the infrastructure and the proposed standards the running prototype will have to be populated with a selection of language resources and technologies for all participating languages. This population process will normally take place by adaptation and integration of existing resources to the CLARIN requirements although in a number of cases the creation of specific essential resources will be necessary as part of the CLARIN preparatory phase. It is estimated that for most of the circa 100 relevant European languages, even the basic resources are not yet available. It will be the task of the construction phase to fill the gaps identified in the preapratory phase.

The objective is to deliver a sufficiently populated and thoroughly tested prototype that demonstrates the adequacy of the approach for all participating languages, a prototype that can be used to bootstrap the construction phase.
The fourth and most important dimension is the user dimension. The intended users are the humanities and social sciences research communities.
In order to fully exploit the potential of what language technology has to offer, a number of actions have to be undertaken: (i) an analysis of current practice in the use of language technology in the humanities will help to ensure that the specifications take into account the needs of the humanities, (ii) the execution of a number of typical humanities projects will help validating the prototype and its specifications, (iii) less advanced sectors of the humanities and social sciences communities have to be made aware of the potential of the use of language resources and technology (LRT) to improve or even innovate their research, (iv) the humanities and language technology communities have to be brought together in order to ensure lasting synergies between the communities. The objective of this cluster of activities is to ensure that the infrastructure has been demonstrated to serve the humanities and social sciences users, and that we create a joint, informed community that is capable of exploiting and further developing the infrastructure.
Finally, a rich LRT domain as intended by CLARIN will inevitably include protected material and therefore we will have to build the necessary legal and ethical agreement patterns into CLARIN. During the preparatory phase we will need to develop a thorough understanding of these problems and need to work out first such patterns to prepare the construction phase. Agreements and licenses are needed for successful cooperation among the various actors and users of CLARIN, and for achieving and maintaining sufficient levels of trust. A network of agreements, licenses and auditing is needed to relate the actors to each other and to avoid or reduce risks incurred in possible violations of intellectual property rights (IPR) or basic ethical rules.

6.1. Construction and Exploitation Agreement

Even though a large proportion of the work to be carried out in this project has to do with the specification of the infrastructure, interoperability standards and an IPR framework, the main deliverable of this project is a single draft agreement between all participating countries and players to move on to the Construction Phase. This is by no means simple, as such an agreement will have to include a significant long term financial commitment: We envisage a construction phase of circa 5 years, followed by an exploitation phase that might easily cover up to 10 years. Other issues to be addressed include governance, management, and coordination of national programmes related to language resources, all of which are crucial for the success and sustainability of CLARIN.

6.2. The Technical Dimension

The underpinnings of the CLARIN infrastructure is conceived of as a federation of archives where the existing resources and tools are made available mostly in the form of web services. The preparatory phase has the task of working out the technical specification of the infrastructure, which will be tested in the form of a working prototype involving a number of strong centres across Europre. The prototype will be validated on a rich variety of languages (at least 20 in number), resources and services.

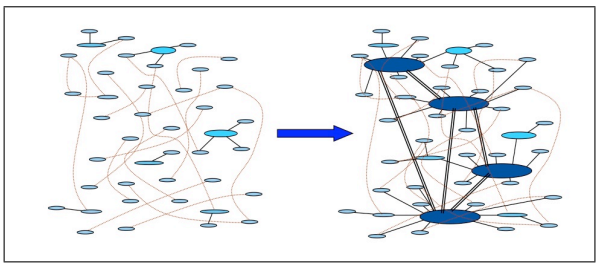

Figure 1: Federation of strong centres in CLARIN

For a federation of archives to work together and provide seamless access to the resources and tools, it is essential that the centres meet some technical requirements. In particular, they should join their metadata registry for resources and tools, provide support for virtual collections of resources located at different sites based on metadata and for combining services to build more complex operations. They should work out a unique way of referencing electronic resources and fragments in federation and they should provide a single sign-on system, all based on trusted and signed certificates.
Any proposal to organize the access to an integrated domain of European language resources and technology as suggested by CLARIN in the coming decades would fail, if it would not be built on the pillars of a new platform for authentication and authorization in a distributed scenario, which we will call "Authentication and Authorization Infrastructure" (AAI). AAIs are currently driven forward by big national and university libraries, by eLearning institutions and big publishers to replace the existing inflexible access granting solutions. However, such an LRT Federation needs to integrate with all efforts that are currently taken. The experiences from the small-scale DAM-LR[4] project are excellent, since they have shown us what the requirements are to set up an AAI, what the state of the technology is and what the requirements for the participating service providers are. As indicated in the following figure the LRT Federation is just part of a bigger game, i.e. it has to carefully synchronize with all relevant players and activities. Since CLARIN is a European effort, it has to talk to all national federations.

6.3. The Language Dimension

CLARIN intends to cover all languages spoken or studied in participating countries. Ideally, we'd like to see the same minimal coverage of basic resources and tools for all languages concerned. This lies behind the concept of *BLARK* (Basic Language Resources Toolkit) [5], which will be defined in the preparatory phase. For a well-documented language, the BLARK must consist of two types of lexica, one form-based and one lexical-semantic, a manually annotated corpus (treebank) and an automatically annotated

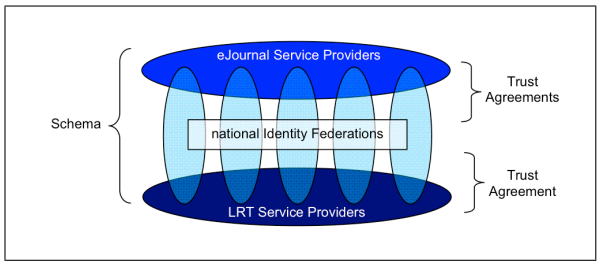

Figure 2: LRT Federation in its context

large-scale corpus as well as standard analysis tools like taggers parsers and speech recognition systems. Potential gaps in the BALRKs of individual languages will be identified by a coordinated action involving CLARIN members from the particular countries concerned.

The full implementation of BLARK, i.e. development of particular missing resources and tools lies outside the scope of the present phase. It can only be undertaken if other sources for funding are available. It will be left to the national decision boards whether they will fill the identified gaps in the BLARK.

Activities within the preparatory phase will concentrate on the comprehensive survey of resources and tools containing a detailed analysis of the structural and encoding characteristics of the resources and the interfaces of the tools that will serve to design a service oriented architecture. Based on this broad and detailed investigation, a comprehensive taxonomy of language resources and tools will be specified.

Interoperability of language resources is a key concern to CLARIN. To achieve this, we intend to draw on the use of existing standards defined by international organizations like the International Organization for Standardization (ISO)[6], The World Wide Web Consortium (W3C)[7] and the Text Encoding Initiative (TEI)[8] and, where necessary, develop proposals for new standards within these standardization organizations.

A vital aspect of handling language resources and tools is the inherent intellectual property rights (IPR) issues, which can be complex and delicate. While CLARIN is committed to the notion of open source tools and resources the IPR issues involved with existing and future non-open resources must also be tackled. We intend to work out a limited number of template licence agreements that would suffice for most of the common situations. Our proposal must respect relevant national legislation and address ethical issues as well.

6.4. The User Dimension

Support for Social Sciences and the Humanities (SSH) is the cornerstone of the CLARIN mission. It is important to realize that while the CLARIN community have a clear vision about the potential value of language technology in the area of Humanities and Social Sciences (along with any field, in fact, where intelligent analysis of text and speech is important) we do not have large-scale experience in exploiting the obvious benefits. The language technology community do not necessarily have a clear picture of the needs of the humanities researchers and, for their part, they may not be aware of the advantages of employing language technology in their research.

To meet the above challenge, we intend to undertake a large-scale analysis of past and on-going SSH projects and a scoping and impact study collecting data through traditional survey techniques as well as up-to-date cybertechnology. The most practical means of gaining direct experience about user needs, data and methods used in SSH research is through actual collaboration with SSH colleagues in some well chosen areas. Actual collaborative work with a handful of publicly invited projects would enable us to assess the technical, methodological, organizational etc. requirements involved in serving the SSH field in the later phases of CLARIN. A key component in the envisaged CLARIN support to the selected projects and to the SSH community in general is the advisory and counselling activity that is dispensed through virtual help-desk centers.

Another important strand of planned activities relate to building links with the SSH communities, exploring the potential stakeholders in the CLARIN mission, raising awareness and promoting the use of language technologies and resources. To reduce redundancy and optimally exploit available expertise and resources this activity will be carried out in close collaboration with the DARIAH project[6].

7. The CLARIN Consortium

7.1. Selection critera

The CLARIN consortium consists of 32 partners from 22 countries. At present count, the overall CLARIN community consists of a little over a hundred members. CLARIN members were encouraged to form a national network with one centre acting as national CLARIN representative. All the national representatives that could present a letter of support from their national funding agents were invited to join the CLARIN consortium. In addition, a number of CLARIN members were selected because their expertise or their assets in tools and resources made their contribution indispensable to the success of the CLARIN project.

7.2. Budget issues

The increasing number of partners were not matched with an equally expanding budget – on the contrary, the proposed project budget was severely cut back. This resulted in a great number of partners being allocated what amounts to a merely token person month share in the overall effort. However, even in the preparatory phase, CLARIN cannot rely on funding by the European Commission alone. There is plenty of scope for national contribution as regards development of new national resources and tools, funding of participation of non-partner members in CLARIN working groups (see below), running one of the main hubs in the future federation of archives etc.

Along with all other infrastructure projects, EC support for CLARIN extends to the preparatory phase only. The expectation is that the construction and the operation phase of the project will be financed entirely by the participating member countries. In the light of this prospect it is all the more important to make national funding agencies aware that the customary fund-matching principle cannot be applied to a

project like CLARIN with so many participants representing a country or a language having such a small share in the work (in person months) without jeopardizing the success of the whole enterprise.

7.3. The Structure of CLARIN

7.3.1. Work Packages

Work in CLARIN is organized around 7 work packages dealing with the following areas:

- WP1: Management and coordination
- WP2: Technical aspects of infrastructure
- WP3: Humanities overview
- WP5: Language resources and technology overview
- WP6: Dissemination
- WP7: IPR issues
- WP8: Construction and exploitation agreement

7.3.2. Working Groups

Most tasks are executed in working groups. The idea of working groups allows participation of non-partners in CLARIN project work. Working groups consists of project partners, the so-called *core WG members*, who are contractually bound to do the work and *associate WG members*. Participation in working groups is open and voluntary for associate working group members, compulsory for core working group members. The contribution of associate working group members is seen as essential, not just to create a critical mass of involvement in the project by the CLARIN community but as a means of safeguarding the concensus that is required within the community. Standards, for example, cannot be imposed by a small group of experts.

7.3.3. Governance

Operative work is coordinated by the Executive Board (EB), consisting of the seven work package leaders plus a special representative to liaise with the Humanities Community (especially the DARIAH sister project). The Work of the EB is helped by three Boards:

- *The Scientific Board* consisting of leading language technology experts delegated by the national funding agencies (typically at the suggestion of the national CLARIN community)
- *Strategic Coordination Board* consisting of policy makers, stakeholders nominated by the national funding agencies
- *International Advisory Board* consisting of leading international figures from the field of language technology.

Each year, the CLARIN community will hold a Member's Meeting at which the wide CLARIN community will be invited to discuss strategic issues with the Consortium. The governance structure of CLARIN is displayed in Figure 3.

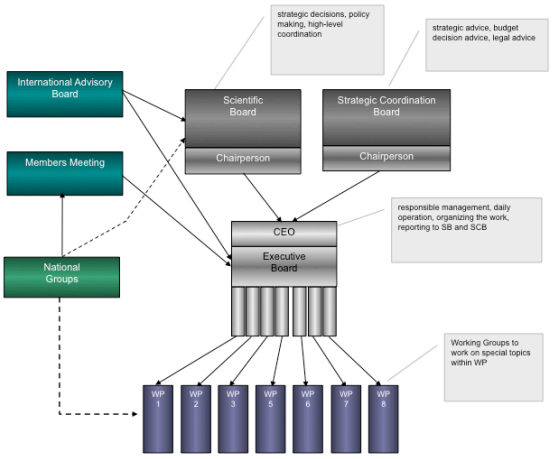

Figure 3: The governance structure of CLARIN

8. Concluding remarks

CLARIN has still a long way to go but it offers an exciting opportunity to fully exploit the achievements of especially language (and even speech) technology over the last decade to the benefit of communities that traditionally do not maintain a close relationship with human language technologies. Contrary to many EU programmes the main beneficiaries of this project are not expected to be the big ICT-oriented industries or the bigger language communities in Europe: CLARIN addresses the whole humanities and social sciences research community, and it very explicitly addresses all languages of the EU and associated states, both majority and minority languages, both languages spoken and languages studied in the participating countries.

9. References

[1] http://www.clarin.eu
[2] http://cordis.europa.eu/esfri
[3] http://ftp.cordis.europa.eu/pub/esfri/docs/esfri-roadmap−report−26092006en.pdf
[4] http://www.mpi.nl/DAM-LR/
[5] http://www.elda.org/blark/
[6] http://www.iso.org/
[7] http://www.w3.org/
[8] http://www.tei-c.org/
[9] http://www.dariah.eu/