Pregledni članak

CLARIN: Common Language Resources and Technology Infrastructure

Steven Krauwer
Utrecht University – Faculty of Humanities
CLARIN/ELSNET/UiL OTS
Trans 10, 3512 JK Utrecht, NL
Phone: +31 30 2536050, fax: +31 30 2536000
steven.krauwer@let.uu.nl
http://www-sk.let.uu.nl/

Summary

This paper gives an overview of the CLARIN project, which aims at the creation of a common language resources and technology infrastructure in Europe, to serve the humanities and social sciences research communities.

Key words: available language resources, technology infrastructure, humanities, social sciences, access.

The mission

The CLARIN [1] mission is to create an 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). In our age we are presented by many challenges as we deal with language in electronic formats, in spoken, written, and multimodal forms, as a carrier of information, as an object of study, and otherwise. The volume of texts and recorded spoken texts is enormous, and it is growing exponentially. The sheer size of this material makes the use of computer-aided methods indispensable for many scholars in the humanities and in neighbouring areas who are concerned with language material.

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 carrier of cultural content and knowledge, instrument of communication, component of identity and object of study. 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 become available only when a critical mass of coordinated effort is invested in building an enabling infrastructure, which can then provide services in the form of provision of all the existing tools and resources as well as training and advice across a wide span of domains. Making resources and tools easily accessible and usable is the mission of the CLARIN infrastructure initiative.

The technological challenge

The purpose of the infrastructure is to offer persistent services that are secure and provide easy access to language processing resources. As language, speech and vision technology improve, it should be commonplace to carry out tasks such as: *“summarize Le Monde from 11th March 2007” “list all uses of ‘enthusiasm’ in 19th century English novels written by women”, “find all video clips of Tony Blair on the BBC in 2007”.* But without the proper infrastructure, these technologies to make these tasks possible will only be available to a few specialists. At present one needs to find an appropriate program (to do translation, summarization, or extraction of information, etc.), download the program, make sure it is compatible with the computer that will execute the program, understand the form of input it takes, download the data (e.g. novels, newspapers, corpus, videos), and convert them to the correct format for the programs, and all this before one can get started. For most researchers outside of computer science, at least one of these tasks will be an insurmountable barrier. Our vision is that both the resources for processing language and the data to be processed be 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 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”.

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 to the user they present themselves as a large scale facility with (at least virtually) one single entry point offering access to a broad variety of data and services.

Unlike in fields such as nuclear physics, climate research, environmental research, energy, space research and many others, where shared infrastructures with transnational access are quite common, nothing of this kind has ever existed for the humanities and social sciences. Only 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 [2], has recently led to a report describing 35 essential research infrastructures for Europe. This report, called the ESFRI Roadmap, 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. As a first step the EU has launched a call for proposals whereby the selected infrastructures were invited to come up with a proposal for a Preparatory Phase for each of the envisaged infrastructures.

The CLARIN infrastructure is one of them and the contract negotiations between the participants and the EC are now ongoing. The start-up of the project is expected at the end of this year or early next year. It will have 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 31 partners from 22 EU and associated countries. In addition there is a wider community of CLARIN members, ca 90 institutions with specific expertise in language resources, spread over 33 countries.

Objectives of the Preparatory Phase

According to the EC call for proposals for the preparatory phase it 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 (now 22) participating countries 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 both technical aspects, linguistic aspects (see below) and user aspects (see further 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. 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) 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.

Technical Infrastructure

The overall objective of the CLARIN research infrastructure project is to turn existing, fragmented technology and resources into accessible and stable services so that users can use them the way they want it. This type of advanced infrastructure enabling eScience scenarios is comparatively new, we can only refer to a few projects, such as for example the Physiome project, that are working on similar goals facing similar degrees of complexity, i.e. before the real construction work can be done prototypical work needs to be carried out to study

- all aspects of applying modern architecture concepts such as Service Oriented Architectures to the field of LRT;
- the level of standardization that is necessary to implement workflow concepts making it more easy for users to use and combine language resources and technology components to solve their advanced problems;
- the various requirements to deploy a distributed service and server landscape in Europe with a broad geographic coverage;
- the problems of integration and interoperability in our domain with a more comprehensive approach.

In building the ingredients of a modern distributed digital infrastructure, we can rely on the activities of many different fields and standardization bodies such as W3C, ISO, OASIS etc. and we can look back on a long formation process in our discipline. Standardization in the area of linguistic resource management has a long tradition (EAGLES, ISLE, MILE, TEI) that recently led to the formation of ISO TC37/SC4, for example. However, the level of unification and the range of acceptance achieved is not yet sufficient to implement a service oriented architecture that would guarantee the required degree of integration and interoperability. Nonetheless, the conformity achieved already now in the field will allow us to take the next steps towards automated workflows. In contrast to earlier attempts CLARIN has the broad coverage of institutes and well-known experts that is needed to overcome the barriers that we are faced with.

Building an infrastructure offering a rich domain of services requires work at different levels. All of them will be based on the fast European network and the goal must be to build a foundation rich and powerful enough to generate new types of complex eScience applications. Grid-like functionality will ensure that communication occurs between trusted servers, that all resources in the domain have unique and persistent identifiers, that authentication and authorization is working seamlessly in distributed scenarios, that users of organizations participating for example in national identity federations are accepted with a single identity etc. We need a network of powerful centres that can offer stable and highly available services of a great variety. These ranges from archiving services allowing others to store data resources with a guarantee of long-term accessibility to advanced ontology services that offer widely accepted and well defined domain concepts. These centres need to offer portals to register all types of resources (data, tools, knowledge components) such that they can be accessed and interpreted by humans and algorithms. Standards need to be worked out that are flexible enough to cope with resource type differences and sub-discipline terminology and that enable automated workflows. Services will be needed that help to overcome the syntactic and semantic differences that we are

faced with when applying a certain tool to a certain resource or when concatenating two tools to a new more complex operation.
For all these different layers different steps are at hand. The middleware technology to establish a federated domain of data repositories has been tested already by projects such as DAM-LR [3]. Here the next step is to test its scalability in a pan-European network with national hubs and to train young people to fully understand and to maintain such basic infrastructures. At other levels such as for achieving structural and semantic interoperability still much analysis and specification work needs to be done to come to a set of generic frameworks, before adaptation, construction and implementation work in large quantities can be carried out. Still for a number of services such as for the ISO TC37/SC4 data category registry for example we have to create prototypes to test out principles such as high availability by distributed and synchronized services and automatic workload distribution.

Landscape of Services and Centres

The CLARIN infrastructure will establish a rich landscape of services and centres. As indicated above, the types of services are very heterogeneous and they will be hosted by certain centres. Centres that want to offer specific services at the European level need to fulfil a number of criteria such as long-term support by funding agencies, competence and cost effectiveness. In addition, centres will be selected based on the criterion of geographical distribution. A set of open criteria will be published together with the specification of the services. Interested parties can apply for taking over these services at the European level and will be selected by means of an evaluation process.

Activities

We will do theory-based analysis work by making use of the results that have been achieved in many standardization projects (EAGLES, ISLE/IMDI, TEI, ISO, W3C, OASIS, OAI/OLAC, OAIS, etc) and related initiatives and projects (OGF, Internet2, D-Grid, HAKA, EGEE, TERENA, DEISA, DAM-LR, etc).
We will do design studies with a limited vertical and horizontal coverage to deeply understand the problems that need to be solved. The vertical studies include various layers ranging from the basic grid services to the integration applications, the horizontal studies investigate one layer and the interplay and scalability issues at a pan-European level. Driven by top-down decisions prototyping work needs to be done to test out elements of the infrastructure. We will keep a close and continuous eye on the needs of our customers, i.e. the humanities and social sciences research community. To meet the needs and requirements of our users prototypical infrastructure elements - be it services or centers - will be set up not only to demonstrate the functionality of the design choices, but also to learn from the experiences and to make realistic effort and cost estimates. We will work closely together with IPR experts to specify the types of

agreements that will be necessary from a purely technological point of view on the one hand and to understand the requirements from IPR considerations on the infrastructure on the other hand.

User needs

We will start out analyzing past and existent projects that were carried out in the domain of the humanities such as the Bricks project to get a deeper understanding about the concrete use of LRT in the humanities and about the state of technology in that area. We will also stimulate a few new concrete projects to help pointing to gaps on the one hand and to develop requirements. The projects to be carried out can be extensions of existing or finished projects to achieve compliance with CLARIN or completely new ones tailored to test the emerging infrastructure concepts. To gain a detailed understanding of the requirements for LRT infrastructure in the humanities domain the humanities projects need to be selected very carefully, i.e. criteria need to be set up that in particular include chains of LRT operations, multilingual scope and addressing the increasing interest of young people in multimedia presentations. CLARIN is committed to boosting humanities research in a multicultural and multilingual Europe, by facilitating access to language resources and technology for researchers and scholars across a wide spectrum of domains in the humanities and social sciences (HSS). To reach this goal we will establish an active interaction with the research communities in HSS and to contribute to overcome the traditional gap between the Humanities and the Language Technology communities. We will achieve these goals by gathering information about: potential collaborators in network building, the current impact of LRTs in these fields, and user needs.

LRT Overview

We will collect detailed information about existing language resources and technology, specify and prototypically implement representational standards and strategies for achieving interoperability, and validate the proposed technical standards and service specifications of the CLARIN infrastructure. It is obvious that much effort will be required to integrate the speech/ multimodality resource and technology sub-community which is not yet well-represented in CLARIN and that strong links need to be established with the community working on ontological issues.

The term language resource subsumes the whole range of linguistic data types such as text, speech and multimodal corpora, lexica, treebanks, typological databases, grammars, ontologies, schemas, and the term language technology covers a wide range of processing and annotation components such as taggers, parsers, semantic extractors, manual annotation tools, speech alignment tools, etc. The survey needs to include a detailed analysis of the structural and encoding characteristics of the resources and the interfacing and import/export characteristics of the tools that will serve to design service oriented architecture.

Based on this broad and detailed investigation a comprehensive taxonomy of language resources and tools will be created which will be the basis for the classification and standardization work.

In addition, a Basic Language Resource Kit (BLARK [4]) for the major languages will be specified and gaps for individual languages will be identified by a coordinated action involving the CLARIN members from the various countries. This will include criteria for the quality assessment of resources and tools, since many resources and tools lack basic characteristics to make them effectively available in a service oriented architecture serving the humanities needs. It will be left to the national decision boards whether they will fill the gaps that are identified. Depending on the requirements some new resources and tools may be developed already in the preparatory phase with high priority.

Five activities will be in the focus: (1) investigating all aspects that have to do with the integration of resources and tools into the infrastructure; (2) studying the problems of interoperability in detail; (3) studying usage scenarios including chains of operation in detail; (4) creating missing resources and tools where they are required for the success of the preparatory phase; (5) validate the specifications and the prototypical implementations.

IPR Issues

Despite the clear commitment of CLARIN to open access and open source principles for all resources and code developed in CLARIN, legal and ethical aspects cannot be ignored. A rich LRT domain as intended by CLARIN will be bound to 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. CLARIN may also include new business and accounting models. The opportunities and threats of the inclusion of commercial services have to be studied carefully in order to understand the positive and negative effects for the CLARIN infrastructure that will be constructed. We will have to create sufficient model agreements for the operational prototype and test beds that will be built in CLARIN during the preparatory phase and to find out what their coverage and level of acceptance is. The work can build on earlier work such as in the LE-PAROLE project whose licensing scheme consisting of four parties can be adapted to the CLARIN situation. Also for distributed authentication and authorization scenarios CLARIN can build on experience made in a number of national identity federations and in the DAM-LR project where the goal was to establish a federation of archives. Keys to all these federations are trusted organizations that can offer (1) a reliable identification of users and (2) a reliable certification of the user reading the license and having signed it. Methods for auditing the reliability of trusted organizations must be put in place and provided to the resource collectors, for routine periodic re-assurance of trust. Since CLARIN wants to be open to guest researchers worldwide it has to be analyzed how the strict rules can be applied. Only reliable distributed authentication schemes open the way to providing licenses for large numbers of users to access materials and tools, and to migrating legacy language resources into wide use within CLARIN.

A number of additional issues need to be taken up such as (a) licensing templates for language materials and tool usage, (b) establishing model licenses to be used for new CLARIN materials, (c) migrating legacy licenses into CLARIN, (d) licensing Templates for Language Technology Software and (e) a common code of conduct covering widely agreed ethical rules.

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 by no means simple, as such an agreement will have to include a significant long term financial commitment: We envisage a construction phase of ca 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. None of them are hot topics from a scientific point of view, but they are crucial for the success and sustainability of CLARIN.

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.

Acknowledgements and references

This paper is an abridged version of a document written by Peter Wittenburg (Max Planck Institute, Nijmegen), Tamas Varadi (Hungarian Academy of Sciences, Budapest), Martin Wynne (Oxford Text Archive, Oxford), Kimmo Koskenniemi (University of Helsinki) and Steven Krauwer, with contributions by many other members of the CLARIN community.

References

http://www.clarin.eu
http://cordis.europa.eu/esfri/
http://www.mpi.nl/DAM-LR/
http://www.elsnet.org/blark.html