CLARIN: Where We Stand and Where We Need *Your* Input

Steven Krauwer
Utrecht institute of Linguistics UiL-OTS
Trans 10, 3512 JK Utrecht, Netherlands
s.krauwer@uu.nl

Summary

In this paper we give a brief overview of what the CLARIN Research Infrastructure is, and where we stand in the process towards its construction. We present the current CLARIN position on a number of issues and we identify a number of challenges, highlighting the ones where we feel more input from our communities, both users and providers, is still needed.

Key words: research infrastructures, language resources and tools, humanities, social sciences

What is CLARIN

CLARIN is the short name for Common Language Resources and Technology Infrastructure. This is one of the proposed research infrastructures for Europe that have been selected by the ESFRI Roadmap process as a candidate for inclusion in the European research infrastructure landscape [6].

The objective is to create a European federation of digital archives containing language-based material (e.g. text and speech corpora, dictionaries, language descriptions, multimodal data, etc, etc) and tools. This federation should provide our target audience, which consists of scholars in the humanities and social sciences, easy access to data and tools, independent of location. In addition to this CLARIN will also provide access to language and speech technology tools through web services, so that –ideally- all these tools can operate on all the data types they were designed for, irrespective of location and origin of data and tools.

In the CLARIN philosophy all languages spoken or studied in the participating countries are equally important, irrespective of size or commercial potential. The CLARIN infrastructure should eventually cover all EU and associated countries. More information can be found on our website [1].

To illustrate what CLARIN wants to achieve we give some examples of what the researcher should be able to ask:

- give me digital copies of all contemporary documents that discuss the Great Plague of England (1348-1350)

- give me all negative remarks about Islam or about soccer in the 2008 proceedings of the European Parliament
- find TV interviews that involve German speakers with a Spanish accent
- summarize all articles in Le Figaro of August 2009 about Mr. Barroso – in Polish

Some of these examples may still sound futuristic, but others can already be realized on the basis of existing technology.

Who are CLARIN

At this moment the CLARIN infrastructure as such doesn’t exist yet, but there is already a large and active community working on its design and construction. With financial support from the EC (grant EC-FP7-212230, under the FP7 Capacities programme) a consortium of 33 partners in 23 EU and associated countries is now working on the CLARIN Preparatory Phase Project, aimed at defining the future infrastructure and lining up the funding agencies and other stakeholders in the participating countries. In addition to that over 140 other institutions from 32 countries are actively involved in laying the foundations for the infrastructure (see [2] for a full list). The large majority of the participants are academic institutions or data repositories. Contributions from these participants to the project typically consist of data, technology and expertise.

In the Balkans region we have two participants in Croatia (University of Zagreb as a consortium partner [contact is Marko Tadic], and the Institute of Croatian Language and Linguistics as a member [contact is Damir Cavar]), and we have 5 participants in Bulgaria, 1 in Greece, 6 in Romania, 1 in Serbia, 2 in Slovenia, and 2 in Turkey.

What is the time schedule

When I spoke at the INFuture conference in 2007 we had just been informed by the EC that our Preparatory Phase Project had been approved and that we could go ahead early 2008. On January 1st 2008 the preparatory phase started, and it will last until the end of 2010. The EC contribution to this phase is 4.1 M€. Governments from participating countries have contributed ca 14 M€ to this phase or to parallel, related projects. Contributions range from 50 K€ to 5 M€. If all goes well the next phase, the Construction Phase, will start on Jan 1st 2011. This phase will last until 2014 and will be used to build the infrastructure. Contrary to the construction of physical infrastructures such as space telescopes or particle accelerators we anticipate that the construction of CLARIN will be an evolutionary process, so that we expect to be able to deliver (initially limited) services already after the first year. We expect the cost of this phase to be ca 100 M€. Funding for this phase will have to come from the national governments in the participating countries. This looks like an awful lot of money but it has to be kept in mind that (i) this is not all ‘new’ money because many of the

existing operations by digital archives and language technology centers can be included in CLARIN at no or little extra cost, and that (ii) with 23 countries and a duration of 3 years the actual average cost per country per year will be less than 1.5 M€ (and even less per language). The Exploitation Phase, when the infrastructure will be fully up and running, is expected to start at the latest in 2015, but as we said before, we expect initial services to start much earlier than that. As the digital world is very dynamic we expect that even during the Exploitation Phase CLARIN will go through a continuous evolution process and take up new technologies as they emerge. At this moment we estimate that the total cost until 2018 will be 146 M€, but more precise estimations can be made towards the end of the current Preparatory Phase.
As part of our design activities we are now in the process of building a small experimental prototype, but this is not intended for end users. It will help us to check soundness and consistency of our design.
To what extent and by when users will be able to start using the services will mainly depend on what is going to happen in the various countries: every country is responsible for its own content and its own language. As some countries need more time for their internal preparations we expect some countries to start later than others.

Main challenges

We see a number of challenges ahead of us. We will discuss a number of them, not just because we want to tell you about what we are doing but rather as an invitation to join our discussions aimed at taking away the main obstacles.

Technical challenges

Interconnecting existing archives in such a way that the researcher who visits the archive will move seamlessly from one archive to the other and will be able to create virtual collections of documents without being bothered by differences in the ways different archives encode and describe their data is by no means an easy task. This type of task is not unique to CLARIN. In many countries libraries have interconnected their catalogues in similar ways and the customer doesn't see more than one catalogue. The technical problems that have to be solved in order to make this happen are far from trivial, but solvable. The real problem is the lack of standards in the way creators of digital data and owners of archives encode and describe their material.
What makes CLARIN unique is the service infrastructure that will allow researchers to use existing technology tools to retrieve, explore, analyse, exploit or transform their data. The challenge for CLARIN is not to design or build the tools. This happens in other programmes, aimed at technology development. The real challenge is to make tools work on data collections different from the ones they were designed for, and to ensure that they can be chained together to perform more complex operations. Here too the real problems are not techno-

logical in nature, but follow from the fact that every tool builder has his own ideas about the input on which it should operate and about the format of the output. The only possible way to address this problem is through better standards: if tool developers can agree on using a limited number of agreed upon standards the tools and the data can be combined just like Lego bricks. To get a quick overview of the CLARIN position on a number of technological (and linguistic) issues you can check our Short Guides [7].
Standards cannot be dictated or imposed: they can only be effective if they are based on best practice and widely supported by the user community. In CLARIN we have organized the discussions about standards in such a way that the whole community can participate, and through this paper we would like to strongly encourage the members of our community to join the discussion, if only to ensure that the standards that will eventually be supported by CLARIN are suitable for YOUR language and for YOUR research purposes. See our standardization action plan on [5] and join the discussions!

Linguistic challenges
CLARIN has the ambition to cover all languages relevant for the European research community, irrespective of market potential, status, or number of speakers. With the limited budget available, and the relatively small consortium (33 partners in 32 countries) there is no way we can ensure complete and adequate coverage for all languages. Broad consultation with the community at large outside the project consortium is necessary to make sure that the approach we adopt will fit all the languages that need to be covered.
We are also in the process of collecting as much information as we can about available language resources and technology and storing this information in our language resources technology registry, so that the existence of material (data collections and tools) can easily be discovered by researchers looking for facilities to support their research. The current registry is now accessible through our portal, the Virtual Language World [3].
Invitation: We invite and encourage the whole community to participate actively in this work, so that we can be sure that YOUR language, research interests and resources and tools will be covered by CLARIN.

Take-up
CLARIN’s target audiences are humanities and social sciences (HSS) scholars without technical background. Unfortunately in most of the HSS sub-disciplines there is very little tradition in making use of technological tools. There seems to be a wide gap between ‘the converted’ (as reflected by e.g. the annual Digital Humanities conferences organized by the ALLC [8], and the traditional pencil-and-paper scholars. Like in the case of standards the use of digital techniques can not and should not be imposed on people. It is our task to discover their potential needs and to make them aware of the possible benefits. This is mis-

sionary work that should be carried out on the work floor. Our invitation to you: talk to your analogue colleagues and show them the delights of going digital – and tell us about their needs.

Legal and ethical issues

Intellectual property rights (IPR) constitute a very hard problem, especially in the case of language-based material. We see three main issues here.

First of all we can observe that IPR legislation within Europe is far from uniform, which means that access to material –even if technically without any obstacles- may be constrained by the legislation of the country where data or researcher may be located. This is very hard to reconcile with the concept of the European Research Area, where there should be no obstacles for researchers and knowledge to move around. This legislation should be harmonized.

Secondly the CLARIN position is that data and other digital resources created with public funding (regional, national, or EU) should be completely free for other researchers to use (taking into account ethical considerations and a reasonable protection of the creator's interests).

The third problem is special for CLARIN: 're-purposed data'. This refers to data created for other purposes than research (e.g. novels made to entertain, newspapers and TV news made to inform, or recordings of telephone dialogues intended to communicate). There is a wealth of such material available, but very few creators of such data are willing to share it with the research community because they are afraid that giving it away might do damage to their interests. The CLARIN position is that legislation at the EU and/or national level should be adapted to ensure that such data can be used for research purposes and that at the same time the legitimate interests of the owners are sufficiently protected.

Within CLARIN we hope to be able to set up a light but secure licensing system, based on a small number of templates that should cover most of the cases, based on current best practice.

We urge all members of our community to communicate these messages to their legislative bodies at national and EU level and to participate in the modeling of the templates in order to ensure that these templates cover their needs.

Business models

Building and maintaining an infrastructure such as CLARIN costs money. Where should the money come from? The EU position is simple: research infrastructures are the responsibility of the national governments. The EU has contributed to the CLARIN Preparatory Phase, although even there the full responsibility to ensure that the design of the CLARIN infrastructure covers national needs lies with the national governments. If, e.g. Croatia or Slovenia or any other country do not participate actively (and at their own expenses) in the CLARIN design process no one will take care of the interests of their languages and their research communities. For the next phase, the so-called Construction

Phase no EC contribution is foreseen. It is not until the final phase, the Exploitation Phase, that a possible contribution from the EC (the running figure is up to 20% of the operational costs) is being discussed as an option for the 8th Framework Programme.
But apart from these global financing aspects there are other financial questions to be considered: who pays what to whom for what.
Expectations depend on your role in life (guess who says what):

- Everything should be available for free
- I want to be reimbursed for the extra effort to make my data and tools accessible through CLARIN
- I don’t want others to use my results to make a profit
- Funders should not pay for the creation of tools and data that can be bought on the market
- Funding infrastructures is primarily a national responsibility
- We fund you for now but we expect you to become self-sustaining in the future
- Creation of data and tools is the responsibility of the infrastructure

The current CLARIN position is the following:

- CLARIN is not a creator of digital data or technological tools: its role is to facilitate sharing and interconnecting existing and future resources and tools
- The creation and operation of the local (national) part of the infrastructure will be fully financed by the national authorities
- The overall coordination and management of the infrastructure are a joint responsibility (also financially) of the countries participating in the ERIC
- Standard use of the whole infrastructure for research purposes should be free to all research institutions in countries that have joined the ERIC
- For special services (through third parties) the extra cost may be charged to the user
- Research institutions from sites outside ERIC countries, or researchers with commercial objectives may be charged on a subscription or case by case basis

Given our commitment to the principle that results of public funding should be freely available to the research community we do not envisage facilities for e.g. university institutes to generate extra income through CLARIN by charging users for their services.

Shape
We see the future CLARIN infrastructure as a networked structure with one or more centers in most participating countries. These centers may be data centers or service centers (both with guaranteed 24/7 availability, guaranteed for a longer period of time), centers of expertise, and other centers (more loosely

connected to the infrastructure), as well as a small office to accommodate the organizational headquarters. We anticipate that all centers will be based on existing centers, which may have to extend their scope of activities or their service level. We do not anticipate any major investments in physical installations or network facilities.
In our documents we have specified the requirements for becoming a CLARIN center. If you are interested in hosting such a center we recommend that you first of all talk to your national CLARIN coordinator, and then read our documents about center types (see [4]).

Governance
The main challenge here is to find a legal form that allows 23 (or possible even more) countries to jointly build and operate an infrastructure distributed over all these countries, and to jointly fund and operate this infrastructure in a sustainable way (i.e. based on a firm long term commitment rather than on ad hoc grants for a couple of years).
The current CLARIN position is that we aim at the creation of an ERIC (European Research Infrastructure Consortium), which is a new type of international legal entity, created by the EC for the specific purpose of operating research infrastructures. Participants in such a consortium are governments (i.e. not research institutions), because only governments can make long-term commitments.
CLARIN will approach the relevant ministries in the participating countries with more details about this in the coming months.

Sharing

What can be shared through CLARIN? The answer is simple: you can use CLAROIN to share anything that (i) might be relevant for our target user community, that (ii) satisfies certain quality criteria, and that (iii) you are legally allowed to share. This can apply to data (raw or enriched), tools, programs, expertise, etc.
Why would one want to share at all? For researchers there are a number of possible motives: idealism, hope for fame, hope that others will share with you or simply because your funder tells you to share. For the funder the obvious motive to insist on sharing is of course that it will ensure a better return on their (i.e. the tax payers') investment.
Why would one not want to share? One reason might be that it may involve an extra effort (adapting it to representation or interoperability standards, creating metadata, writing documentation). Another might be that it is possible that others do brilliant things with your material, which you had never thought of doing. Or others might criticize it because it isn't as good as it should have been.
The CLARIN position here is that the key to this issue is in the hand of the public funding bodies: they should make sharing through CLARIN (including

the extra efforts it might require) a contractual obligation in their funding contract.

How can you share through CLARIN? The best way to do it is to register and deposit your material at one of the CLARIN centers. These centers will be especially equipped to handle your material, keep it accessible in a sustainable way on a 24/7 basis, and handle all the licensing that should ensure that only authorized people can put their hands on it.

Alternatively you can make it available in a more traditional way (e.g. through your university’s web servers), but it should be noted that this situation is far from ideal as universities tend to be extremely unstable and unreliable in this respect: they change URLs whenever they have contracted a new company to create a new web presence, and there is no guarantee that you material will remain accessible after your project is over or when you leave for another job or for retirement. Your resources are definitely better off in the hands of specialized digital repositories.

Concluding remarks

I have referred to a number of CLARIN documents (see links below), and you may wonder how you can get access to them. If your organization is a CLARIN member you can apply for an account on the member space of our website [9], which will give you access to all the documents produced and discussed in our project working groups. If they are not a member and if they qualify for membership you should ask them to join [2].

In this paper I have tried to give you a more or less up-to-date picture of where CLARIN stands, but it is important to stress that CLARIN doesn’t exist yet and is still full of challenges that need to be addressed and for which we urgently need your input, especially if you want to make sure that your own needs and requirements are taken into account. Be selfish and join CLARIN!

We urgently need better connections with the humanities and social sciences communities at large, because we don’t want CLARIN to be based on the needs and priorities of the language and speech technology community alone. Here too we need your help!

Finally, to avoid any misunderstandings or false expectations: CLARIN is not about content creation (which should be the responsibility of other national or EU programmes) but about sharing, and it aims at doing this by providing service-based access to what exists and to what will exist in the future.

Links

[1] The CLARIN website. http://www.clarin.eu (16.10.2009)
[2] List of CLARIN members and how to join. http://www.clarin.eu/members (16.10.2009)
[3] CLARIN’s Virtual Language World portal. http://www.clarin.eu/vlw (16.10.2009)
[4] CLARIN centers. http://www.clarin.eu/files/wg2-1-centers-doc-v8.pdf (16.10.2009)

[5] Standardisation action plan.
http://www.clarin.eu/system/files/private/Standardisation%20action%20plan-v8.pdf
(16.10.2009)
[6] ESFRI Roadmap. http://cordis.europa.eu/esfri/roadmap.htm (16.10.2009)
[7] CLARIN Short Guides. http://www.clarin.eu/documents/short-guides (16.10.2009)
[8] ALLC (Digital Humanities Conferences). http://www.allc.org (16.10.2009)
[9] Join a Working Group. http://www.clarin.eu/join-a-working-group (16.10.2009)