CLARIN: how to make it all fit together?

Steven Krauwer
Utrecht institute of Linguistics UiL-OTS
CLARIN Coordinator

Background

- ESFRI: EU initiative to identify essential research infrastructures for Europe in all areas of science (hard and soft)
- First report in 2006, with 35 candidate research infrastructures (the ESFRI Roadmap)
- EU member states to decide which ones to create and support
- CLARIN is one out of 5 selected infrastructures for Humanities and Social Sciences

What is CLARIN

- Common Language Resources and Technology Infrastructure (http://www.clarin.eu)
- Basic idea:
  - European federation of digital archives with language data and tools (text, speech, multimodal, gesture …)
  - with access to language and speech technology tools through web services to retrieve, manipulate, enhance, explore and exploit data
  - with uniform single sign-on access to the archives
  - target audience humanities and social sciences scholars
  - to cover all EU and associated countries
  - all languages are equally important

What should the user be able to ask?

- *give me digital copies of all contemporary documents in European archives that discuss the Great Plague of England (1348-1350)*
- *give me all negative remarks about Islam or about soccer in the 2009 proceedings of the European Parliament*
- *find TV interviews that involve German speakers with a Latvian accent*
- *summarize all articles in European newspapers of August 2010 about Estonia – in Lithuanian*
- *Show me the pronoun systems in the languages of Nepal*

Who are the people

- At this moment a core consortium of 36 partners in 26 EU and associated countries (and more to join)
- LV, LT and EE are all included
- Outside the consortium over 160 contributing institutions in 33 countries in Europe
- Mostly academic institutions active in language and speech technology, and a number of digital archives
- Contributions consist typically of data, tools, or expertise

Schedule & who pays

2008-11: *Preparatory phase*, funded by the EU (grant 212230, 4.1 M€) with additional funding from national governments

2011-13: *Construction phase* to be jointly funded by the national governments in the participating countries

2013-…: *Exploitation phase* to be jointly funded by national governments, possibly with some extra funding from the EU

Total estimated budget 2008-18: 146 M€

Do we really have to wait until 2013?

- First small experimental prototype during this preparatory phase, but no real end user services
- Already in the next phase (construction) we may gradually start operations in 2011-2012
- Every country responsible for its own content, no central funding for content creation foreseen
- What will be available (content and services) will depend on what countries and EC do
- CLARIN integrates it and makes it available via web services

Important features

- CLARIN is not about technology development or content creation, but aims at integrating what is available and making it accessible, BUT: without content (data and tools) no CLARIN!!!!!
- CLARIN is not oriented towards markets, but serves the Humanities and Social Sciences research communities
- CLARIN covers both historical and contemporary language material in all modalities
- CLARIN is interested in both linguistic data and its content
- CLARIN finds all languages equally important

What are the main challenges or obstacles?

- We look at a just few
  - technical
  - linguistic
  - take-up
  - legal
  - organisational

Main challenges
Technical

- Interoperability:
  - Interconnecting existing archives across Europe that may use very different ways to encode and describe data
  - Ensuring that existing language technology tools made for material in archive A will also work for material in archive B, and will work together
- Single sign-on access
  - Transnational scheme for access and authentication

Main challenges

Linguistic

- Linguistic challenges:
  - Ensure that all languages are sufficiently covered in terms of data and tools
  - Ensure that we know what exists and where to find it
  - Ensure that approach adopted fits for all languages
  - Needed: broad consultation (e.g. about standards) and verification (for each language)

Main challenges

Take-up

- Take-up by target audience:
  - aim at humanities and social sciences scholars
  - who have no technical background and who have very little tradition in using technological tools
- Special challenges:
  - discovering what they need
  - making them aware of the potential benefits of the infrastructure, e.g. to speed up or innovate their research

Main challenges

Legal and ethical

- Legal challenges:
  - making a light access and licensing system for the users
  - protecting owners' rights and interests
  - respecting national IPR legislation
- Special problems:
  - transnational access and diversity of national legislation
  - repurposed data (e.g. using novels or TV news for linguistic studies)
  - ethical & privacy considerations (e.g. use of recorded phone calls to build railway information systems)

Organisational challenges

Future shape of the infrastructure

Some features of the RI as we see it

- networked digital infrastructure with one or more centers in most participating countries:
  - data centers (24/7 availability, long term preservation)
  - service centers (24/7 availability)
  - centers of expertise
  - other centers (more loosely connected to the infrastructure)
- all based on or hosted by existing centers
- sustainability to be ensured by long term commitments from governments
- national consortia to be responsible for the creation of data and tools according to national programmes
- how to shape this financially and organisationally?

Structure of CLARIN

Three layers:

- Governed by CLARIN ERIC, an international legal entity which is a consortium of governments (not universities)
- Two operational levels:
  - Infrastructure level, consisting of centres (one or more per country, fully funded by own government), coordinated by ERIC
  - In each country a national consortium responsible for creation of data and tools compliant with CLARIN, nationally funded

State of affairs

Ongoing:

- Discussions with and between ministries about the creation of the ERIC for CLARIN
- Memorandum of Understanding to be signed next month
- ERIC application to be submitted end 2010 (EC approval needed to set up an ERIC)
- Expected to be up and running summer 2011 with initial consortium of those who are ready
- Other countries can join in later

What is going to happen now

- The CLARIN Preparatory Phase project will end June 30th 2011
- We hope/expect that the CLARIN ERIC will be up and running by then, to take over the responsibility and start building and operating the CLARIN infrastructure
- … but there is more

The SSH Infrastructures

- CLARIN is one out of five Social Sciences and Humanities Research Infrastructures: CLARIN, DARIAH, CESSDA, ESS and SHARE
- Joint project to be launched, addressing
  - Architecture
  - Data quality
  - Archiving
  - Shared access
  - Legal and ethical issues

Integration of existing data

- New EC project proposal for countries that are likely to join the CLARIN ERIC, aiming at integrating existing key resources (data and tools)
- Focus on
  - Language variation (geographic, social, historical)
  - News (written, audio, video)
  - Parliament records
- Collaboration with libraries
- Number of (funded) participants limited (max 10-15 countries)
- But open to others (on self paying basis)

International collaboration in this project:

- Network of regional endangered languages archives all over the world
- Reaching out to related language communities (e.g. Brazil)
- Reaching out to related initiatives in other parts of the world

Collaboration with other relevant initiatives

- META-NET and META-SHARE (see Georg Rehm's talk)
- Many common features (sharing language resources and tools)
- With different audience and objectives
- But often with the same players
- Lots of opportunities for close collaboration, formal agreement made

Concluding remarks

- CLARIN is not a project, but a long term endeavour, based on long term commitments at the government level
- CLARIN will fail if only a few countries decide to give their long term commitment
- CLARIN will fail if we don’t manage to reach the users
- CLARIN is cheap compared to other research infrastructures
- Your task: create a national consortium and talk tt your funders!

More info

- More info http://www.clarin.eu
- Read the CLARIN Newsletter!
- Next week SDH2010 in Vienna
- Institutions can join as CLARIN members during the current phase and participate in working groups
- This network will continue to exist, also after June 2011 (when the current preparatory phase project ends)