



CLARIN: going international – why, and what to do?

Steven Krauwer

Utrecht institute of Linguistics UiL-OTS

CLARIN Coordinator

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 ...)
 - target audience humanities and social sciences scholars
 - with uniform single sign-on access to the archives
 - with access to language and speech technology tools through web services to retrieve, manipulate, enhance, explore and exploit data
 - to cover all EU and associated countries

Who are the people?



- At this moment a core consortium of 36 partners in 26 EU and associated countries (and more to join)
- Outside the consortium over 160 contributing institutions in 33 countries in Europe
- Mostly academic institutions active in language and speech technology
- 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

2013-....: *Exploitation phase* to be jointly funded by national governments, possibly with some EC contribution.

Important features



- CLARIN is not about technology development or content creation, but aims at integrating what is available and making it accessible
- CLARIN is not oriented towards markets, but serves the Humanities and Social Sciences (HSS) research communities
- CLARIN covers both historical and contemporary material in all modalities
- CLARIN is interested in both language data and its content
- CLARIN finds all languages equally important

Why am I here?



- CLARIN is all about language resources and technology, couldn't exist without
- CLARIN is not a stand-alone incident but part of an evolving European infrastructure landscape
 - Diachronically: Eagles, Elsnet, Telri, Enabler, Dam-Ir
 - Synchronically: Elda/Elra, Flarenet, Metanet, Cocosda-Write, e-Infrastructures
 - Thematically: sister HSS infrastructures

Why think international?



- Research is global: ‘useful’ resources and tools can be found anywhere
- Many European infrastructure initiatives to bring players, resources and technologies together and to make them work together and share what they have
- But very little to support this on an international basis

My perception



- What we have
 - Many personal international collaboration contacts
 - Many institutional collaboration contacts
- All ok for doing research but not for infrastructure level cooperation

What we need



- Collaboration initiatives at the infrastructure level, I.e. between infrastructures
- They won't happen automatically
- Will be shaped differently from country to country
- Existing infrastructure initiatives in excellent position to act as a launching forum
- Coordination between them necessary to avoid fragmentation and duplication

Our priorities (1)



- Provide access to data and tools wherever they are, which presupposes
 - That their existence is known (unfragmented and properly maintained catalogs)
 - That they can be found and referred to (persistent identifiers, preservation)
 - That they are well-described (metadata)
 - That they can operate together (standards)
 - Simplification of the IPR jungle (licensing)

Our priorities (2)



- Ability to cross language barriers
- Ensure minimal coverage for all languages (BLARKs, BLARKettes)

Domain and audience specific:

- Specialized BLARKs
- Training, education, awareness
- Access through national research networks

My conclusions



- All priorities probably shared by the other infrastructure initiatives
- All lend themselves for international collaboration
- Some can be initiated by bottom-up actions
- Some require support from funders and policymakers
- All can be more effective if we act together
- Let's start, make a plan and do something!