Language Technology for the soft sciences: opportunities and challenges

Steven Krauwer

CLARIN ERIC / Utrecht University

Overview

- What is all this about
- CLARIN in a nutshell
- The dream
- The vision
- Phasing
- CLARIN ERIC
- The nightmares
- Action lines
- Concluding remarks

What is all this about

- Wealth of digital language data, spread all over Europe in archives, repositories, libraries
- Reflects human behaviour, communication, knowledge, culture etc
- Rich source of data, information and knowledge for HSS scholars (historians, philosophers, social scientists, …)
- In addition results of 30 years of European HLT efforts
- In brief: a great opportunity for HSS to innovate itself and to become world leaders, especially because of our multilinguality

BUT

- …….

What is all this about

BUT …

- How do HSS scholars know what data exists
- How can they get access to data from all over Europe
- How do they know what tools exist to retrieve, explore and exploit these data
- How do they know how to decompose their HSS research questions into sub-questions that can be answered by digital methods

OUR ANSWER:

- CLARIN: the Common Language Resources and Technology Infrastructure for the Humanities and Social Sciences

CLARIN in a nutshell

- Common Language Resources and Technology Infrastructure (http://www.clarin.eu)
- Basic idea:
  - European federation of digital repositories 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 archives and tools
  - target audience humanities and social sciences scholars
  - to cover all EU and associated countries
  - and all languages relevant for target audience

The CLARIN dream

- *give me digital copies of all contemporary documents in European archives that discuss the Great Plague of England (1348-1350)*
- *give me all negative articles about Islam or about soccer in the Slovenski Narod daily newspaper (1868-1943)*
- *find European TV news interviews that involve speakers with a Bavarian accent*
- *summarize all articles in European newspapers of August 2012 about OCR – in Polish*
- *show me the pronoun systems of the languages of Nepal*

The vision: the role of language

- Language is at the heart of many disciplines in the Humanities and Social Sciences (HSS), e.g.
  - as an object of study
  - as a means of human communication
  - as a means of human expression
  - as a record of our history
  - as part of one's cultural identity
  - as carrier of knowledge and information
- CLARIN wants to support them all
- Language and speech technology are part of this (e.g. in the form of computational linguistics or speech science) – but just a part!

The vision:
what CLARIN wants to offer

- CLARIN makes it possible for the researcher to find resources (metadata search), and to refer to them in a persistent way (persistent identifiers)
- CLARIN allows for content search in and across collections
- CLARIN offers access to web services and workflows to perform complex linguistic & content operations and visualisations
- CLARIN covers both historical and contemporary language material in all modalities
- CLARIN serves both expert and non-expert users
- CLARIN offers access to depositing and long term preservation services

Phasing of CLARIN

- Does CLARIN exist? Yes and no.
- 2008-2011: CLARIN Preparatory Phase Project, EC funded
  Goal: *designing the infrastructure technically and organisationally, and lining up the players*
- 2012-2015 Construction Phase, jointly funded by the participating countries, no EC funding
  Goal: *building the European infrastructure*
- 2015-…: Exploitation Phase, jointly funded by the participating countries, no EC funding
  Goal: *making and keeping it running, populating it, and ensuring that it follows new trends in technology and research*

CLARIN ERIC

- CLARIN ERIC is the governance and coordination body, but will not run or fund operational data services
- An ERIC is new type of intergovernmental legal entity, created by the EC, essentially a consortium of countries, with no end point
- CLARIN ERIC member countries pay a modest annual fee
- Countries will each set up a national CLARIN consortium, that will provide data and linguistic services and create data and tools
- It is up to the countries to decide how to shape and fund their CLARIN consortia and how to relate them to other activities at the national level (e.g. research programmes, digitisation programmes, etc)
- CLARIN ERIC established by the EC on Feb 29th 2012, with 9 founding members: AT, BG, CZ, DE, DK, EE, NL, **PL**, DLU
- More in the pipeline – but we want all European countries in!

What is so nice about ERICs?

- They are legal entities, not projects, which helps to make them more sustainable
- Members are governments, committing themselves for longer periods of time (min. 5 years)
- CLARIN ERIC is a sign of recognition by governments and EC of the importance of sharing language resources
- Closeness to funding agencies may help to enforce use of standards and sharing of data in projects they fund
- Good starting point for international collaboration as third countries can join or make collaboration agreements (e.g. through agencies or data centres)
- ERICs may submit proposals for EC funding

But: bulk of the funding dependent on funding mechanisms and cycles in participating countries

The CLARIN nightmare

- *give me digital copies of all contemporary documents in European archives that discuss the Great Plague of England (1348-1350)*
- *give me all negative articles about Islam or about soccer in the Slovenski Narod daily newspaper (1868-1943)*
- *find European TV news interviews that involve speakers with a Bavarian accent*
- *summarize all articles in European newspapers of August 2010 about OCR – in Finnish*
- *Show me the pronoun systems of the languages of Nepal*

The CLARIN nightmare, example1

- *give me digital copies of all contemporary documents in European archives that discuss the Great Plague of England (1348-1350)*
- "All" means from all countries and all archives, not just some archives in some (9) countries
- If contemporary docs exist in digital form at all they are probably pictures – how do we get access to the content? Is OCR doable?
- Can we rely on standardized metadata to find them?
- Many of the docs may be in Latin, can we handle that, and what about the other languages?
- How would a non-technical scholar know how to formulate this query?

The CLARIN nightmare

- Do HSS scholars realize at all that they should be interested in these things?
  - Some do, most don't; we should make an effort to show them the potential benefits of adopting these new methods
  - Showcases and visualisation tools are indispensable
  - Distinguish between lost and future generation
- Are the tools offered by language and speech technology the answers to the problems of HSS scholars as they see them?
  - Technologists have a strong tendency to offer more and better gearboxes to people who are just waiting for a bus with comfortable seats
  - Technologies that work for modern versions of big languages may not work for older versions of digitally less favoured languages
  - Use and adaptation of existing tools to specific HSS questions may always require intervention by technologically skilled people

CLARIN's answer: Action lines (1)

- Coverage: consolidate 9 members, reach out to others, 15 members in 3 years, 20 in 5 years
- Legal: common license templates promoted for new and legacy data, collaborate with others, talk to legislators about IPR, establish Access and Authentication for single sign-on
- Integration of data: standards action plan, tools for mapping, tools for curation; identify priority areas for cross border research
- Integration of services: interoperability, identify chainable services, work on showcases that convince potential users
- Preservation: identify at least 1 centre per country, work on change of culture, follow broader data initiatives; in 3 years all data and tools from funded projects deposited

CLARIN's answer: Action lines (2)

- Ease of access: Knowledge Sharing Infrastructure to support ease of access, awareness, training & support, curricula development, centres of expertise; Portal targeting different audiences; emphasis on interfaces and visualization
- Crossing borders: use language as vehicle to collaborate with other disciplines; inter-Research Infrastructure and international collaboration; explore industrial collaboration models
- Sustainability: demonstrate societal impact; review sustainability models; after 3 years vision and strategy

Concluding remarks

- There is still a lot of work to do – but it is good to realize that CLARIN is not a project: it has a start but no fixed end
- Legacy resources need to be upgraded, new resources should comply with community standards from the start
- Further development of language and speech technology for all languages (from big to small) is essential, but it should be kept in mind that proven technologies may not work for older variants of languages, and require adaptation
- Much effort needed to ensure adoption of digital methods in the humanities and social sciences (education, showcases, visualisation)
- Collaboration with the HLT community is crucial and we are looking forward to continuing the collaboration with META-NET that was already initiated during the CLARIN Preparatory Phase project.