A multilingual research infrastructure for the humanities

Steven Krauwer

CLARIN ERIC / Utrecht University

Overview

- Introductory remarks
- What is the problem
- CLARIN in a nutshell
- The dream
- The vision
- Phasing
- CLARIN ERIC
- The nightmares
- The gearbox syndrome
- Action lines
- Coming up
- Concluding remarks

Introductory remarks

- This is me using Bente Maegaard’s slot
- We believe in the same truths, but our perspectives may be different
- She was going to tell you about things that CLARIN has already done, I will tell you what I think you have to do!

What is the problem

- 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 the problem

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 Finnish*
- *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 in 6 sleepless nights

- *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 in 6 sleepless nights – night 1

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

1. "All" means from all countries and all archives, not just some archives in some (9) countries that happen to be in CLARIN
2. If contemporary docs exist in digital form at all they are probably pictures – how do we get access to the content?
3. Can we rely on standardized metadata to find them?
4. Many of the docs may be in Latin – can we handle that, and what about the other languages?
5. How would a scholar know how to formulate this query?
6. How to present results?

The CLARIN nightmare in 6 sleepless nights – night 2

- *give me all negative articles about Islam or about soccer in the Slovenski Narod daily newspaper (1868-1943)*

1. Not all old newspapers exist in digital form yet
2. Many digitized newspapers are just pictures – how can we analyze their structure, and do we have usable OCR to read them?
3. Topic and attitude extraction tools exist – but do they exist for Slovenian, do they fit together and will the same tools still be available in 5 years time?
4. What if the scholar does not read Slovenian?

The CLARIN nightmare in 6 sleepless nights – night 3

- *find European TV news interviews that involve speakers with a Bavarian accent*

1. Is any TV channel (public or commercial) willing to give us access to their TV news recordings?
2. Would we need permission from the Bavarians to analyse their speech at all?
3. Do we have Bavarian accent detection for other languages than German?
4. Are our accent detection tools good enough?
5. If they exist, do we have to pay to use them?

The CLARIN nightmare in 6 sleepless nights – night 4

- *summarize all articles in European newspapers of August 2010 about OCR – in Finnish*

1. How many European newspapers would give us access to their digital versions for research purposes?
2. Do we have good quality and sustainable topic detection at all for all languages?
3. Do we have summarizers for all languages (if we summarize before we translate)?
4. Do we have other tools than Google Translate to translate them?

The CLARIN nightmare in 6 sleepless nights – night 5

- *Show me the pronoun systems of the languages of Nepal*

1. Are field linguists willing to share their findings at all?
2. Are their results available in digital form
3. Are they described in a language independent form?
4. If so, is there any common structured format that can be used to extract data and to present their findings?

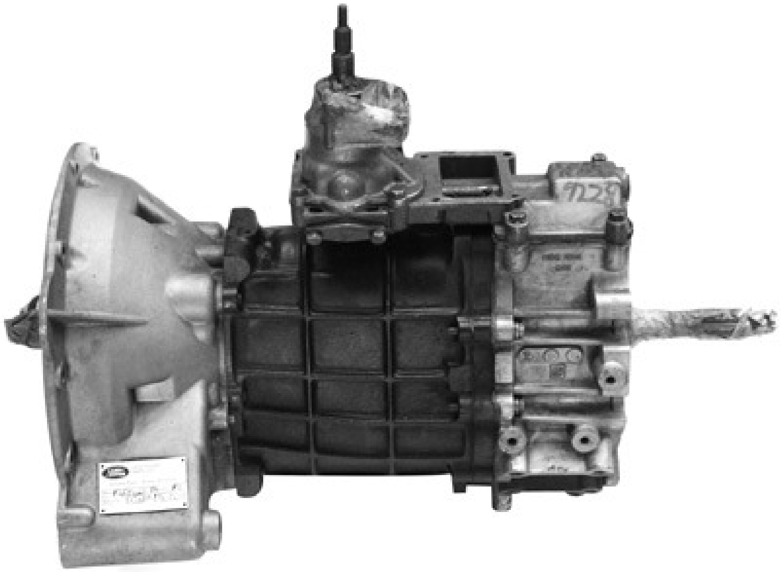
The CLARIN nightmare in 6 sleepless nights – night 6

- 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?
  - 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
- The gearbox syndrome

The gearbox syndrome

The gearbox syndrome

The gearbox syndrome

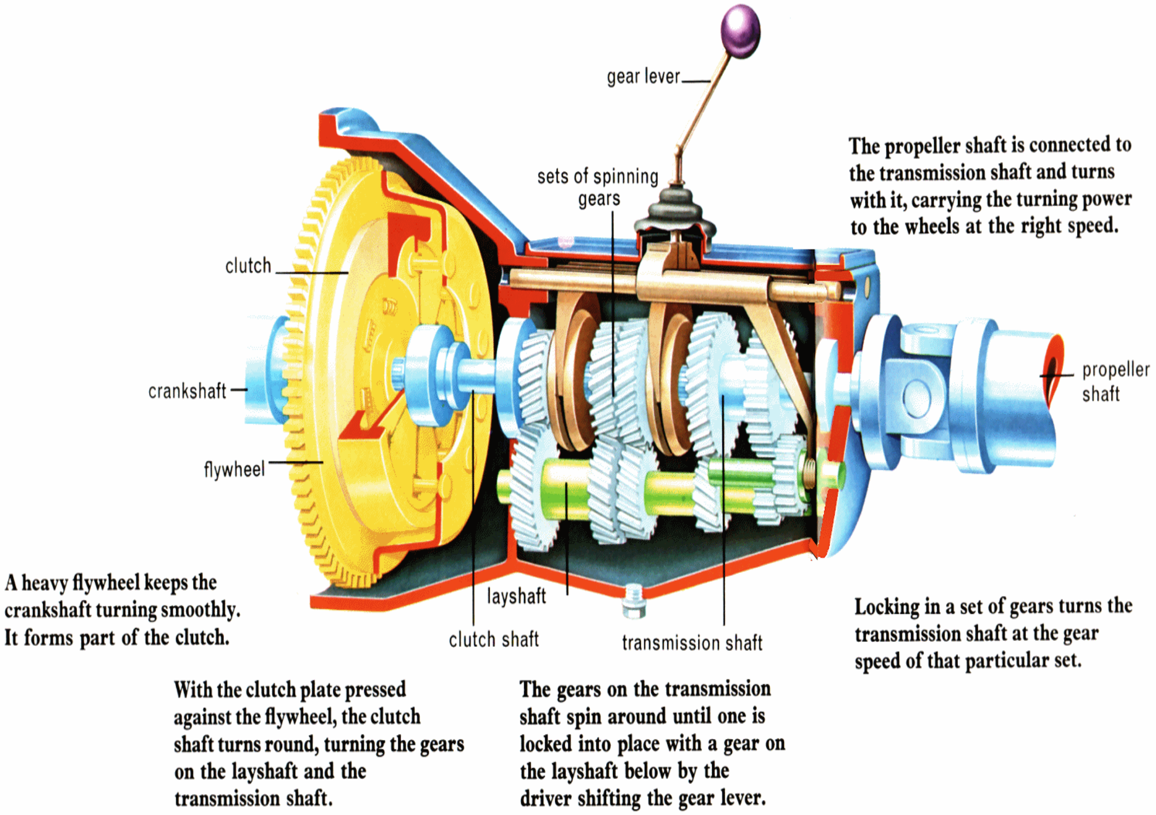


The gearbox syndrome

The gearbox syndrome

The gearbox syndrome

The gearbox syndrome



The propeller shaft is connected to the transmission shaft and turns with it, carrying the turning power to the wheels at the right speed.

A heavy flywheel keeps the crankshaft turning smoothly. It forms part of the clutch.

Locking in a set of gears turns the transmission shaft at the gear speed of that particular set.

With the clutch plate pressed against the flywheel, the clutch shaft turns round, turning the gears on the layshaft and the transmission shaft.

The gears on the transmission shaft spin around until one is locked into place with a gear on the layshaft below by the driver shifting the gear lever.

Where do we come in?

The gearbox syndrome explained

Humanities scholar with a problem, waiting for a solution

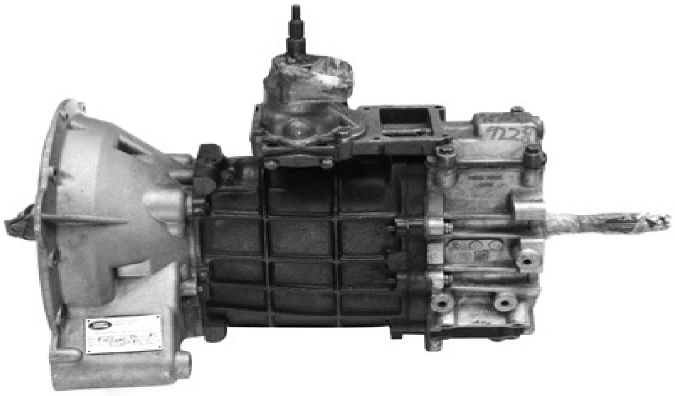
The gearbox syndrome explained

Humanities scholar with a problem, waiting for a solution

First HLT researcher offering help

The gearbox syndrome explained

Humanities scholar with a problem, waiting for a solution



First generation named entity recognizer (rule based)

The gearbox syndrome explained

Humanities scholar with a problem, waiting for a solution

Second HLT researcher offering help

The gearbox syndrome explained

Humanities scholar with a problem, waiting for a solution

Second generation named entity recognizer (statistics based)

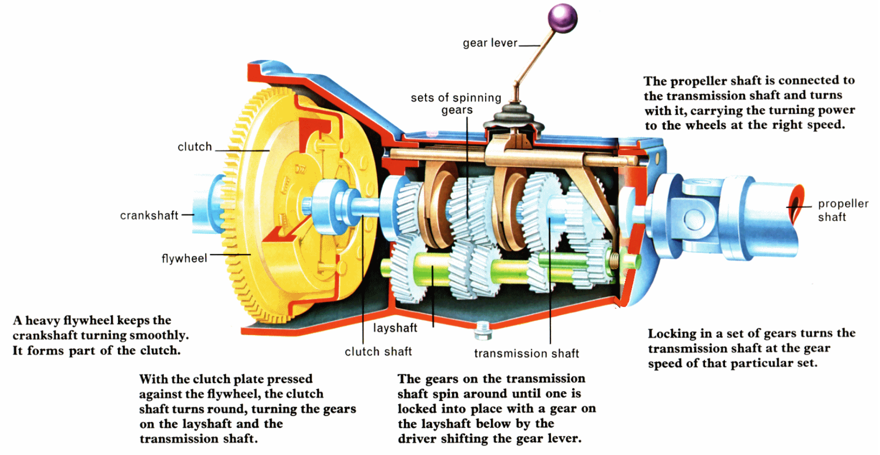
The gearbox syndrome explained

Humanities scholar with a problem, waiting for a solution

Third HLT researcher offering help

The gearbox syndrome explained

Humanities scholar with a problem, waiting for a solution



The propeller shaft is connected to the transmission shaft and turns with it, carrying the turning power to the wheels at the right speed.

A heavy flywheel keeps the crankshaft turning smoothly. It forms part of the clutch.

Locking in a set of gears turns the transmission shaft at the gear speed of that particular set.

With the clutch plate pressed against the flywheel, the clutch shaft turns round, turning the gears on the layshaft and the transmission shaft.

The gears on the transmission shaft spin around until one is locked into place with a gear on the layshaft below by the driver shifting the gear lever.

LREC 2012 paper about next generation named entity recognizer

Are you gearbox makers?

- Look at the proceedings of this conference, listen to the talks and posters, and draw your own conclusions …

Does CLARIN need gearboxes?

Yes and no

Yes:

- They are indispensable because they provide the main building bricks needed to solve HSS problems

No:

- Just having gearboxes is not sufficient: we also need (examples of) complete buses to take passengers from where they are to where they want to be

If you are not in CLARIN you can just spend the rest of your lives making better gearboxes

But if you are in CLARIN this is not enough!

What needs to be done?

Three possible ways forward:

- HLT researchers should not just focus on their engines and tools but also think about complete workflows
- HLT researchers should team up with HSS scholars and help them to solve their problems
- HSS scholars should be educated in how to apply HLT results in their daily research practice

All three directions should be pursued, and it is one of CLARIN’s main tasks to make this happen.

The obstacles (1)

Recap of the nightmare issues

- CLARIN will only work if all countries and languages are covered
- Not every relevant source is available in digital form
- Many digital sources are just pictures
- OCR for manuscripts and old scripts not mature
- Metadata poor and not standardized
- HLT coverage for languages uneven
- Language barriers for the researchers have to be overcome
- Scholars may not be able to use the existing technological building bricks to build their research castles

The obstacles (2)

- There is no guarantee that your data and my tools will fit together (interoperability standards)
- Sustainability of data and tools is not guaranteed (reproducibility!)
- IPR issues, especially for re-purposed data
- Ethical issues
- Immaturity of HLT tools (for all or some languages), especially MT
- Researchers not always willing to share data and tools
- HLT community affected by gear box syndrome

CLARIN's answer: Action lines (1)

- Coverage: consolidate 9 members, reach out to others, 15 members in 3 years, 20 in 5 years
- Legal: license templates promoted for new and legacy data, talk to legislators about IPR, establish Access and Authentication in old and new countries
- 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
- 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, centres of expertise; Portal targeting different audiences; Interfaces and visualization
- Crossing borders: develop long term vision, but start from existing collaborations; use language as vehicle to collaborate with other disciplines; inter Research Infrastructure, international collaboration; explore industrial collaboration models
- Sustainability: demonstrate societal impact; review sustainability models; after 3 years vision and strategy

Coming up

CLARIN Conference in Sofia (Oct 26-28)

- Aimed at integration of CLARIN infrastructures in CLARIN ERIC member countries
- Synchronizing our agendas
- Defining the rules of the game for the coming years

Participants:

- Members of national consortia in CLARIN ERIC countries
- Representatives from candidate countries

Expected outcome:

- Integration plan
- Agreement on strategy and priorities

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
- Interoperability of data and tools from different sources is crucial
- Language barriers need to be crossed
- Much effort needed to ensure adoption of digital methods in the humanities and social sciences (education, showcases, visualisation) – just giving them gearboxes won't help!