Dreams and nightmares: what to tell the user?

Steven Krauwer
CLARIN ERIC / Utrecht University

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 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 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 2012 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 2012 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 sufficient to be of help to HSS scholars?
  - 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 (see later presentation)

My questions to you

- What are we going to tell prospective users other than computational linguists (whose only purpose in life seems to be creating data and tools that should help them to create yet more data and tools)?
- My longer-term examples may sound sexy but some of them are still highly problematic – do you have any suggestions for better ones that I could use in my sales talks?
- Do you have any suggestions for shorter-term examples that are both realistic and convincing to others than computational linguists?