



# **Working with older texts: wishes, needs, requests of researchers in the Human and Social Sciences**

**Ineke Schuurman**

Coordinator CLARIN-Vlaanderen

GALATEA II (18-02-2011)

# Structure



- European project: CLARIN
- Researchers Human and Social Sciences: Cold feet !
- Wishes, needs, requests of HSS community

# CLARIN



- European ESFRI project
- Most European countries/languages:
  - Flanders, the Netherlands
- Make resources/tools available for HSS
  - often computer 'illiterates'
- Modern language as well as older forms

GALATEA addresses historical language

# Contrast old/new



## Dutch language and NLP

- STEVIN-programme
- Lots of tools and resources available for modern Dutch, especially 1980-now
  - TTNWW (CLARIN pilot)

## Older forms: relatively little available

- CATCH programme (NWO), CLARIN NL (tools Adelheid, Inpolder)

# Historical Dutch texts



## Issue 1

- Many digitized texts are just ‘images’
  - Manual transcription very time consuming
  - OCR problematic for manuscripts
    - IMPACT project

## DigiHist workshops

⇒ Paleographic Workbench

# Historical Dutch texts



## Issue 2

- Not that many tools and resources available to be made accessible for HSS researchers
  - Taggers
  - Parsers
  - Balanced and/or annotated corpora

# Historical Dutch texts



## Issue 3

- ‘Raw’ texts are problematic for HSS as well
  - Variation in spelling
    - Adapt search tool

GALATEA workshops: make historical texts accessible for HSS as well

# Goals CLARIN - HSS



1. To enable HSS researchers to address new topics of research, and/or
  2. To address 'old' ones in a more convenient way
- Modern language
  - (old)er language:
    - Fields: history, linguistics, literature, ...

# Cold feet



- Not just HSS researchers dealing with historical texts
  - Especially the over-40 set
    - Accustomed to ‘manual work’
- ⇒ Rather wary to adopt “completely new”, “very complicated” approaches they don’t control
- ⇒ Spontaneity is lost (lucky strikes)

# Cold feet



- Affects behaviour younger staff, PhD-students etc

⇒ We should try to convince the over-40 set as well!

For example wrt 'complexity'

# Complexity



Request GALATEA I (linguists)

## **Keep it simple**

- Homogeneous tools for possibly heterogeneous corpora  
NOT: x query languages for x corpora
- Easy to learn, in keeping with 'their' language (NOT theoretical linguists!)

# DBNL



- Most older documents not machine readable, just images
- Annotation impossible
- Only partially downloadable
- Search functionality too limited: one word, unreliable, no wild cards (*ghel\*t*), ...
- Mix original text, comments etc
- Limited indexation (Google)

# Corpora



- Balanced, annotated, diachronic corpus 1250-now
- Corpus/tools 15<sup>th</sup> and 16<sup>th</sup> century, several types of text

## “Nobody Knows”

- Central registry of machine readable corpora
- Cottage farming: *ad hoc* collaborations

# More challenging



- Tool to ‘align’ image of manuscript and machine readable version
  - For example to verify correctness of transcript
- Parallel text in modern Dutch (not just lemmata)
- Links with electronic dictionaries (like WNT)

# Tools



- Not just for one platform:
  - Request for tools that also work in Microsoft Word environment (taggers etc)
  - ‘Please make everything work under Unix as well’
- Open access

# GALATEA II



What can WE do for THEM ??