Infrastructure building as a research task and a necessity for language and speech R&D

Janne Bondi Johannessen og Ruth Vatvedt Fjeld
Text Lab and Lexicography, ILN, UiO

Clarin meeting, Bergen 15.-16. December 2008

Developing infrastucture resources for a language bank

- Take care of existing resources
  - Harmonize formats
  - Make available for research and development
- **Develop new resources**
  - Who is to decide what to develop?
  - Who is to decide purpose of resource?
  - Who is to decide how to develop resource?
  - Who is to develop resource?

Who is to develop resources?

- Choose people and institutions who have experience and ideas!
- Who are they? Research institutions!
- Why researchers?
    - Have knowledge of the field
    - Ask the appropriate questions
    - Find appropriate solutions (material, methods)
    - Are critical to previous solutions
    - Will experience acquisition of new insight based on development issues
    - Will share knowledge (research publications)
    - Have experience in development of new methods and resources
    - Have experience in big projects

What if the LT resources were to be developed by non-researchers?

- Premise: No Language Bank administration would be able to set up a full specification set in advance
- Premise: No Language Bank administration would be able to foresee all possible questions arising daily through a development process

=> Important decisions would be made without a scientific research perspective

=> Any specification would be based on yesterday's solutions, i.e. old.

=> Important new developments would be missed.

Success: research org. as developers, commercial org. as users

- UiO: (*Bokmålsordboka* + *Nynorskordboka* Norw.BM + NN dictionary). Used by:
  - IBM to develop their own lexicon
  - NorKompLeks (NTNU+Telenor)
  - Nyno translation system
  - Several bilingual dictionaries: (No-Bulgarian, No-Lithuanian)
  - Connexor (Finnish language technology)
  - Lingsoft (Finnish language technology)
  - Gule sider (Yellow pages)
  - Mikroværkstedet (Danish language technology)

- UiO: Oslo-Bergen Tagger. Used by:
  - Nyno translation system, Lingsoft, Microsoft grammar checker
- UiO: Frequency lists. Used by:
  - Several European and American mobile phone companies
- UiO: Tagged texts. Used by:
  - European and American software companies
- UiO: Transcriptions of speech. Used by:
  - Some big software companies

Why do private enterprises want research-developed material?

- They know they get good quality
- They know there is public documentation available
- They know there are critical scientific publications available
- They trust the judgments researchers have made at various points in the development.

Resources developed by research are good for R&D

- Wordnet
- Constraint Grammar
- Lexical-functional Grammar
- Head-driven Phrase-structure Grammar
- TreeTagger
- Brill-tagger
- Algorithms in models such as Maximum Entropy, Memory-Based Learning

A Language Bank

- Is a golden opportunity to develop high quality resources
- Is a golden opportunity to enhance research, insight and knowledge not just for infastructure users, but earlier, for infrastructure developers
- Is a golden opportunity to combine the needs from technological industry and research on the one hand with the needs from LT research on the other hand
  - => two for the price of one!

Summing up

- Who is to decide what to develop?
  - Researchers must be a central proportion of the group
- Who is to decide purpose of resource?
  - Researchers must be a central proportion of the group
- Who is to decide how to develop resource?
  - Researchers
- Who is to develop resource?
  - Researchers