Metadata in ASK – a learner corpus of Norwegian as a second language

Workshop on Interoperability of Second Language Resources and Tools

University of Gothenburg, December 7, 2017

Silje Ragnhildstveit and Kari Tenfjord

siljera@hvl.no

Western Norway University of Applied Sciences and University of Bergen

The ASK corpus

Language data

Texts from two L2 tests

1700 in the mainkorpus (appr.)

Norwegian control group

Metadata

Text relatd metadata

Personal metadata

Annotation

Manual error tagging

Automatic grammatical tagging

Semi-automatic POS tagging

The language data

- is written essays compiled from standardized tests at two different levels (intermediate- and higher level)
  - Narratives, argumentatives

- is homogeneous as far as the circumstances of the production
  - the same time framing
  - no use of reference tools
  - the tests have been scored by assessors with the same kind of training
  - the personal metadata have been recorded in conjunction with the same test occasion

Text related metadata

- Text ID
- Test level (intermediate- and higher level)
- The prompts
- Prompts grouped in themes

- CEFR levels
  - the essays were reassessed by a group of trained assessors according to CEFR levels (Carlsen 2010)

Personal metadata

LEARNER

- L1: Albanian, Bosnian-Croatian-Serbian, Dutch, English, German, Polish, Russian, Somali, Spanish, Vietnamese
- Homeland
- Age
- Gender
- Utdanning
- Profession
- What they are doing in Norway
- Length of stay in Norway

LANGUAGE

- Use of Norwegian
  - in sparetime
  - at work
  - daily, seldom, never
- Knowledge of other languages:
  - English proficiency
  - 3rd language
  - 4th language

LEARNING NORWEGIAN

- Type of Norwegian course
- Motivation
- Hours of Norwegian lessons
- Level
  - beginner
  - intermediate
  - advanced
- Time since starting at the Norwegian course
- Social contact with Norwegians

Building a learner corpus

Two crucial principles:

1. A clear decision is needed regarding the research purpose(s) of the corpus

   → design criteria
2. Acknowledge the interdisciplinarity of the corpus building enterprise, the necessity of dialogues between SLA- and language resource experts

… and a good advice:

Choose a **userfriendly** interface