CLARIN
Common Language Resources and Technology Infrastructure

Bridging the Gap between First Language Acquisition and Historical Dialectology with the Help of Digital Humanities

Folkert de Vriend & Martin Snijders

18/11/2011

Understanding Society

Time and team

- Project duration: 1 year (may 2011 - may 2012)
- Multi-disciplinairy team:
  - Leonie Cornips
  - Wilbert Heeringa
  - Marc Kemps-Snijders
  - Martin Snijders
  - Student assistants: Yvonne Cruijssen, Gertruud Hoff, Anke Meevissen
  - Jos Swanenberg
  - Folkert de Vriend

General

- **COAVA**: **CO**gnition, **A**cquisition and **VA**riation Tool

- Aims of COAVA:
  A) Curation of resources from two separate linguistic subdisciplines: first language acquisition and dialect geography.
  B) Development of a demonstrator tool for interdisciplinary research into the lexical characteristics of concepts

A) Curation

Resources in COAVA

- Seven corpora from CHILDES
  - The Netherlands and Flanders
  - Children (mostly between 2 and 3,5 years)

- Part III of WBD/WLD
  - (Dutch and Flemmish) Brabant and Limburg
  - Adults

CLARIN-compliance

Dialect data and CHILDES data

- CMDI-metadata
- Persistent identifiers
- ISOcat

Dialect data

- Lexical Markup Framework (LMF)

B) Demonstrator

Lexical characteristics

- First language acquisition:
  For some concepts the lexical form typically is acquired early (‘dog’ for instance) while for other concepts the lexical form typically is acquired later (‘blue titmouse’ for instance.).’
- Dialect geography:
  For some concepts there is lot of lexical variation while for other concepts there is very little variation.

Value of combined interpretation

- For researchers in **both** disciplines these characteristics are interesting for at least two reasons:
  - Research into the ‘basic level vocabulary’ of a community
  - Research into the relation between age of acquisition and (dialect)variation

Implementation

- A concept taxonomy is constructed. This taxonomy will only contain concepts for which lexical forms can be found in **both** resources
- Since the Dutch CHILDES data mostly contain data for children aged between 2 and 3,5 years of age we focus on lexical forms that are **nouns**.
- To enable linking from this taxonomy to the CHILDES data, these first need to be lemmatised and tagged for their POS (Lexicon by Gilles)

Demo

Technology

- Client server application
- Search services

- Java/Google Web Toolkit
- Apache/Tomcat
- Solr search server
- Open Source

Solr

- Indices, multi core
- Facetted search
- Fast

Demo

Thank you