ICLFI
Overview of the project and corpus

Jarmo Harri Jantunen
ASKeladden netwok meeting on learner corpus
March 25 – 26, 2010

Project
• Corpus study on language-specific and universal features in learner language
• Nordic Council of Ministers, Riksbankens Jubileumsfond, Universities of Oulu and Umeå
• 2006 pilot stage
• 2007 –

Project: members
• research group
  – University of Oulu, Umeå, Tallinn, Petrozavodsk State University, Karelian Pedagogical University
• teacher group
  – teachers at 20 universities collecting material

Project: research questions
• Learner language universals (LLU)
  – LLUs are features that exist in learner language in spite of the learner’s target and source language.
  – e.g. untypical phraseological units, simplified language usage, atypical frequencies of items (cf. Jantunen 2008)
• Language-specific features (LSF)
  – LSFs are elements that are present only in one specific learner language but cannot be found in other LLs.
  – Features that are typical for learners of Finnish

Project: members and their research projects
• Annekatrin Kaivapalu (Tallinn)
  – cross-linguistic influence in second and foreign language learning
• Tuija Määttä (Umeå)
  – Transfer and local cases in Finnish learner language
• Marianne Spoelman (Oulu/Groningen)
  – The use of the partitive case in Finnish learner language
• Sisko Brunni (Oulu)
  – Phraseological units in language learning and teaching
• Jarmo Jantunen (Oulu)
  – Phraseological units in learner language (beyond collocations), corpus-driven approach
  + master and bachelor theses in several Finnish and foreign universities

International Corpus of Learner Finnish (ICLFI)
Typology of ICLFI
• general
• written
• synchronous (partly longitudinal)
• monolingual
• multi-L1
• multi-text type
• multi-proficiency level
• whole text
• medium: electronic and manually written texts
• raw text
ICLFI: general

- Size 600,000 tokens
- Mother tongues: 22
  - large subcorpora (c. 100,000): Russian, Estonian, Polish
  - medium-sized (30,000 – 50,000): Swedish, Dutch, Chinese, German, Czech
  - small (< 10,000): Spanish, Icelandic, English, Italian...
- Proficiency levels: beginners, intermediates, advanced
- Text types:
  - large subcorpora: essays, narratives, diaries, argumentative texts, letters
  - small: newspaper articles, postcards, applications...

ICLFI: variables

- Learner variables
  - age, place of birth, gender, place of residence
  - mother tongue, other languages, proficiency level
- Task variables
  - text type, timing, exam, reference tools, writing place, topic
- Learning context variables
  - language exposure:
    - parents' mother tongue
    - Finnish spoken at home, relatives taught Finnish, teacher's mother tongue, stay in Finland
  - text book
- Other:
  - collecting place and time
  - medium (electronic/manual text)

ICLFI: future

- Lemmatization
- POS tagging
- Syntactic parsing
- Problem-oriented tagging (error annotation)
- Downloading
- Advance notice: Conference on comparative analyses of Finnish, Estonian, and other learner languages 2012 (ICLFI + VirSu)

Co-operation

- Estonian Interlanguage Corpus (EVKK)
  - Pille Eslon and Annekatrin Kaivapalu
  - two related language variants: learner Finnish vs. learner Estonian
- Cefling Project
  - National Certificates test performance corpus
  - Finnish as a foreign vs. second language
- University of Turku: Corpus of advanced learners (theses)
- VirSu, Estonian and Finnish as foreign languages

Web page


Literature

Bruni, S. 2009. Verbirakenteet L2-opetussa. [Verb structures in L2 teaching.]
Presentation in Kielitieteen päivät, Jyväskylä 5.5.2009.