New corpora in World Englishes and Learner English

Heli Paulasto and Lea Meriläinen

University of Eastern Finland

GlobE seminar, Helsinki 13-14 June 2012

Corpus plans at the UEF

- ENO International Corpus of Student English (ENOICSE)
    - New corpus, collected and coordinated by Paulasto (& Meriläinen)
    - Funded 1.9.2012-31.8.2015 by Academy of Finland
- Louvain International Database of Spoken English Interlanguage (LINDSEI)
    - Directed by Granger & Gilquin, CECL
    - Finnish component collected by Meriläinen (& Paulasto)

ENO International Corpus of Student English

Background

- A longstanding ‘paradigm gap’ between World Englishes and Learner English research (Sridhar & Sridhar 1986)
- Recent efforts to fill it, e.g. Nesselhauf (2009), Mukherjee & Hundt (2011), Paulasto et al. (2011)
- There is a need for comparable data representing both types

Present corpora, e.g.

- ICLE:
  - Learner English, 16 mother tongue backgrounds
  - 3.7 M words, (mainly) argumentative essays written by EFL learners, university students of English
  - Expanding Circle
- ICE corpora:
  - national or regional varieties of English
  - 1 M words, standard variety of spoken and written English
  - Inner and Outer Circles
- ICNALE (International Corpus Network of Asian Learners of English); Ishikawa (2011)
  - Learner English, focus on Asia
  - 1 M words, essays written by college students
  - Inner, Outer and Expanding Circles

Aims of the ENO corpus

- Corpus of
  - upper secondary level student writing
  - representing Inner, Outer & Expanding Circles
- 150-200,000 words from 15-25 countries
  - min. 2,250,000 words in total
- To be made available for public academic use

ENO – Environment Online

- Global virtual school and network for sustainable development and environmental education: http://www.enoprogramme.org/
- Founded by Mika Vanhanen (Eno/Joensuu) in 2000
- Today, nearly 10,000 schools in 150 countries
- Three cornerstones: technology, structure and empowerment (Vanhanen 2012):
  - ICT enabling schools to network and learn online
  - Co-operation between teachers instead of a hierarchy
  - Learners empowered to act in the community by taking part in campaigns and activities (e.g. tree planting)

ENO countries

- Focus on the developing countries in Africa, Asia, Middle-East, South America, Eastern Europe
- 25 potential countries for the ENO corpus:
  - *Inner C:* USA
  - *Outer C:* Malaysia, India, Philippines; Kenia, Tanzania, Nigeria, Ghana, Uganda, Namibia, South Africa (Black SAfE)
  - *Expanding C:* Finland, Russia, Slovenia, Romania, Turkey; Brazil, Argentina; Indonesia, Taiwan, Thailand
- Several mother tongue backgrounds in one country?
  - “If Nigerian English is ‘A Variety’, does it matter?”

The essays

- 150-200,000 words from 15-25 countries
  - 20 school classes x 20 students x 400 words = 160,000
- The essay topics to be tied with ENO aims of environment, sustainability and peace

→ The corpus project to be a part of the ENO educational programme

→ Essay topics to be planned together with Mika Vanhanen, ENO teachers and global/environmental educators

→ The corpus to be of interest also for teachers and scholars in global and environmental education

The teachers

- Contacting ENO country coordinators to find and enlist English teachers in the above countries interested in the project
    - Preliminary survey (Paulasto 2011) is positive
    - ENO is not actively integrated with English teaching in the schools at the moment, although it's a brilliant platform for teaching EIL
- Creating a network of ENO (upper secondary) English teachers, i.e. a team of research assistants, to plan the essay topics and the questionnaire needed for informant data
- Creating and utilizing online discussion forums

The questionnaire

- Follows the ICLE model (Granger et al. 2009), but relevant variables also need to be planned together with local experts, i.e. the teachers
- Task variables:
  - Common ones: Medium (writing), Genre (essay), Field (general English), Length (e.g. 300-800 words)
  - Varying ones: Topic (from a set list), Timing, Essay writing conditions, Use of reference tools, Other?
- Learner variables:
  - Age, Gender, Mother tongue, Region, Other languages, Stays in English-speaking countries (where relevant), Learning context (e.g. medium of instruction, role of English – ENL/ESL/EFL, years of learning English), Proficiency level (how?), Other?

The questionnaire

- Should answer the specific needs of each country and/or learning environment
- Each variable should be encoded in the questionnaire, with each questionnaire tuned according to the needs of each country/ environment
- Needs to be designed *extremely carefully!*

Collecting the data

- Questionnaire and essays to be collected using an online platform, e.g. the UEF E-lomake
- Essays to be written without a word processor(?), e.g. in Notepad (or by hand), and added (copied) into the questionnaire (or digitised in Joensuu)
- Teachers responsible for
  1) Filling in the correct variable data for each student
  2) Making sure that the process of writing the essays follows the guidelines
- They mustn’t edit the texts in any way.

And then what?

- Tagging the essays for the background data
- Markup in the text (minimal)
- Linguistic annotation: Which tagger? How detailed?

→ Assistance will be needed: local, national and/or international

Collaboration with CECL

- The final aim: publishing the corpus in collaboration with the Centre for English Corpus Linguistics, Louvain-la-Neuve
- Funding required. Possibilities:

1. University partnership: outsourcing, a part of the work delegated to CECL, who will then employ a researcher to carry out the job
2. Researcher mobility: CECL will have a consulting role, while I(?) do the work
3. Submitting a joint project between UEF and CECL to be funded by Finnish and/or Belgian research academies or foundations

Schedule

- 1.9.2012 --- 31.8.2015 -----
- Cf. ICLE was begun in 1990, edition 1 published in 2002...
- 1st stage: finding the teachers and creating the network in fall-winter 2012-13
- Sizeable amount of the data to be collected by August 2015
- Obtaining funding for collaboration with CECL in 2015 (or earlier)

The potential of the project

Research potential

- Linguistics: the corpus data, the compilation process
- Language education: the corpus data, global English teacher networking, EIL in ENO (see Paulasto 2011), language education in/ through ENO, promoting sustainability and peace in the English classroom (see Birch 2009)
- Global and environmental education: the corpus contents (impact on the questionnaire?), the compilation process

Applied potential

- Creating an online resource for English teachers (cf. *Backbone: Pedagogic Corpora for Content & Language Integrated Learning*, http://u-002-segsv001.uni-tuebingen.de/backbone/moodle/)

Louvain International Database of Spoken English Interlanguage (LINDSEI)

- A corpus consisting of informal interviews with intermediate to advanced learners of English from various L1 backgrounds http://www.uclouvain.be/en-cecl-lindsei.html
- Project directors: Sylviane Granger and Gaëtanelle Gilquin, the Centre for English Corpus Linguistics, Louvain, Belgium
  - Collaborators in various countries
- A sister corpus to the *International Corpus of Learner English* (ICLE)
  - 3.7 million words of written learner English by learners representing 16 different mother tongue backgrounds

- LINDSEI v1 (Gilquin *et al*. 2010)
  - c. one million words
  - Learners from 11 mother tongue backgrounds: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish
- LINDSEI v2 (forthcoming)
  - Additional components from L1 Arabic, Basque, Brazilian Portuguese, Finnish, Lithuanian, Norwegian and Turkish
  - Altogether 18 L1 backgrounds

LINDSEI components

- 50 informal interviews (c. 100,000 words) of students majoring in the English language
- 3 tasks: discussion on a set topic, free discussion and picture description
- Each interview: c. 15 minutes; 2000 words
- Transcribed and marked-up according to the same conventions
- Profile of background information
- Built according to similar principles as the ICLE
- LOCNEC: comparison corpus of NS interviews

The significance of LINDSEI

- Enables more reliable comparisons between learner English and other Englishes
  - The commonalities and the developments observed in Englishes across the globe more typical of spoken language
  - Earlier studies into L2 and contact varieties of English as well as ELF largely based on spoken language corpora

References

Birch, B. M. 2009. *The English Language Teacher in Global Civil Society*. New York and London: Routledge.

Gilquin, G., De Cock, S. & Granger, S. (2010) *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM.* Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S. et al. 2009. *International Corpus of Learner English: Version 2.* Louvain-la-Neuve: Presses universitaires de Louvain.

Ishikawa, S. 2011. 'A new horizon in learner corpus studies: the aim of the ICNALE project.' In G. Weir, S. Ishikawa & K. Poonpon (eds), *Corpora and Language Technologies in Teaching, Learning and Research*. Glasgow: University of Strathclyde Press, 3-11.

Mukherjee, J. and M. Hundt (eds), 2011. *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap.* Amsterdam/ Philadelphia: John Benjamins.

Nesselhauf, N. 2009. 'Co-selection phenomena across new Englishes: Parallels (and differences) to foreign learner varieties.' *English World-Wide* 30(1): 1-26.

Paulasto, H. 2011. English as an international language in the ENO virtual school. Paper presented at the AFinLA fall symposium, 11-12.11.1011.

Paulasto, H. Meriläinen, L. and E. Ranta 2011. Syntactic features in Global Englishes: how 'global' are they? Paper presented at International Society for the Linguistics of English 2011 Conference (ISLE 2), Boston, USA, 17.-21.6.2011.

Sridhar, K.K. and S.N. Sridhar 1986. 'Bridging the paradigm gap: Second language acquisition research and indigenized varieties of English.' *World Englishes* 5(1): 3-14.

Vanhanen, M. 2012. 'Foreword.' In M. Vanhanen & H. Paulasto (eds), *Planting Seeds of Action: The Environmental Learning Process of ENO Schools since 2000*. Joensuu: ENO Programme Association.