Construction of a multi-modal learner corpus of STEM student language production: A pilot study

Ralph ROSE and Hinako MASUDA
Center for English Language Education (CELESE)
Faculty of Science and Engineering
Waseda University – Tokyo, Japan

Acknowledgments: Maiko Serizawa, Akari Nakai, Hideyuki Nagao, & Kenta Nozaki

8th International Conference on Corpus Linguistics
2-4 March 2016
Malaga, Spain
Learner Corpora

• Corpus definition
  - A "collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety" (McEnery et al 2005, p. 5)

• Learner corpus definition
  - "[E]lectronic collections of natural or near-natural data produced by foreign or second language (L2) learners and assembled according to explicit design criteria." (Granger et al 2015, Loc 328)

• Learner corpora construction boom
Learner Corpora

- LC enable focus on a narrow(er) population
  - STEM students (HKUST: Wen et al 2005)

- LC allow contrastive interlanguage analysis (Granger, 1996)
  - Learning Prosody in a Foreign Language (LeaP) Corpus (Gut 2012): Contrast spoken German across age, language background, etc.
  - Japanese English as a Foreign Language Learner (JEFLL) Corpus (Tono 2007): Contrast written English across school level, topic, etc.
L1 and L2 contrast

- **Growing need for L1 data**
  - "Very few learner corpora incorporate L1 data as an integral part of the design. This will become more important in future learner corpora projects as we are beginning to realise the need to identify specific features of L1-related errors or over/underuse patterns." (Tono 2003: 803)

- **Corpus Escrito del Español como L2 (CEDEL2; Lozano and Mendikoetxea 2013):** corpus of L1 English and L2 Spanish writing
Speech and writing contrast

- Spoken and Written English Corpus of Chinese Learners (SWECCL; Wen et al 2005)
- The Santiago University Learner of English Corpus (SULEC; http://www.sulec.es/)
- Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al 2010) and International Corpus of Learner English (ICLE; Granger et al 2009) complement each other for spoken-written contrast
Writing process data

- Written corpus data typically includes end product, or occasionally the drafts (cf., Mäkinen and Hiltunen 2014, Kreyer 2014).

- There is a growing research interest in fine-grained observations of the writing process.
  - Keystroke logging and language production (Sullivan and Lindgren 2006)
  - Inputlog tool for keystroke logging in Word (Leijten and Van Waes 2013; http://www.inputlog.net)
  - Analysis of pauses during writing (Braaksma et al 2010, Hoàng 2015)
A corpus for our needs

- Desired:
  - Constructed from production by English learners who are science, technology, engineering, and mathematics (STEM) students
  - Contains samples of the functional uses of language our students expect to use (research-oriented)
  - Includes both first and second language data
  - Includes both speech and writing
  - Writing data shows writing process
  - Is available

- None found...
SELCor Objectives

- Create a resource to answer several research questions
  - What kind of linguistic patterns do STEM learners (mis)use for certain communicative functions?
  - What relationships can be observed between STEM learners' writing and speech productive behaviors?
  - Which aspects of the learners' L1 behavior are predictive/not predictive of their L2 behavior?
  - What kind of relationship can be observed between use of linguistic patterns and STEM learners' proficiency?
  - How does the nature of the speaking tasks influence STEM learners' fluency?
SELCor Objectives

- Create a resource that is useful to our teaching staff
  - Answering their own specific pedagogical questions (e.g., do STEM students have difficulty with a certain English phoneme, stress pattern, lexical item, grammatical structure, or communicative function?)
  - Examining how STEM students approach the tasks
  - Providing sample materials to use in instruction

- Create a resource that might be useful outside our institution
  - Other corpus researchers
  - ESP practitioners
SELCor design

• Participants
  – Undergraduate and graduate students in the Faculty of Science and Engineering at Waseda University
  – Recruited through the Waseda part-time job list (commonly used by researchers within Waseda for recruiting experimental participants)

• Demographic info
  – Personal: age, gender, academic orientation, dominant hand, hearing difficulty
  – Linguistic: native language, other languages, English test results (e.g., TOEIC, TOEFL), living abroad experience
SELCor design

- **Japanese (L1)**
  - Reading aloud (「狼と男」)
- **Participants spoke for about two mins per task.**
- **Speech was recorded in sound-attenuated room.**
- **15 min. writing task**
  - Expository or argumentative
  - Keylog info captured via Java application

- **English (L2)**
  - Reading aloud ("The boy who cried wolf")
  - Picture description
  - Diapix task (spot the difference)
  - Map task (giving directions)
  - Topic narrative
  - Problem-solving
SELCor design

- Speech: Transcripts were created and annotated by four native speakers of Japanese.
  - All fully spoken words
  - All clipped words
  - Utterance units
  - Filled pauses (um/uh)
  - Use of Japanese (L1)
  - Non-linguistic data (laughter, ingresses, etc.)
  - Each recording was transcribed by one individual. At present, transcripts have not been cross-validated.

- Writing: Keylog info on inserts, removes, and cursor movements saved in XML format
Results: L1-L2 contrast

Fluency measures observed in reading aloud

Articulation rate (syll/sec)

Silent pause rate (per min)

Silent pause duration (sec)

Statistical tests with \textit{lme} in \textit{R} using \textit{language} as fixed and \textit{participant} as random factors

Participants’ L2 speech behavior patterns after their L1 speech behavior. This is consistent with previous findings for unplanned speech (Derwing et al 2009, Cox and Baker-Smemoe 2012, De Jong et al 2015, Rose 2015)
Results: speech-writing contrast

Relationship between L2 speech rate and writing activity

![Graph showing the relationship between speech rate and writing activity.]

Participants who speak faster show a higher keyboard activity (inserts + deletes) rate but also a larger proportion of inserts to deletes: They type faster but backtrack more.

- For speech rate: $t(11) = 5.6, p<0.001$, $R^2 = 0.58$
- For writing activity ratio: $t(11) = 2.4, p<0.05$, $R^2 = 0.30$
Results: speech-writing contrast

Relationship between L2 speech and writing pauses

Participants who pause (silently) more in speech also pause more when writing. Participants who use *um/uh* more in speech, pause longer when writing.
Discussion: SELCor design issues

- SELCor shows patterns and trends consistent with previous observations.
- SELCor allows observation of intra-learner variation.
- SELCor is informative on some novel research questions.
  - Relationship between on-line speech and writing processes
- Some problems and limitations remain
  - Participant interest level in tasks was not high, hence concerns about naturalness (cf., Gilquin 2015)
  - Lack of L1 spontaneous speech sample prevents some desired contrasts.
Summary

- SELCor design is characterized by several key features.
  - Focuses on STEM learners
    - As participants
    - With respect to developmental needs
  - Contains L1 data for baseline comparisons
  - Contains both speech and writing samples
  - Allows intra-speaker comparisons
Future work

- Expand corpus with more participants
- Incorporate non-STEM student data for baseline comparison
- Collaborate with other universities in Japan
- Incorporate other L1-L2 language pairings
References


Gilquin, G. 2015. From design to collection of learner corpora In Granger, Gilquin, and Meunier (Eds.), pp. 9-34.


Kreyer, Rolf. 2014. "The people on the island steal all the fish": What we can learn from deletions in authentic learner texts? In *Proceedings of the 35th ICAME Conference*. Nottingham, p. 56.


