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

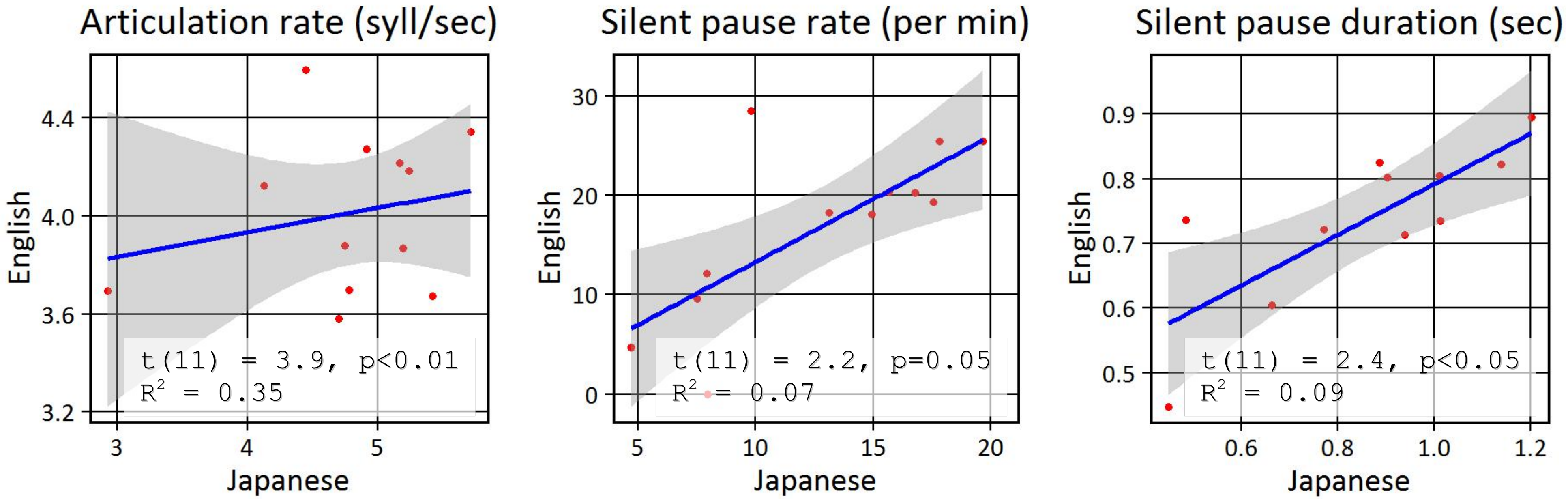
- English (L2)
  - Reading aloud ("The boy who cried wolf")
  - Picture description
  - Diapix task (spot the difference)
  - Map task (giving directions)
  - Topic narrative
  - Problem-solving

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

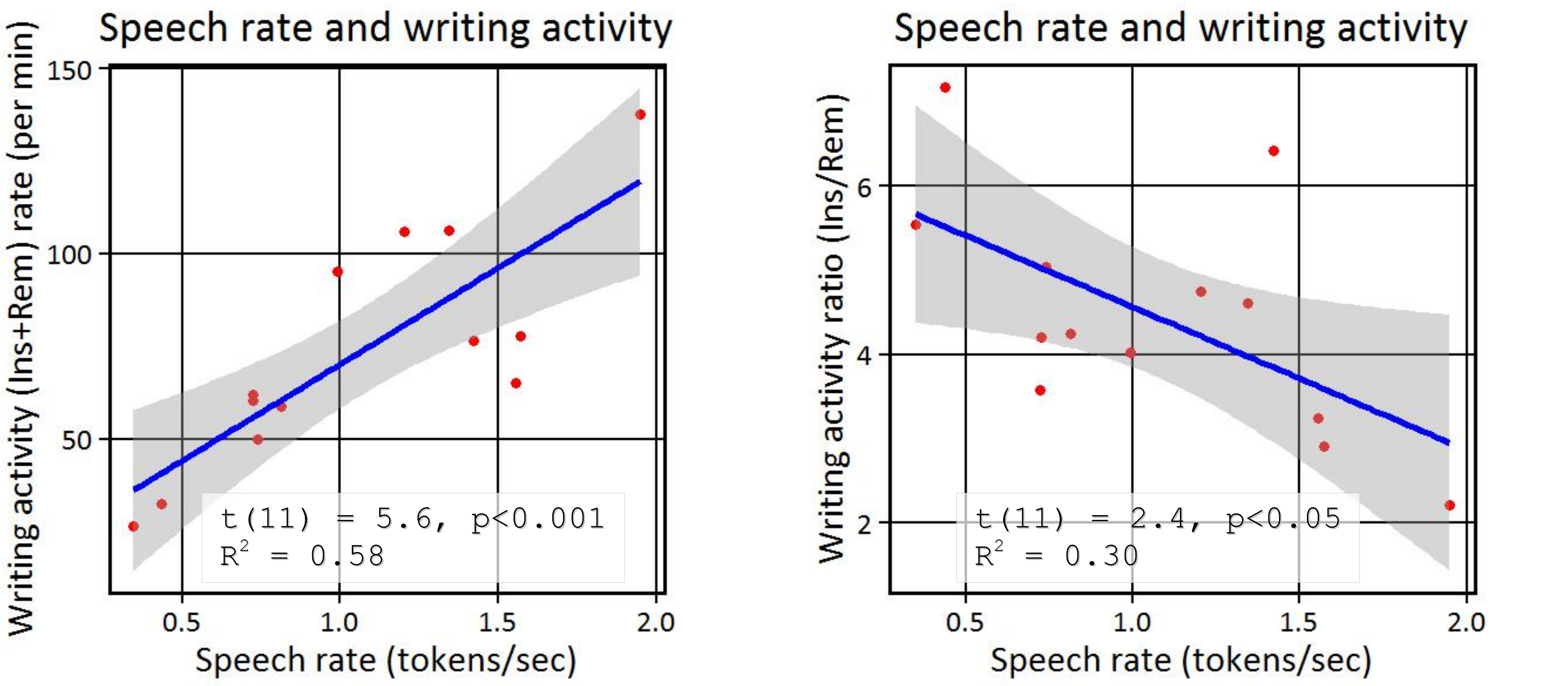


Statistical tests with `lme` in `R` using `language` as fixed and `participant` as random factors

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

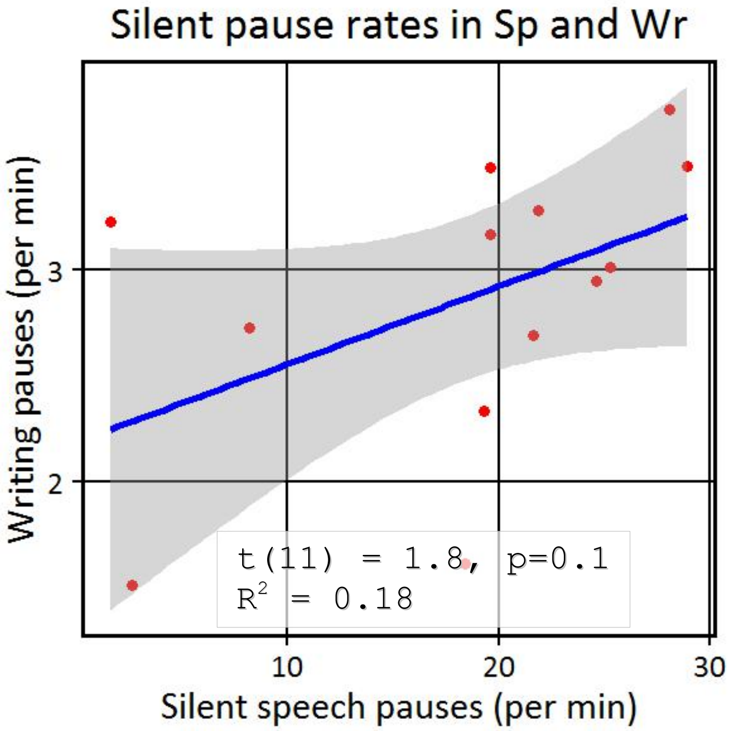
Relationship between L2 speech rate and writing activity

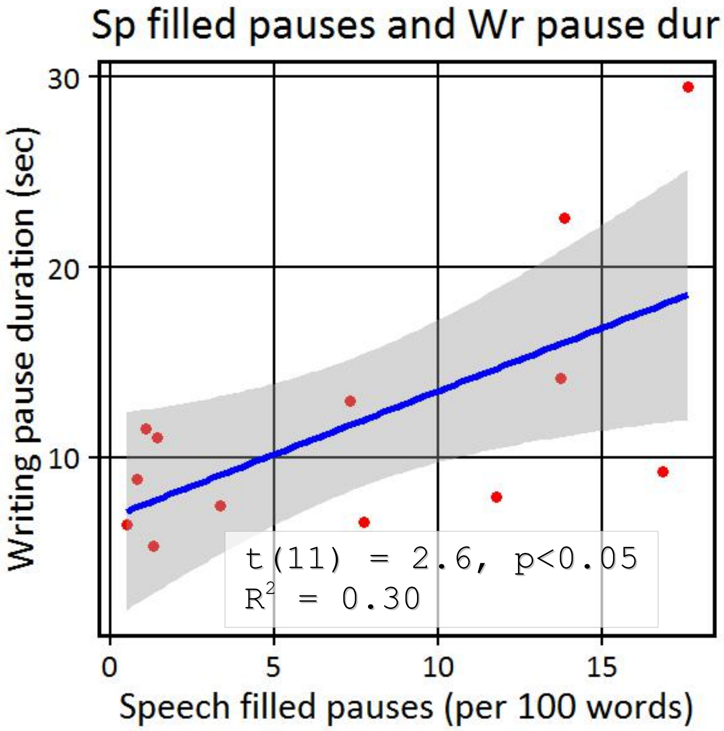


Participants who speak faster show a higher keyboard activity (inserts + deletes) rate but also a larger proportion of inserts to deletes: They type faster but backtrack more.

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

Braaksma, M., Rijlaarsdam, G., and van den Bergh, H. 2010. Hypertext writing versus linear writing: Effects on pause locations and production activities and its relation with text quality. Paper presented at the SIG Writing Conference, Heidelberg, Germany.

Cox, T. and Baker-Smemoe, W. (2012). The relationship between L1 fluency and L2 fluency across different proficiency levels and L1s. Presentation at Workshop Fluent Speech (Utrecht University, The Netherlands).

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, 31(4), 533-557.

De Jong, N., Groenhout, R., Schoonen, R., Hulstijn, J.H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behaviour. *Applied Psycholinguistics*, 36(2): 223-243.

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

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

Granger, Sylviane (ed). 1998. *Learner English on Computer*. London: Addison Wesley Longman Limited.

Granger, S., Dagneaux, E. Meunier, F. and Paquot, M. 2009. *International Corpus of Learner English. Handbook and CD-ROM. Version* 2. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, Sylviane, Gilquin, Gaëtanelle, and Meunier, Fanny. (2015) *The Cambridge Handbook of Learner Corpus Research*. London: Cambridge University Press

Gut, Ulrike. 2012. The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In Schmidt, Thomas and Wörner, Kai (eds.), p. 3-23. *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: John Benjamins.

Hoàng Thi Đoan Ha. 2015. *Metaphorical Language in Second Language Learners' Essays: Products and Processes*. Ph.D. Dissertation. Victoria University of Wellington, Australia.

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

Leijten, Mariëlle and Van Waes, Luuk. 2013. Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, 30: 358-392.

Lozano, Cristóbal and Mendikoetxea, Amaya. 2013. Learner corpora and second language acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data* (pp. 249-264). Amsterdam/Philadelphia: John Benjamins.

Mäkinen, M. and Hiltunen, T. 2014. Approximating the norm: Exploring EFL students' use of formulaic expressions in economics papers. In *Proceedings of the 35th ICAME Conference*. Nottingham, p. 133.

McEnery, A., Xiao, R., and Tono, Y. (2005) *Corpus-Based Language Studies: An Advanced Resource Book*. London, U.K.: Routledge.

Rose, R. (2015). Temporal Variables in First and Second Language Speech and Perception of Fluency. *Proceedings of the 18th International Congress of Phonetic Sciences* (ICPhS 2015), p. 0405.1-5. the University of Glasgow: Glasgow, UK.

Sullivan, K.P.H. and Lindgren, E. (Eds.). 2006. Studies in Writing: Vol. 18. *Computer Key-Stroke Logging and Writing: Methods and Applications*. Oxford: Elsevier. Glasgow, UK.

Tono, Yukio. 2003. Learner Corpora: design, development and applications. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the 2003 Corpus Linguistics Conference UCREL*. Lancaster University: United Kingdom, UCREL Technical Paper #16.

Tono, Yukio. (ed.) 2007. *Nihonjin Chukousei 10,000-nin no Eigo Corpus* [JEFLL Corpus: A Corpus of 10,000 Japanese EFL Learners]. Tokyo: Shogakukan.

Wen, Q., Wang, L., & Liang, M. 2005. *Spoken and written English Corpus of Chinese learners*. Beijing: Foreign Language Teaching and Research Press.