The significance of learner corpus research

Yukio Tono
Tokyo University of Foreign Studies

Workshop on Spanish Learner Corpus Research
Organized by Group LyS
14 July 2015, Universidade de Coruna

Learner corpora

- LC = language resources
  Thus LC serve different roles, depending on the purpose of their developers/users
- LCR = Intersection of corpus linguistics, second language acquisition and foreign language teaching/learning
  - LC as resources for SLA research
  - LC as resources for foreign language teaching
  - LC as resources for corpus linguistics

Historical perspectives

- Interest in learner language:
  → S. Pit Corder (1967) "The significance of learner's errors" IRAL 5: 161-170.
  - Study learner language in its own right to understand the learner's interim state of grammar system (Interlanguage)
- Various studies were conducted from late 1960s to early 1980s to collect learner errors
- But most of the data were discarded after collecting error samples.
  → No idea of sharing the data with others

Two commercial LC

- Longman Learner's Corpus
  - Originally developed by Michael Rundell for Longman dictionaries
  - The size was big (10 million words) back in 1990s
  - The data was not used until 1995, when major monolingual English learner's dictionaries were revised using corpora (OALD5, LDOCE3, COBUILD2, CIDE).
- Cambridge Learner Corpus
  - This is also in-house resources for dictionaries originally.
  - Later more widely used for materials development
- These two corpora are for lexicographical purposes, thus the size does matter.
- These two corpora were compiled with pedagogical applications in mind, not for SLA research.

International Corpus of Learner English (ICLE)

- The project was launched as an additional component of the International Corpus of English (ICE) in the early 1990s.
- The original purpose of the ICE project was to compare regional varieties of English (e.g. BrE, AmE, AusE, etc.).
- A corpus of "learner" English was added to this to compare it against NS English, which is why advanced learners were selected.

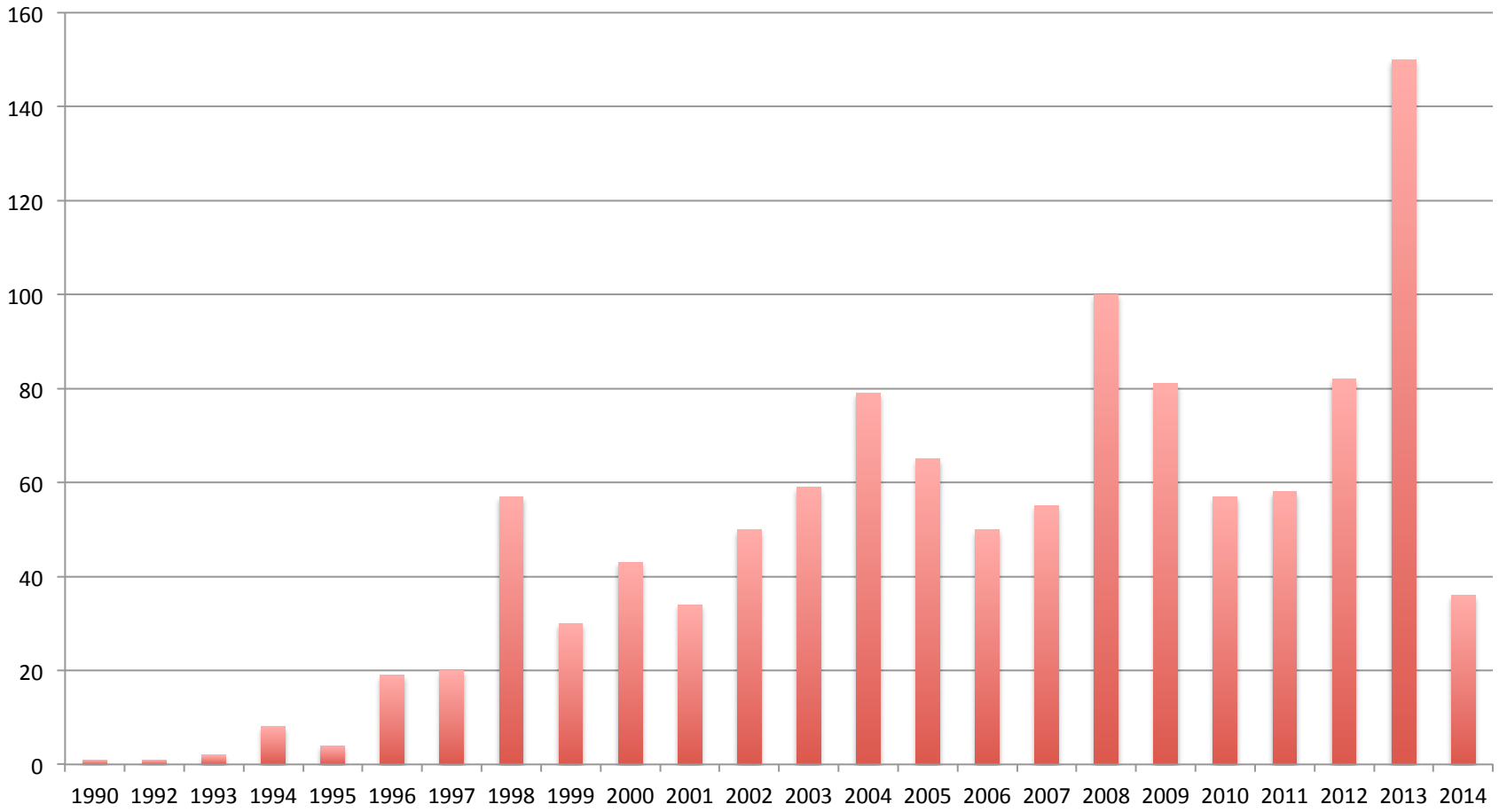
ICLE's contribution

- Proposed strict design criteria for compiling learner corpora
- Formed an international team of contributors just like representatives of ICE
- Proposed the potential impact of corpus linguistics in the study of learner language:
  - Contrastive Interlanguage Analysis (Granger 1996)
  - Computer-aided Error Analysis
- Lead to other projects to supplement ICLE:
- Spoken (LINDSEI), longitudinal (LONGDALE), and other L2 (FRIDA), etc.

ICLE's contribution (2)

- Provide the opportunities for evaluating the quality of LC from various perspectives
- ICLE has become a good example of what is missing:
  - Need for bigger data (size: cf. CLC)
  - Need for more control (cf. essay task)
  - Need for more developmental perspectives
  - Need for spoken data
  - Need for more varieties of written tasks
  - Need for longitudinal data
  - Need for publicly available error-tagged corpora
  - Need for more data tuned to specific SLA hypotheses

LC bibliography (n=1141)



Various LC constructed

- List of learner corpora at Louvain (n = 140):
  - More varieties of spoken vs. written data
  - More varieties of developmental/longitudinal data
  - More varieties of elicitation tasks
  - More varieties of target languages
  - More complex error annotation schemes
- Spanish learners → 9 projects listed
- The development looks healthy, overall, but the impact of LC in SLA research is yet to be seen (Tono 2015).

Mere replications…?

- Some camps use a very small corpus of learners with detailed error annotations.
- This reminds me of the old times when error analysis people did all sorts of error taxonomies and diagnosis.
- If the data analysis does not show the strength of corpus linguistic approach, then what people are doing now is the same as 40 years ago.

Back in the 60s & 70s

- Duskova (1969):
  - 50 Czech learners of English; each wrote 3 essays
  - Distinguish “errors” from “mistakes”
  - Classify errors into 9 categories with frequencies
  - A small scale study, but a very similar approach of what we are doing today using LC.
- Etherton (1977): How much data is needed?
  - 4,000 -6,000 examples to get the overall impression of performance
  - 20,000 examples will provide reliable sources of information.
  - No clear empirical evidence

New perspectives: More sophisticated data analysis

- Traditional approach:
  - Simply count frequencies between NS vs. NNS or between different NNS groups
  - Compare the frequencies across groups using significance tests
- Recent approaches
  - Overuse/underuse/misuse → class to be explained
  - Linguistics/task/learner variables as predictors
  - Various statistical approaches are used to build a model of cause-effect relationship or the best predictive model (e.g. regression, discriminant analysis, support vector machine, random forest). (Gries & Deshores 2014; Tono 2013)

New perspectives: More fine-grained error annotation

- Diaz-Negrillo (2007)
- Lozano & Mendikoetxia (2013) : CEDEL2
  - ILA Workshop in Poznan, 2014
- Association rule mining (Tono 2014)
  - If X, then Y. → association rule
  - Association rule mining between the knowledge of grammatical items as prerequisites to other items

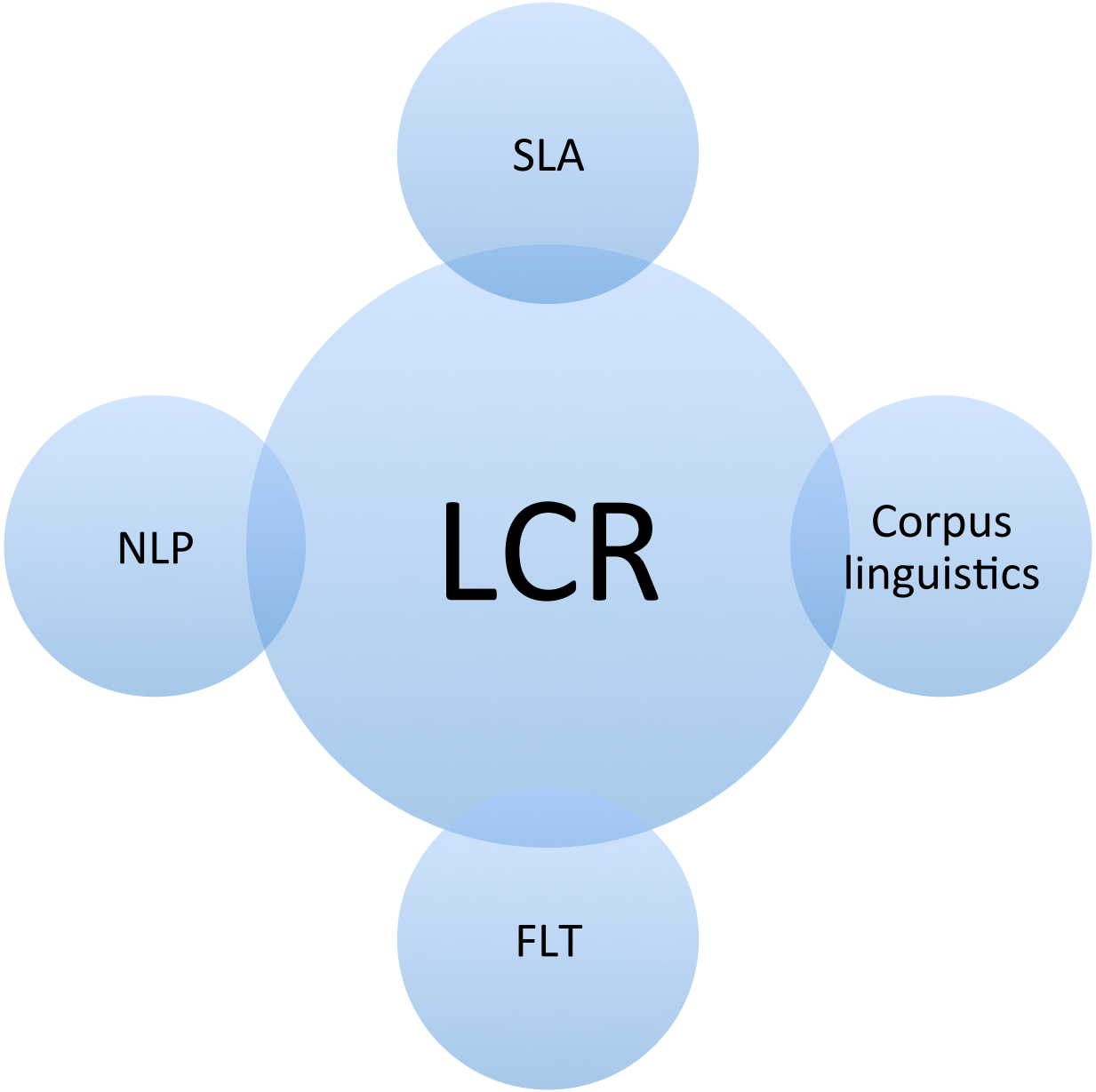
New perspectives: Using big data for LCR

- Lang8 (http://lang-8.com)
  - Free SNS
  - More than 90 target languages from 190 countries
  - Posting writing and corrections made by NS
- EFCamDat (http://www.ling.cam.ac.uk/ef-unit/corpus.html)
  - EF Education First (English language school)
  - 30 million words
  - Learners of English with various L1s
- NLP communities use these big data to do machine learning of automatic error identification and correction

New perspectives: Involvement of NLP communities

- Growing interest in NLP applications in language learning, especially language testing
- CALICO workshop on automatic analysis of learner language
- ETS/ Cambridge English Assessment/ Pearson
- Commercially-led NLP shared tasks (ACL)
  - Automatic error detection & correction
  - Automatic classification of CEFR-level texts
  - Automatic detection of NS vs. NNS texts
  - Automatic identification of writers' L1s

LCR: multiple use of the data



LCR for teaching

- Syllabus design: “criterial features” (English Profile) for CEFR levels
- Materials design: “Common learner errors”
  - Dictionaries (Macmillan/ Longman/ Cambridge)
  - Coursebooks (Touchstone)
- Local learner corpora (Mukherjee 2007)
  - Action research-oriented use of LC

LCR for learning

- Online writing/speaking session
  - Possibility of data mining
- Analysis of LC for individuals can be integrated into the e-portfolio
  - Quantitative & qualitative assessment of progress
- ICALL
  - Integrating learner data to monitor the progress, identify & diagnose the problems, provide the necessary remedial tasks

LCR for assessment

- Automatic scoring of speech & writing
- Automatic error detection & correction
- Data mining of exam data
- Longitudinal analysis of an individual
- Multi-modal analysis of an individual's competence in line with the CEFR descriptors

Conclusions

- LCR has been growing into an independent research discipline, but needs further effort in terms of its relevance to existing SLA theories and methodology.
- LCR will continue to influence areas such as foreign language learning/teaching and language assessment.
- The new approaches in LCR show a promising direction.
- The independent volume for learners of L2 Spanish shows a clear indication of positive aspects of the growth of LCR.

THANK YOU!