Latest Trends in Learner Corpus Research

Elizaveta Smirnova

Plan

- Literature
- Objects of research
- Approaches to complexity assessment
- Data
- Methods
- Recommendations

Literature

- Works on LCR published over the last five years
- Mainly authored by scholars from the Centre for English Corpus Linguistics (CECL), Belgium
- See References

Objects of Research

- Multiword units typical of academic language (Granger, 2017)
- Lexical bundles (Huang, 2015)
- Subject-specific markers (Flowerdew, 2019)

Approaches to Complexity Assessment

- *assessment of formulaic sequences in learner texts*: a technique that assigns to each pair of contiguous words in a learner text two association scores (mutual information and t-score) computed on the basis of a large native speaker reference corpus.
- "Correlation and hierarchical regression analyses, conducted on two datasets of English-*as*-a-foreign-language texts, showed that formulaic measures were the best predictors of text quality and provided a much higher specific contribution to the prediction than single-word lexical measures of diversity and sophistication" (Bestgen, 2017).

Approaches to Complexity Assessment (2)

However

- sentence length
- TTR and
- MTLD (Measure of Textual Lexical Diversity)

are still used to assess L2 proficiency (Bulon et al., 2017)

There is today an urgent need for more text-based or internal methods to assess proficiency level in LCR (Paquot & Plonsky, 2017).

Data

- Most LC studies focus on written corpora, spoken data are explored in a third of works a small number of studies (about 3%) analysed both (Paquot & Plonsky, 2017)
- There is a trend to analyse LC without any reference corpus, i.e. without comparing results with corpus data sampled from native/expert speakers (Paquot & Plonsky, 2017)
- Diachronic approach: learner language across different years of study (Flowerdew, 2019)

Methods

- Relevant shortcomings of LCR as reported by Paquot and Plonsky (2017):
- Corpus linguists often report results for complete (sub-)corpora and rarely inspect by-speaker or by-text results (Brezina & Meyerhoff 2014; Gries 2006a).
- Corpus linguists rarely provide information concerning dispersion as a supplement to frequency data (e.g. Baayen 2001; Gries 2014).
- Corpus linguists often fail to report whether the assumptions of statistical tests have been checked and met (Baroni & Evert 2008; Köhler 2013; Gries 2015b).

Methods (2)

- Chi-square and log-likelihood are considered to be not valid for description of lexical variations between corpora as they produce too many significant results. The author proposes using an easy procedure in R to perform the significance tests (Bestgen, 2017)
- R, alongside with AntConc and Coh-Metrix, is currently gaining popularity among LC researchers with WordSmith Tools being used most frequently (Paquot & Plonsky, 2017).

Recommendations (Paquot & Plonsky, 2017)

1. Substantive areas in need of further attention are Pragmatics and Pronunciation

2 Investigate a greater variety of learner production (i.e. speech in its various forms, more varied genres and tasks).

3 Resort to text-based methods to assess proficiency.

4 Carry out more cross-sectional and longitudinal studies.

5 Check the assumptions of statistical tests.

6. Conduct fewer tests of statistical significance and correct for the alpha level. Be skeptical of p values.

Recommendations (2)

7 Consider multivariate statistics

8 Formulate research questions

9 Identify each software tool used, report the settings employed, and describe each methodological step.

10 Report precision and recall rates for any automatic annotation tool (e.g. POS-tagger, parser) used.

11 Report more thoroughly descriptive statistics, including standard deviations with all means.

References

- Granger, S. (2017). Academic phraseology: A key ingredient in successful L2 academic literacy. *Oslo Studies in Language*, *9*(3)
- Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65-78
- Bulon, A., Hendrikx, I., Meunier, F., & Van Goethem, K. (2017). Using global complexity measures to assess second language proficiency: Comparing CLIL and non-CLIL learners of English and Dutch in French-speaking Belgium. *Travaux du CBL 11 (1)*, 1-25.
- Bestgen, Y. (2017). Getting rid of the Chi-square and Log-likelihood tests for analysing vocabulary differences between corpora. *Quaderns de Filologia-Estudis Lingüístics*, *22*(22), 33-56.
- Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3*(1), 61-94.
- Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. (can share the book)
- Huang, K. (2015). More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. *System*, *53*, 13-23.
- Flowerdew, L. (2019). English as a lingua franca and learner English in disciplinary writing. *Specialised English: New Directions in ESP and EAP Research and Practice*.