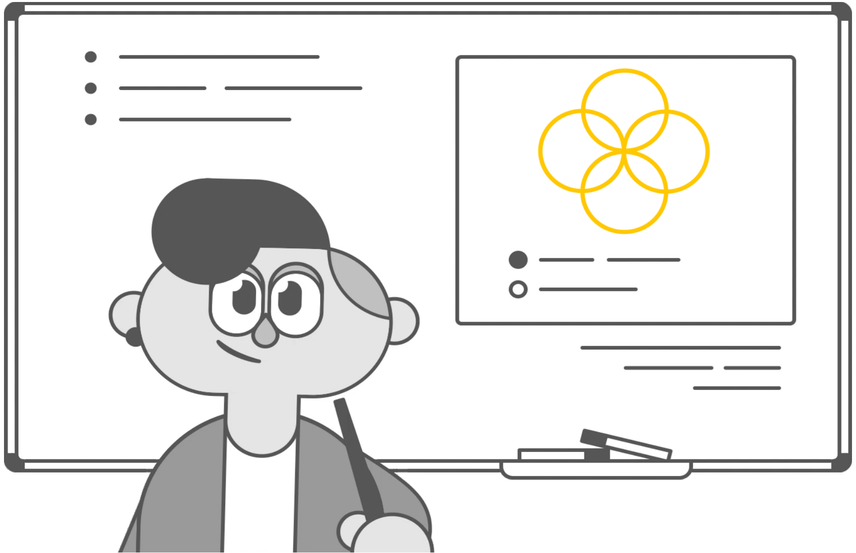
Reading group: Using corpus linguistics to understand the academic domain

Ben Naismith & Ramsey Cardwell
Duolingo English Test

April 21, 2023
BALEAP 2023 Conference

Agenda

- Intros & paper summary (15 mins)
- Methods questions (15 mins)
- Implication questions (25 mins)
- Wrap up (5 mins)



Introductions

Ben Naismith

Ramsey Cardwell

Duolingo English Test
Assessment Scientists

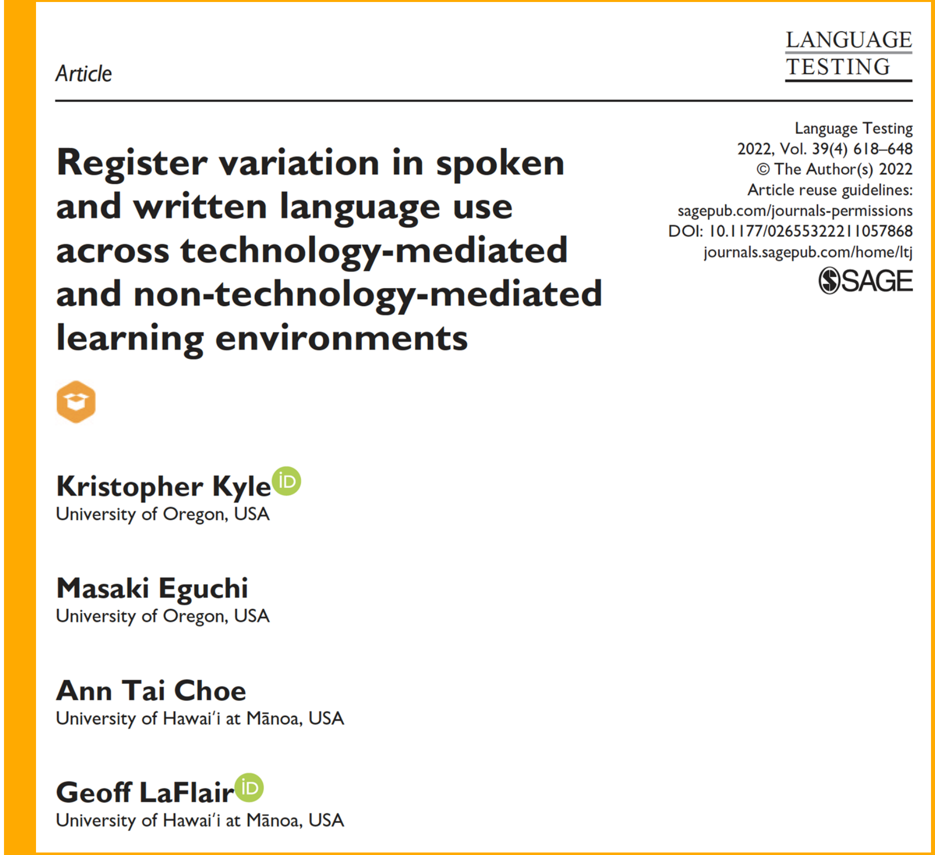
paper summary

PAPER SUMMARY

Focus

- *Changes to the language in current university settings compared to the early 2000s*

- ***RQ:** How (dis)similar are the features of spoken and written language use within and across registers represented in technology-mediated and non-technology-mediated learning environments?*



Article

LANGUAGE
TESTING

Language Testing
2022, Vol. 39(4) 618–648
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322211057868
journals.sagepub.com/home/ltj
SAGE

Register variation in spoken and written language use across technology-mediated and non-technology-mediated learning environments

Kristopher Kyle
University of Oregon, USA

Masaki Eguchi
University of Oregon, USA

Ann Tai Choe
University of Hawai'i at Mānoa, USA

Geoff LaFlair
University of Hawai'i at Mānoa, USA

PAPER SUMMARY

Background: Corpora to describe TLU domains

- Used to identify linguistic features specific to a language use domain
- Register affects the distribution of linguistic features (e.g., Biber 1988, 2004)
- Biber et al. (2004)
  - university **writing** has features that increase informational density
  - university **speaking** has more interactional language features
  - features vary by register across modalities, e.g., syllabi and service encounters share similarities that syllabi and textbooks don’t in terms of amount of procedural discourse

PAPER SUMMARY

Background: Corpora and validity arguments

- Descriptions of linguistic features in a TLU can support validity for language proficiency assessments (e.g., Chapelle et al., 2008)

KANE

  - **Domain description inference** (Is the language elicited authentic in the TLU domain?)
  - **Extrapolation inference** (Are the constructs assessed representative of successful language use in the TLU?)
- Corpora can provide this information, assuming they are representative
- TOEFL validation used T2K-SWAL corpus, but what changed in last 20 years?

PAPER SUMMARY

Background: Technology-mediated learning environments

- Increase in TMLEs, accelerated by COVID-19 (online courses, course management systems, etc.)
- Corpus analysis can identify differences between TMLEs and non-TMLEs, e.g., TMLE project (Kyle et al., 2021)
    - Meaningful differences between TMLEs and non-TMLEs
        - TMLE spoken input more challenging, written input less
    - Still a need to explore register differences

PAPER SUMMARY

Methods

- Comparing texts from T2K-SWAL and TMLE corpora:
  - **T2K-SWAL:** 2.8M words, US unis, 1990s-early 2000s
  - **TMLE corpus:** 4.5M words, TMLE environments in US unis, 2018-2020

- Multi-dimensional analysis (MDA; Biber et al., 2004)
  - TAASSC to analyze lexicogrammatical features (Kyle, 2016)
    Tool for the automatic analysis of syntactic sophistication and complexity
  - Exploratory Factor Analysis to find common latent dimensions

PAPER SUMMARY

Main results

- Dimension 1 (Oral vs Literate Discourse):
    - TMLE speech more ‘writing-like’ than T2KSWAL speech
    - TMLE - syllabi and slides least ‘speech-like’ of registers
- Dimension 2 (Lexical and Phrasal complexity)
    - TMLE registers varied, instructional videos least complex, instructional readings more complex than other written registers
- Dimension 3 (Procedural discourse)
    - More in speech than writing (both corpora)
    - Variation in written TMLE registers
        - announcements & discussions = most
        - presentation slides = least

PAPER SUMMARY

Additional results

- “Few” sig differences between TMLEs & non-TMLEs on Dimensions 4-6
  - Elaborated Discourse—Clausal Complements
  - Narrative Orientation
  - Elaborated Discourse—Relative Clauses

clarification
questions?

methods questions

METHODS QUESTIONS

1. Do you think the TMLE corpus is representative of the target language use (TLU) domain? Would you have included any other text types?
2. Do you think corpus data from this context is generalizable to the TLU domain in the UK?
3. Are there any other corpora you would like to see compared using these methods?
4. How do you feel about focusing on lexicogrammatical features for describing the TLU domain?
5. Have you ever used TAASSC or tools of this nature to analyze texts? Why/Why not?

implication questions

IMPLICATION QUESTIONS

1. Did you find any of the results surprising? If so, why?
2. How might the findings inform EAP assessment practices?
3. How might the findings inform EAP curriculum design practices?
4. How might the findings inform EAP pedagogic practices?
5. What might be some other differences between TMLEs and non TMLEs not captured by corpus data?
6. What further studies in this same vein would be useful for understanding the current TLU domain?

additional readings

ADDITIONAL READINGS

Staples et al. (2022)

Example of similar lexicogrammatical feature analysis, but in a UK context

- Analysis of linguistic complexity development over 1 year in UK EAP context
- Significant differences for most complexity features
- Differences between L1 English and L2 English writers
- Differences across disciplines

ADDITIONAL READINGS

Yan et al. (2020)

Example of dimensionality analysis of **speech** features

- Focus on speaking performances on the Aptis test
- Grouping of features into macro and micro fluencies
- Application of these methods to a Complexity/Accuracy/Fluency (CAF) framework

References

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus. *TOEFL monograph series.* http://www.ets.org/Media/Research/pdf/RM-04-03.pdf

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language.* Routledge.

Kane, M. T. (2006). Validation. In R. L. Brennan, National Council on Measurement in Education, & American Council on Education (Eds.), *Educational Measurement.* Praeger Publishers.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

Kyle, K. D., Choe, A. T., Eguchi, M., LaFlair, G. T., & Ziegler, N. (2021). A comparison of spoken and written language use in traditional and technology mediated learning environments. *ETS Research Report No. RR–21-16*. Educational Testing Service. https://doi.org/10.1002/Ets2.12329

Kyle, K. D. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication. Georgia State University. http://scholarworks.gsu.edu/alesl_diss/35/

Staples, S., Gray, B., Biber, D., & Egbert, J. (2022). Writing trajectories of grammatical complexity at the university: Comparing L1 and L2 English writers in BAWE. *Applied Linguistics*. https://doi.org/10.1093/applin/amac047

Yan, X., Kim, H. R., & Kim, J. Y. (2021). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing, 38*(4), 485–510. https://doi.org/10.1177/0265532220951508

Thank you for attending!
Any questions?

✉ ben.naismith@duolingo.com

🐦 @BenNaismithELT

🌐 www.bennaismith.com

✉ ramsey@duolingo.com

🐦 @RamseyCardwell

🌐 linkedin.com/in/ramseycardwell