Big data in language education and research

Hayo Reinders, King Mongkut's University of Technology Thonburi

Yu-Ju Lan, National Taiwan Normal University

Introduction

The staggering increase in both types and amount of data in recent years is starting to impact many aspects of our lives, from politics to transportation to health care. It is also changing the way we carry out research and make sense of our environment. As Anderson envisioned in the early days of the convergence between increasing data generation, the technological infrastructure to process it, and the artificial intelligence to extract meaning from it: "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves" (2008). Hyperbole aside, significant changes in how research is conceived of are already visible. As Halfpenny and Procter (2015) suggest, "It is possible that it will promote the use of new computational social science methods in place of more traditional quantitative and qualitative research methods" (p. 18).

In language education research, the availability of large sets of data (from corpora to social media posts, and from attendance data to the ways and frequency with which learners interact with online resources) presents intriguing opportunities. If we can track large groups of learners over long periods of time, could we identify common patterns, facilitative and inhibitory variables, and possibly even predict future performance? Could we identify possible problems more easily and intervene more quickly? Could we observe what our learners do beyond the classroom—even after their course finishes? And could we then provide ongoing support for genuine life-long learning (Thomas et al., 2017)?

These questions may seem broad, even far-fetched, but they are starting to become relevant even in regular classrooms—even more so now with the increase in online courses—where teachers have access to learners' data beyond what is observable in class, from grade point averages across school subjects, to comparisons in performance between classes, courses, and teachers (Lan et al., 2017). The availability and often considerable transparency of the data are providing teachers and administrators alike with unprecedented opportunities (Hsiao et al., 2017; Kuromiya et al., 2020; Wen & Song, 2021).

These opportunities come with significant challenges. Data is meaningless unless it can serve a pedagogic purpose, unless it can be mined, and unless it can be validated and interpreted. This may require a rethinking of the role of teachers—as well as everyone else who works in education, from researchers to managers. Researchers are likely to need to develop new skills (Godwin-Jones, 2017). New big data analytical techniques must be adopted to deal with "the five Vs" (i.e. the characteristics of big data, volume, velocity, variety, veracity, and value; see Gandomi & Haider, 2015). All of this is challenging enough, and then there are considerable privacy, security, and ethical concerns. Who owns the data? Who should have access to them? Who safeguards them?

These questions are becoming more pertinent as more data is becoming available. One source is the growing number of devices connected to the internet. It is estimated that by 2023 there will be over 50 billion connected devices. This will go beyond cellphones and computers, to include objects such as cars,

household appliances, and—as the technology improves—clothes, utensils and all manner of everyday items. The "Internet of Things" will also enable physical educational resources (e.g. classrooms, libraries, and also textbooks and worksheets) to be connected to other datasets (e.g. student information, attendance, test scores, etc), offering opportunities for identifying patterns previously difficult or impossible to observe. Another source of the growing availability of data is the increase in online language education due to Covid-19. Some of that will return to face-to-face instruction, but a long-term increase in blended and hybrid forms of education is likely.

All this means that teachers increasingly—and as is clear from the experiences of 2020-2021—urgently need the skills and tools to recognize the types and scope of data available to them, their potential uses, and the ways of putting the insights to pedagogical use (Reinders, 2018).

This Issue

Given the excitement in many scientific fields about the impact of big data, and its relevance to our current educational context, when we were asked by the editors of *Language Learning & Technology* to guest-edit a special issue on this topic. We anticipated a large number of submissions. Interestingly, this did not turn out to be the case. While previous special issues on "What are the digital wilds?" (Sauro & Zourou, 2019) and "Corpora in language learning and teaching" (Vyatkina & Boulton, 2017) received 49 and 67 abstract submissions respectively, we only received 30. Out of these submissions, some did not report on research, and others did not report on big data. Despite the extensions in deadline and an active call for articles, we ended up with just two articles.

In the first paper, "Lexical complexity, writing proficiency and task effects in Spanish dual language immersion," Schnur and Rubio investigated the effect of task types on three measures of lexical complexity, diversity, density, and sophistication. By using a large selection the of written Spanish subsection of the Corpus of Utah Dual Language Immersion, they confirmed that the effect of the tasks indicated that text genre impacts learners' lexical density, while tasks that were more complex elicited higher lexical sophistication. Meanwhile, a broad and deep lexical repertoire is a key feature of more advanced proficiency levels.

The second paper, "Item-level learning analytics: Ensuring quality in an online French course," written by Youngs, reports on the development of a visualization tool aimed at using learning analytics to help the instructor make informed decisions about students' learning. The results of piloting the tool in an online French course showed that it led to identifying low-achieving students and in particular issues with poor course materials, in order to be able to intervene earlier on.

It is difficult to speculate on the reasons for the lackluster response to our call for papers. Perhaps we did not disseminate our call widely enough, or perhaps competing publications diminished the pool of available research. However, given the relatively few publications that have appeared in our field that draw on big data, it is our impression that language education is only just starting to wake up to its potential for learning, teaching and research. It is our hope that the current circumstances will provide an impetus (or a wake-up call) for our field to catch up. The computational turn in language education is coming and it is our duty to ensure it is pedagogically—and not technologically—driven and relevant.

Acknowledgements

References

Anderson, C. (June 23, 2008). *The end of theory: The data deluge makes the scientific method obsolete*. *Wired*. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Gandomi, A., Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Godwin-Jones, R. (2017). Scaling up and zooming in: Big data and personalization in language learning. *Language Learning & Technology, 21*(1), 4–15. Retrieved from https://scholarspace.manoa.hawaii.edu/bitstream/10125/44592/1/21_01_emerging.pdf

Halfpenny, P., & Procter, R. (Eds.). (2015). *Innovations in digital research methods*. Sage Publishing.

Hsiao, I. Y. T., Lan, Y.-J., & Kao C.-L., & Li, P. (2017). Visualization analytics for second language vocabulary learning in virtual worlds. *Educational Technology & Society, 20* (2), 161–175.

Kuromiya, H., Majumdar, R., & Ogata, H. (2020). Fostering evidence-based education with learning analytics: Capturing teaching-learning cases from log data. *Educational Technology & Society, 23*(4), 14–29.

Lan, Y.-J., Chen, N. S., & Sung, Y.-T. (2017). Guest editorial: Learning analytics in technology enhanced language learning. *Educational Technology & Society, 20*(2), 158–160.

Reinders, H. (2018). Learning analytics for language learning and teaching. *JALT CALL Journal, 14*(1), 35–44.

Sauro, S., & Zourou, K. (2019). What are the digital wilds? *Language Learning & Technology, 23*(1), 1–7. https://doi.org/10125/44666

Thomas, M., Reinders, H., & Gelan, A. (2017). Learning analytics in online language learning: Challenges and future directions. In: L. Wong & K. Hyland (Eds.), *Faces of English* (pp. 197–212). Routledge.

Vyatkina, N., & Boulton, A. (2017). Corpora in language learning and teaching. *Language Learning & Technology, 21*(3), 1–8. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44750/1/21_03_commentary.pdf

Wen, Y., & Song, Y. (2021). Learning analytics for collaborative language learning in classrooms: From the holistic perspective of learning analytics, learning design and teacher inquiry. *Educational Technology & Society, 24*(1) 1–15.

About the Authors

Hayo Reinders (www.innovationinteaching.org) is TESOL Professor and Director of Research at Anaheim University, USA, and Professor of Applied Linguistics at KMUTT in Thailand. His interests are in CALL, teacher leadership and autonomy. He is editor of *Innovation in Language Learning & Teaching*.

E-mail: info@innovationinteaching.org

Yu-Ju Lan is a Professor in the Department of Chinese as a Second Language at National Taiwan Normal University. She is currently the editor-in-chief of *Educational Technology & Society*. Her research interests include technology-enhanced foreign language learning, language learning in virtual worlds, mobile learning, and online synchronous teacher training.

E-mail: yujulan@gmail.com