Language Learning & Technology
ISSN 1094-3501
February 2021, Volume 25, Issue 1
pp. 4–19

EMERGING TECHNOLOGIES

Big data and language learning: Opportunities and challenges

Robert Godwin-Jones, *Virginia Commonwealth University*

Abstract

Data collection and analysis is nothing new in computer-assisted language learning, but with the phenomenon of massive sets of human language collected into corpora, and especially integrated into systems driven by artificial intelligence, new opportunities have arisen for language teaching and learning. We are now seeing powerful artificial neural networks with impressive language capabilities. In education, data provides means to track learner performance and improve learning, especially through the application of data mining to expose hidden patterns of learner behavior. Massive data collection also raises issues of transparency and fairness. Human monitoring is essential in applying data analysis equitably. Big data may have as powerful an impact in language learning as it is having in society generally; it is an important resource to have available, but one to use with care and caution.

Keywords: *Big Data, Artificial Intelligence, Neural Networks, Learning Analytics, Data Ethics*

Language(s) Learned in This Study: *English*

APA Citation: Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology*, *25*(1), 4–19. http://hdl.handle.net/10125/44747

Introduction

For many of us, crises in public health and political institutions have absorbed much of our attention in recent times. All the while, however, developments in advanced technologies have taken place—especially in massive data collection, analysis, and application—, which are likely to have a substantial impact on our lives in many ways and in multiple domains. That includes education and language learning. Machine learning, based on advances in artificial intelligence (AI) and applied to huge sets of collected data, has led to breakthroughs in the creation of powerful artificial neural networks. The output of these systems blurs the lines between human and machine-generated speech. At the same time, concerns over bias and inequity in data collection and use have increased. In education, those concerns have become more urgent at a time when more and more teaching and learning is taking place online, raising obvious issues of equitable access, but also of fairness and accountability in the use of automatically generated student data. In this column we will be looking at some of the ramifications of developments in big data collection, including the use of learning analytics, data mining, and data ethics as related to language learning.

Data Flows: From Corpora to Artificial Neural Networks

Learner data has been collected from the earliest days of CALL (Computer-Assisted Language Learning). The analysis of that data can show the effectiveness of experimental treatments, the results of integrating new technologies, or the effect of applications of theories of SLA (Second Language Acquisition). While data collected was limited in scope, analysis could be done simply, for example, through the use of spreadsheets. With the rise of more complex CALL applications, more data necessitated the use of statistical software, such as the R language. That has been particularly the case with the proliferation of online learning and the arrival of MOOCs (Massive Open Online Courses). One of the major developments in data

collection related to language learning has been in the area of corpora. The availability/use of corpora has been one of the signature developments in language technologies in recent decades, to the point that corpus use has become mainstream in many areas, from basic linguistic research to writing textbooks and compiling dictionaries. More slowly, corpus use is taking place in instructed language learning as well (Boulton & Cobb, 2017; Godwin-Jones, 2017a). Chambers (2019) has called for closing the gap between research in DDL (Data-Driven Learning: direct use of corpora by students) and its use in the classroom, with teachers encouraged to engage in action research. A recent article on the use of corpora with student teachers (Meunier, 2019) recommends demonstrating the advantages in the use of DDL, but also suggests avoiding using that or similar technical terms, instead focusing on the practical learning benefits of studying language in context and in demonstrating the confluence of lexis and syntax.

In fact, learners directly consulting corpora has issues beyond its professional jargon. Students accustomed to traditional ways in textbooks and classrooms of working with sentences and of learning grammar patterns, may struggle initially with how language samples are presented in a concordance, as well as with the expectation that they induce patterns of usage themselves. A recent study using think-aloud protocols for analyzing college student use of corpora discussed the heavy cognitive load involved: "Learners may need to autonomously search materials for target linguistic items while being immersed in language data, some of which may be beyond their comprehension" (Lee et al., 2020, p. 346). Lexical inferencing in corpus-based language learning is a complex process requiring active observation and hypothesizing. Lee et al. (2019) found that consequently DDL may be most effectively used with learners at higher proficiency levels, although other studies have shown success working with intermediate or even lower proficiency learners (e.g., Karras, 2016; Mueller & Jacobsen, 2016).

Traditionally, one of the impediments in the use of corpora in instruction has been the lack of user friendliness of corpus access mechanisms, especially KWIC (Keyword in Context), which is especially problematic on mobile devices. Alternative approaches have been shown to be effective, including pre-selection of examples or the use of paper printouts (see Vyatkina 2020a, 2020b); more experimentation in presentation and visualization would be welcome. Flowerdew (2015) has noted the lack of theoretical underpinnings in DDL use and research, while O'Keefe (2020) suggests that DDL be more widely explored in the context of different theories of SLA. A meta-analysis (Boulton & Cobb, 2017) identifies a need as well for delayed post-testing on the long-range benefits of DDL. Boulton (2017) calls for studies examining the effectiveness of corpus use "not just for learning specific items but in helping users to become better language learners, more sensitive to language as a whole" (p. 189). That, after all, is one of the goals we can logically expect from working with corpora, greater understanding of how language works and insight therefore in optimizing learning. In fact, Lee et al.'s study (2020), as have others, demonstrates significant metalinguistic gains on the part of students using DDL.

One of the major advances made possible by corpus linguistics has been in our understanding of the nature of human speech, in particular the importance of formulaic language, multiword units frequently used/reused. The viability of a usage-based understanding of language, reinforced by studies in psycholinguistics (see Millar, 2011), has led to profound insights into language, namely that it "consists of regular overlapping sequences in the form of chunks that are processed, stored, and retrieved as wholes rather than being constructed bottom-up from grammar 'rules' as traditionally thought" (Boulton, 2017, p. 189). That has significant ramifications for language learning; rather than focusing on individual words and grammar rules, learners should pay attention to how words are used in conjunction with other words, from collocations and idioms to syntactical patterns of frame and slots (Christiansen & Arnon, 2017).

The reality of this model of language is further demonstrated by findings arising out of computational linguistics. That is a fairly recent development. Computer science has had difficulty historically in dealing with multiword sequences (Sag et al., 2002) in that they are "highly idiosyncratic, making them difficult to capture using standard machine learning techniques" (Christiansen & Arnon, 2017, p. 544). That idiosyncrasy makes learning to use multiword patterns difficult for L2 learners as well (Christiansen & Arnon, 2017). One of the findings in recent studies on informal language learning has reinforced the idea

that acquiring formulaic language is aided by frequent exposure to examples in context, for instance, through repeated viewing of television shows or movies (Sockett, 2014; Sundqvist & Sylvén, 2016). Now, AI researchers and data scientists are in a way duplicating that process on a massive scale, collecting an enormous volume of language and using a series of powerful processors to analyze the data for statistical regularities. In the process, the system creates an artificial neural network (imitating human brain activity) with weighted, frequency-based connections linking nodes together. Together these multiple "parameters", i.e. mathematical representations of patterns, provide a model of language, based not on rules, but on actual language usage. That enables such systems to predict speech sequencing, based on regularly occurring constructions of words and phrases, thereby enabling the machine production of natural-sounding language utterances, incorporating lexico-grammatical and collocational sequences. One should note that a statistical approach to understanding how language works, through the importance, for example, of word frequency and probabilistic co-occurrence, has been a given in linguistics for some time (Markov, 1913; Zipf, 1949).

The current leader in size and power among artificial neural networks is GPT-3 from OpenAI (short for Generative Pre-trained Transformer, the third version of the system). On the surface GPT-3 is "disarmingly simple" (Vincent, 2020, para. 1). Its interface is familiar from online search, namely text-in, text-out: in essence it is an auto-complete program. But that simplicity belies its power and versatility. GPT-3 can not only complete a phrase or sentence coherently, it can generate connected discourse of considerable length. Its demonstrated capabilities are being seen as a giant step towards the realization of "artificial general intelligence" (AGI), the ability of a system to use language in virtually any domain of human activity (Vincent, 2020). Normally, an AI system will be able to deal effectively only within a narrowly defined domain, for which the system has been trained, so as to expect specific language patterns typically used in that context. Google Duplex, for example, does a remarkable job in conversing over the phone with human operators in making dinner reservations or reserving a ride on Uber (Godwin-Jones, 2019a). GPT-3, in contrast, has been shown to interact through language in a wide variety of genres and content areas: creative writing, journalism, essays, poetry, text-based gaming, and even writing software code (Vincent, 2020). The *Guardian* newspaper ran an article written by the program, while the *New York Times* asked it to write about love. A blogger used GPT-3 to write multiple blog posts, subsequently receiving numerous subscribers and notice on tech websites (Hao, 2020). The fact that many readers were not able to tell that the GPT-3 generated texts were written by an AI system raises questions of trust and authenticity, mirroring the concerns raised about audio and video "deepfakes", based on training an artificial neural network on many hours of real audio or video footage of the targeted individual.

Examples of GPT-3's output are, in fact, remarkable in terms of natural sounding language (idiomaticity, flow, reasonableness), currently in English only. Its ability to use natural language does not derive from any training or reliance on a rule-based language model, but rather from mining its huge collection of data (45 terabytes or 45,000 GB), drawn from crawling the Internet, including all of Wikipedia. While GPT-3 can provide textual responses to any query, its effectiveness can be enhanced by providing both a prompt and just a few examples (known as "few-shot generalization"). The sematic analysis of the system calculates probabilities of text sequencing, looking at what words and expressions are used adjacently within the (175 billion) parameters it has stored. One can imagine how powerful GPT-3 could be integrated into a chatbot. We are already seeing interesting uses of chatbots and intelligent assistants in language learning (Fryer et al., 2020; Godwin-Jones, 2019a). A company called LearnFromAnyone is building on top of GPT-3 a kind of automated tutor, which can take on the identity of famous scientists or writers.

The achievement represented by GPT-3 is substantial. On the other hand, although the texts it generates are coherent and idiomatic, they don't always make sense in terms of lived human experience (Marcus & Davis, 2020). The system is very good at words, but not so good at understanding social norms and cultural practices. It has not lived in our world and therefore lacks the pragmatic abilities we develop through living and communicating with others that allows us to generate language that is contingent on human interactions and appropriate to the context. Not available to GPT-3 as well are the tools of embodied communication (gesture, gaze, posture), dependent as they are on physical existence and lived experience. GPT-3 and other advanced AI systems also raise a whole host of ethical issues, as discussed below.

Learning Analytics and Data Mining

Big data, such as that used by GPT-3, is different from traditional collected data in significant ways. Big data is understood to be "huge in volume, high in velocity, diverse in variety, exhaustive in scope" (Tsai et al., 2020, p. 555; based on Kitchin, 2014). In contrast to partial, sampled data, big data is comprehensive and not collected in order to address a specific issue or to answer a set of research questions. As Tsai et al. (2020) comment, the perceived objectivity and comprehensiveness of big data have "led to an assertive, optimistic form of empiricism" (p. 556) with the implication that simply due to its size, such systems can offer powerful and effective insights. The term "techno-solutionism" has been coined to describe that perspective (Prinsloo & Slade, 2017). The idea that the scale of big data combined with AI can provide effective solutions is evident in domains such as e-commerce (recommendation systems) or online search.

That confidence can often be found in education as well, although the scale of data generated is not at a level comparable to that collected by AI systems or through consumer online services. Data use in education generally manifests itself in the form of *learning analytics*. Learning analytics has been defined as "the measurement, collection, analysis and reporting of data about learners in their context, for purposes of understanding and optimizing learning and the environment in which it occurs" (Siemens & Long, 2011, p. 33). With the widespread use of learning management systems (LMS) and the surge in online learning, large amounts of data on student learning performance are being collected. That processed data is available to learners, typically through an LMS dashboard or online gradebook, showing assignment completion status and grades. Instructors have access to the results of individual student performance as well as to group statistics. The learning analytics provided are intended to help students and instructors monitor learning and to take action as indicated. That might involve automatic alerts in the form of email or other forms of notification. Given the high dropout rate in online courses, such actions can be helpful. The data provided by an LMS, however, focuses primarily on administrative kinds of reporting data, such as frequency of sign-ons or number of clicks on particular resources. The visualization tools used tend to simplify the learning process and may not provide a full picture of its complexity (Gelan et al., 2018; Reinders, 2018).

The effectiveness (and fairness) of applying learning analytics to actual learning depends on a variety of factors, including what and how data is collected and visualized, the degree of individual performance information included, and the transparency of the process. Concerns over the use of learning analytics include privacy and security, as well as the danger of the data reports contributing to a student profile that is standardized and geared towards administrative efficiency (Roberts-Mahoney et al., 2016). Learning analytics holds the unfortunate potential of viewing students as numerical representations, not as flesh and blood humans with complex lives and individualized experiences. Reinders (2018) comments: "If a student did poorly in one class because his pet dog died and had a hard time as a result, is he going to be classified as a poor learner for the rest of his school career?" (pp. 84). A human teacher is able to make allowances for life experiences, but an algorithm cannot. I have argued recently for the importance of an active human teaching presence in online language learning (Godwin-Jones, 2020). That humanizing element takes on an even greater weight at a time when resources at educational institutions are dwindling, resulting in a push towards efficiency, measured by metrics like graduation rates or course enrollments.

Reinders (2018) argues that data analysis, applied to education, may encourage "a mechanistic, behaviorist view of learning" (p. 84). That may not be a major concern in disciplines in which there are clear right or wrong answers and linear learning pathways. That reductionism is a problem when it comes to language learning. Given the clear evidence that language constitutes a complex dynamic system (Larsen-Freeman, 1997, 2019), language learning must consequently be seen as well as an open, dynamic system and unpredictable in its course, dependent on initial conditions and the subsequent interactions between learner and environment (Godwin-Jones, 2018a, 2019b). Prinsloo and Slade's (2017) warning concerning the reliance on big data applies in spades to SLA: "Student success is the result of mostly non-linear, multidimensional, interdependent interactions at different phases in the nexus between student, institution and broader societal factors" (p. 113). Just as the use of an LMS can suggest that learning can be

compartmentalized in terms of discrete semester courses—not as an individualized, real-world lifelong endeavor—learning analytics implies that collected performance data defines learning. That is particularly pernicious in SLA, where opportunities for informal, non-academic learning abound.

Many studies of learning analytics have largely supplied answers to questions that are already known or are commonsensical. One study, for example, found that students completing assignments on time were more successful, as were students completing more exercises or accessing explanatory information more frequently (Gelan et al., 2018); in other words, conscientious students performed better. Other studies showed that regular submission of assignments was a "robust indicator towards course success" (Martín-Monje et al., 2018, p. 1), incidentally reported to be a "novel perspective into students' language learning" (p. 1). A frequent problem with learning analytics is a tendency to look at group data in cause-and-effect terms and as revelatory of individual student performance.

One approach that can deliver more specific information is to look beyond group averages and focus on the evidence the data supplies for marked individual or small group patterns. This is the goal of educational *data mining* (Warschauer et al., 2019), which is seeing wider use in CALL. In a study on the effectiveness of DDL, Lee et al. (2019) applied data mining to uncover hidden clusters of learners not apparent from analysis of the whole group. Such an approach enables drilling down into data on specific learning activities. Youngs et al. (2018) supply an example using arc diagrams to visualize data from students in an online French course interacting with a course video and its attendant questions. Visualizations can be helpful in differentiating groups of learners. The text mining tool DocuViz was used, for example, in uncovering different patterns of collaboration among undergraduates' synchronous writing in Google Docs (Yim & Warschauer, 2017). In Hsiao et al. (2017), visualization analysis revealed a link between the learning paths/strategies and learners' outcomes in three separate groups. For social network analysis, tools such as UCINET and Pajek (Warschauer et al., 2019) are frequently used, which visually depict interactions in online exchanges. A study by Zhu (2016), uncovered distinct configurations in multi-party exchanges, namely star (centralized) or network (interconnected) patterns, easily distinguished when mapped graphically. Butler & Liu (2019) used social network analysis to identify usage and learning patterns among young learners collaborating.

While such data can be highly informative in guiding instructional approaches to different kinds of learners, the process of obtaining that information is more challenging than is the case with learning analytics. Compared to the "low hanging fruit" approach of LMS assessment data and visualizations, for data mining, technical skills in statistics (R language), databases (SQL), and programming (Python) are likely necessary (Youngs et al., 2018, p. 17). Tools for educational data mining are evolving rapidly with a "current move from the lab to the general market" of educational institutions (Romero & Ventura, 2020, p. 16). One promising model is represented by the Pittsburgh Science of Learning Center DataShop (Carnegie-Mellon University), which incorporates a variety of necessary tools and steps into one integrated app (Slater et al., 2017). Given the benefits of identifying individual/group learning patterns, easier applications of data mining to real-world settings has the potential to transform learning analytics into *learner* analytics. This is particularly important for SLA, which is such a dynamic and individualized process. That process occurs over time, making it also necessary to collect data at different points in a learner's trajectory:

> If we are interested in grand sweep effects that may be generalizable to large populations of learners, we will have to carry out group studies with representative samples that can be analyzed using Gaussian statistics based on the normal distribution. But if we are interested in how an individual learner progresses over time as a result of changing variables in a changing context, we will have to conduct longitudinal studies and use nonlinear methods of analysis (Lowie & Verspoor, 2015, p. 63)

Both group studies and longitudinal case studies are necessary, to determine overall effectiveness of treatments, as well as to be able to focus in on individual learners' development (Lowie & Verspoor, 2019). Quantitative measures can often be usefully supplemented/individualized through qualitative approaches.

Data Openness

A major issue with LMS data is its closed status. Data typically is available only within the proprietary system with limited import/export options and generally only accessible during the term of use. As discussed above, the model of learning presented by LMS data to learner and teacher provides a limited conception of the learning process. The generic nature of the feedback and assessment data is not as likely to be helpful or motivating to the learner as would be information that corresponds to a more inclusive account of learning activities, many of which are likely to occur outside the LMS. There are both practical and technical impediments to that possibility. On the technical side, data would need to be collected from diverse sources and platforms, including multiple mobile apps, online services, and informal learning resources, such as participation in social media. While some data may be accessible through export options or APIs (Application Programming Interface), much will be inaccessible or available as unstructured text. Practical issues are in the areas of security and privacy, including setting permissions for access and ensuring that the data is authenticated for the named individual.

One of the options available for gathering and displaying learner data from multiple devices and services is the use of an "open learner model", which establishes a personal profile of the learner in terms of language proficiency, specific skills, learning style, language learning goals, etc. (Bull, 2020; Godwin-Jones, 2017b). Learner models are an essential element in intelligent tutoring systems, although they are typically used and accessible only within that system. Rosell-Aguilar (2018) advocates for the emergence of a universal and open language learning monitor, envisioning a unifying dashboard-style app such as that in Apple's Health app. Particularly useful would be the ability to integrate informal learning activities, both online and in person. That could involve automatic recording and the option to enter data manually. The collected data would be stored in the cloud and available by default only to the user, with the option to share with others, for possible use in formal educational settings. Bull (2020) discusses different examples and characteristics of open learner models, along with the different reporting and visualization tools available. That includes models which have been used in language learning, including Next-TELL (Next generation Teaching, Education, and Learning for Life) and Lea's Box (Learner analysis toolbox; Bull & Wasson, 2016).

Studies listed in Bull (2020) point to a range of benefits for the users of open learning models, including increased motivation, more effective self-monitoring, enhanced metalinguistic skills, and a greater sense of learner control/autonomy. Of particular interest are "persuadable learning models," which provide the user the option to challenge the learner status the model indicates (Bull, 2020). That might occur through the user supplying additional data or through an automatic process in which the learner responds to queries. Examples are the SMILI project (Bull & Kay, 2007) and STyLE-OLM (Dimitrova & Brna, 2016). Yousuf & Conlan (2020) found that the use of a persuadable open learner model in online learning enhanced student motivation and engagement. In order to provide maximum flexibility and user-friendliness, some systems allow learners to choose among input methods, using language rules, example sentences, or grammatical statements (Bull, 2020). While models typically use an algorithm assigning predefined weighting for items based on factors such as time (recent entries weighted more heavily), readability index of materials, or other parameters, some allow for teacher editing when used in a formal educational setting (Bull & Wasson, 2016). At this point, open learner models have been developed largely as academic projects, with limited distribution and sustainability. One of the hopeful developments for greater usage is the integration of open learner models into e-portfolios (see Raybourn & Regan, 2011, for an example).

The movement towards openness, transparency, and interoperability runs counter to mainstream business models in online and educational tools and services, for which proprietary systems ensure exclusivity and hence profitability. At the same time, there are increasing voices for greater openness and accessibility. That has been evident in North America in recent years in the growth of open access textbooks. Within academia, there has been considerable interest in open access to research data and publications, given issues of equitable access and rising cost. The interest in sharing teaching materials through open educational resources (OER) continues, despite barriers such as low funding and lack of professional credit/recognition (MacWhinney, 2017). One of the impediments to open and sharable resources is the reality of content being

trapped in outdated or proprietary formats (Colpaert, 2016). Enabling content sustainability necessitates the use of standards and interoperability services. An example is LTI (Learning Tools Interoperability), a standard developed by IMS Global, which allows third-party tool integration into an LMS, now widely used in higher education. One of the standards which is seen growing acceptance recently is the ExperienceAPI (or xAPI). It is used, for example, for data import/export in the VITAL project (Visualisation Tools and Analytics to monitor Online Language Learning & Teaching; Gelan et al., 2018) as well as in Next-TELL (Bull & Wasson, 2016). The xAPI standard has a simple, familiar format (subject-verb-object) which is human readable (using JASON, a language-independent data format) and can be used to exchange data from formal and informal sources, either synchronously or asynchronously (Johnson et al., 2016).

Open data is not a panacea for assuring universal access to data collection and distribution. Smith and Seward (2020) write: "Given that the contributions of technology to social change are a function of a myriad of contextual factors, it could be that innovations in openness are contributing to and exacerbating existing inequalities" (p. 8). Studies have shown that openness practices, such as peer collaboration, may be challenging in low resource settings (Berdou, 2017). There are disparities in contributions to open access resources, with contributors coming overwhelmingly from developed and Western economies. Contributors to Wikipedia, for example, who are unpaid volunteers, need both access and leisure time to edit entries (Graham, 2020): "The underlying expectation of spare-time contributions probably exacerbates economic inequities, and may be a significant contributor to global information imbalances (Graham et al., 2019, p. 74). This situation has led to uneven contributions to public information about locations across the globe. The "uneven geographies of information" (Graham et al., 2019, p. 62) are evident in the fact that information about Western countries is produced overwhelmingly in those locations, while information elsewhere is shaped by outside perspectives. That is especially the case for African countries, where approximately 5% of contributions are produced locally (Graham et al., 2019). Geospatial content is becoming more important, as augmented realities become increasingly prevalent, with layers of digital code and content which serve to shape our perceptions of social and cultural spaces (Graham, 2020). Promising in that regard are projects such as that described in Ashikoto et al. (2018), which allows for local users in Namibia to add content to an AR app. Another project along those lines is the creation of localized OER in Afghanistan, described in Oates et al. (2017). Similar open access and local projects are discussed in Hodgkinson-Williams et al. (2017).

Social Concerns

The imbalance in geographical and cultural contributions to open data about people and places indexes unequal power distribution globally, reflecting profound disparities in education opportunities, socio-economic status, and resource allocation. Those disparities range from a lack of basic human necessities (food, shelter) to available Internet access. Wikipedia has recently launched a project to address the imbalance among its contributors by providing support for underrepresented groups and seeking to eliminate barriers to contribution (Graham, 2020). The disparities are evident not just in underdeveloped economies, but also within developed economies, where social and economic inequality is prevalent in underserved communities, often consisting of black or brown populations in urban centers.

To provide more informed and balanced representations of underprivileged communities, enlisting local contributions is essential. D'Ignazio & Klein (2020) in their study of inequities in data science discuss several examples of what they label "co-liberation" in data-informed projects, understood to be "a commitment to and a belief in mutual benefit, from members of both dominant groups and minoritized groups" (chap. 2, Principle: Challenge Power, para. 9). *Local Lotto* demonstrates that approach (Rubel & Nicol, 2020). Created by a math teacher, this "mathematics for spatial justice" project combines math, Spanish, and community engagement. High school students are tasked to collect data through interviews with local inhabitants of a Latinx urban community, addressing the question of whether a lottery is good for the community. Students learn practical lessons in money management and in community relations. The

project has transformational goals in terms of reimagining urban space and highlighting its importance in lived human experience.

Another project which combines language, data science, and community engagement was developed by SAFElab, a research lab at Columbia university (Blandfort et al., 2019). It involved analyzing Twitter data to understand and prevent gang violence in Chicago. The team working on the project discovered that they needed help in understanding the youth language used in the tweets. They hired former gang members as domain experts, who coded and categorized a subset of the tweets collected. That provided the cultural and linguistic knowledge needed to interpret non-standard uses of terms:

> For example, a tweet like "aint kill yo mans & ion kno ya homie" would likely have been classified as aggressive or violent, reflecting its use of the word "kill." But drawing on the knowledge provided by the young Black men they hired for the project, Frey and Patton [project PIs] were able to show that many tweets like this one were references to song lyrics, in this case the Chicago rapper Lil Durk. In other words, these tweets are about sharing culture, not communicating threats (D'Ignazio & Klein, 2020, chap. 6, Raw Data, para. 10).

The next step was to train a machine learning classifier to label tweets accordingly. In the words of co-PI Patton, "We trained the algorithm to think like a young African-American man on the south side of Chicago." (cited in D'Ignazio & Klein, 2020, chap. 6, Raw Data, para. 8). This is a striking example of enlarging the language scope in data collection. D'Ignazio and Klein (2020) argue that data science can be a means "to remake the world" (Introduction, para. 11). Equity in the representation of language can be a step in that direction.

The projects above demonstrate one of the main points of D'Ignazio & Klein's study (2020) – and a guiding principle of data ethics—namely to consider collected data from the perspective of its contextual origins. That concept corresponds to the finding in applied linguistics that language use is *situated*, that its manifestations in real-world usage need to be reviewed and understood in relevant social, cultural, historical, institutional, and material conditions, as well as in consideration of the personal identities of the speakers (Ellis, 2015; Lantolf, 2006). In the creation of language models, that translates into inclusive practices that collect language data from diverse communities.

That process is not always followed in language data collection. Mayfield et al. (2019) discuss the *algorithmic bias* in data collection and in natural language processing (NLP). The authors point out that because most language models are trained on standard written professional English, "NLP performance is degraded for underrepresented groups," such as African-Americans (p. 445). Such exclusions have real-world consequences, in that NLP-based materials such as texts with comprehension questions or graphic representations through avatars (in immersive games/tutorials) fail to represent some cultures. Representations generally do not account for conditions such as autism, deafness, or disability. In addition, a binary definition of gender is assumed (Mayfield et al., 2019). Language learning systems "make numerous design choices to implicitly or explicitly reject the grammar and lexicon of minority dialects. Typically, code-switching is neither taught as a skill nor supported as input" (Mayfield et al., 2019, p. 449). In other words, language learning systems do not generally allow for real world language usage:

> An equitable language treats cultural knowledge instead as an asset, and allows students to build on what they know. This extends to technologies used in the everyday lives, homes, and communities of students—influencing their ability to impact student learning outcomes. (Mayfield et al., 2019, pp. 447–8).

Mayfield et al. (2019) argue that, as in the concept of collaboration discussed above, cultural representation should be built into NLP systems, for example in the development process "through teams with 'cultural competence' through lived experience and group membership shared with the students they are building applications for" (pp. 450–1). Inviting wider participation in projects, including by local inhabitants, can build confidence, pride, and a sense of ownership, leading to greater acceptance and motivation. A model in that direction is a storytelling project involving migrant youth in the Netherlands (Vandommele et al.,

2017). The authors found that multimodal L2 writing was an effective way to reach adolescents, often disengaged from traditional literacy practices, a problem likely to be exacerbated by struggles to communicate within a different culture. Multimodal writing allows struggling writers (L2 learners and migrants) to take on identities as productive students, sharing their personal backgrounds in a positive environment (Godwin-Jones, 2018b). The recognition of learners' personal linguistic background offers validation of their identities. A similar positive effect was reported in Smith et al. (2017), involving bilingual students creating multimodal projects built around personal, local heroes.

Inclusive practices in language data collection and analysis may interfere with convenience and efficiency, but they reflect a necessary redirection in language technologies. Data science evinces a clear preference for "clean" data (D'Ignazio & Klein, 2020), but the reality is that the world is messy and complicated. Ignoring that reality can lead to a distorted view of language use as well as the reinforcement of stereotypes. Clean data has the tendency to minimize outliers, hide contradictory realities, and discount uncertainty. D'Ignazio and Klein (2020) discuss how difficult it is to convey uncertainty in data visualizations in a way that is understandable to viewers. That was demonstrated dramatically in the last two US presidential elections, in which the representation of poll and election results led to widespread confusion and misinterpretation.

Data visualization aims for clarity of messaging, uncluttered presentation, and an impression of rational objectivity. That tends to eliminate pluralism and ambiguity in the messaging. At the same time, the big picture presented can obscure important details, such as historical-cultural contexts. A counter-example is the map of Canada created for the *Come Home to Indigenous Places* project. In creating the map, in place of the normal placing of markers indicating population size, this project maps the location of indigenous tribes and leaves out urban centers all together, including those located in tribal areas. The effect is an attention getter, with an emotional appeal based on the visual evidence of historical injustice through colonization. A similar emotional resonance is achieved in the *Tech Bus Stop Eviction Map*, which shows the overlap of technology company bus stops (San Francisco technology workers commuting to Silicon Valley) and gentrification and, consequently, to evictions (D'Ignazio & Klein, 2020). Data representation is often seen as an objective process, a value-neutral phenomenon, but data involving humans always has a social dimension. Visualizations are based on choices in terms of what data to use and how to display it. Using data visualization as a motivator for social action may need to elicit an emotional response, as in the examples cited above.

Conclusion: Pedagogical Implications

The articles in this issue of *LLT*, as well as other recent studies involving data-informed language learning and teaching demonstrate the potential for using statistical analysis and visualizations of performance records to inform the effectiveness of specific pedagogical approaches and implementations. Learning analytics can provide specific feedback to learners themselves that has the potential to provide more guidance on learning strategies and behaviors (Tsai et al., 2020; Youngs et al., 2018). The individualized learning data in turn sets up the possibility of using that information to tailor learning to the needs, preferences, and abilities of individual learners (Godwin-Jones, 2017b). In addition, data mining provides the opportunity to identify hidden groups of learners who may profit from differentiated instructional approaches or alternative learning materials (Warschauer et al., 2019).

At the same time, caution is advisable in pedagogical practices, when it comes to the collection and use of personal learning data. Reliance on student performance data has the potential to depersonalize the learning experience, reducing students to statistical data (Phuong et al., 2017). There is the tendency as well in relying on data to view the learning process as linear and predictable, a particularly untrue proposition when applied to language learning. Data collection should not be viewed as neutral and objective; socio-economic factors and personal histories are rarely considered in learning analytics. In addition, the crucial role of the teacher may be viewed as less important, compared to the information and guidance supplied by the analysis of learning data (Papi & Hiver, 2020; Roberts-Mahoney et al., 2016).

A human monitoring role is important as well in the use of AI-based systems such as GPT-3, built on huge collection and analysis of language data. The growing presence of such powerful and effective text-generating machines in society generally and within education specifically, raises crucial issues of authenticity, attribution, and authorship. Experimentation with pedagogical uses will need to be done with that awareness, just as many language teachers have been doing in discussing the use of machine translation. While language learning uses of GPT-3 have been demonstrated, including its ability to provide samples of language patterns or to identify stylistic traits and provide pertinent examples (Vincent, 2020), such uses are less instructive than what this AI system reveals simply through the limitations in its text production. Those limits point to aspects of human language important to convey to students, namely its essentially cultural and pragmatic nature. While language is a statistical phenomenon, based on probability (Ellis, 2015), its use in real-world exchanges is personal, social, and dynamic, based on, but not constrained by, patterns of usage. In that way, language is an *emergentist* process which is profoundly human (MacWhinney, 2001), a reality reinforced by the nature of GPT-3's texts which lack real-world coherence, with a "tunnel-vision understanding of how words relate to one another" (Marcus & David, 2020, Non sequiturs, para. 11).

One of the principal issues associated with data collection and its use in education is that of *personal agency*, for both learner and instructor. Learning analytics treats students as "rational agents," endowing them with the responsibility for understanding and acting on system-supplied feedback (Tsai et al., 2020). This view of agency, based on individual cognition, runs counter to the reality of students as social beings, whose actions are bounded by personal, social, and institutional factors. That complicates considerably the ability of the learner to respond to the expected model of changed behavior the system's algorithm may advocate:

> The emphasis on action informed by predictive models has, for its critics, a tendency to prioritise effects and indicators (signals) over causes, thus leading to narrow remedial strategies in which students and teachers are channelled along predefined trajectories of educational performance that, paradoxically, leave little room for agency (Tsai et al., 2020, p. 557).

In fact, one of the impediments to implementing recommended "actionable insights" (Siemens, 2013) provided by learning analytics may be quite practical, namely insufficient specificity in the feedback or uncertainty over how to understand/interpret assessment visualizations (Youngs et al., 2018). Added to that is the larger issue of the inherent imbalance of power relationships involved in data collection/analysis which "poses questions about the extent to which individuals can truly make informed decisions about the use of their data" (Tsai et al., 2020, p. 558). That is particularly the case when one considers that AI systems do not take into consideration socio-cultural and economic backgrounds.

That deficit has recently been recognized in the AI community itself. The approach to affecting change, however, has been predominantly computational, adjusting parameters in data collection and filtering input for evidence of bias or intolerance. GPT-3, for example, in its indiscriminate crawling of Internet data inevitably collects abundant hate speech and biased language, which then may be included in text production. OpenAI addresses this issue in their usage guidelines for GPT-3:

> Our API models do exhibit biases that will be reflected in generated text. Here are the steps we're taking to address these issues: We've developed usage guidelines that help developers understand and address potential safety issues. We're working closely with users to understand their use cases and develop tools to surface and intervene to mitigate harmful bias. We're conducting our own research into manifestations of harmful bias and broader issues in fairness and representation, which will help inform our work via improved documentation of existing models as well as various improvements to future models (OpenAI API, 2020, FAQ: How will OpenAI mitigate harmful bias, para. 1).

The steps outlined here include developing guidelines, producing more research, and generating "improved documentation." Absent is one of the measures that would likely be most beneficial, namely increasing the diversity of researchers on the project. Technology companies tend to view issues of bias and inequity in their products as solvable through technology, rather than through human actors or societal change.

Google's firing of Timnit Gebru in 2020, a leading voice in AI ethics, does not bode well for the industry in that regard.

Treating students as rational agents is problematic in all academic areas, but is particularly so for second language learning, which engages so intensely learners' emotions, often affecting profoundly issues of personal identity (Swain, 2013). Increasingly in applied linguistics, language learning is seen as complex and multifaceted, with non-cognitive factors playing essential roles in the process (Butler, 2019), especially in the vital area of learner motivation. Motivation research in SLA increasingly seeks "to understand the motivational dynamics of language learners as whole persons situated within a broad and dynamic social context" (Papi & Hiver, 2020, p. 209). Viewing learners as real persons, not as numerical abstractions places SLA within a broader context of human activity and growth (Larsen-Freeman, 2019). From that perspective, relying on learning analytics to guide learner behavior restricts considerably the options available for enhanced language learning. It may be, for example, that for some students, particularly those from underprivileged backgrounds, alternative approaches, such as online gaming, can prove effective and motivating. Those are not likely to be included in algorithm-generated feedback (see Phuong et al., 2017).

Reinders (2018) points out that learning analytics "may lead to over-monitoring and micro-managing of students" (p. 84). The danger of "spoon feeding students" (Tsai et al., 2020, p. 561) can be counteracted by the monitoring and mentoring roles of a teaching presence, whether that be online or in person (Godwin-Jones, 2020). The concept of "structured unpredictability" (Little & Thorne, 2017) captures well the advantages of combining structured learning, which may include automated feedback from learning analytics, along with the nurturing presence of a teacher, with the unstructured experiences of informal language learning. The latter is occurring increasingly today online, through social media, streaming services, and participation in Internet affinity groups (Godwin-Jones, 2018a).

Building in access to informal language learning can set the stage for a "pedagogy of serendipity" (Anwaruddin, 2019, p. 12), the opportunity for learners to explore unexpected resources and encounter cultural Others, including some who may be beyond learners' normal sphere of contact. That allows instructors to provide "contingent scaffolding" that helps students to engage with alternative viewpoints and with individuals who may be outside their comfort zones (Anwaruddin, 2019, p. 12). This can additionally have an impact on the "development of a student's problem-solving abilities, which is sometimes gained through a painful learning process" (Tsai et al., 2020, p. 562; see also Levine, 2020, on conflict resolution). Social media often work in the opposite direction, built as they are around "networked individualism" (Rainie & Wellman, 2019), namely, encouraging users through algorithm-based recommendation systems, to stay within their established bubbles, leading to system-reinforcing echo chambers. Learners seeking to expand their horizons are taking on the difficult task of swimming against the tide. Such efforts may bring peer group disapproval and challenge facets of personal identity. At the same time, stepping out of one's lane can, like the process of second language learning itself, expand personal identity, create solidarity with other communities, and encourage a joint sense of global citizenship.

References

Anwaruddin, S. M. (2019). Teaching language, promoting social justice: A dialogic approach to using social media. *CALICO Journal*, *36*(1), 1–18.

Ashikoto, L. I., Ajibola, D., & Virmasalo, V. (2018, December). A room for social justice: Affective and interactive augmented reality exploration. In *Proceedings of the Second African Conference for Human Computer Interaction: Thriving Communities* (pp. 1–4). *ACM*. https://dl.acm.org/doi/pdf/10.1145/3283458.3283505?casa_token=XD5L4NQEZskAAAAA:p55rlXlPUZvdf979RSAvSf1XeLKecYdoE8PkseS6uMF0wpd-3E8jvZYE_XbA0nljTjU-IUIqknbAag

Berdou, E. (2017). Open development in poor communities: Opportunities, tensions, and dilemmas. *Information Technologies & International Development, 13*, 18–32.

Blandfort, P., Patton, D.U., Frey, W.R., Karaman, S., Bhargava, S., Lee, F.T., Varia, S., Kedzie, C., Gaskell, M.B., Schifanella, R. and McKeown, K. (2019). Multimodal social media analysis for gang violence prevention. *Proceedings of the International AAAI conference on web and social media*, *13*, 114–124.

Boulton, A. (2017). Data-driven learning and language pedagogy. In S. Thorne & S. May (Eds.). *Language, Education and Technology: Encyclopedia of Language and Education 3* (pp.181–192). Berlin: Springer.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta‐analysis. *Language Learning*, *67*(2), 348–393.

Bull, S. (2020). There are open learner models about!. *IEEE Transactions on Learning Technologies, 13*(2), 425–448.

Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI: Open learner modelling framework. *International Journal of Artificial Intelligence in Education*, *17*(2), 89–120.

Bull, S., & Wasson, B. (2016). Competence visualisation: Making sense of data from 21st-century technologies in language learning. *ReCALL, 28*(2), 147–165.

Butler, Y. G. (2019). Linking noncognitive factors back to second language learning: New theoretical directions. *System*, *86*, 102–127.

Butler, Y. G., & Liu, Y. (2019). The role of peers in young learners' English learning: A longitudinal case study in China. In M. Sato, & S. Loewen (Eds.), *Evidence–based second language pedagogy: A collection of instructed second language acquisition studies* (pp. 145–167). New York: Routledge.

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, *52*(4), 460–475.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in cognitive science*, *9*(3), 542–551.

Colpaert, J. (2016). Big content in an educational engineering approach. *Journal of Technology and Chinese Language Teaching*, *7*(1), 1–14.

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Dimitrova, V., & Brna, P. (2016). From interactive open learner modelling to intelligent mentoring: STyLE–OLM and beyond. *International Journal of Artificial Intelligence in Education*, *26*(1), 332–349.

Ellis, R. (2015). *Understanding second language acquisition*. Oxford University Press.

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15-36). Amsterdam: John Benjamins.

Fryer, L. K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology, 24*(2), 8–22. http://hdl.handle.net/10125/44719

Gelan, A., Fastré, G., Verjans, M., Martin, N., Janssenswillen, G., Creemers, M., Lieben, J., Depaire, B., & Thomas, M. (2018). Affordances and limitations of learning analytics for computer–assisted language learning: A case study of the VITAL project. *Computer Assisted Language Learning*, *31*(3), 294–319.

Godwin-Jones, R. (2017a). Data-informed language learning. *Language Learning & Technology, 21*(3), 9–27. http:// www.lltjournal.org/item/3012

Godwin-Jones, R. (2017b). Scaling up and zooming in: Big data and personalization in language learning. *Language Learning & Technology, 21*(1), 4–15. http://llt.msu.edu/issues/february2017/emerging.pdf

Godwin-Jones, R. (2018a). Chasing the butterfly effect: Informal language learning online as a complex system. *Language Learning & Technology, 22*(2), 8–27. https://doi.org/10125/44643

Godwin-Jones, R. (2018b). Second language writing online: An update. *Language Learning & Technology, 22*(1), 1–15. https://dx.doi.org/10125/44574

Godwin-Jones, R. (2019a). In a World of SMART Technology, Why Learn Another Language?. *Journal of Educational Technology & Society*, *22*(2), 4–13.

Godwin-Jones, R. (2019b). Riding the digital wilds: Learner autonomy and informal language learning. *Language Learning & Technology, 23*(1), 8–25. https://doi.org/10125/44667

Godwin-Jones, R. (2020). Building the porous classroom: An expanded model for blended language learning. *Language Learning & Technology, 24*(3), 1–18. http://hdl.handle.net/10125/44731

Graham, M. (2020). Regulate, replicate, and resist–the conjunctural geographies of platform urbanism. *Urban Geography*, *4*1(3), 453–457.

Graham, M., Ojanperä, S., & Dittus, M. (2019). Internet geographies: Data shadows and digital divisions of labor. In M. Graham, W. Dutton, & M. Castrells (Eds.), *Society and the internet: How networks of information and communication are changing our lives* (pp. 58–79). Oxford University Press.

Hao, K. (2020, August 14). A college kid's fake, AI-generated blog fooled tens of thousands. This is how he made it. *MIT Technology Review*. https://www.technologyreview.com/2020/08/14/1006780/ai-gpt-3-fake-blog-reached-top-of-hacker-news/

Hodgkinson-Williams, C. & Arinto, P. (Eds.) (2017), *Adoption and impact of OER in the global south*. https://idl-bnc-idrc.dspacedirect.org/bitstream/handle/10625/56823/IDL-56823.pdf?sequence=2&isAllowed=y

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

Johnson, A., Nye, B., Zapata-Riversa, D., & Hu, X. (2016). Enabling intelligent tutoring system tracking with the experience application programming interface (xAPI). In R. Sottilare, A. Graesser, X. Hu, & G. Goodwin (Eds.), *Design Recommendations for Intelligent Tutoring Systems* (pp. 41–45). U.S. Army Research Laboratory.

Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL, 28*(2), 166–186.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big data & society*, *1*(1), 1–12.

Lantolf, J. P. (2006). Language emergence: Implications for applied linguistics—A sociocultural perspective. *Applied linguistics*, *27*(4), 717–728.

Larsen-Freeman, D. (1997). Chaos/complexity science and second language acquisition. *Applied linguistics*, *18*(2), 141–165.

Larsen–Freeman, D. (2019). On language learner agency: A complex dynamic systems theory perspective. *Modern Language Journal*, *103*(Supplement 2019), 61–79.

Lee, H., Warschauer, M., & Lee, J. H. (2019). Advancing CALL research via data mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. *ReCALL, 31*(2), 135–149.

Lee, H., Warschauer, M., & Lee, J. H. (2020). Toward the establishment of a data-driven learning model: Role of learner factors in corpus-based second language vocabulary learning. *The Modern Language Journal, 104*(2), 345–362.

Levine, G. (2020). A human ecological language pedagogy. *Modern Language Journal, 104* (S1), 1–130.

Little, D. & Thorne, S. (2017). From learner autonomy to rewilding: A discussion. In M. Cappellini, T. Lewis, & A. Rivens Mompean (Eds.), *Learner Autonomy and Web 2.0.* (pp. 12–35). Equinox.

MacWhinney, B. (2001). Emergentist approaches to language. *Typological studies in language*, *45*, 449-470.

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, *67*(S1), 254–275.

Marcus, G., & Davis, E. (2020, August 22). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Markov, A. A. (1913). An example of statistical investigation of the text ‘‘Eugene Onegin’’concerning the connection of samples in chains, trans. into English by G. Custance and D. Link. *Science in Context*, *19*(4), 591–600.

Martín-Monje, E., Castrillo, M. D., & Mañana-Rodríguez, J. (2018). Understanding online interaction in language MOOCs through learning analytics. *Computer Assisted Language Learning*, *31*(3), 251–272.

Mayfield, E., Madaio, M., Prabhumoye, S., Gerritsen, D., McLaughlin, B., Dixon-Román, E., & Black, A. W. (2019). Equity beyond bias in language technologies for education. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, & T. Zesch (Eds.), *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 444–460). Florence: ACL. https://www.aclweb.org/anthology/W19-4400.pdf

Meunier, F. (2019). Data–driven learning: From classroom scaffolding to sustainable practices. *EL.LE, 8*(2), 423–434.

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, *32*(2), 129–148.

Mueller, C. M., & Jacobsen, N. D. (2016). A comparison of the effectiveness of EFL students' use of dictionaries and an online corpus for the enhancement of revision skills. *ReCALL, 28*(1), 3–21.

Oates, L., Goger, L. K., Hashimi, J., & Farahmand, M. (2017). An early stage impact study of localised OER in Afghanistan. In C. Hodgkinson-Williams & P. Arinto (Eds.), *Adoption and impact of OER in the Global South* (pp. 549–573). Cape Town: African Minds.

O'Keeffe, A. (2020). Data-driven learning: A call for a broader research gaze. *Language Teaching*.

OpenAI API (2020, June 11). *OpenAI Blog*. https://openai.com/blog/openai-api/

Papi, M., & Hiver, P. (2020). Language learning motivation as a complex dynamic system: A global perspective of truth, control, and value. *The Modern Language Journal*, *104*(1), 209–232.

Phuong, A., Nguyen, J. & Marie, D. (2017). Conceptualizing an adaptive and data-driven equity-oriented pedagogy. *Transformative Dialogues: Teaching & Learning Journal, 10*(2), 1–20.

Prinsloo, P., & Slade, S. (2017). Big data, higher education and learning analytics: Beyond justice, towards an ethics of care. In B. Daniel (Ed.), *Big data and learning analytics in higher education* (pp. 109–124). Springer.

Rainie, L., & Wellman, B. (2019). The internet in daily life: The turn to networked individualism. In M. Graham & W. H. Dutton (Eds.), *Society and the Internet* (pp. 27-42). Oxford University Press.

Raybourn, E. M., & Regan, D. (2011). *exploring e-portfolios as independent open learner models: Toward army learning concept 2015* (No. SAND2011-5944C). Sandia National Lab.(SNL-NM), https://www.osti.gov/servlets/purl/1120257

Reinders, H. (2018). Learning analytics for language learning and teaching. *JALT CALL Journal*, *14*(1), 77–86.

Roberts-Mahoney, H., Means, A. J., & Garrison, M. J. (2016). Netflixing human capital development: Personalized learning technology and the corporatization of K-12 education. *Journal of Education Policy*, *31*(4), 405–420.

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), 1–21.

Rosell-Aguilar, F. (2018). Autonomous language learning through a mobile application: a user evaluation of the Busuu app. *Computer Assisted Language Learning*, *31*(8), 854–881.

Rubel, L. H., & Nicol, C. (2020). The power of place: Spatializing critical mathematics education. *Mathematical Thinking and Learning*, *22*(3), 173–194.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 1–15). Berlin: Springer.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist 57*(10), 1380–1400.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, *46*(5), 30–40.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106.

Smith, B. E., Pacheco, M. B., & de Almeida, C. R. (2017). Multimodal codemeshing: Bilingual adolescents' processes composing across modes and languages. *Journal of Second Language Writing, 36*, 6–22.

Smith, M. & Seward, R. (2020). Updating open development: Open practices in inclusive development. In M. Smith & R. Seward (Eds.), *Making Open Development Inclusive* (pp. 1–22). MIT Press.

Sockett, G. (2014). *The online informal learning of English*. New York: Palgrave Macmillan.

Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in teaching and learning*. Palgrave Macmillan.

Swain, M. (2013). The inseparability of cognition and emotion in second language learning. *Language Teaching, 46*(2), 195–207.

Tsai, Y. S., Perrotta, C., & Gašević, D. (2020). Empowering learners with personalised learning approaches? Agency, equity and transparency in the context of learning analytics. *Assessment & Evaluation in Higher Education*, *45*(4), 554–567.

Vandommele, G., Van den Branden, K., Van Gorp, K., & De Maeyer, S. (2017). In-school and out-of-school multimodal writing as an L2 writing resource for beginner learners of Dutch. *Journal of Second Language Writing, 36*, 23–36.

Vincent, J. (2020, June 30). OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws. *The Verge*. https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential

Vyatkina, N. (2020a). Corpora as open educational resources for language teaching. *Foreign Language Annals, 53*(2), 359–370.

Vyatkina, N. (2020b). Corpus‐informed pedagogy in a language course: Design, implementation, and evaluation. In M. Kruk & M. Peterson (Eds.), *New technological applications for foreign and second language learning and teaching* (pp. 306–335). Hershey, PA: IGI Global.

Warschauer, M., Yim, S., Lee, H., & Zheng, B. (2019). Recent contributions of data mining to language learning research. *Annual Review of Applied Linguistics*, *39*, 93–112.

Yim, S., & Warschauer, M. (2017). Web-based collaborative writing in L2 contexts: Methodological insights from text mining. *Language Learning & Technology, 21*(1), 146–165. https://dx.doi.org/10125/44599

Youngs, B. L., Prakash, A., & Nugent, R. (2018). Statistically-driven visualizations of student interactions with a French online course video. *Computer Assisted Language Learning*, *31*(3), 206–225.

Yousuf, B., & Conlan, O. (2020, July). Assessing the impact of controllable open learner models on student engagement. In *2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT)* (pp. 47–49). *IEEE*. https://ieeexplore.ieee.org/document/9156066

Zhu, E. (2016). Interaction and cognitive engagement: An analysis of four asynchronous online discussions. *Instructional Science, 34*(6), 451–480. https://doi.org/10.1007/s11251-006-0004-0

Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley.