Big Data in SLA: advances in methodology and analysis

Theodora Alexopoulou, Detmar Meurers and Akira Murakami

Introduction

With computers being used in all areas of life, the Internet connecting computers and their users, and the Internet of Things with its sensors providing information about the world, there is a vast amount of digital data. The data is generated in many different contexts and for many different purposes. As such, it provides extensive, often readily accessible sources of information that can be analyzed and interlinked for a wide range of secondary purposes, usually referred to as Big Data. An early example of the use of such Big Data was Google's use of the search queries entered into their web search tool to determine the current flu prevalence (Butler, 2008). Search queries and many other types of Big Data are expressed using language. This makes them a potential treasure trove for people interested in language and its use in context.

While there is no consensus on a definition of what constitutes Big Data, the term typically is used in connection with the four Vs: when there is a lot of it (volume), it takes many different forms (variety), the data is frequently updated (velocity), and the data is of unknown quality (veracity). The high volume and variety of language data accessible through the Internet, from relatively static web sites to highly dynamic news and social media platforms, potentially makes this kind of big data particularly attractive for language scientists interested in representative samples of language use in authentic contexts. In addition, high velocity in this context may provide a longitudinal perspective on language development, from a fine-grained perspective on individual trajectories to overall language change in progress.

At the same time, the attractiveness of Big Data, with its high volume and high variety, comes at a substantial cost. Since Big Data is typically collected and aggregated from a range of sources, where they were produced for diverse purposes, whether a given data set can validly be used to empirically explore a given language research purpose needs to be carefully established. The veracity issue caused by the reuse of Big Data for language research thus differs fundamentally from data elicited under controlled conditions for the primary goal of language analysis. This concerns both the quality of the language data, as well as the availability and quality of meta data on the people who produced the language and the tasks and contexts for which it was produced.

The high volume, variety, and velocity of the language data also defines clear desiderata that need to be met to make use of such Big Data for language research. Large volumes of data can only be processed automatically, which generally requires scripting or full-blown programming, Natural Language Processing (NLP) methods for the identification of the relevant linguistic properties, and statistical methods or machine learning techniques for the identification of reliable patterns in the data. On the conceptual side, the variety and velocity of the data makes it crucial to identify properties of the task to be able to validly interpret the data in variationist terms. The same holds for meta-information on who produced the language and when, which is needed to support interpretation related to individual and group development. This is particularly relevant for language learning research, where task factors and the individual differences of learners are primary factors to be considered by the research.

Following a historical overview of the use of Big Data, we focus on the critical issues, present the current contributions and research methods, before concluding with suggestions for best practice in using Big Data for second language acquisition research.

Historical perspectives

Big Data is systematically being harnessed for research at different levels of granularity. In the health sciences, for example, large-scale data at molecular (e.g., gene expression), tissue (e.g., brain image), patient (e.g., severity of disease based on physiological variables), and population levels (e.g., social media data) have been exploited to address biological, clinical, and epidemic questions (Herland et al., 2014). Big Data has also advanced research in sociology (Edelmann et al., 2020) and Psychology (Harlow & Oswald, 2016). In the language sciences, corpora are frequently being used as a source of naturally occurring language (McEnery & Hardie, 2012) and in many cases corpora can be argued to constitute Big Data. Complementing carefully designed collections such as the British National Corpus, where data curation eases some of the veracity challenges posed by more typical Big Data, the use of noisier web and social media data is becoming increasingly common. Such data tends to be noisier in terms of information about the author, the consistency of writing conventions and style. Nevertheless, Grieve et al. (2019) compared 180 million geo-tagged Tweets and the BBC Voices dialect survey, and demonstrated that the two data types align well, suggesting that Twitter data can be used to study regional lexical variation in the UK. Twitter data has also been used in the investigation of language contact (Trye et al., 2020). Web-crawled data and related sources have been employed to analyze regional syntactic variation (Dunn, 2019), and Sagi (2019) used digitized books from Project Gutenberg[1] to investigate diachronic semantic change and the phonological correlates of meaning. Frey et al. (2019) exploited interactional data using 250 million chat messages from one million children to examine how the emotional valence of a chat message affects the valence of subsequent messages. Though such corpora do not focus on SLA per se, they highlight aspects of the target language learners aim for. More generally this research has contributed to methods for investigating language-related questions that can be adapted to corpus-based SLA research.

In First Language Acquisition research, the Child Language Data Exchange System[2] (CHILDES) has supported the collection of 59 million words in 41 languages by 3,056 speakers (MacWhinney, 2019). By standardizing data representation and analytical tools, CHILDES readily supports secondary use of substantial, heterogeneous data so that arguably it satisfies core characteristics of Big Data.

Data collected in a formal educational setting across primary and secondary school levels makes it possible to study academic language development in terms of accuracy and complexity (Weiss & Meurers, 2019a). Big Data from high stakes testing can also been used to inform our understanding of human grading biases, as illustrated by Weiss et al. (2019) analyzing student writing from the German school leaving examination ("Abitur").

Corpora also play an important role in Second Language Acquisition (SLA) research. While learner writing has long been collected in learner corpora (Granger, 2008)[3], the systematic design for language learning research and controlled corpus collection, often using the same prompts to elicit the data, makes such curated resources less typical for Big Data. At the same time, different from data collected in experimental settings designed to address a specific hypothesis, learner corpora are compiled and annotated to support a broader range of uses, and many of the methodological issues, such as the need for automated linguistic and statistical analysis, apply to

[1] https://gutenberg.org
[2] https://childes.talkbank.org
[3] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

large learner corpora in the same way as they do to more typical Big Data collections – issues we will turn to in the Main Research Methods section.

With an increasing number of digital tools being used in education, there is a growing body of digital data that could provide Big Data of relevance to SLA research. The potential can be illustrated with the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2014) providing access to 1.2 million writings that were submitted by 175,000 learners from around the world to the online school of EF Education First, which is used in various blended learning packages fo different contexts. The online school is a Computer-Assisted Language Learning (CALL) system that covers the full spectrum of proficiency levels so that the corpus resulting from system use over several years also includes longitudinal records for individual learners. EFCAMDAT has been used to investigate language-focused issues such as relative clause development (Alexopoulou et al., 2015), which requires the automatic identification of complex language constructions. EF teacher feedback included as meta-information in the corpus also supports the analysis of complexity and accuracy development (Murakami, 2020) or the prediction of spelling difficulties (Beinborn et al., 2016). Meta-information about the learners and task characteristics supports the investigation of L1 influence (Murakami, 2016) and task effects (Alexopoulou et al., 2017; Michel et al., 2019). Big data are particularly helpful for empirical investigations of the complex interactions of standard SLA variables, which smaller scale lab or field-based studies typically cannot support.

Large-scale language testing can also provide Big Data attractive for SLA research. For example, the Cambridge Learner Corpus (CLC; Nicholls, 2003) is composed of essays written for the exams administered by Cambridge Assessment English. Part of the corpus has been manually error-annotated, making it suitable to cross-sectionally analyze linguistic accuracy (Murakami & Alexopoulou, 2016a). While most research has targeted English, data from official standardized foreign language certification tests is also starting to be analyzed in terms of linguistic complexity for other languages, such as Estonian (Vajjala & Lõo, 2013) or German (Weiss & Meurers, 2019b). A study exploiting the State Examination of Dutch as a Second Language will be discussed in the next section (Schepens et al., 2020). While the vast majority of learner corpora include written production (Paquot and Plonsky, 2017), the NICT Japanese Learner English (JLE) Corpus[4] is a spoken learner corpus including 1,281 manually error-annotated speech samples collected through an oral proficiency interview test. The Trinity Lancaster Corpus (Gablasova et al., 2019) includes spoken interactional data of 2,000 learners and examiners in the speaking tasks administered by Trinity College London examination board. Interactional data allows researchers to study phenomena such as filled pauses (Götz, 2019) and lexical backchannels (Castello & Gesuato, 2019).

Native language corpora such as the Corpus of Contemporary American English can also be relevant for SLA research, for example to approximate the input L2 learners have received (Wolter and Yamashita, 2018) or to derive word and formulaic sequence lists for L2 research and instruction (e.g., Simpson-Vlach & Ellis, 2010). Based on movie and TV subtitles, Brysbaert and New (2009) compiled the SUBTLEX-US corpus and showed that the frequencies obtained on that basis appear more representative of typical language experience (as measured by a lexical decision task) than other frequency norms.

Complementing the corpus-based research, other Big Data research in SLA has exploited citation data to perform meta-science research. Chen (2018), for instance, carried out bibliometric analyses on 11,381 research papers in SLA, while Meara (2012) analyzed citations in relation to

[4] https://alaginrc.nict.go.jp/nict jle/index E.html

vocabulary research. Lei and Liu (2019) similarly analyzed citations in 10,028 articles in applied linguistics, while at the same time examining word sequences in abstracts. This research revealed trend patterns in SLA and related areas, such as the rise of sociocultural and functional theories over the years. The U.S. census data have also been exploited in SLA in order to investigate the critical period hypothesis (Hakuta et al., 2003). While less often used than corpus data, there is a variety of Big Data types that can in principle contribute to our understanding of L2 acquisition and SLA as a discipline.

Critical issues and topics

The critical issues and topics for the use of Big Data for SLA derive from the need to link the requirements of SLA research questions to the characteristics of the datasets and the methods available to exploit the volume and velocity of the data, while also addressing the veracity challenges.

Big Data that lends itself to SLA research often arises from digital language teaching and examinations in authentic learning and assessment contexts. Such data supports *high ecological validity* for studying how people learn languages in real-life, instructed SLA (ISLA) contexts. This can potentially provide a treasure trove of data across the proficiency spectrum, elicited by activities or tasks covering different skills (oral, writing) in different teaching and assessment contexts. Responses are provided by learners from a range of linguistic, educational, and social backgrounds around the world. Such Big Data may thereby provide a bridge for taking foundational SLA research questions previously only explored in the lab to post-hoc analysis of already collected data. The structure and volume of the data lends itself to *multi-factorial designs and analyses* to reflect the range of proficiency levels, tasks, modes and linguistic backgrounds and potentially examining their complex interactions. Such analyses are beyond small-size lab or field data sets, yet they are crucial to empirically investigate the impact and relevance of foundational phenomena observed under lab conditions under ecologically valid conditions. The data volume can help clarify the *scalability of empirical generalizations*, both for less frequently occurring instantiations of variables and for combinations and interactions of different variables. Finally, the high velocity of Big Data can support research on longitudinal individual development, which is very difficult and costly to collect using traditional means, but arguably is critical for understanding L2 development, in particular in light of the important role of individual differences. In sum, Big Data provides new opportunities for investigating language learning questions in an applied context under perspectives that can contribute to developmental SLA research and support empirically grounded investigations of aspects that are beyond the scope of smaller-size lab or field-based studies.

These opportunities co-exist with non-trivial challenges regarding the *representativeness* and *veracity* of the data. Representativeness is not (just) an issue of size but of the reach of the institution providing the data. For example,, data in EFCAMDAT is skewed towards beginner and intermediate learners as the largest groups engaged in learning, whereas the CLC compiled from high stakes tests is skewed towards advanced levels. Data will also be biased towards the particular age groups or national markets that a particular institution is targeting. While sampling (e.g., stratified sampling) and matching techniques (e.g., propensity score matching) can be used to overcome such biases, they depend on metadata or related approximations to identify the relevant data subsets. Metadata regarding relevant linguistic, educational and socioeconomic backgrounds often is unavailable in Big Data, which is only natural given that it was not originally collected with SLA research in mind. So, researchers often need to rely on proxy information such as the

nationality of learners and general assumptions about the educational levels of learners in a country. For example, researchers interested in L1 effects may need to combine information from various sources (e.g., nationality, country of residence, self-referencing information in text) to compensate for the absence of L1 metadata. In addition to the veracity challenge arising from missing or approximated metadata, there also can be substantial noise in the language data itself. The term noise here contrasts with the true signal one is trying to identify and refers to encoding problems and unknown or uncontrolled factors influencing the data. While even in highly controlled experimental research, outlier removal techniques are used to reduce noise, the lack of control is intrinsic to Big Data. Careful selection of data subsets, data preparation and cleaning therefore is crucial to identify interpretable data-subsets (e.g., eliminating texts that are too short or too long, texts written in a language other than the one targeted) and to address encoding and formatting inconsistencies. Noise in Big Data may also arise from the typical multi-site data collection process, easily resulting in repeated data or mismatches between different tasks or task versions and the submitted learner answers.

To be able to exploit the large volume offered by Big Data for addressing SLA research questions, it is necessary to *automate data processing and linguistic analysis*. As a result, our capacity to automatically analyze Big Data accurately and reliably for a wide range of language features of relevance to SLA delimits the kind of research questions we can answer. Typically, the large, structured nature of the data will also require appropriate *statistical methods* to interpret the findings, possibly combined with qualitative manual interpretation of data subsets. We will focus on these issues in the Main Research Methods section below.

Turning to the topic of *ethical use of Big Data* (Barocas & Nissenbaum, 2014), two important notions to be discussed are consent and anonymity. Obtaining consent will generally require the collaboration of the institution collecting the data, e.g., the education company or examination board, to ensure consent to use the data for research was granted. For Big Data use that is more distant from the original context it was collected in, such as the use of social media data (e.g., Twitter), the use of anonymized data has traditionally been handled less restrictively, though there is increasing concern about data use and laws such as the EU General Data Protection Regulation provide explicit guidelines. Depending on the nature of the data, anonymization may concern personally identifying information in the language as well as in the metadata. Just as for the analysis, for large data sets anonymization will often necessitate the use of automated methods for the identification and systematic data anonymization.

A final important topic connected to Big Data is its (un)availability for research. The Big Data of interest to SLA tends to come from commercial teaching and examining institutions and therefore is not originally public. Across the sciences, access to experimental, field, and digital datasets is opening up in the interests of research transparency and reproducibility of results. Open access to Big Data for the research community is a prerequisite for its use in a sustainable research program.

Current contributions and research

A basic empirical question in SLA is when learners start using a particular linguistic feature, or how their linguistic inventory expands as they advance in their proficiency. Big Data can provide a very rich empirical source to answer this question for a comprehensive list of features across proficiency levels. A case in point is the English Profile Project, which used the CLC to identify *criterial features* for each proficiency level in the Common European Framework of Reference

(CEFR; Hawkins & Filipovic, 2012). The English Vocabulary Profile[5] and the English Grammar Profile[6] are two resources aligning vocabulary and grammatical features with proficiency based on the CLC.

Documenting when learners start using a feature of the linguistic system is also the starting point for asking questions about how instructional input and tasks support or constrain its use and how this may interact with learner characteristics, such as the L1. This is particularly important for features that are not obligatory and therefore possibly infrequent, such as relative clauses (RCs). The acquisition of RCs has been widely researched in SLA due to the relevance of this structure in the study of syntactic complexity, typological variation and sentence processing (Gass and Lee, 2007). What makes Big Data particularly relevant to the study of RCs is that large amounts of data are needed to address data sparseness and support solid generalizations. As already highlighted by Schachter (1974), the choice between use and avoidance is central to the study of RCs as it is affected by the L1 of the learner – yet this influential study was based on a total of 250 compositions from four L1 backgrounds. A Big Data set such as EFCAMDAT allows significant scaling up of such empirical investigations, for a larger number of native language backgrounds, across proficiency levels, and for a variety of tasks. The collection of the data in a CALL system makes it possible to tease apart the impact of L1, proficiency, and task while controlling for the teaching input offered by the system. Alexopoulou et al. (2015) investigate the use of RCs by learners from six native language backgrounds, selecting learners who have completed at least eight lessons. They find that RCs become productive at late elementary level (CEFR A2) – interestingly, well before the explicit introduction of RCs in the course, at the end of the B1 level. While the timing of RCs is not affected by L1 or particular tasks, there are nevertheless tasks that elicit far more RCs than average. In addition, Chinese, Russian and German learners avoid using *that* RCs with animate heads (e.g., "a song writer that brought your feelings to your music") which results in underuse of *that* RCs in comparison to Brazilians, Italians and Mexicans. The study also shows that productive and formulaic use of RCs co-exist throughout the proficiency spectrum.

The richer empirical base available through Big Data is directly relevant to long-standing strands of SLA research, such as crosslinguistic influence. Big Data makes it possible to include a wider number of L1s and thereby can go well beyond contrasting language pairs to broader typological investigations. For example, Schepens et al. (2020) compare the test scores of thousands of immigrants taking the official test of Dutch as a second language. They investigate 56 L1s, including a large number of typologically diverse languages, and show that the linguistic distance (in lexical, morphological, and phonological terms) between the L1 and Dutch predicts attainment in L3 Dutch. Importantly, they are able to compare the effect of linguistic distance while also taking into account variables such as years and quality of education, age of arrival, length of residence in the Netherlands, geographical region of origin, and gender. Linguistic distance remains a significant predictor of test scores even when taking these variables into account.

Using written exams from the CLC, Murakami and Alexopoulou (2016a) show that the L1 influences the order of acquisition of English morphemes as well as the accuracy of their use until advanced levels of proficiency. Making use of the error annotation included in the corpus, they consider whether the presence of a specific L2 morpheme in the L1 has an effect on the accurate realization of different morphemes. Analyzing seven distinct L1s, they show that the L1 type correlates with accuracy, with morphemes differing in how vulnerable they are to L1 influence. The study questions the long-held assumption that the acquisition of morphemes follows a

[5] https://www.englishprofile.org/wordlists
[6] https://www.englishprofile.org/english-grammar-profile

universal order – while at the same time highlighting a hierarchy of morpheme difficulty depending on L1 background. Such findings and those of Schepens et al. provide empirically rich insights in classical SLA domains with relevance for teaching practice that would be hard to obtain without the use of Big Data.

Turning from the impact of learner characteristics to that of the context in which the learner language to be interpreted was produced, there is a wide range of potentially relevant contextual factors, from the general educational setting to the specific instructional input and task prompt. Such variables are central to Task-Based Language Teaching research aimed at understanding the effect of instructional tasks on learning. For Big Data, and learner corpora in general, the question is crucial to interpreting the learner language. The cognitive complexity of the task and task type (e.g. description, narrative, argumentation) is well known to impact on the complexity and accuracy of the elicited language (Loschky & Bley-Vroman, 1993; Yoon & Polio, 2016, and references therein). A corpus like EFCAMDAT with 128 task prompts across proficiency allows us to investigate how task complexity and type shape learner language and at the same time understand how varying tasks are distributed across the curriculum. Michel et al. (2019) and Alexopoulou et al. (2017) show that task type has a strong effect on the linguistic complexity of the elicited language and that large corpora investigations can help identify natural or essential features associated with a particular task. Descriptive tasks tend to be characterized by lower complexity in comparison to narrative and argumentative tasks. Michel et al. (2019) also point out that descriptive tasks tend to dominate in beginner and early intermediate activities, with more challenging task types such as argumentative or narrative tasks appearing more frequently in more advanced levels. Is this distribution reflecting a meaningful alignment of tasks with learner abilities or is there scope for introducing more challenging tasks earlier to enhance learning at lower levels? Big Data open the possibility for approaching such questions of task alignment and adaptation to the learner's level.

Finally, Big Data present an opportunity for conducting more longitudinal investigations of individual development. Researchers in Complexity Theory (e.g., Larsen-Freeman, 1997) and Dynamic Systems Theory (DST; e.g., Verspoor et al., 2011) have long argued for studying L2 development at the level of individual learners (e.g., de Bot et al., 2007; Lowie & Verspoor, 2019). Indeed, a discrepancy between group-level and individual-level developmental patterns has been demonstrated in the L2 development of article accuracy (Murakami & Alexopoulou, 2016b) and syntactic complexity (Bulté & Housen, 2018). While the research in (C)DST so far was mostly based on small numbers of learners, large-scale longitudinal data can greatly facilitate this line of research. This makes it possible to disentangle intra-learner and inter-learner variability and to evaluate the magnitude of individual variation (Murakami, 2016). Connecting this with the issues discussed above, Big Data can make it possible to explore individual development taking into account learner-external variables such as task prompts and learner-internal variables like L1.

Main Research Methods

The high volume and variety that are characteristic of Big Data require the analysis to be automated. In addition, the interpretation of the results is often supported using statistical methods that help us to take the noise and complex structure of the data into account. But arguably the biggest challenge for using Big Data usually lies in finding a way to let the data speak to the research questions one wants to investigate. After identifying the data source and meta-information on the task that elicited the language and the authors, we are ready to carry out analyses based directly on the collected language forms. This is meaningful for some research issues, such as strands of work investigating

the role of token frequencies for language learning (e.g., Ellis, 2002), where the surface forms collected in corpora are of interest as representative examples of language use.

Going beyond the language forms as such, most research into language refers to linguistic classes (e.g., parts-of-speech, morphological characteristics, word senses, phrase and sentence types) and relations between them (e.g., grammatical functions, semantic or discourse relations). For corpora to be able to play a role in empirically validating theories and developing them further, the notions used to describe the subset of data that the theories and research questions talk about must be mapped to the relevant corpus instances, typically by using annotated corpora. There is a battery of NLP tools for annotating the lexical units in terms of parts-of-speech and morphological properties. It is often possible to express or approximate the notions used in common syntactic generalizations in those terms (e.g., Meurers, 2005), though the NLP tools do not take into account the particular characteristics of learner language, such as the fact that distributional, morphological, and lemma evidence can systematically diverge (Díaz Negrillo et al., 2010). Even where the particular conceptual challenges of analyzing learner language (Meurers & Dickinson, 2017) are only of secondary interest, the fact that learner language differs from the edited newspaper language most NLP tools are trained on, results in analyses with lower accuracy. Accuracy generally decreases for representations later in the NLP pipeline and for representations further from the surface forms. As a result, annotations of semantic and discourse features are particularly scarce and error prone.

For the strand of SLA research interested in learner errors, the challenge of identifying them in Big Data is substantial. Since the current state of the art for automatic grammatical error correction only reaches around 70% precision and recall (Bryant et al., 2019), high quality accuracy analyses still depend on human annotation, which is only realistic when the Big Data derives from a large-scale assessment context (e.g., CLC) or digital schools integrating human error feedback (e.g., EFCAMDAT).

Big Data poses unique challenges for their statistical analysis. Firstly, the variety and veracity of Big Data means that such data often have complex structure and contain a lot of noise. Whereas controlled experiments often isolate the effects of interest by design, for Big Data identifying and interpreting the effects of interest requires sophisticated statistical models to disentangle the effects of different factors in a post-hoc manner. Hartshorne et al. (2018), for example, modeled learning trajectories and the developmental change in learning rate in order to separate the effects of the age of onset, the current age, and the lengths of exposure (Frank, 2018). Hierarchical models such as mixed-effects regression models are particularly useful for teasing apart the sources of variation at different layers (e.g., Murakami, 2016; Michel et al., 2019). Murakami (2016), for instance, employed mixed-effects regression models to partition the variance of the accuracy of grammatical morphemes into three levels (between-L1 variability, between-learner variability, and within-learner variability). This makes it possible to quantify individual variation (i.e., between-learner variability) and investigate the longitudinal development of individual learners (i.e., within-learner variability) while controlling for variability from higher levels.

Analyses based on Big Data typically start with data exploration to identify the properties of relevance to the research question. Starting with such exploratory analysis brings with it the danger that the findings derived from a given data set may not generalize. For instance, even if one finds that L2 proficiency can be modelled by lexical complexity indices and their interaction, this may be the result of sampling variability and may not hold in another data sample. Similarly, if the characteristics of the data set at hand (e.g., properties of the writing prompts, proficiency tests

guiding activity selection) gives rise to a particular pattern, this will not generalize. The danger is particularly pronounced in Big Data due to their nature of secondary use of the data, where data collection generally was not optimized to obtain a representative random sample with respect to the questions researchers want to ask and the primary context in which data arose may well introduce an undesirable bias or systematicity. Overfit models capture not only signal but also sampling noise in the data, i.e., properties that only happen to be the case in this particular data set and therefore fail to generalize. One can reduce the chance of overfitting by splitting the available data into separate training, development, and test sets and to only use the test set at the very end to validate the results obtained on the development set. To make the most use of the available data and gain insights into the variability of the results for different data sets, an alternative approach is to perform *k*-fold cross-validation (Baayen, 2008). The available data is split into *k* equally-sized data sets, a model is trained on *k–1* data subsets, and the performance of the model is tested on the held-out remaining subset. This is repeated *k* times so that each of the *k* data subsets is once used as a test set. High variability between the *k* results indicates low generalizability. For example, Eguchi and Kyle (2020) performed stepwise variable selection in building a multiple regression model to predict learner proficiency based on lexical sophistication features. Since the predictors were selected in a data-driven manner, they performed 10-fold cross-validation to examine whether the performance (R^2) of their regression model generalizes to unseen data. At the same time, drawing subsets from one Big Data set cannot protect against the idiosyncratic characteristics of that data set. To avoid overfitting and ensure generalizability of the results obtained on a given Big Data set, it is essential to validate the results on different data sets, ideally collected by different people, for different purposes, in different context (Vajjala & Meurers, 2014; Harlow & Oswald, 2016; Yarkoni & Westfall, 2017).

The surge of automated text analytical tools (e.g., Graesser et al., 2004; Lu, 2010; Chen & Meurers, 2016; Kyle et al., 2018) has made it possible to obtain a large number of variables (e.g., linguistic complexity measures) for a given corpus. This opens up new possibilities – though, as Polio and Yoon (2018) point out, one needs to investigate the reliability and consistency of the different systems, and it is essential for such tools to transparently document what exactly is being calculated and on what basis. Rich data with a large number of variables can cause challenges in statistical modeling and interpretation. Modeling an outcome with a large number of predictors more easily results in overfitting and uncertainties in parameter estimation. Interpreting so many variables also becomes a formidable task, especially when many variables are highly correlated for a given data set. This may happen, when variables capture the same or similar constructs (e.g., various measures of lexical diversity), but also when conceptually distinct variables happen to be correlated in a given data set (e.g., the use of complex noun phrases and dependent clause, which both are characteristic of academic language). A useful technique in this kind of situation is latent variable models such as factor analyses and structural equation modeling, which reduce dimensions of data by identifying (typically a few) latent factors based on correlational patterns among observed variables. The identified factors are often interpreted as constructs giving rise to the correlations. Eguchi and Kyle (2020), for example, identified ten factors underlying 78 lexical sophistication indices in L2 speech, where the factors distinguished abstract common content words (e.g., concreteness), frequent function words (e.g., frequency and familiarity of function words), strongly associated common trigrams (e.g., range and T-score of trigrams), among other characteristics.

It can also be helpful to identify a small number of prototypical patterns and cluster the data according to those patterns. In L2 research, clustering techniques have been used widely to group learners according to, for example, their motivational profiles (Csizér & Dörnyei, 2005), their

ability/aptitude profiles (Rysiewicz, 2008), and their developmental patterns (Murakami & Alexopoulou, 2016b).

The scale, structure, and noise of Big Data pose unique challenges for data retrieval and analysis. Automated linguistic annotation is essential for identifying the part of the corpus and the particular corpus instances that are relevant under a particular research perspective. Once the data are retrieved, sophisticated statistical modeling and machine learning is generally needed to let the data speak to the given research questions, and care must be taken to avoid overfitting and ensure generalizability of the findings.

Recommendations for Practice

Training in SLA today often includes a substantial background in research design intended to maximize the reliability and internal validity of the empirical findings. The large-scale, post-hoc nature of analyses carried out using Big Data arguably requires an extension of this methodological toolkit. Given a particular research question, the first step is to identify potential existing data sets, to develop an understanding of potential sources of confounding noise or systematicity, and to explore how the variables needed to address the question can be identified in or derived from the data using automatic means. Informal exploration needs to be followed up with the appropriate computational and statistical methods sketched in the previous section.

SLA research is characterized by a combination of hypothesis-driven and data-driven investigations. For Big Data research, maintaining a productive balance in that respect will require a conscious effort. Big Data favors inductive research, in particular, where data-driven techniques are applied to cluster and discover patterns. Some variables are more easily accessible in or automatically derivable from Big Data than others. To ensure researchers can ask the research questions they want to ask – rather than simply the ones that are easy to ask given the tools provided by computational linguists and statisticians – will require training in and further development of the methods that are needed to carry out this research using Big Data. This also remains true for cross-disciplinary collaborations. While collaboration supports some division of labor in Big Data research, it crucially does not invalidate the need to jointly care about connecting genuine SLA research questions with the methods needed to answer them using Big Data. Only then will the results be relevant and replicable across data sets, and the methods and research questions can grow to suit the needs of the field.

The absence of a controlled experimental design means that many hypotheses cannot be tested (e.g., hypotheses requiring production of data in very controlled conditions with tasks or context varying in fine ways). Furthermore, some features of interest may not be elicited at all by the available tasks (Tracy-Ventura & Myles, 2015). This limitation may be reduced in the future since data from teaching and assessment institutions that Big Data sets can be derived from involve a broader range of tasks than traditional learner corpora.

The increasingly sophisticated statistical techniques also play an important role in helping identify and isolate the factors of interest. To obtain a complete picture at different levels of granularity, when Big Data research is genuinely connected to SLA research questions it can and arguably should also be complemented with research using lab-experiments or fieldwork data. While it is necessary to guard against the danger of research questions shifting to what available data sets offer, it is also beneficial to identify and pursue the areas where Big Data allow us to ask new questions or open new ways to approach long-standing questions. We highlighted individual variation in longitudinal development as one such area and offer more suggestions in the next section.

Future Directions

The current Big Data corpora represent a small fraction of the resources that can become available for SLA research given the increasing levels of online learning. As digital foreign language learning keeps expanding worldwide, there is clear potential for developing data sources for a variety of foreign languages in a broad range of educational contexts. This will be crucial to better understand the impact of varying curricula and confirm how findings generalize across contexts.

The foreign language learning of children is another potentially very attractive growth area for ISLA research. In 2016 there were almost 22 million school children in upper secondary schools (ISCED level 3, age 14–18) in Europe, with 94% learning English and almost 60% of them studying two or more foreign languages. With school education becoming increasingly digital, there is substantial opportunity for Big Data to inform long standing SLA questions regarding age effects on learning.

It also becomes possible to scale up intervention research on SLA topics such as feedback to authentic school settings when engaging with the development of digital learning tools for the school context. For example, a full-year randomized controlled field study (Meurers et al., 2019) confirmed the effectiveness of the feedback and produced a wide range of data on the learning process and products (e.g., interaction logs, exercises, pre-/posttests, free writing). The FeedBook tutoring system used in this study provides scaffolding feedback to support students in 7th grade English classes in secondary schools in Germany while doing their homework. Developing such setups further will make it possible to conduct research on a wider range of topics (e.g., motivation) in ecologically valid ISLA settings. Current data sources primarily involve written answers by learners to a specific prompt. Teacher-student interaction data as well as speech data will be valuable for empirical investigations of teacher-learner interaction in digital learning platforms – though such data also brings with it more technical challenges and privacy concerns that will need to be properly addressed.

There is clear potential in developing a dynamic relation between SLA and Big Data, where research based on Big Data informs teaching and learning methods, which in turn offers a new cycle of Big Data in which the effect of further variables can be evaluated. Learner-input alignment, or more broadly speaking, the adaptive choice of input or activities at different levels of complexity depending on the individual learner ability is a domain where this may be particularly fruitful for activity and curriculum development. Initial research targeting the parametrization of activity complexity (Pandarova et al., 2019) and the automatic identification of reading material that is individually adaptive in complexity (Chen & Meurers, 2019) provide first illustrations of the viability of such an approach.

Going beyond SLA, connecting SLA research using Big Data with the rapidly growing fields of educational data mining and learning analytics may provide substantial synergy. Learning analytics currently is mostly focused on broad outcomes (e.g., overall test scores, or identifying students at risk of failing) and considers language independent variables (e.g., number of interactions). SLA research can enrich such analytics with crucial information about the content of learning. Clearly the wealth of SLA insights into language learning should enhance our ability to interpret data in learning analytics, which will also enable us to further validate those insights. In carrying out such research, close collaboration with empirical educational science will be significant for building Big Data, for analyzing and interpreting it, and for ensuring impact of the research results on real-life learning.

Acknowledgements

We are grateful to the Editors and two anonymous reviewers for their comments and suggestions. Alexopoulou gratefully acknowledges support by EF Education First. We also thank Yutaka Ishii for pointing us to some of the relevant literature.

References

Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: perspectives on relative clauses. *International Journal of Learner Corpus Research*, 1(1):96–129. https://doi.org/10.1075/ijlcr.1.1.04ale

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task Effects on Linguistic Complexity and Accuracy: a large-scale Learner Corpus Analysis Employing Natural Language Processing. *Language Learning*, 67(1):180–208. https://doi.org/10.1111/lang.12232

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. CUP.

Barocas, S. & Nissenbaum, H. (2014). Big data's end run around anonymity and consent. In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, page 44–75. CUP.

Beinborn, L., Zesch, T., & Gurevych, I. (2016). Predicting the spelling difficulty of words for language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 73–83. https://aclweb.org/anthology/W16-0508

Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA 2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 52–75. https://aclweb.org/anthology/W19-4406

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Bulté, B. & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, 28(1):147–164.

Butler, D. (2008). Web data predict flu. *Nature*, 456(7220):287–289.

Castello, E. & Gesuato, S. (2019). Holding up one's end of the conversation in spoken English: Lexical backchannels in L2 examination discourse. *International Journal of Learner Corpus Research*, 5(2):231–252.

Chen, M.-L. (2018). A data-driven critical review of second language acquisition in the past 30 years. *Publications*, 6(3):33.

Chen, X. & Meurers, D. (2016). CTAP: A Web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC),* pages 113–116. https://aclweb.org/anthology/W16-4113

Chen, X. & Meurers, D. (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer-Assisted Language Learning*. 32(4):418–447. https://doi.org/10.1080/09588221.2018.1527358

Csizér, K. & Dörnyei, Z. (2005). Language learners' motivational profiles and their motivated learning behavior. *Language Learning*, 55(4):613–659.

de Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, 10(1):7–21.

Díaz Negrillo, A., Meurers, D., Valera, S., & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154. http://purl.org/dm/papers/diaz-negrillo-et-al-09.html

Dunn, J. (2019). Global syntactic variation in seven languages: Toward a computational dialectology. *Frontiers in Artificial Intelligence*.

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, 46:61–81.

Eguchi, M. & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of Oral Proficiency Interviews. *The Modern Language Journal*, 104(2):381–400.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, 24(2):143–188.

Frank, M. (2018). With great data comes great (theoretical) opportunity. *Trends in Cognitive Sciences*, 22(8):669–671.

Frey, S., Donnay, K., Helbing, D., Sumner, R. W., & Bos, M. W. (2019). The rippling dynamics of valenced messages in naturalistic youth chat. *Behavior Research Methods*, 51:1737–1753.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, 5(2):126–158.

Gass, S. & Lee, J. (2007). Second language acquisition of relative clauses. *Studies in Second Language Acquisition*, 29(2):329–335.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Selected proceedings of the 2012 Second Language Research Forum. Building bridges between disciplines*, pages 240–254. Cascadilla Proceedings Project.

Götz, S. (2019). Filled pauses across proficiency levels, L1s and learning context variables: A multivariate exploration of the Trinity Lancaster Corpus Sample. *International Journal of Learner Corpus Research*, 5(2):159–180.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). CohMetrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Granger, S. (2008). Learner corpora. In *Corpus linguistics. An international handbook*. de Gruyter.

Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence*.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical-period hypothesis for second-language acquisition. *Psychological Science*, 14(1):31–38.

Harlow, L. L. & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, 21(4):447–457.

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, 177:263 – 277.

Hawkins, J. & Filipovic, L. (2012). *Criterial Features in L2 English.* Cambridge University Press., Cambridge.

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, 1.

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3):1030–1046.

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

Lei, L. & Liu, D. (2019). Research trends in applied linguistics from 2005 to 2016: A bibliometric analysis and its implications. *Applied Linguistics*, 40(3):540–561.

Loschky, L. & Bley-Vroman, R. (1993). Grammar and task-based methodology. In *Tasks and Language Learning: Integrating Theory and Practice*. Multilingual Matters, Philadelphia.

Lowie, W. M. & Verspoor, M. H. (2019). Individual differences and the ergodicity problem. *Language Learning*, 69(S1):184–206.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods*, 51(4):1919–1927.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics*. CUP.

Meara, P. (2012). The bibliometrics of vocabulary acquisition: An exploratory study. *RELC Journal*, 43(1):7–22.

Meurers, D. (2005). On the use of electronic corpora for theoretical linguistics. case studies from the syntax of German. *Lingua*, 115(11):1619–1639.

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B., & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, 39:161–188. https://doi.org/10.1017/S0267190519000126

Meurers, D. & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(2). https://doi.org/10.1111/lang.12233

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of

A1 to C2 writings. *Instructed Second Language Acquisition*, 3(2):124–152. https://doi.org/10.1558/isla.38248

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, 66(4):834–871.

Murakami, A. (2020). On the sample size required to identify the longitudinal L2 development of complexity and accuracy indices. In *Usage-based dynamics in second language development*, pages 20–49. Multilingual Matters.

Murakami, A. & Alexopoulou, T. (2016a). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, 38(3):365–401.

Murakami, A. & Alexopoulou, T. (2016b). Longitudinal L2 development of the English article in individual learners. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, pages 1050–1055.

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003*, pages 572–581.

Pandarova, I., Schmidt, T., Hartig, J., Boubekki, A., Jones, R. D., & Brefeld, U. (2019). Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*.

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.

Polio, C. & Yoon, H.-J. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*, 28(1):165–188.

Rysiewicz, J. (2008). Cognitive profiles of (un)successful FL learners: A cluster analytical study. *Modern Language Journal*, 92(1):87–99.

Sagi, E. (2019). Taming big data: Applying the experimental method to naturalistic data sets. *Behavior Research Methods*, 51:1619–1635.

Schachter, J. (1974). An error in error analysis. *Language Learning*, 24(2):205–214.

Schepens, J., van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, 194:104056.

Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4):487–512.

Tracy-Ventura, N. & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1(1):58–95.

Trye, D., Calude, A. A., Bravo-Marquez, F., & Keegan, T. T. (2020). Hybrid hashtags: #YouKnowYoureAKiwiWhen your Tweet contains Māori and English. *Frontiers in Artificial Intelligence*.

Vajjala, S. & Lõo, K. (2013). Role of morpho-syntactic features in Estonian proficiency classification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building*

Educational Applications (BEA). https://aclweb.org/anthology/W13-1708.pdf

Vajjala, S. & Meurers, D. (2014). Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics*, 165(2):142–222. https://doi.org/10.1075/itl.165.2.04vaj

Verspoor, M., de Bot, K., & Lowie, W., editors (2011). *A dynamic approach to second language development: Methods and techniques*. John Benjamins.

Weiss, Z. & Meurers, D. (2019a). Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. https://aclweb.org/anthology/W19-4440

Weiss, Z. & Meurers, D. (2019b). Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research*. Presses Universitaires de Louvain. http://purl.org/dm/papers/Weiss.Meurers-19LCR.pdf

Weiss, Z., Riemenschneider, A., Schröter, P., & Meurers, D. (2019). Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. https://aclweb.org/anthology/W19-4404

Wolter, B. & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance? *Studies in Second Language Acquisition, 40*(2): 395-416

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Yoon, H.-J. & Polio, C. (2016). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, 51: 275–301.