The EuroCALL Review
Volume 30, No. 1, 2023, 52-62
ISSN: 1695-2618
https://doi.org/10.4995/eurocall.2023.14950

European Association for Computer-Assisted Language Learning

THE EUROCALL REVIEW

RESEARCH PAPER

Giving Students the Tools: Looking at Teaching and Learning using Corpora

Aidan Carter*; Matt Absalom**
The University of Melbourne, Australia

*aidan.carter@unimelb.edu.au | **mabsalom@unimelb.edu.au

How to cite this article:

Carter, A. and Absalom, M. (2020). Giving Students the Tools: Looking at Teaching and Learning using Corpora. *The EuroCALL Review, 30*(1), 52-62. https://doi.org/10.4995/eurocall.2023.14950

Abstract

This article discusses a pilot project aimed at giving tertiary students a wider repertoire of resources to use in language learning, with a particular focus on Italian. This project responds to the exponential increase in and access to online data and the potential value such data represent for students studying additional languages at tertiary level. By examining whether current language students are aware of online resources, such as linguistic corpora and other potential applications of big data, we aim to provide an insight into the possible uses of corpus-assisted learning in the language classroom. In this paper, we detail a project undertaken in 2017 with undergraduate students of Italian in a major metropolitan university. Our project directed students to complete a translation task using corpora-based resources and assessed their experience through a post-assessment survey. Subsequently, we present our initial findings in relation to the possibilities of a corpus-based approach to language teaching and learning. While today's students are already predisposed to relying on online resources as part of their language studies, our results suggest students are not aware of emerging online resources such as corpora. Moreover, even when these resources are presented to students, the complex nature of the software programs used to interrogate corpora often results in their underutilisation.

Keywords

Big data, computer assisted language learning (CALL), teaching and language corpora (TALC), language learning, L2 pedagogy.

1. Introduction

Technology has fundamentally changed the way languages are taught worldwide (Jarvis & Krashen, 2014; Godwin-Jones, 2017b). One of the most significant advancements in technology and language learning relates to the sheer volume of data created due to continuous online activity. This is particularly true for naturally occurring language which is readily available to teachers and students of additional languages (L2). This wealth of data has expanded exponentially due to increased reliance on the internet across multiple platforms, including social media, online newspapers, online blogs, websites, and similar online platforms. Consequently, L2 teachers and students can easily access texts and online databases, which provide them with naturally occurring language data that can bridge the gap between the classroom and the real world (Godwin-Jones, 2017a; Liu & Curran, 2005; MacWhinney, 2019). Many Second Language Acquisition (SLA) researchers consider naturally occurring language as key to supporting learning both within the classroom and beyond the classroom (Potter, 2013). Using this type of language, teachers are optimally positioned to enhance their teaching (Braun, 2006). Whilst these data represent an indispensable resource for L2 students and teachers, their applications, particularly in the classroom, remain relatively untested. Moreover, as every computer keystroke and swipe on smartphones and tablets is recorded and stored in "big data" banks, it is becoming increasingly difficult to determine which online data are reliable and how to process them (Godwin-Jones, 2017a). Nevertheless, the sheer volume of information available to the world cannot be overlooked. If used correctly, these ever-increasing sets of instantaneously available online data represent a goldmine for L2 students and teachers.

Panichi (2015) emphasises that L2 students are already predisposed to online resources because of their easy access and currency, particularly those pertaining to social media networking sites and internet search engines. This is further supported by Abrams and Schiesti (2017) who note the limitations of printed L2 resources and their inability to reflect the varieties present in a language, as well as emerging trends that may be present in the language. Access to natural language through online data can provide L2 teachers and students with a remedy to this. Nevertheless, it is imperative to check if students are using online resources to the best of their ability and if big data has a future role in the L2 classroom.

2. Big data and language learning

Definitions relating to big data (the immense volumes of data that have emerged in the last decade) have evolved rapidly, often creating confusion regarding what exactly is indicated by "big" and how this is measured (Gandomi & Haider, 2015). When seeking to explain the challenges relating to the management of these data sets, initial definitions, such as those provided by Laney (2001), focused on three dimensions i.e., volume, variety and velocity (the three Vs) as a framework for describing big data and the challenges they posed to existing data-management systems (Gandomi & Haider, 2015). This definition has evolved significantly alongside the ever-increasing frequency of big data mentions in academic literature. Nowadays, the definition has expanded to include veracity, variability (and complexity) and value (Gandomi & Haider, 2015). Each dimension accounts for a specific characteristic of big data. For example, volume refers to the magnitude of the datasets; velocity to the rate at which data are generated; and variety to the structural heterogeneity of the datasets. Newer dimensions consider the unreliability presented by data (veracity).

One of the key tools now available to teachers and students to engage with these data are online corpora banks which Li (2017) defines as "of great value to language teaching." The term corpus bank refers to a collection of written and/or spoken texts by (native) speakers that provide "both a rich source of attested language and an authentic learning

context" (Li, 2017). Corpora represent a steadily growing resource that provides instant, online access to natural language and various processing tools with which L2 students can engage to guide their learning. Similarly, Boulton and Cobb (2017) suggest that the applications of corpora in applied linguistics and L2 instruction has led to a "corpus revolution" across both fields in which researchers are introducing new studies to assess the validity of corpora in L2 classrooms (Godwin Jones, 2017a). However, as these datasets continue to expand in volume, emerging problems relating to them include concerns that not all data are authentic and trustworthy (veracity). Privacy issues are one of the biggest concerns when utilising big data, both in education and in other contexts as the data volume rapidly increases (Feldstein, 2013; National Academy of Education, 2016).

Language teachers have consistently placed a strong emphasis on the use of authentic resources in both spoken and written texts (Godwin-Jones, 2017a). Consequently, it is vital to guide students in sourcing and using online authentic language materials. Our contention is not that students are unable to access the benefits of the broad repertoire of online sources resulting from big data. In fact, a simple Google search under the right conditions may lead students to authentic materials. However, this is not always the case, and it can prove more difficult than imagined. Moreover, while applications for processing big data such as corpora are continually developing and becoming easier to use, at present they tend to require a level of computer literacy which is often beyond the competence of the casual user (Li, 2017).

Most research into big data focuses on learning analytics and the ability to monitor students' progress via online learning platforms (Godwin-Jones, 2017b; Kei Daniel, 2017; Peters 2017). Instead, we are interested in how we can equip students with the ability to access and interrogate linguistic corpora and encourage them to draw upon these resources to further their own learning. This is particularly important for data-driven learning (DDL), defined by Boulton (2010, p. 535) as "encouraging noticing and consciousness-raising" in L2 instruction, ultimately with the goal of "leading to greater autonomy and better language skills in the long term" (Han & Shin, 2017, pp. 173–4). When combining DDL and corpus linguistics, Boulton (2017) proffers an inductive approach in which students are encouraged to query large collections of online texts by looking at frequencies, distributions and multiple occurrences of target items in context. Consequently, students are encouraged to reach conclusions that are meaningful to them individually. Moreover, the cognitive processes required of them purportedly lead to longer retention of information studied as opposed to when information is "simply taught via traditional language instruction" (Boulton, 2017, p. 2; Han & Shin, 2017). Therefore, the combination of DDL and corpora in the language classroom offers students more independence and greater inductive learning and problem-solving skills. It also helps students to become better language learners and users. However, one of the major concerns regarding DDL and corpora lies in its perceived unsuitability for all students (Flowerdew 2015; Han & Shin, 2017). Often, advanced computer skills that go beyond the competence of the average user are required (Li, 2017). This presents further obstacles to be overcome when engaging students at lower levels. For example, Liu and Jiang (2009) only targeted intermediate and advanced L2 learners as corpus-based learning was deemed too difficult for lower-level students. Nevertheless, Boulton (2017) asserts that the skills obtained by learning to deal with a variety of texts, noticing forms and variation and inferring meaning i.e., self-directed processes of interpreting data, far outweigh the time-consuming nature of overcoming these obstacles. Notably, the rapid increase in online data in the past decade reveals an ever-increasing usage of online tools, indicating that there is likely a higher benchmark of computer literacy worldwide. In acknowledgement of this, our research design targeted students at an intermediate level (of whom the majority were in their first year of tertiary study) and investigated the ways in which they could engage corpora resources with minimal guidance from the language instructor.

Existing research relating to corpora and Italian is limited. In their investigation of corpora as a writing apprenticeship in an intermediate Italian university program, Kennedy and

Miceli (2010) highlight the importance of promoting both skills to L2 students. This includes the specific functions of corpora, as well as the principles that underpin reference-resource consultation and their implications. This research reinforces previous findings that the introduction of corpora in L2 classrooms is not a straightforward process, but one that will require extended guidance to support L2 students in their use of corpora. Additionally, corpora-based translation activities have been introduced to Italian students in English as a Second Language classrooms as a way of evaluating and revising translations by engaging with diverse corpus resources (Zanettin, 2009). In this, Zanettin (2009) underscores the beneficial applications of teaching students to engage corpora resources in authentic communicative tasks. Furthermore, Zanettin posits the usefulness of corpora-based resources even if students possess limited formal linguistic knowledge, or if language teachers are not native speakers. Drawing upon both Kennedy and Miceli (2010) and Zanettin (2009), it is evident that one of the continuing hurdles to be overcome in teaching and learning using corpora relates to the complexities associated with corpus tools and methodologies. Whilst the increased access to naturally occurring language allows L2 students direct access to examples of language used by native speakers (which Zanettin posits may further support non-native L2 teachers), the success of corpora resources is dependent on the time devoted to teaching students how to engage with and interpret corpus data (Kennedy & Miceli, 2010; Zanettin, 2009). Finally, it is worth noting that engaging with corpora also represents a significant barrier for L2 teachers themselves (Zanettin, 2009).

Whilst this article focuses on the importance of guiding L2 students to access natural language resources via corpora, previous studies have indicated that high demands on hardware and computing skills (alongside the affordability, reliability and user-friendly nature of resources), has led to a reluctance of L2 teachers to incorporate them into their teaching (Leńko-Szymańska, 2017; Römer, 2015; Tribble 2015). In what follows, we detail a first attempt to introduce corpus-assisted language learning via a small translation project. This project was included as part of the teaching and learning program of an undergraduate Italian subject at a major metropolitan university. In doing so, we present our initial findings and offer a simple model of how L2 teachers may guide students towards corpora resources and increase data-driven learning.

3 The task

3.1 Using corpora for translation

The participants for this research project were 41 undergraduate students at The University of Melbourne, completing the subject entitled 'Italian 6' in 2017. Typically, students in this subject have a level of Italian comparable to CEFR B1 and are in their first year of tertiary study. We developed a small assessment task focused on translation which was distributed online via the Blackboard Learning Management System (LMS) and made available for completion over a seven-day period. This assessment task (detailed below) was worth 5% of the students' overall grade for the subject. It assessed both the ways in which students went about the translation task and the validity of the translations provided. Following the assessment task, all students were invited to complete a short, voluntary questionnaire relating to the assignment (Appendix 2).

The student participants were randomly allocated to two groups: Group A (control) and Group B (experimental). This was facilitated by the random block function in Blackboard Learn. Both groups were provided with the same translation task. Importantly, participants were provided with one of two sets of instructions for going about the task based on their group membership. The control group received no extra guidance or resources, whereas the experimental group was provided with additional information highlighting extra resources (including corpora) for completing the task (Appendix 1).

The assessment task required the students to find and verify translations for five words or phrases from English to Italian solely using online resources. These were: "to photoshop," "last minute," "to scan" [as in to scan a document], "to google" and "to text." Participants were tasked with providing three legitimate translations for each word or phrase in Italian. To assess the resources they engaged with and the methods by which

the translations were obtained, participants were required to provide links (or references) to examples of those translations from an authentic source and to comment on the steps they followed to locate them (See Appendices 1 and 2 available online here). This design provided insight into the students' thought-process and allowed us to contrast the differences in approach employed across the two groups.

As noted, the experimental group had supplementary information to assist them in locating extra resources. This had the secondary purpose of promoting consideration of existing corpora resources and offering students a way to go beyond their normal translation techniques. Both Godwin Jones (2017) and Li (2017) attest to the benefits of corpus-assisted learning at a tertiary level, however, both note the difficulties associated with locating and utilising corpora online. Thus, these extra instructions were provided to guide participants in overcoming the difficulties associated with locating authentic corpora resources such as the Italian corpus Paisà and the multilingual Sketch Engine (multilingual corpora with an Italian database). These instructions provided students with details on accessing the databases (links, processes for signing up) as well as defining briefly what the term "corpora" referred to. Whilst corpora were a primary focus of the extra instructions provided to the experimental group, the resources also included various online monolingual dictionaries such as Treccani or Il Sansoni, as well as other big data analysis programs such as Google Ngram (an online search that charts frequencies of words sources from printed materials, including an Italian database between 1500 and 2008), and Twitter's advanced search function to promote strategies outside their normal repertoire. Whilst it is possible to conduct advanced searches across multiple social media platforms, Twitter's advanced search option was identified as one of the easiest programs to operate and it does not require an account. Participants were also incentivised to utilise the resources provided as instructions stated extra marks would be awarded for engaging with these resources.

3.2 Designing the assessment task: selecting the translations

To ensure that students needed to use online resources rather than traditional print materials, each word or phrase selected for this translation task was studied beforehand and selected due to its contemporary usage over a relatively recent time period. Several of the terms chosen, such as "to text," "to google" or "to photoshop" emerged in the 2000s and early 2010s and their prevalence in print in this period is confirmed via Google's Ngram. The recent increase in usage of these terms limited their presence in print sources and required students to engage critically with several online sources to determine whether their translations were accurate and reliable. Whilst phrases such as "last minute" or "to scan" were often easier to find (also appearing in sites such as Word Reference), others such as "to photoshop" were more difficult to locate. Possible translations were limited, thereby requiring students to interrogate new resources. This requirement was reinforced by the fact that students were required to locate three translations for each word or phrase.

3.3 Assessment task feedback

Following the assessment task, a short questionnaire (Appendix 2) was administered to determine how participants went about the task. Whilst assessment submissions indicated which tools and resources the participants used, the survey was crucial in determining their approaches to translation. This provided invaluable insight into the thought process behind assessing the appropriateness of online resources and whether this assignment had changed the way students might approach translation tasks in the future. Following the collation of both the assessment tasks and the post-assessment questionnaire, content analysis (including counts) was then completed in order to assess the possible applications of big data in the L2 classroom.

4. Results

4.1 Assessment task

The random block allocation by Blackboard Learn placed 23 of the 41 participants in the control group. The remaining 18 were allocated to the experimental group.

Table 1

Participant demographics and group membership.

	Male	Female	Total Participants
Control Group	8	15	23
Experimental	5	13	18
Total *n*			41

Of the 18 students allocated to the experimental group, 16 either stated that they had used at least one of the resources listed in the supplementary material provided, or the use of these resources was evident in the work submitted. Of the 16 participants that engaged with at least one resource suggested to them via the document, seven accessed and utilised either the Sketch Engine or Paisà corpora listed. Only two students allocated to the experimental group chose not to use any of the resources provided in the document. One student misinterpreted the instructions and believed that using these resources would *not* result in bonus points being awarded to their individual grade, whilst the second participant admitted to opening the document and immediately closing it as the resources seemed "too difficult." Even though extra resources were provided, 13 of the 18 students (72.2%) in the experimental group stated that they had used Google searches to complete the translation task, often leading them to resources such as Reverso Context. This is an online search engine for translations in context (Reverso, 2021) based on a corpora approach using concordancing of search terms (online language blogs/language forums, newspaper articles, and YouTube videos amongst other resources). Using Google to search for websites and other online resources for information to aid the translation exercises was also a popular approach for participants within the control group. 18 of 23 students utilised Google as a resource to complete the assignments. No students from the control group either demonstrated or claimed to have used corpora to complete the assignment.

4.2 Post-assessment questionnaire (Appendix 2)

Of the 28 participants (68.3%) who responded to the post-assessment questionnaire, 14 belonged to the control group and 14 to the experimental group. Overall, 20 out of 28 female participants responded (71.4%) whilst 8 of 13 (61.5%) of males responded to the survey.

When participants were asked to explain their process and the resources they typically used when approaching translation, an overwhelming majority of the participants, 24 of 28 (85.7%), cited using Word Reference as their main source of translation. When approaching tasks similar to the one presented to them, hard copy dictionaries and Google Translate were the second and third most frequent sources, with 9 and 6 counts respectively. Other resources cited included Reverso Context, iTranslate, and Linguee. No students claimed to have previously used corpora to complete translations.

Table 2 details the impact of the translation task on the students' future approaches to similar tasks.

Table 2

Q11: Has this task changed the way you might approach translation in the future?

	Control	Experimental
Discovered new sources; improved their translation skills; had altered their approach to the task in the future	21.4% (3 of 14 respondents)	64.2% (9 of 14 respondents)
Sample responses (reported as provided by respondents)	**Student A:** Not really **Student B:** I didn't find any translations that weren't legitimate using my normal way after checking with backup sources, so not really.	**Student C:** I definitely will be using Facebook and corpus italiano in the future as resources, because I found them really useful in making sure my translations were authentic. **Student F:** Yes, sites that not only translate and define words but also put them into context are very helpful. **Student E:** Yes, I hadn't considered using social media platforms to understand how words are used

When asked whether this assignment had changed the way they would approach translation tasks in the future, 9 out of the 14 participants (64.2%) allocated to "Experimental" who responded to the questionnaire stated that they had "discovered new sources" to aid translation tasks in the future or believed they had "improved their translation skills." Conversely, only 3 students (21.4%) in the control group indicated that they had discovered new resources or techniques that would change the way they approach translation tasks in the future. These data were further emphasised by the tendency of the control group to rely on the same resources (such as Word Reference or Reverso Context) provided in response to Question 1, whereas resources utilised by participants in the experimental group indicated greater variance due to the extra resources provided.

Whilst seven students from the experimental group used the corpora banks (listed in Appendix 1) when completing the assessment task, three of them did not respond to the online questionnaire. Nevertheless, the four that did respond all noted that the Paisà and Sketch Engine corpora were "difficult to use," "time consuming" and "confusing." Notwithstanding these observations, three of the four students stated that they would engage with these resources in the future noting the "more reliable information" provided and their ability to locate some of the more difficult translations.

Experimental group members utilised several other resources listed in the extra resources document provided. Four participants used the advanced Twitter search function to interrogate the validity of their translations on a social media platform. When completing

the survey, only two of four participants who used the advanced Twitter search referred to it in the online questionnaire, stating that it provided a useful insight into the Italian language in a social context and the way in which the language is evolving.

Both the assignment and post-assessment questionnaire highlight a significant difference between the two groups regarding the resources used, individual attitudes, approaches to completing the translation task and potential future approaches.

5. Discussion

Consistent with findings by Backus (2008), Godwin-Jones (2017a) and Li (2017), the often-confusing nature of corpora and the difficulties in accessing such resources played a significant role in discouraging participants from using them in this assessment task. Of the 18 participants allocated to the experimental group, 16 consulted the supplementary material provided but only seven accessed corpora while completing the assessment task. Participants who opened the document but did not make use of corpora declared that these resources were “too confusing” or “difficult to use.” The complexities involved in using online corpora were also evident in the decisions of nine participants who refrained from using corpora-based resources when more straightforward resources were available. Despite the limited number of participants accessing corpora as part of their assignment and the general concerns relating to their accessibility, three of the four participants who used Paisà or Sketch Engine as a resource (and had responded to the post-assessment questionnaire) recognised the future applications of corpora as authentic resources for Italian.

When noting the difficulties surrounding the workings of corpora, time was a significant factor in determining whether or not a resource is used, a factor exacerbated by the pressure of an assessment task with a limited timeframe. Students were more likely to engage with resources which seemed more user-friendly, even if the translation may not have been exact, rather than attempting to learn new software.

The primary objective of the translation task was to prompt students to discover new online resources. Participant responses from both the questionnaire and the individual assignments indicate that some words were easier to translate and discoverable via traditional methods (online dictionaries and Google searches) and that this approach did in fact furnish legitimate translations. The most difficult phrases to translate were “to google” and “to photoshop”, these required participants to forego traditional resources and seek alternatives in order to complete the assignment. The requirement to provide *three* translations which the participants deemed to be authentic often required engaging with multiple online resources. Further research into the integration of corpora in the L2 classroom may approach tasks in an analogous manner, encouraging students to analyse naturally occurring data as opposed to the didactised texts often found in L2 textbooks. This should further highlight the currency and adaptability of corpora resources.

Answers to Question 11 (Has this changed the way you might approach translation in the future?) reveal that the students’ default position is to rely on the same resources every time. When given as an option, however, students will attempt to use alternatives. 9 of 14 participants (Experimental) who completed the questionnaire expressed that they had discovered new resources that they would incorporate in future translation tasks. This contrast underpins our contention that L2 students are able to engage with alternate resources when prompted, but that they are unlikely to discover them on their own.

Moreover, this research suggests that current L2 students are engaging with few of the extensive online resources currently available to them, and that they consistently favour the same resources without attempting to explore alternatives, even when prompted. We also note the limited knowledge of and engagement with online resources amongst current L2 students. In response to Question 15 (Were you aware of resources such as corpora beforehand?) from the post-assessment questionnaire, all participants responded “No.” With regard to corpora and the L2 classroom, it is evident that highlighting the presence of resources is not sufficient, and that harnessing these resources requires direct

(teacher) guidance if L2 students are to recognise the potential of corpora and utilise these and other emerging online resources to their full potential.

Finally, in recognition of the advantages of DDL, one of the predominant hurdles to be overcome continues to be finding an appropriate balance between DDL and effectively training students to use corpora. With multiple participants providing inaccurate translations from "unreliable" resources such as Google Translate, particularly when the plethora of resources containing naturally occurring data continues to grow, emphasis must be placed on teaching students to think critically when accessing online L2 resources. Ultimately, the problems of validity and reliability in relation to big data and corpora remain a significant obstacle in corpora assisted language learning (Wang, 2016; Godwin-Jones, 2017a).

6. Conclusion

In this paper we have detailed a preliminary project which exploits advances in big data through corpus-assisted language learning in the L2 classroom. We conducted a pilot study in an undergraduate Italian language classroom at a major metropolitan university. Despite the limited sample size and the limitations of this small project, along with a single language and geographical focus, the results of our work indicate that L2 pedagogy could usefully take advantage of the rapid increase in the volume of online data and its value as a modern language resource for students engaged in the SLA process. As students are, for the most part, unaware of the extensive, authentic resources available to them online, it is essential that teacher-led assistance be provided to help L2 students make use of them.

Whilst this represents a first step towards the possibilities of incorporating big data and corpora into the language classroom, as always, more research must be conducted in the area with a focus on promoting critical thinking regarding online resources and guiding students though the process of accessing and utilising emerging resources.

Ethical statement

Student participants in this pilot study undertook the task and evaluation as a routine part of the teaching program in the semester in question. In accordance with the University of Melbourne's policy on research ethics at the time, such projects did not require separate ethical approval. Students' responses were anonymous. There are no conflicts of interest to declare.

References

Abrams, Z., & Schiesti, S.B. (2017). Using Authentic Materials to Teach Varieties of German: Reflections on a Pedagogical Experiment. *Unterrichtspraxis/Teaching German, 50*(2), 136–150. https://doi.org/10.1111/tger.12038

Backus, A. (2008). Data Banks and Corpora. In Wei, L. and Moyer, M.G. (Eds.), *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism* (pp. 232–248). Wiley-Blackwell.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, *60*(3), 534–572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

Boulton, A. (2017). Data-driven learning and language pedagogy. In Thorne, S. & May, S. (Eds.), *Language, Education and Technology: Encyclopedia of Language and Education* (pp. 1–12). Springer. https://doi.org/10.1007/978-3-319-02328-1_15-1

Braun, S. (2006). ELISA—A pedagogically enriched corpus for language learning purposes. In Braun, S., Kohn, K. & Mukherjee, J. (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 25–47). Peter Lang.

Feldstein, M. (2013). Why Big Data (Mostly) Can't Help Improve Teaching. *E-Literate*, http://mfeldstein.com/why-big-data-mostly-cant-help-improve-teaching/

Flowerdew, L. (2015). Data-driven learning and language learning theories. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15–36). John Benjamins. https://doi.org/10.1075/scl.69.02flo

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods and analytics. *International Journal of Information Management, 35*(2), 137–44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Godwin-Jones, R. (2017a). Data-informed language learning. *Language Learning & Technology, 21*(3), 9–27. https://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. https://dx.doi.org/10125/44592

Han, S., & Shin, J. (2017). Teaching Google search techniques in an L2 academic writing context. *Language Learning and Technology, 21*(3), 172–196. https://www.lltjournal.org/item/3015

Jarvis, H. & Krashen, S. (2014). Is CALL Obsolete? Language Acquisition and Language Learning Revisited in a Digital Age. *TESL-EJ, 17*(4), 1–6. http://www.tesl-ej.org/pdf/ej68/a1.pdf

Kei Daniel, B. (Ed.). (2017). *Big Data and Learning Analytics in Higher Education: Current Theory and Practice.* Springer International.

Kennedy, C. & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology, 14*(1), 28–44. https://www.lltjournal.org/item/2709

Leńko-Szymańska, A. (2017). Training Teachers in data-driven learning: Tackling the challenge. *Language Learning and Technology, 21*(3), 217–241. https://www.lltjournal.org/item/3017

Li, S. (2017). Using Corpora to Develop Learner's Collocational Competence. *Language Learning & Technology, 21*(3), 153–171. https://www.lltjournal.org/item/3018

Liu, D. & Jiang, P. (2009). Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *Modern Language Journal, 93*(1), 61–78. https://doi.org/10.1111/j.1540-4781.2009.00828.x

Liu, V. & Curran, J.R. (2005). Web Text Corpus for Natural Language Processing. *Association for Computational Linguistics*, 11th Conference of the European Chapter of the Association for Computational Linguistics, (April 2005), 233–240. https://www.aclweb.org/anthology/E06-1030.pdf

MacWhinney, B. (2019). Understanding Spoken Language through TalkBank. *Behaviour Research Methods, 51*, 1919–2197. https://doi.org/10.3758/s13428-018-1174-9

National Academy of Education. (2016). Big Data in Education: Balancing the Benefits of Educational Research and Student Privacy. http://naeducation.org/wp-content/uploads/2017/05/NAEd_BD_Booklet_FINAL_051717_3.pdf

Panichi, L. (2015). The employment of Social Networking for Language Learning and Teaching: Insights and Issues. In R. Hernández & P. Rankin (Eds.), *Higher education and*

second language learning: promoting self-directed learning in new technologies and education contexts (pp. 159–180). Peter Lang. https://doi.org/10.3726/978-3-0353-0685-9

Peters, J. (2016). *How Big Data Can Improve Student Performance and Learning Approaches.* Dataconomy http://dataconomy.com/2016/10/big-data-can-improve-student-performance-learning-approaches/

Potter, J. (2002). Two kinds of natural. *Discourse Studies, 4*, 539–542. https://doi.org/10.1177/14614456020040040901

Römer, U. (2015). Corpus research and practice; What help do teachers need and what can we offer? In K. Aijmer (Ed.) *Corpora and language teaching* (pp. 83–98). John Benjamins. https://doi.org/10.1075/scl.33.09rom

Tribble, C. (2015). Teaching and language corpora: Perspective from a personal journey. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 37–62). John Benjamins. https://doi.org/10.1075/scl.69.03tri

Wang, Y. (2016). Big Opportunities and Big Concerns of Big Data in Education. *TechTrends: Linking Research & Practice to Improve Learning, 60*(4), 381–384. https://doi.org/10.1007/s11528-016-0072-1

Zanettin, F. (2009). Corpora-based Translation Activities for Language Learners. *The Interpreter and Translator Trainer, 3*(2), 209–224. https://doi.org/10.1080/1750399X.2009.10798789

Online References

Appendices 1 & 2
https://osf.io/ncyuh/?\view_only=ec2873db967d44299b6bc8ce3038dc67

Dizionario Internazionale https://dizionario.internazionale.it

Google Books (Ngram) https://books.google.com/ngrams

Paisà (Corpus Italiano) https://www.corpusitaliano.it/en/

Reverso http://context.reverso.net/translation/

Sketch Engine https://www.sketchengine.eu/#blue

Treccani http://www.treccani.it

Twitter https://twitter.com/