Data-Driven Learning: Tools, Approaches, and Next Steps

Rachel Allan, Terry Walker (Mid-Sweden University),

and Virginia Langum (Umeå University)

1. Introduction

This special issue of the *Nordic Journal of English Studies* is the combination of two events. It is the outcome of a symposium held at Mid-Sweden University on 23 October 2020 organised by Terry Walker and Rachel Allan. Moreover, it is intended to reflect the research interests of the new editors of the *Nordic Journal of English Studies*, who took over the role at the beginning of 2021, Virginia Langum and Terry Walker, and their goal to highlight research areas that bring together the disciplines of literary studies and linguistics. In the first special issue (2022) edited by Virginia and Terry, the focus was on the medical humanities; this second special issue, together with guest editor Rachel Allan, concentrates on a number of key areas within data-driven learning (DDL).

We first give a background on the symposium (2.1), and the aims of this special issue on DDL (2.2), before describing the state of the art in the field of DDL as regards language and literary education and research (section 3). We include a brief overview of the articles as regards tools and approaches (section 4). Section 5 focuses on the results of the articles and how these contribute to the next steps.

2. Background

2.1 The 'Incorporating Corpora in Teaching' symposium

This symposium hosted by the English subject at Mid-Sweden University was organised by Terry and Rachel in 2020, originally as an on-site event, to highlight the research profiles of subject didactics and corpus linguistics, with invited experts from Sweden and internationally. The advent of Covid-19 meant that the planned event was reorganized into an online format via Zoom. The anticipated networking opportunities were somewhat lessened by this format, but a positive outcome was that the four

Allan, Rachel, Terry Walker, and Virginia Langum. 2023. 'Data-driven Learning: Tools, Approaches, and Next Steps.' *Nordic Journal of English Studies* 22(1): 1–12.

55-minute sessions (each with three presentations) were recorded—with permission from those involved—and are available upon request. Zoom also allowed for much greater participation, with c. 150 students, teachers and researchers attending from all over the world.

In Session 1, Alex Boulton, Fanny Meunier, and Pascual Pérez-Paredes, covered 'Empirical studies in data-driven learning', 'Revamping Data-Driven Learning: Making actual use of the affordances of digital technology' and 'Developing a critical agenda for learning-driven DDL' respectively. Fanny has contributed to this volume (see section 4). In Session 2, Christer Geisler and Christine Johansson demonstrated 'Uses of a learner corpus for student teachers', while the contributions of both Anne O'Keeffe and Erik Smitterberg are included in this special issue. Session 3 was a demonstration of materials session: Vaclav Brezina presented on '#LancsBox in language teaching', Ana-Frankenberg Garcia discussed 'Data-driven learning from a text editor', and Tatyana Karpenko-Seccombe presented her book 'Academic writing with corpora: A resource book for data-driven learning'. Tatyana has also contributed to this issue. In the final session Randi Reppen spoke on 'Using corpora to inform instruction: three practical approaches', Rachel Allan's contribution can be found in this special issue, and Dana Gablasova focused on 'Teaching about spoken English with the British National Corpus 2014: introducing BNClab'. Please contact the editors if you would like access to any or all of these recorded sessions and the accompanying abstracts.

The symposium was the initial springboard for the current special issue, which was further inspired by the editors' desire for a special issue on an area that cuts across disciplines within English Studies, and considers a variety of digital tools.

2.2. Aims of this Special Issue

The present issue is intended to promote and evaluate the use of corpora and other digital tools in the classroom, with a focus on current methodology and research and how this can be exploited effectively in teaching, including a range of perspectives, encompassing both historical and contemporary corpora, learner corpora, as well as a range of other digital tools. The aim is to highlight methods for incorporating the use of these in the classroom, whether for research purposes, as a tool for teaching, and/or to encourage future teachers in the use of such tools, and

to evaluate the use of data-driven learning (DDL) in the classroom. A further aim is to offer new insights and promote new methodologies related to corpora and other digital tools in teaching.

3. DDL: The state of the art

3.1 Data-driven learning: Language teaching and learning

The term DDL is generally associated with Tim Johns, who began working with an inductive approach to language exploration using concordances drawn from corpora in the 1980s (Johns 1986, 1988, among others). A distinguishing feature of Johns' DDL approach was that learners would assume control of the learning process. The language learner became 'a research worker whose learning needs to be driven by access to linguistic data—hence the term "data-driven learning"' (Johns 1991: 2). The approach was an 'attempt to cut out the middleman as far as possible' (Johns 1994: 297), the learner interacting directly with the corpus to 'identify—classify—generalize' (Johns 1991: 4); *identify* key words pertinent to their query and carry out associated searches, *classify* the results, organizing concordance lines by sorting to the left or right to discover patterns of use, and *generalize*, derive their own hypotheses based on the results. DDL was thus initially envisaged as a hands-on, computer-based activity, putting the learner at the helm, with the teacher providing direction as needed.

In the intervening years, what we have come to understand as DDL has broadened considerably. As Boulton and Vyatkina (2021: 66) point out, although Johns' approach may be the prototype, its boundaries have become much fuzzier. Corpus consultation remains at the core, with perhaps the most commonly adopted definition now being 'using the tools and techniques of corpus linguistics for pedagogical purposes' (Gilquin and Granger 2010: 359). This, however, has been interpreted in many and diverse ways. In the early days, one way of making DDL more accessible to a wider range of learners was through the use of paper-based tasks and teacher-curated concordances. However, more widespread familiarity with computers along with rapid advances in technology have led to alternative approaches to navigating corpora for learning. Mediating DDL through the data used has become commonplace now that graded, learner and specialist corpora are readily available. Developments in software have resulted in a new generation of more intuitive, freely available data-processing tools such as LancsBox (Brezina, Weill-Tessier and McEnery

2021), with features such as automatic annotation of the corpus, visualization of data, and easy comparison of corpora. Other advances have led to the mediation of the corpus so that the resulting data bears little to no resemblance to the traditional concordance, such as through video clips (e.g., PlayPhrase.me, Youglish), graphs (e.g., Google n-gram viewer) or word clouds, automatically generated on many different websites.

Despite these advances, and although DDL has generally received an enthusiastic response from the research community, it is uncommon to find it used in language learning in mainstream, non-tertiary educational contexts (Chambers 2019; Allan, this issue). Boulton and Vyatkina's (2021) meta-analysis of published corpus studies into DDL (1989–2019) demonstrates this, finding that studies with university level students accounted for 85% of the studies, with just 9% with younger learners in schools or other pre-university courses in the 477 studies included that indicated institutional context. From its inception, the research-orientated nature of DDL has appealed to teachers and researchers at university level, and it seems that this is still the case. Some encouragement can be found in the fact that a higher proportion of studies involving younger learners took place in the latter period of the meta-analysis, suggesting an upturn in the level of interest in the approach at school level. The more appealing interfaces with DDL discussed above may contribute to this in the future (see Crosthwaite's 2019 volume for further examples and discussion), but until mainstream language teacher education incorporates training in DDL both at a conceptual and practical level, it seems unlikely that the approach will be more widely adopted (Crosthwaite 2019; Allan, this issue).

Over three decades have passed since Johns' initial writings on DDL, and in that time, many empirical studies into DDL have been carried out. Meta-analyses of these studies, of which there have been several in recent years (e.g., Boulton and Cobb 2017; Lee, Warschauer and Lee 2019; Boulton and Vyatkina 2021), have found corpus use to have a positive effect on language learning. What remains unclear is *why* it is effective. These studies observe that most theoretical justification for DDL has rested on the broad concepts of constructivism and/or socio-cultural theory, with some mention of Second Language Acquisition (SLA) theories such as the Noticing Hypothesis (Schmidt 1990), the Usage-Based (UB) model (Ellis 2002) and the Involvement Load Hypothesis (Laufer and Hulstjin 2001). While each of these can be said to offer a valid rationale for the approach, without a better understanding of the interface

between theory and practice it is difficult to know *how* to use DDL effectively. When should teachers use which approach, in what context, for what purpose, and with which learner profiles—in other words, how does DDL work within language pedagogy? Such questions have been asked from the outset, but they are becoming more prominent as DDL matures (e.g., Pérez-Paredes 2019; O'Keeffe 2021). Boulton and Vyatkina (2022) demonstrate the apparent reluctance to commit to a theory in their corpus-based overview of the theoretical underpinnings of DDL, and find that theories are rarely intentionally and empirically tested. However, the landscape is changing, and there is broader acknowledgement of the UB model to inform the use of DDL, aligning it with acquisition processes. O'Keeffe and Mark (2022), for example, draw on SLA theory, notably UB research, to propose a framework of key principles for DDL design, consisting of the acquisition principle, the complexity principle and the formulaicity principle. They use this to make a case for refining DDL, differentiating tasks and data by level. In her contribution to this issue, O'Keeffe presents a case study exploring the UB model of acquisition as a theoretical basis for DDL (see sections 4 and 5). Theory-building of this kind is essential to move DDL forward into a more mainstream role.

3.2 Data-driven learning and literary studies

The application of computational methods and digital tools has been a fraught issue in literary studies. Some have worried that digital tools and methods take readers away from the core of literary studies—e.g., close reading—and further towards the clutches of a neoliberal agenda. Teaching literature becomes a means of learning competences for the market rather than critical thinking, according to this line of thought (Allington et al. 2016). One aspect of digital literary studies that has been particularly divisive is 'distant reading'. Coined by Franco Moretti, 'distant reading' refers to the application of computational methods to literary corpora. Provocatively, Moretti invests distance reading with the means of 'a more rational literary history' (Moretti 2005: 4), as well as the capacity of studying more works excluded by the literary canon (Moretti 2013). Rather than reading texts laboriously one by one, Moretti advocates a wider-scale counting, graphing and mapping (Moretti 2005).

In his recent book, *The Digital Humanities and Literary Studies*, Martin Paul Eve plods a middle ground for digital literary studies, suggesting that 'digital methods […] can give us a route to viewing a text

anew, seeing with fresh eyes what was always there to know, just never before calculable' (Eve 2022: 154). Eve gives several examples of how digital tools can challenge 'quantitative and empirical assertions in literary studies', for example, about periodization and genre (Eve 2022: 142). Rather than replacing traditional literary methods, such as argumentation and close reading, digital tools should be used in tandem with such methods. The data can only be processed and understood through critics and their analysis: '[i]t is, though, in this synthesis that we best see the merits of both approaches—of distance and depth' (Eve 2022: 142).

How do these new modes of reading and working with texts function in the literature classroom? What are the implications of the critiques raised about digital literary studies? Is literature merely content, a vehicle for preparing students to enter the market? While there is not nearly so much scholarship on digital tools in teaching literature as linguistics, several case studies have been published in recent years. For example, one case study of a digital humanities course uses the Sherlock Holmes stories as a corpus to practice numerous digital humanities tools, such as visualization, edition making with Text Encoding Initiative (TEI), topic modelling, Graphic Information System (GIS) mapping, and distant reading (Swafford 2016). However, rather than merely using the literature as content or data as a means to learn these digital tools, the students gain a greater understanding of the Holmes stories. For example, the teacher aims to 'help students historicize their own technological moment and better understand both the Victorian period and the discourses around modern technology' (Swafford 2016). Such a learning outcome would not be out of place with 'traditional' literary studies. Another teacher uses the TEI to have students make digital editions of literary texts. In the process, students must read and reread the text, note formal properties, choose a critical lens, and other processes normally part of close reading. The teacher finds that 'the technology drove them deeper into the words, developing even richer views of the text' (Gailey 2014: 195). The teacher concludes that teaching such digital literary skills both offers 'professional currency to students', as well as textual engagement and criticism (Gailey 2014: 198).

What these case studies in digital literary studies have in common is that they emphasize both the advantage of the tools themselves (for further applications) and that the tools deepen the study of literature. So, too, does the study in this special issue: Ida Margrethe Rask Krogh and Ruben Moi's

'Literature and Data-Driven Learning'. They present a pilot study employing digital storytelling in the classroom. Students read Mark Haddon's *The Curious Incident of the Dog in the Night-Time* (2003) and create their own multimodal responses using iMovie or Microsoft Bilder. In so doing, the intervention not only 'provides pupils with a strong foundation of twenty-first century skills' but also trains traditional literary skills, such as interpretation and critical thinking in an imaginative way.

4. Contributions to the special issue: Tools and approaches

The first contribution is 'In-depth Data-driven Learning: At Least Eight Reasons to Rejoice!' by **Fanny Meunier**, a renowned expert in the field of DDL. Fanny generously agreed to write a commentary on the contents of this special issue, and concludes that the aims of this special issue (see 2.2) have been fulfilled. She neatly sums up the key areas addressed by the authors, and some key findings; therefore, here the editors focus on the tools and approaches adopted in the articles. **Anne O'Keeffe** opens the discussion by considering how UB theory can be applied to DDL and help our understanding of how to use it effectively, as mentioned in 3.1. She examines one language pattern that UB-based research identified as used by beginner-level learners of German (Römer 2019), and tests this using the Cambridge Learner Corpus across learners of English from over 150 different first language backgrounds. This is followed by **Daniel Ihrmark**'s study, which draws on socio-cultural theory to support a teacher-mediated approach to implementing DDL. His study investigates the routines of a small sample of English language teachers at upper secondary level in Sweden in providing feedback to their students. He uses this to inform the development and fit of a tool (a language analytics suite) intended to alleviate the workload experienced by teachers assessing student texts. With the long-term aim to find ways to normalise corpus use in the classroom, **Rachel Allan** exploits the web-based interface Voyant Tools, which allows users to upload their own corpora, in this case a corpus of learner essays and a graded corpus of texts matching the pupils' level of English: the tool displays a lot of visual information from the corpus selected as well as the usual concordance lines. The practical approach adopted was to demonstrate to student teachers how this tool could be used in the upper secondary school classroom, using both written and audio-visual material, including how to use the tool and the corpus files, and two hands-on tasks. A key element to the study was the student

teachers' responses to the question 'to what extent is it realistic to expect teachers to develop corpus skills and use corpora with learners?'. Similarly, **Erik Smitterberg** adopts a learning-by-doing approach to encouraging the use of corpora for learning about the subject—the history of English (1500–1945)—and for learning about doing research, supported of course by coursebooks and instruction. In the course described, students gain an understanding of the development of the language and of corpus linguistics methodology through practical corpus-based tasks, culminating in a small-scale corpus study on an aspect of language (grammar, lexis, etc.). The tools he uses to this end are 'very large corpora' (corpora of over 50 million words), such as the Corpus of Historical American English (COHA). **Mats Deutschmann** and **Anders Steinvall** consider 'the development of pedagogical DDL tools for raising awareness of matters related to language bias and stereotyping': the tools in question being the open-guise technique (participants are informed of the purpose of the exercise) and the undisclosed, matched-guise, technique (participants are informed after the exercise). These techniques have hitherto been used to measure attitudes, but have not been exploited as *pedagogical* tools, which is the focus here. The approach taken is to test and critique these two different techniques by applying each to a different group of students: the students respond to voice-manipulated recordings of the same dialogue, and later evaluate the exercise. **Tatyana Karpenko-Seccombe** responds to the lack of pre-prepared DDL materials for teachers by presenting a series of practical tasks that can be integrated into lessons. She demonstrates how corpus materials can be used to introduce learners (post-graduate students) to rhetorical features typical of academic writing and to develop students' awareness of the role rhetorical features play in a discourse. **Ida Margrethe Rask Krogh** and **Ruben Moi**'s approach using digital storytelling is described in 3.2.

5. Next steps

This special issue set out to explore the use of corpora and related digital tools to promote learning, focusing on classroom settings. As described above, the articles included examine this from a variety of perspectives using a range of tools and approaches, and in doing so they meet our aims in offering original and innovative insights. We here highlight the results of the contributions to this special issue and how these inform the next steps in the development and application of DDL. Three distinct themes

can be identified; the importance of scaffolding, the extension of DDL to new domains, and the need for a greater understanding of the relationship between DDL and SLA.

First, in keeping with the theme of mediating DDL, discussed in 3.1, the concept of scaffolding, i.e., providing support that is gradually removed as the learner becomes more independent, is referenced in several of the papers. **Daniel Ihrmark** bases his DDL tool on the concept of contingent scaffolding, operationalized within a framework of corrective feedback. He argues that by using a data-driven approach via the tool in the initial phase of feedback, the teacher can quickly and easily present curated data to the students, and gradually withdraw scaffolding in subsequent text revisions. Scaffolding is also referenced in terms of task-design. **Rachel Allan** found that initial scaffolding through using structured, modelled tasks with small corpora was important for the successful introduction of DDL to teacher trainees online. In another context, preparing research students to use large-scale historical corpora, **Erik Smitterberg** outlines how smaller-scale tasks are used to support students new to corpus linguistics, building the skills needed for a larger project. Finally, **Tatyana Karpenko-Seccombe**'s contribution gives a step-by-step account of how corpus tools and tasks can be used to highlight the use of rhetorical features in written discourse. The tasks can be used as they stand, but could also be tailored to examine discourse features relevant to other student profiles, providing a springboard for teachers looking to experiment with DDL. A common theme in the research literature for DDL is the need for mediation, and accessibility to tasks and models like these is important for the way ahead.

Secondly, this collection of articles shows how DDL is being extended to diverse contexts and in novel ways, departing from its traditional association with lexico-grammatical features in written discourse. **Tatyana Karpenko-Seccombe** goes beyond sentence level and explores its use in relation to higher level discourse features, while **Ida Margrethe Rask Krogh** and **Ruben Moi** apply it to digital story-telling, and demonstrate how it promotes the learning of a wide range of skills. As regards next steps, their work is being refined and pursued further by Krogh. In the school year 2021/2022, a new group of pupils were introduced to the project and made their own digital stories: the experience from the tentative *The Curious Incident*-LDST pilot led to improvements in the criteria for assessment and changes to the presented tasks. At this

point in time, the project is being introduced once more to a new English class. Further improvements have been made regarding which programmes to use when making a digital story. This year the pupils are using PowerPoint and its 3D animations to introduce new ways of creativity and critical thinking. The pupils can add animation effects to 3D graphics which help them practice critical and reflected behaviour using digital skills (Krogh pers. comm.). Another departure is **Daniel Ihrmark**'s use of DDL to inform the construction of a digital tool to assist in feedback on writing. Furthermore, **Mats Deutschmann** and **Anders Steinvall** take an innovative approach to DDL, using it to inform a technique to raise awareness of sociolinguistic features such as gender stereotypes. This is a potentially promising direction for DDL; as noted in 3.1, audio-visual applications of DDL are becoming increasingly sophisticated, and it seems likely that use of the approach with spoken language is one area that will offer many new opportunities for research and practice.

Finally, along with the examples of DDL in practice, this collection of articles includes a much-needed contribution to theory-building. **Anne O'Keeffe** demonstrates the potential of the UB model of acquisition as a theoretical basis for DDL. Her results found the language pattern investigated to be important not just at beginner level but across all levels, with learners building on the initial basic use to incorporate it in increasingly complex patterns. She points out the value of this for DDL design, as it shows that working with language patterns is important from an early stage of learning, but that a more structured approach may be needed. Use of the UB model, she argues, can help us understand how DDL can be useable with language learners at all stages.

It is hoped that theory will inform the next phase of empirical studies and practical applications, helping us to understand DDL's potential as a pedagogical tool, and how to realise it. It is also hoped that further innovative applications and approaches like those described in this collection of articles, along with the scaffolding required to implement them, will emerge as technology develops. DDL has come a long way in the last three decades, although there is still plenty of road ahead and, as Meunier points out in this issue, there are reasons to rejoice in this.

References

Allington, Daniel, Sarah Brouillette, and David Golumbia. 2016. Neoliberal tools (and archives): A political history of digital

humanities. *Los Angles Review of Books*. 1 May. Accessed 24 February 2023. https://lareviewofbooks.org/article/neoliberal-tools-archives-political-history-digital-humanities/.

Boulton, Alex, and Tom Cobb. 2017. Corpus use in language learning: A meta-analysis. *Language Learning* 67(2): 348–393.

Boulton, Alex, and Nina Vyatkina. 2021. Thirty years of Data-Driven Learning: Taking stock and charting new directions over time. *Language Learning & Technology* 25(3): 66–89. doi:/10125/73450.

Boulton, Alex, and Nina Vyatkina. 2022. Atheoretical or agnostic? A corpus-based overview of theoretical underpinnings of data-driven learning. Paper presented at Teaching and Learning Corpora (TaLC), Limerick, Ireland, July 13–16.

Brezina, Vaclav, Pierre Weill-Tessier, and Anthony McEnery. 2021. #LancsBox v. 6.x. [software package].

Chambers, Angela. 2019. Towards the corpus revolution: Bridging the research-practice gap. *Language Teaching* 52: 460–475. doi:10.1017/S0261444819000089.

COHA = The Corpus of Historical American English: 400 Million Words, 1810–2009. 2010–. Compiled by Mark Davies.

Crosthwaite, Peter (ed.). 2019. *Data-Driven Learning for the next generation: Corpora and DDL for pre-tertiary learners*. London: Routledge.

Ellis, Nick 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: 143–188.

Eve, Martin Paul. 2022. *The digital humanities and literary studies*. Oxford: Oxford University Press.

Gailey, Amanda. 2014. Teaching attentive reading and motivated writing through digital editing. *The CEA Critic* 76(2): 191–199.

Gilquin, Gaëtanelle, and Sylviane Granger. 2010. How can Data-Driven Learning be used in language teaching? In *The Routledge handbook of corpus linguistics*, edited by Anne O'Keeffe and Michael McCarthy, 359–370. London: Routledge.

Johns, Tim. 1986. Micro-Concord: a language learner's research tool. *System* 14(2): 151–162.

Johns, Tim. 1988. Whence and whither classroom concordancing? In *Computer applications in language learning*, edited by Theo

Bongaerts, Pieter de Haan, Sylvia Lobbe, and Herman Wekker, 9–27. Dordrecht: Foris.

Johns, Tim. 1991. Should you be persuaded: two examples of data-driven learning. In *Classroom Concordancing*, edited by Tim Johns and Philip King, 1–13. Birmingham: ELR.

Johns, Tim. 1994. From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In *Perspectives on pedagogical grammar*, edited by Terence Odlin, 293–313. Cambridge: Cambridge University Press.

Laufer, Batia, and Jan Hulstijn. 2001. Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics* 22: 1–26.

Lee, Hansol, Mark Warschauer, and Jang Ho Lee. 2019. The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics* 40(5): 721–753.

Moretti, Franco. 2005. *Graphs, maps, trees: Abstract models of a literary history*. London: Verso.

Moretti, Franco. 2013. *Distant reading.* London: Verso.

O'Keeffe, Anne. 2021. Data-Driven Learning: A call for a broader research gaze. *Language Teaching* 54(2): 259–272.

O'Keeffe, Anne, and Geraldine Mark. 2022. Principled pattern curation to guide data-driven learning design. *Applied Corpus Linguistics* 2(3). doi:10.1016/j.acorp.2022.100028.

Pérez-Paredes, Pascual. 2019. A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning* 35(1/2): 36–61.

Römer, Ute. 2019. A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics* 24(3): 268–290.

Schmidt, Richard. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11: 129–158.

Swafford, Joanna. 2016. Teaching literature through technology: Sherlock Holmes and digital humanities. *The Journal of Interactive Technology Pedagogy* 9. Accessed 24 February 2023. https://jitp.commons.gc.cuny.edu/teaching-literature-through-technology-sherlock-holmes-and-digital-humanities/.