

Language awareness and medium-term benefits of corpus consultation

Alex Boulton

CRAPEL-ATILF/CNRS, Nancy Université

Abstract

Data-driven learning (DDL) involves learners exploring corpora to discover language rather than 'being taught'. This is claimed to have a number of advantages, and while empirical evidence to date is encouraging, it is less conclusive than might be hoped – partly, it is argued, as it tends to focus on short-term goals by testing knowledge and retention of selected language items covered. It may be, however, that the real advantages of DDL lie in longer-term benefits, cognitive/constructivist as well as purely linguistic; in addition to 'incidental' learning and greater learner autonomy, these include language awareness and noticing ability, all of which are more difficult to assess.

This paper reports on a medium-term experiment comparing noticing skills between experimental and control groups of lower-intermediate learners of English in an architecture school in France. The control groups were taught in the usual way, while the last 15 minutes of the experimental classes were given over to exploring the British National Corpus on line, with learners working in pairs on specific language points. Both experimental and control groups were tested after 12 weeks – not on the language points covered, but on their sensitivity to other language items in a previously unseen text.

The time-frame of the present experiment is not enough to allow the full effects of DDL to be revealed, but does provide some indication that its main advantages go well beyond the short-term learning outcomes generally examined in current research. If further long-term work can confirm these findings, it might go some way towards helping DDL to reach a wider audience.

1. Introduction

For over two decades, researchers (e.g. Johns 1991) have been promoting the use of electronic corpora in language learning. In its most common form, learners use concordance lines such as in Figure 1 to detect the common, probable patterns of usage in context, concentrating on collocation, colligation, frequency, distribution, and so on. A number of studies have attempted to evaluate some aspect of this 'data-driven learning' (DDL) approach: in a survey of 20 such research projects, Boulton (forthcoming) finds that the results are generally encouraging, but rarely overwhelmingly positive. It is notable that the research is often designed around short-term returns on predefined language items; in other words, many of them

attempt to show that practical and logistical barriers (cf. Boulton 2009a) can be overcome under the right circumstances – especially given enthusiastic teachers with sufficient resources. However, it might be that the main advantages of DDL lie more in its long-term effects (Chan & Liou 2005: 241; Yoon 2008): developing a range of essential cognitive processes, increasing language sensitivity and ability to cope with authentic language, enhancing motivation due to greater learner-centeredness, and adopting a more ‘naturalistic’ approach to patterns (as opposed to rules) coupled with an interactive, discovery approach fosters better language learning ability on the whole, leading to greater autonomisation for life-long learning. If aspects such as these do represent the main strengths of DDL, then studies that do not address them directly clearly have their limitations.

with an ounce of sense knows that results **depend** on factors other than staff efficiency. --, T. Baines, Oxford. AS a
 ael Fishes, Suzanne Charltons and Alex Hills of TV **depend** upon for their forecasts. For the predictions that have m
 for him. "But we can't **depend** upon Alan getting goals like that every week. "The lads are looking forward
 enough. But that may **depend** on who he is fighting and, anyway, Razor Ruddock isn't aggressive,
 fairly small. Discs make sounds which **depend** on how much space they have. " When you use the mouse to drag
 is defined as attractive. It would also **depend** on incentives and on some serious stimulus from the conservation lobb
 families and individuals this means that they **depend** for their electricity on what they can generate for themselves, gro
 ic modern" prints, bring sums which **depend** both on the actual rarity of the image and its attractiveness. A beautiful
 ctacularly decorative seventeenth-century bronzes easy to move. AMERICAN buyers **depend** on advisers, whether
 re things that we take for granted, **depend** on individuals, their skills, and a surprising amount of physical effort. Fac
 ith a series of coalition governments that have had to **depend** on several minor parties whose influence has been out
 amah oil-for-weapons programme on which 30,000 British jobs **depend**. The deal, released yesterday, will boost T

Figure 1. Concordance of *depend*. <http://corpus.byu.edu/bnc/>

This paper focuses on one potential long-term advantage, namely in promoting noticing skills, a feature frequently mentioned in the DDL literature¹ (e.g. Allan 2009; Aston 2001; Bernardini 1997; Bondi 2001; Gabrielatos 2005; Johns 1997; Johns et al. 2008; Mauranen 2004; Meunier & Gouverneur 2009; O'Sullivan 2007; Pérez-Paredes & Cantos-Gomez 2004; Thompson 2002; Yoon 2008). Noticing is also present in many DDL materials (e.g. McKay 1980; Johns 1991; Bowker & Pearson 2002; Thurstun & Candlin 1997; Tribble & Jones 1997), whereas Meunier and Gouverneur (2009: 195) find it is often absent in more traditional materials. Indeed, Carter (1998: 51) claims that communicative teaching on the whole “has not encouraged in students habits of observation, noticing, or conscious exploration.” While traditional deductive approaches may allow the teacher to ‘do the noticing for the learner’, inductive approaches are entirely dependent upon

¹ In this context, it refers mainly to spontaneous noticing by the learners, rather than teacher-directed focus on form.

noticing (see Schaffer [1989] for a comparison), and DDL is largely inductive in nature. Of course, DDL has no monopoly on noticing: on the whole, it is implicit at some level in virtually all approaches, especially where the focus is on learning (featuring problem-solving, discovery learning, task-based or process-oriented approaches, for example) rather than teaching. Cook (2001) even claims that it is an advantage of invented sentences that they do promote noticing. Noticing is closely related to a number of other features, such as focus on form, consciousness, language awareness, sensitisation, and so on. It is particularly associated with Schmidt (1990), who claimed that it preceded understanding; but unlike understanding, “noticing is the necessary and sufficient condition for converting input to intake” (p. 129). While we may learn, know and use language effectively without conscious understanding, noticing is essential – or at least, “more noticing leads to more learning” (Schmidt 1994: 18). Also Schmidt (2001); Gass (1997); though see Robinson (1997) for a contrary view.

2. Method

This study involved the participation of 59 students (average age just over 20; 48% women) in their second year of architecture studies in France – 34 in experimental groups (EG), 25 in control groups (CG). Their main motivations are for architecture and not languages: despite eight years of English, their levels of proficiency are fairly low (averaging about 450 points on an in-house TOEIC, i.e. about A2–B1 on the CEFR) – hence the relevance of a new approach, a claim made by Boulton (2010). The primary objective is to score at least 650 points on the TOEIC by the end of the following year. To this end, the common language programme focuses on listening skills (in the computer room) as well as reading (mainly grammar and vocabulary). Conditions for both were essentially the same: a total of 30 hours of English in 90-minute weekly classes on the same days and times, with the same materials, facilities and evaluations, but different teachers.

For the EG only, the last 10-20 minutes of class time was devoted to corpus-based activities in the computer room; due to various constraints, this was on an irregular basis for only 12 of the classes, a total time of about 3 hours. Students worked in pairs of their own choice, using Davies’ (2004) interface to the British National Corpus. In an attempt to integrate the DDL activities and make them relevant to the learners’ immediate concerns as recommended by Somogyi (1996), all DDL language points covered arose during class, either in the current week, or as revision / reinforcement from the previous week, or preparation for the following week. As time was at a premium and the same core syllabus had to be completed as for the other classes, it was not possible to offer extensive training. Rather, after a brief 5-minute introduction the first time, detailed

instructions for each DDL activity were printed out; examples are given in Figure 2. This of course limits the students' autonomy, but means they can start corpus consultation on relevant points immediately, and by the end of the course should be sufficiently familiar with the procedures to undertake further corpus work on their own should they so wish.

HUNDRED(S) (OF)

- a) Is it possible to have a number with a plural (e.g. *hundreds*)? Try the NEWS section for *hundred* and *hundreds*. Is either followed by *of*? What do you conclude?
- b) Compare with *thousand*, *million*, *billion*; then *ten* and *dozen* – any differences? How would you translate each in French?

PASSIVES

- a) Search only in NEWS for *is said*; how would you translate this in French? What structure(s) typically follow it?
- b) What would you call *is said*: present continuous, present perfect, present passive, present simple?
- c) Try for the following verbs (*be* + past participle); do they all exhibit the same usage? Which are the most frequent? *believe*, *expect*, *know*, *report*, *suppose*, *think*.

Figure 2. Sample activities.

At the end of the course, the EG learners completed a questionnaire, the results of which are reported in Boulton (2009b) where the focus was on learning styles. Additionally, learners in both CG and EG were given a short test to assess their spontaneous noticing skills. As there is no standard procedure for assessing noticing (Williams 2005: 681), the activities were designed around a familiar support, a short (345-word) newspaper article with a business focus on a topic previously covered by all students.² The participants were told they would have 5 minutes to read the text, which would subsequently be retrieved before they answered questions on it. The first focus-on-form (FOF) question type simply required them to remember which of two words had been used in the text, each pair having similar form (e.g. *achievable* / *available*) or similar meaning (e.g. *March* / *January*). The second focus-on-meaning (FOM) question type asked them to choose the best translation of a word from the text (e.g. SQUEEZE: *satisfaire*, *contraindre*, *commencer*, *accroître*). Crucial for the purpose of noticing here is the fact that none of the language points selected had been overtly covered during the course, and the learners did not know what type of questions they would be asked (cf. Hulstijn 1997).

² "Mortgage lending increases but remains historically low". Sandra Haurant. 14/04/09.
<http://www.guardian.co.uk>

The results of the experiment are given in Table 1, with the mean scores, standard deviation and results of a two-tailed p -test for the experimental and control groups for each question type. The EG performed better than the CG on both questions types, especially for FOM. However, the differences in neither case were statistically significant. Combining both sets of data increases the statistical difference, but this is still not significant at the usual levels on a two-tailed paired t -test ($p > 0.05$).

		mean		SD		p
		EG	CG	EG	CG	
FOF	(/20)	14.91	14.04	2.25	2.57	0.17
FOM	(/10)	5.59	4.80	1.89	2.14	0.14
TOTAL		20.50	18.84	3.37	3.78	0.08

Table 1. Experiment results.

3. Conclusions

This paper reports on a simple experiment designed to test whether the processes involved in basic corpus consultation lead to medium-term benefits in noticing skills. The experimental group does indeed seem to improve more than the control group, but the difference does not meet the standard levels of significance. The constraints of the course limited the corpus work possible, in terms of overall time spent (about 3 hours), the length of each session (10-20 minutes including time lost for various reasons) and the irregular rhythm (roughly once every two weeks on average). Although Flowerdew (2008) claims short sessions may be preferable, the learners in the study by Gan et al. (1996) would have preferred full two-hour classes. The time constraints also ruled out significant training, and it may be that “noticing things in corpus data is an acquired skill even for linguistically relatively sophisticated learners” (Mauranen 2004: 99). The lack of training also inhibited more autonomous work (though see Boulton 2009c), and the activities conducted thus had to be tightly controlled; this further limits student input, with (one might suppose) a concomitant reduction in motivation during the process and reduced impact on noticing skills. Although these limitations are likely to be typical of many language classrooms where limited time can be devoted to alternative approaches, especially in a computer laboratory, the modest improvement does suggest that greater benefit would be recorded from longer and more in-depth corpus work. In short, if a little DDL leads to a modest improvement, then more DDL might lead to more substantial differences.

The overall evidence to date shows that DDL can bring immediate benefits, but these are relatively small, and there is a case to be made that its real advantages lie in longer-term outcomes. The present study suggests that for DDL to fulfil its potential in promoting substantial long-term benefits, occasional short sessions are not enough. While it seems reasonable to suppose that more time spent in corpus work would lead to more significant improvement, DDL is not an entire method, and in most contexts will vie for class time with other materials and techniques. The question then turns from whether it is effective, to whether it is an efficient use of learners' time over long periods. An alternative but equally important step forward will be to see how DDL can be used outside the classroom, a possibility largely untested to date (Chambers 2007: 13).

References

- Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal*, 63, 23–32.
- Aston, G. (2001). Learning with corpora: an overview. In G. Aston (Ed.), *Learning with corpora* (pp. 7–45). Houston: Athelstan.
- Bernardini, S. (1997). A 'trainee' translator's perspective on corpora. In G. Aston, L. Gavioli, & F. Zanettin (Eds.), *Proceedings of corpus use and learning to translate*. <http://www.sslmit.unibo.it/cultpaps/paps.htm>
- Bondi, M. (2001). Small corpora and language variation: reflexivity across genres. In M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small corpus studies and ELT: theory and practice* (pp. 135–174). Amsterdam: Benjamins.
- Boulton, (2009a). Data-driven learning: reasonable fears and rational reassurance. *CALL in Second Language Acquisition: New Approaches for Teaching and Testing. Indian Journal of Applied Linguistics*, 35/1, 81–106.
- Boulton, A. (2009b). Corpora for all? Learning styles and data-driven learning. *5th Corpus Linguistics Conference*. University of Liverpool, 20-23 July.
- Boulton, A. (2009c). Testing the limits of data-driven learning: language proficiency and training. *ReCALL*, 21/1, 37–51.
- Boulton, A. (2010). Data-driven learning: taking the computer out of the equation. *Language Learning*, 60/3.
- Boulton, A. (forthcoming). Learning outcomes from corpus consultation. In F. Serrano, M. Calzada, & M. Moreno Jaén (Eds.), *Exploring new paths in language pedagogy: lexis and corpus-based language teaching*. London: Equinox.
- Bowker, L. & Pearson, J. (2002). Working with specialized language: a practical guide to using corpora. London: Routledge.
- Carter, R. (1998). Orders of reality: CANCODE, communication, and culture. *ELT Journal*, 52/1, 43–56.

- Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 3–16). Amsterdam: Rodopi.
- Chan, P-T., & Liou, H-C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning*, 18/3, 231–251.
- Cook, G. (2001). 'The philosopher pulled the lower jaw of the hen.' Ludicrous invented sentences in language teaching. *Applied Linguistics*, 22/3, 366–387.
- Davies, M. (2004). *BYU-BNC: The British National Corpus*. <http://corpus.byu.edu/bnc>
- Flowerdew, L. (2008). Corpus linguistics for academic literacies mediated through discussion activities. In D. Belcher, & A. Hirvela (Eds.), *The oral-literate connection: perspectives on L2 speaking, writing and other media interactions* (pp. 268–287). Ann Arbor: University of Michigan Press.
- Gabrielatos, C. (2005). Corpora and language teaching: just a fling or wedding bells? *Teaching English as a Second Language – Electronic Journal*, 8/4, 1–35.
- Gan, S-L., Low, F., & Yaakub, N. (1996). Modeling teaching with a computer-based concordancer in a TESL preservice teacher education program. *Journal of Computing in Teacher Education*, 12/4, 28–32.
- Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Erlbaum.
- Hulstijn, J. (1997). Retention of inferred and given word meanings: experiments in incidental vocabulary learning. In P. Arnaud, & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 113–125). London: Macmillan.
- Johns, T. (1991). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In T. Johns, & P. King (Eds.), *Classroom Concordancing. English Language Research Journal*, 4, 27–45.
- Johns, T., L. Hsingchin, & W. Lixun. (2008). Integrating corpus-based CALL programs and teaching English through children's literature. *Computer Assisted Language Learning*, 21/5, 483–506.
- Mauranen, A. (2004). Spoken corpus for an ordinary learner. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 89–105). Amsterdam: John Benjamins.
- McKay, S. (1980). Teaching the syntactic, semantic and pragmatic dimensions of verbs. *TESOL Quarterly*, 14/1, 17–26.
- Meunier, F., & Gouverneur, C. (2009). New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 179–201). Amsterdam: John Benjamins.

- O'Sullivan, I. (2007). Enhancing a process-oriented approach to literacy and language learning: the role of corpus consultation literacy. *ReCALL*, 19/3, 269–286.
- Pérez-Paredes, P., & Cantos-Gomez, P. (2004). Some lessons students learn: self-discovery and corpora. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 247–257). Amsterdam: Benjamins.
- Robinson, P. (1997). Individual differences and the fundamental similarity of implicit and explicit adult second language learning. *Language Learning*, 47/1, 45–99.
- Schaffer, C. (1989). A comparison of inductive and deductive approaches to teaching foreign languages. *Modern Language Journal*, 73, 395–403.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11/2, 129–158.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge: Cambridge University Press.
- Somogyi, E. (1996). Using the concordancer in vocabulary development for the Cambridge Advanced English (CAE) course. *ON-CALL*, 10/1. <http://www.cltr.uq.edu.au/oncall/somogyi102.html>
- Thompson, P. (2002). Modal verbs in academic writing. In B. Kettemann, & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 305–325). Amsterdam: Rodopi.
- Thurstun, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes*, 17/3, 267–280.
- Tribble, C., & Jones, G. (1997). *Concordances in the classroom*. (2nd ed.). Houston: Athelstan.
- Williams, J. (2005). Form-focused instruction. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 671–692). Mahwah, NJ: Erlbaum.
- Yoon, H. (2008). More than a linguistic reference: the influence of corpus technology on L2 academic writing. *Language Learning & Technology*, 12/2, 31–49.