ASp
la revue du GERAS

49-50 | 2006
Varia

Habeant Corpus—they *should* have the body. Tools learners have the right to use

Alex Boulton and Stephan Wilhelm

Electronic version
URL: http://journals.openedition.org/asp/661
DOI: 10.4000/asp.661
ISBN: 978-2-8218-0402-9
ISSN: 2108-6354

Publisher
Groupe d'étude et de recherche en anglais de spécialité

Printed version
Date of publication: 1 December 2006
Number of pages: 155-170
ISSN: 1246-8185

Electronic reference
Alex Boulton and Stephan Wilhelm, "Habeant Corpus—they *should* have the body. Tools learners have the right to use", *ASp* [Online], 49-50 | 2006, Online since 03 February 2010, connection on 22 March 2021. URL: http://journals.openedition.org/asp/661 ; DOI: https://doi.org/10.4000/asp.661

Habeant Corpus—they *should* have the body. Tools learners have the right to use

Alex Boulton and Stephan Wilhelm

Introduction

1 Using corpora in language teaching and learning has become "trendy" (Braun 2005: 47) and "fashionable" (Tognini-Bonelli 2006) in many research environments. However, the majority of the many published studies report small-scale operations in universities with well-equipped facilities, using dedicated software which often requires substantial training, and which is in any case prohibitively expensive if available at all to the outside world. Most of these reports tend to concentrate on putting the tools into the hands of the researcher, the materials writer, and occasionally the teacher, but rarely the learners themselves.

2 Given all of this, it is perhaps not too surprising that significant effects have yet to be felt in mainstream teaching in much of continental Europe (Seidlhofer 2000). For example, of the 152 articles published since 2000 in the GERAS journal, *ASp*, only three contain the word corpus in the title—all of them with the same author or co-author.[1]

3 Our position in this paper is that there is little reason for either learners or teachers not to benefit from a corpus linguistics approach. The internet in particular offers a continually expanding and improving source of tools and texts which can be accessed free and used with little training. We report here on a course which attempts to exploit this "free and easy" approach to corpus linguistics.

1. Background

4 The Centre de Télé-enseignement Universitaire at Nancy 2 offers distance degrees in English up to Master's level (five years), with an option in corpus linguistics in the

fourth year. This context means that our learners are generally more advanced and more motivated than many, not to mention older than the average undergraduate. On the other hand, they tend to be less comfortable with information and communication technology (ICT), and the distance learning context puts an extra burden upon them in many cases.

5 Against such a backdrop, offering an option in corpus linguistics may need some justification. First of all, if corpus linguistics is "nothing but a methodology" (McEnery & Wilson 2001: 1), that implies that it might be applied across disciplines (Boulton 2006a, 2006b). Our students are often mainly interested in literature and culture, and corpus linguistics allows a motivating way in, as it can provide "an overarching theme that [can be] recognised as relevant to all aspects of the overall programme of studies" (Seidlhofer 2000: 208-9). This is particularly important for linguistics, as our students come from very varied backgrounds, often with little or no training in linguistics at all; but as none have ever yet done any corpus linguistics, everyone starts off on an equal footing. Furthermore, the course is intended to be practical, with learners defining their own questions and pursuing their own interests, dealing with the very real problems they encounter along the way. They do not, however, need extensive computer skills beyond those required in everyday life. All of this together makes the prospect not only less daunting, but even considerably motivating in most cases.

6 In the course, abstract linguistic and ICT considerations are deliberately kept to a minimum, and the hands-on approach encourages students to research and explore the user of electronic corpora, to experiment with the tools presented, to discover others for themselves, to solve the logistical problems posed by the limitations of the Internet, computers, free on-line tools and everything that implies, as well as their own inexperience in the field. Help is available on demand (usually via e-mail) and students are advised to report in regularly to the teacher, although the flexibility of the distance situation means that this is not an absolute requirement. The mark for the course is based on a 15 to 20-page written assignment produced in students' own time, so they can work at their own pace and within their own capabilities on their own areas of interest. The guidelines are deliberately as open and flexible as possible, requiring only that the final report features some original research on corpus linguistics connected to English. The focus is very much on the process rather than the product and students are asked to be as explicit as possible about their choices at every stage; a lack of concrete results does not necessarily entail a bad mark.

7 As this brief description shows, language may not seem to be the primary focus, but that is not necessarily a problem (Kennedy & Miceli 2001). If the benefits for language learning per se are not self-evident, Chambers' survey of twelve studies using a corpus approach proves optimistic (2005). In our particular case, there is considerable exposure to the target language: learners follow a course in English, use tools in English, select and manipulate a large amount of data in English, and produce the final report in English. As Johns (1991: 30) points out, the underlying assumption behind data-driven learning (DDL) is that "effective language learning is itself a form of linguistic research." Turning this round, linguistic research may in itself be an effective form of language learning, a philosophy not too far from task-based learning. In particular, it may be that some language items are particularly intractable to explicit rule-based learning and amenable only to acquisition or "noticing" and analogy (Boulton 2006c; Bod & Scha 1996).

8 In addition to incidental learning, DDL injects "authenticity" of text, purpose and activity (Stevens 1995), and the process of discovery is consistent with current ideas on autonomisation, learning to learn, and life-long learning—especially where the potential for application to other non-linguistic fields is made clear (Bernardini 2000). The skills involved require deep cognitive processing, which should lead to increased language learning as well as helping learners to reason about language, an important skill for those who go on to teach (Seidlhofer 2000). Despite appearances, a corpus linguistic approach is process- rather than product-oriented, learner- rather than language-centred, meaning- rather than form-focused (Bernardini 2001b).

2. Student papers

9 So much for the underlying philosophy, but given a free rein, what exactly do the learners do? We looked at the finished product for thirty papers submitted between 2002 and 2005; we excluded ones that did not achieve at least 10/20, were not submitted electronically, or which adopted a more theoretical approach.

10 Of the thirty, nineteen used at least one large published corpus;[2] 25 compiled one or more corpora of their own (table 1). This in itself suggests that learners feel the relevance of both large and small corpora, which rather deflates one ongoing debate.[3]

Table 1. Choice of corpora

		own corpora			
		0	1	2	tot
large public corpora	0		4	7	11
	1	3	4	6	13
	2	2	2	2	6
	tot	5	10	15	

11 More relevant perhaps is that only seven learners limited themselves to a single corpus, the others using two or more for comparative purposes. These comparisons frequently focused on differences in style (e.g. between "quality" and "popular" newspapers), genre (e.g. aviation and general English), variety (e.g. British and American English), time (e.g. English today and 100 years ago), language (e.g. French and English for translation purposes), and so on. Almost all the corpora were downloaded from the web, although surprisingly the Internet itself was only used twice as a corpus in its own right despite the enormous potential (see e.g. Bergh 2005).

12 Common choices of corpus included newspapers, governmental and party political websites, literature, and song lyrics. These sources reveal some of the learners' interests, either within the context of their studies or for outside pursuits: culture and politics, news and history, literature and music. This goes against Chambers' (2005) conclusion that learners find the crossover difficult, and would highlight the diversity of applications of a corpus approach. Arguably, in the end, very few student papers limited themselves to a purely language topic. The ones that did were mostly comparisons between pairs or triplets of individual words which students found confusing, or where they were unhappy with traditional dictionary or grammar explanations, such as speak and talk (see Partington 2001b on "confusable words").

13 Whatever the focus, most began with a specific question rather than starting with a corpus to see what it revealed. Although this reduces the "serendipity effect" (Johns

1988), it is perhaps understandable as it makes for a generally more straightforward first project. But in either case (top-down or bottom-up), the process of first-hand interaction with the corpora and the tools allows learners to discover various characteristics of how language works, and not just the target language. For example, various abstract concepts they are familiar with on a theoretical level take on a new dimension, becoming more "real" as their importance becomes clear-part of speech, lemmatisation, word, collocation, homonymy, and so on.

3. From preaching to practice

14 In this section we turn the tables on ourselves and try to put into practice some of what we preach: to create and analyse a small corpus with only free and easy tools as used by our learners. In order to remain within the context of the course, we compiled the corpus from students' past research papers which had been available to all on the course website. The nine that were chosen provided a corpus of around 40,000 words, well within the range of the learners' own home-made corpora. All these papers had been submitted by students enrolled in the course between 2002 and 2004, providing a principled and relatively homogeneous set of texts: they were all written by French postgraduate students of English and received a mark of at least 12/20; they all deal with corpus linguistics and contain language appropriate for a student research paper of this type; they all conform to the same course requirements, and are roughly the same size.

15 The main point of learner corpora, as Pravec points out (2002: 81), is to see how they "provide a deviation from the standard, i.e. the language of the native speakers of a particular language." While not going too far into issues of the desirability of native speakers as a target for language learners (see e.g. Cook 1998), it seems clear that learners welcome native speaker models as a "yardstick" for assessing their own texts (Gabel 2001: 274), and that "the most useful corpus for learners of English is the one which offers a collection of *expert performances* in genres which have relevance to the needs and interests of the learners" (Tribble 1997, original emphasis; see also Williams *et al.* 2002: 49). Such a comparison seems appropriate here since, as we have seen, the majority of the students' papers also involved a comparative approach of some kind.

16 For any comparison to work, the two corpora should be as close as possible in nature so that only one variable is targeted – in this case, learner vs. expert writing. Bearing in mind our aim to use only free and accessible tools, we used Google with the key words *corpus-based* to locate appropriate papers from British and American university sites on the internet; ten were needed for a corpus of equivalent size. These were then manually checked to ensure they were comparable in format and, as far as we could tell, written by native or near-native professionals. Once the texts for the two corpora had been selected, they were then subjected to the same editing treatment: bibliographies, tables, word lists, example sentences, long quotations, definitions and illustrations, etc. were all removed. This gave a final running total of 41,762 words in the edited learner corpus (LC), and 41,358 in the native corpus (NC).

17 Perhaps the majority of studies of learner corpora to date, such as most in Aston (2001a), have concentrated on language errors-indeed, error-analysis seems to have made something of a come-back in recent years. We decided however to take a rather different approach, partly because most of the students' research papers had been

subjected to a certain amount of teacher-correction during the academic year, but partly also because a major concern in specific genres like academic writing appears to be stylistic (Biber *et al.* 2002: 170). As Gabel (2001: 271) points out:

> In the advanced stages of language learning the [interlanguage] is characterized to a lesser extent by overt "errors", but rather by lexical simplifications, semantic vagueness, syntactic monotony and gross pragmatic directness.

18 We therefore decided to concentrate on "foreign-soundingness", to use Pravec's term (2002: 83). This is no easy matter under ordinary circumstances, even for experienced teachers (Cobb 2003). Our intuition, however, was that corpus methodology might allow some insights where vague impressions had failed. Again following Pravec's paper (2002: 83), our starting position was that:

> Aspects of "foreign-soundingness" in non-native essays [...] are usually revealed by the overuse or underuse of words or structures with respect to the target language norms. This is done by means of a comparison between individual L2 sub-corpora and native English corpora.

19 Consequently, a higher or lower frequency of certain forms or structures in the learner corpus is likely to indicate a possible area in which learners' academic writing could be improved. We eventually concentrated on three distinct levels of language structure, beginning with areas most frequently analysed by the students themselves: isolated words, lexicalised patterns and prefabs, and grammatical structures.

4. Isolated words

20 As Ringbom (1998: 49) remarks in his study of advanced EFL writing:

> A frequently voiced view is that learner language is vague and stereotyped. This would be a natural consequence of its vocabulary being more limited than that of native speakers. However, concrete evidence of exactly what constitutes this vagueness has been hard to come by.

21 Corpus methods allowed us to test this view in a number of simple ways. Our first port of call was Cathy Ball's Web Frequency Indexer, which deconstructs a text into its component words.[4] Although the corpora had an equivalent number of tokens (less than 1% difference), the frequency lists revealed a different picture: LC had 3,501 different types, NC 3,920 – a difference of 11.97% (table 2). The learners thus apparently use a more restricted range of vocabulary than the natives, as reported in the overview of learner corpora provided in Granger (1998a). This comparative lack of lexical diversity no doubt accounts in part for the "foreign" quality of their writing – Gabel's (2001) "over-indulgence" of common items vs "under-representation" of less frequent ones.

Table 2. Frequency figures

	LC	NC	NC+LC
tokens	41,762	41,358	– 0.99%
types	3,501	3,920	+ 11.97%

22 Another on-line site used by a number of students was Paul Nation's Vocabulary Profiler,[5] which offers a number of complementary statistical features (table 3). Pasting our two corpora separately into this program revealed that 82.85% of the LC tokens are among the 2,000 most frequent word families of English, compared to only 74.91% of the NC tokens. The learners thus use fewer numbers of off-list (i.e. less frequent) words, and in particular barely half as many items from the Academic Word List. This is a striking conclusion for an academic writing activity with obvious pedagogical implications, but confirms the general finding that learners overuse high frequency items (Granger 1998a).

23 Another useful feature of the program is etymological: classical items (i.e. those of Greek, Latin or French origin) accounted for 33.39% of all on-list tokens in LC compared to 36.41% in NC. It might have been expected that the learners would use a higher percentage, as less frequent words in English are more likely to be of classical origin and thus cognate with French. For example, 56% of the 1,000 most frequent word families of English are classical cognates, but this proportion rises to just over 95% in the Academic Word List.[6]

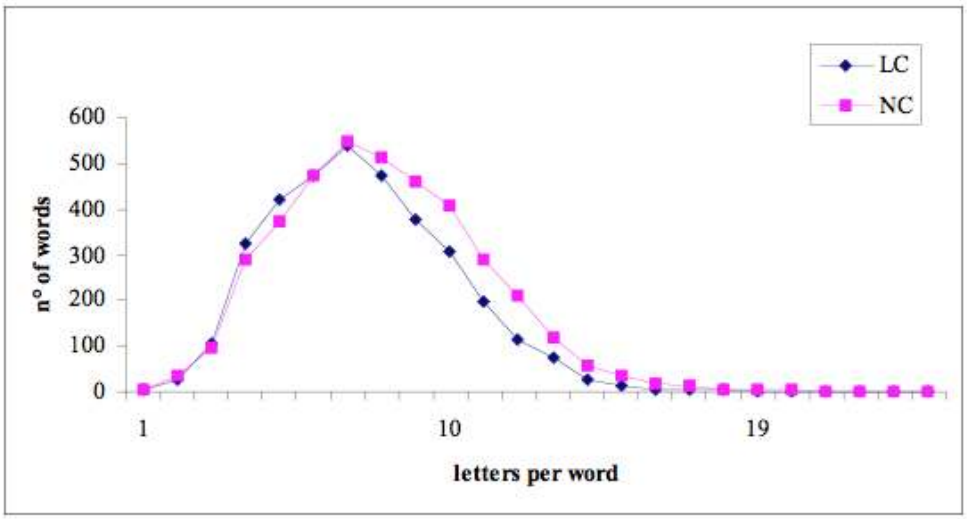
Table 3. Paul Nation's Vocabulary Profiler

	K1	K2	AWL	off-list	lexical density	classical index
LC	76.12%	6.73%	6.48%	10.67%	0.56	33.39%
NC	70.14%	4.77%	12.02%	13.08%	0.58	36.41%

K1 = the most frequent 1000 word families; K2 = the second most frequent 1000 word families; AWL = academic word list; off-list = all other words; lexical density = content words divided by total; classical index = on-list words of French, Latin or Greek origin divided by all on-list tokens; lexical density = content words divided by tokens

24 If learners typically use "simpler" words according to the measures so far, we might also think they would use shorter ones, as the overall complexity of lexical items generally increases with length. This indeed seems to be the case: mean word (type) length is 7.44 in LC, 8.03 in NC (SD 2.62 and 2.91 respectively). It is particularly noticeable that NC features many more words of between 9 and 12 letters (figure 1).

Figure 1. Word length



25 It would clearly be simplistic to suggest that learners should use more or longer words in their written compositions; the question is how they can broaden their range. Corpus linguistics not only encourages noticing (especially once attention has been drawn to certain general features, such as diversity and accuracy of lexis in a particular genre), but also allows exposure to large quantities of appropriate material. This is especially true at higher levels, as the number of encounters needed to assimilate a new word decreases as overall language proficiency increases (Zahar *et al*. 2001).

5. Lexical patterns

26 So far we have looked only at isolated words, which provide a convenient place to start as the analysis is relatively straightforward, and they are prominent in our learners' work. But of course, they do not tell the whole story. Indeed, the advantage of a corpus approach is to explore words in context. Most of our learners do this in a fairly traditional way by generating concordances to look at key words in context (KWIC) or at sentence level, and making collocate lists. Basic concordancing procedures have already been widely reported, so we shall not dwell on them here.

27 Fortunately, as Braun (2005: 52) reminds us, corpus methodology does not end with concordances but lends itself to many other angles of study. For example, the importance of chunking has long been discussed in relation to language teaching and learning (e.g. Pawley & Syder 1983; Nattinger & DeCarrico 1992). In corpus linguistics, this is what Sinclair (1991) calls the "idiom principle" (as opposed to the "open-choice principle"), and the evidence is that learners typically underuse this, contributing again to the "foreign-soundingness" of their productions. Partington (2001a: 53) provides two succinct quotes to illustrate the difference and how the times have changed:

> Our approach to language teaching... is structural. The words we choose to present for use in the structures are only of secondary importance, because once the patterns of English are mastered, it is relatively easy to learn new words to fit into the patterns. (Broughton 1968: 14)
> Syntactic structures and lexical items... are co-selected... It is impossible to look at one independently of the other. Particular syntactic structures tend to co-occur

> with particular lexical items, and—on the other side of the coin—lexical items seem to occur in a limited range of structures. (Francis 1993: 147)

28 If learners do tend to operate more on the "slot-and-filler" principle, we may expect to find fewer collocations and idioms than in native texts, as De Cock *et al.* (1998) and Gabel (2001) have already suggested. To test this on our corpora, we looked first at phrasal verbs, as these are traditionally considered notoriously idiosyncratic and "difficult", at least from a French perspective. They also feature prominently in the corpus literature, starting with Sinclair's 1991 treatment of *set*.[7]

29 For this analysis, we used WordSmith Tools, but restricted ourselves to the limited functions of the free demonstration version, as this is what most students used.[8] The corpora were scanned using WordSmith Concord to find the most commonly used prepositions and adverbs, and then the collocates and patterns associated with them. This allowed us to detect six phrasal verbs which occurred relatively frequently (table 4). Although observations made on the basis of a small number of occurrences must not be overplayed, it is remarkable that all these common phrasal verbs appeared in higher proportion in LC than in NC (on average over four times as often), which again seems to point to a tendency on the part of advanced learners to overuse common items relative to native speakers.

Table 4. Notable phrasal verbs

phrasal verb	**LC**	**NC**	**total**
carry out	18	7	25
look for	10	3	13
point out	8	4	12
sort out	10	0	10
look up	7	0	7
find out	6	0	6
total	59	14	73

30 Paradoxically, overuse of common phrasal verbs in LC may result from learners perceiving them as "native-like" due to the strong emphasis commonly laid on them in traditional teaching. Corpus-based studies might enable them to identify some of the phrasal verbs most likely to be overused and opt for alternative expressions, especially in academic writing.

31 Casting our net rather wider, we were interested to identify larger multi-word patterns in the corpora. We first identified the commonest items for the two corpora together using the Web Frequency Indexer (table 5a). The collocates, pattern and clusters functions of WordSmith Tools were then used to determine what prefabricated phrases occurred in connection with them. The results showed that most set phrases appear in higher proportion in LC than in NC, with ratios of more than 5 to 1 in the case of *in*

order to and 3 to 1 in that of *on the one / the other hand*; expressions like *bear in mind* and *keep in mind* were observed to follow the same pattern (table 5b).

Table 5: a) Commonest function words. b) Commonest prefabricated phrases

rank	word	LC	NC	total		chunk	LC	NC	total
1	the	3197	2455	5652		in order to	53	10	62
2	of	1488	1648	3136		as well as	19	15	34
3	and	1128	1053	2181		the fact that	24	5	29
4	to	1034	1000	2034		in terms of	10	17	27
5	in	1110	922	2032		on the one/the other hand	19	6	25
6	a	860	1038	1898		a number of	7	15	22
7	is	561	751	1312		as far as	9	1	10
8	that	446	524	970		total	141	69	209
9	for	400	449	849					
10	as	438	322	760					

32 These results concord with our other findings so far: our students tend to overuse common items. In the particular case here, we might infer that advanced learners use some prefabricated phrases more frequently than others because they correspond more closely to the logical articulation generally found in French argumentative texts (e.g. *in order to, on the one / the other hand*). For Flowerdew (2000: 153), such "pragmatic errors are more serious than those in the referential area as they can lead to misrepresentation of the content and may also be stylistically inappropriate for the context of the writing." Although numerous textbooks present a variety of prefabricated phrases, it might seem that this is not enough. Again, it has been argued that concordances may serve to draw attention to them in the range of contexts necessary for better assimilation (Zorzi 2001).

6. Structures

33 So far the analyses we have carried out have not required too much imagination, as we have remained largely on the lexical level. More daunting in many cases is a syntactic analysis: it can be almost impossible to collect all and only the target features (Bernardini 2001b: 231), as even such basic items as relative clauses, for example, may lack relative pronouns and thus "have no distinctive feature in their surface form which can be searched for" without complex tagging (Aston 2001b: 17). While our students did use pre-existing tags in large public corpora, they almost never managed to tag their own, no doubt because the tools simply aren't available.[9] Given the

students' reluctance to tag, we decided to follow suit and concentrate on semi-visible surface features in two structures which traditionally pose problems for learners – present perfect and passive constructions. Admittedly, our students, just like those in other studies (e.g. Bernardini 2000), generally avoided such topics as they are less accessible. This has led some to disparage DDL for concentrating on "minute details of the phraseology of particular words, [which] may be difficult to reconcile with the 'big themes' of language teaching, such as 'tenses' or 'articles'" (Hunston 2002: 184). If "big themes" are comparatively rare in our students' papers, they are not impossible and are presented here as an example of the power of corpus methodology even for the user with limited skills (see Hahn 2000).

34 It is a common view among English teachers that, of all the verb forms of English, the present perfect is among the least easily mastered by French learners. This is in part because such aspectual forms are rare in human languages as a whole, but also partly due to the superficially similar *passé composé* in French – the two forms do not reliably have equivalent function or meaning. As for passive forms, a number of comparative studies (famously Vinay & Darbelnet 1977) show that English is more likely to use them where French favours other structures. Our advanced learners might therefore be expected to experience some degree of difficulty with the passive, especially in academic prose where it is purported to be a relatively prominent feature.

35 We used WordSmith's Concord function for both advanced searches: for the present perfect, these were identified using HAVE / HAS + *ed / *en; for the passive, WAS / WERE + *ed / *en. This is far from a perfect search option, and several problems had to be dealt with. Firstly, to cater for such forms as *has been seen* or *were then being taken*, we allowed hits within three words. Secondly, passives with present forms of the auxiliary *be* had to be abandoned due to the over-frequent occurrence of *ed / *en items other than past participles. Thirdly, in the absence of tagging, no satisfactory solution was found for locating irregular past participles; however, the number of hits obtained on regular items was considered sufficient for our purposes. Finally, the lists had to be checked manually to clear the data of irregularities. If such methods seem rather laborious, they are to an extent inevitable with such searches (Flowerdew 2000); their description here serves to highlight the kind of problems facing learners with minimal tools, and how they typically use their ingenuity to overcome the obstacles.

36 The results showed that LC contained significantly fewer examples of the present perfect than NC, but more passives (table 6). Larger corpora and more sophisticated instruments would undoubtedly allow more satisfactory measurements, but even without an in-depth statistical analysis (which our students generally shy away from), the basic trends seem clear.

Table 6. Two structures

	LC	**NC**	**total**
present perfect	103	202	305
passives	139	105	244

37 The question now is to interpret the findings. The under-use of the present perfect supports our hypothesis that even at advanced levels, learners may have trouble using structures that are perceived as "difficult" and resort to the simplest strategy – avoidance (Williams *et al.* 2002).

38 More interesting is the over-use of passive forms. Firstly, the English passive may be perceived as less intimidating if it is related formally and semantically to the *passif* in French; such equivalences for the present perfect / *passé composé* are treated as "howlers" (*barbarismes*) in French classrooms. A further possibility was that the learners might be overusing passive forms to avoid another traditional shibboleth, *viz* the comparative frequency of the pronoun *on* in French compared to one in English, and perhaps also first person pronouns. This appeared all the more probable as Granger (1998b) had previously found that French learners massively overuse such frames as *we/one/you can/cannot*, etc.

39 WordSmith Tools was used to track down occurrences of the various personal pronouns in the two corpora (table 7). Neither natives nor learners made frequent use of *I* or *you* and related forms, presumably to avoid over-personalisation. *One* was also relatively rare in both corpora, especially for the learners who have probably in the past had their knuckles rapped for using it where French has *on*, and subsequently underuse it. However, the proportion of first person plural forms proved consistently higher in LC than in NC, Granger's (1998b) pattern being reproduced in the case of such frames as *we can*: LC 40 vs NC 10.

Table 7. Personal pronouns

	1st pers sing	1st pers pl	2nd pers sing/pl	one
LC	23	607	11	11
NC	32	267	39	13
total	55	874	50	24

40 Learners' overuse of passives is not then just to compensate *we*, which was also overused. This suggests that quite different structures and discourse may be used in the two corpora, with again the learners overusing a limited few. We leave it up to the readers to draw their own conclusions for teaching and learning, but again, it seems to us that learners sensitised to such facts through their own corpus research would probably benefit from it and confront certain choices accordingly.

Conclusion

41 This paper has reported on a course in corpus linguistics in a distance education language degree. The objectives were to sensitise learners to the possibilities of using corpora in their language learning, as well as in their wider studies, both of which seem to have been fulfilled. While learners clearly need guidance in appropriating a methodology radically different from what they are used to, "the difficulties should not be overestimated; learners should quickly acquire the skills needed" (Bernardini 2001b:

243). In the course described, the guidance provided was deliberately kept to a minimum so as to allow the greatest choice: "it would be inherently contradictory to prescribe a methodology when the aim of the approach is to give learners the instruments to develop their own methodologies and make their own discoveries" (Bernardini 2001b: 228). Although "only fairly simple queries can be handled at this stage [...] the results can be illuminating and very helpful" (Sinclair 2004: 288). Indeed, we have argued that "corpus skills constitute a learning task in themselves [...] Once acquired, they facilitate learning greatly and need not be constantly refreshed" (Mauranen 2004: 99).

42 Following on from this, we built and analysed two home-made corpora of student reports and native-speaker papers on corpus linguistics using only the kinds of simple tools freely available on the Internet, all of which had been used by our students. Although this limited the possibilities to fairly simple methods, the results are in line with more sophisticated studies of learner corpora. Briefly, they suggest that even advanced learners could benefit from the kind of noticing that a corpus approach encourages. In particular, attention could usefully be paid to such learners' overuse of highly frequent items, underuse of prefabs, non-nativelike discourse structure, avoidance of certain complex grammatical forms and restricted range in the genre of academic writing. The key point is that such experiments suggest that corpus-based studies, even though based on small corpora (particularly in specialist domains) and carried out with basic computer tools, might provide relatively advanced learners at least with the means of identifying deficiencies in their use of English and direct their efforts accordingly.

BIBLIOGRAPHY

Aston, G. (ed.) 2001a. *Learning with Corpora*. Houston: Athelstan.

Aston, G. 2001b. "Learning with corpora: An overview". In Aston, G. (ed.), *Learning with Corpora*. Houston: Athelstan, p. 7-45.

Bergh, G. 2005. "Min(d)ing English language data on the Web: What can Google tell us?" *ICAME Journal* 29, 25-4, accessed February 2006 <http://gandalf.aksis.uib.no/icame/ij29/ij29-page25-46.pdf>.

Bernardini, S. 2000. "Systematising serendipity: Proposals for concordancing large corpora with language learners". In Burnard, L. & T. McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang, 225-234.

Bernardini, S. 2001a. "Corpora in the classroom: An overview and some reflections on future developments". In Sinclair, J. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 15-36.

Bernardini, S. 2001b. "'Spoilt for choice': A learner explores general language corpora". In Aston, G. (ed.), *Learning with Corpora*. Houston: Athelstan, 220-249.

Biber, D., S. Conrad, R. Reppen, P. Byrd & M. Helt. 2002. "Speaking and writing in the university: A multidimensional comparison". *TESOL Quarterly* 36, 9-48.

Bod, L.W.M. & R.J.H. Scha. 1996. "Data-oriented language processing: An overview". *University of Amsterdam Institute for Logic, Language and Computation, Research Report*, LP-96-13. Reprinted in Sampson, G. & McCarthy, D. (eds.) 2004. *Corpus Linguistics: Readings in a Widening Discipline*. London: Continuum, 304-325.

Boulton, A. 2006a. "When linguistics isn't linguistics: Interdisciplinary approaches to corpus linguistics". Journée d'études IDEA. Université Nancy 2, 18 February.

Boulton, A. 2006b. "Bringing corpora to the masses: Free and easy tools for language learning". Teaching and Language Corpora (TaLC 7). Université Paris 7 / BNF, 1-4 July.

Boulton, A. 2006c. "Tricky to teach, easier to learn: Empirical evidence for corpus use in the language classroom". American Association of Applied Corpus Linguistics (AAACL). Northern Arizona University, 20-22 October.

Braun, S. 2005. "From pedagogically relevant corpora to authentic language learning contents". *ReCALL* 17/1, 47-64.

Chambers, A. 2005. "Integrating corpus consultation in language studies". *Language Learning & Technology* 9/2, 111-12, accessed February 2006 <http://llt.msu.edu/vol9num2/chambers/>.

Cobb, T. 2003. "Analyzing late interlanguage with learner corpora: Quebec replications of three European studies". *Canadian Modern Language Review* 59/3, 393-423.

Cook, G. 1998. "The uses of reality: A reply to Ronald Carter". *ELT Journal* 52/1, 57-63.

De Cock, S., S. Granger, G. Leech, & T. McEnery. 1998. "An automated approach to the phrasicon of EFL learners". In Granger, S. (ed.), *Learner English on Computer*. London: Longman, 67-79.

Flowerdew, L. 2000. "Investigating referential and pragmatic errors in a learner corpus". In Burnard, L. & T. McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt: Peter Lang, 145-154.

Gabel, S. 2001. "Over-indulgence and under-representation in interlanguage: Reflections on the utilization of concordancers in self-directed foreign language learning". *Computer Assisted Language Learning* 14, 269-288.

Granger, S. (ed.) 1998a. *Learner English on Computer*. London: Longman.

Granger, S. 1998b. "Prefabricated patterns in advanced EFL writing: Collocations and formulae". In Cowie, A. (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, 145-160.

Hahn, A. 2000. "Grammar at its best: The development of a rule- and corpus-based grammar of English tenses". In Burnard, L. & T. McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang, 193-205.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johns, T. 1988. "Whence and whither classroom concordancing?" In Bongaerts, P., P. deHaan, S. Lobbe & H. Wekker (eds.), *Computer Applications in Language Learning*. Dordrecht: Foris, 9-27.

Johns, T. 1991. "From printout to handout: Grammar and vocabulary learning in the context of data-driven learning". *ELR Journal* 4, 27-45.

Kennedy, C. & T. Miceli, T. 2001. "An evaluation of intermediate students' approaches to corpus investigation". *Language Learning & Technology* 5/3, 77-90, accessed February 2006 <http://llt.msu.edu/vol5num3/kennedymiceli/>.

Mauranen, A. 2004. "Spoken corpus for an ordinary learner". In J. Sinclair (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 89-105.

McEnery, T. & A. Wilson, A. 2001. *Corpus Linguistics*. (2nd ed.) Edinburgh: Edinburgh University Press.

Nattinger, J.R. & J. S. Decarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Partington, A. 2001a. "Corpora and their uses in language research". In Aston, G. (ed.), *Learning with Corpora*. Houston: Athelstan, 46-62.

Partington, A. 2001b. "Corpus-based description in teaching and learning". In Aston, G. (ed.), *Learning with Corpora*. Houston: Athelstan, 63-84.

Pawley, A. & F. Syder. 1983. "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency". In Richards, J. & R. Schmidt (eds.) *Language and Communication*. London: Longman, 191-226.

Pravec, N.A. 2002. "Survey of learner corpora". *ICAME Journal* 26, 81-114, accessed February 2006 <http://gandalf.aksis.uib.no/icame/ij26/pravec.pdf>.

Ringbom, H. 1998. "Vocabulary frequencies in advanced learner English: A cross-linguistic approach". In Granger, S. (ed.), *Learner English on Computer*. London: Longman, 41-52.

Seidlhofer, B. 2000. "Operationalizing intertextuality: Using learner corpora for learning". In Burnard, L. & T. McEnery (eds.), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang, 207-223.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, J. 2004. "New evidence, new priorities, new attitudes". In Sinclair, J. (ed.), *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins, 271-299.

Stevens, V. 1995. "Concordancing with language learners: Why? When? What?" *CAELL Journal* 6/2, 2-10, accessed February 2006 <http://eisu.bham.ac.uk/johnstf/stevens.htm>.

Tognini-Bonelli, E. 2006. "The role of corpora in linguistic description: Between lexis and grammar". 27e Congrès du GERAS. Lorient, Université de Bretagne Sud, 23-24 March.

Tribble, C. 1997. "Improvising corpora for ELT: Quick and dirty ways of developing corpora for language teaching". In Lewandowska-Tomaszczyk, B. & J. Melia (eds.), *Proceedings of the First International Conference on Practical Applications in Language Corpora*, accessed February 2006 <http://www.eisu.bham.ac.uk/johnstf/palc.htm>.

Vinay, J.P. & J. Darbelnet.1977. *Stylistique comparée du français et de l'anglais*. (2nd edn) Paris: Didier.

Williams, G. 2001. "Mediating between lexis and texts: Collocational networks in specialised corpora". *ASp* 31-33, 63-76.

Williams, G. 2003. "From meaning to words and back: Corpus linguistics and specialised graphy". *ASp* 39-40, 91-106.

Williams, G., C. Sionis & P. Boucher. 2002. "Single language corpus, multilingual background". *ASp* 37-38, 47-57.

Zahar, R., T. Cobb & N. Spada. 2001. "Acquiring vocabulary through reading: Effects of frequency and contextual richness". *Canadian Modern Language Review* 57/4, 541-572, accessed March 2006 <http://www.utpjournals.com/jour.ihtml?lp=product/cmlr/574/574-Zahar.html>.

Zorzi, D. 2001. "The pedagogic use of spoken corpora: Learning discourse markers in Italian". In Aston, G. (ed.), *Learning with Corpora*. Houston: Athelstan, 85-107.

NOTES

1. Williams 2001; Williams 2003; Williams *et al.* 2002.
2. These were overwhelmingly the free but limited demonstration versions of the Bank of English (<http://www.collins.co.uk/Corpus/CorpusSearch.aspx>) and the British National Corpus (<http://sara.natcorp.ox.ac.uk/lookup.html> or more recently <http://view.byu.edu/>)
3. See e.g. Bernardini (2000, 2001a, 2001b) for arguments in favour of large corpora; Braun (2005) for opposing views, bearing in mind that the majority of recent studies tend towards small corpora, such as those in Aston (2001a).
4. This very popular tool can be accesssed from <http://www.georgetown.edu/faculty/ballc/webtools/web_freqs.html>
5. <http://www.lextutor.ca/vp/eng/>
6. Figures compiled from the Compleat Lexical Tutor <http://www.lextutor.ca/ListLearn/>
7. *Pace* Kennedy and Miceli's comment on "the fatal lure of prepositions" (2001: 83).
8. A popular choice among the students, the free but limited demonstration version can be downloaded from <http://www.lexically.net/wordsmith/> courtesy of Mike Scott, one of WordSmith's creators. Several students in the past have purchased a licence for the full version (currently about €80).
9. There are occasional facilities for having short texts tagged by e-mail, such as that at the University of Leeds <http://www.comp.leeds.ac.uk/amalgam/amalgam/amalgtag3.html> One of the most popular professional POS taggers is CLAWS from the University of Lancaster, but this is tremendously complex for learners' needs, and the free demonstration version is restricted to 300 words; the full version costs over €1000 <http://www.comp.lancs.ac.uk/ucrel/claws/purchase.html>

ABSTRACTS

With the advent of fast, powerful, cheap and accessible computer tools, the use of corpora has exploded in the last 20 years. In the field of language learning, however, their use is mainly restricted to researchers, course writers and teachers, while the benefits to the learner are largely second hand: rare is the teacher who allows a class direct access to corpus methodology. This paper argues that there is no reason not to trust at least advanced learners with corpus tools, and that there are significant advantages to encouraging a hands-on approach. After outlining the rationale underpinning this approach, we describe an English course where learners are required to apply corpus techniques to an existing corpus or one of their own devising. We then go on to describe our students' own productions, using only corpus techniques

and tools used by the learners themselves, all freely available on the internet and requiring minimal training.

Grâce à des outils informatiques rapides, puissants, peu onéreux et aisément accessibles, l'utilisation des corpus a vu une véritable explosion au cours des vingt dernières années. Dans le domaine de l'apprentissage des langues étrangères, cependant, l'exploitation des corpus est essentiellement le fait des chercheurs, des auteurs de manuels et des enseignants, tandis que les bénéfices que les apprenants retirent de ces avancées sont la plupart du temps indirects. Rares, en effet, sont les enseignants qui permettent à leurs étudiants un accès direct aux corpus. Cet article défend l'idée que rien ne s'oppose à l'utilisation des corpus au moins par des apprenants « avancés » et que le fait d'encourager cette démarche active comporte des avantages considérables. Après avoir défini brièvement la logique de l'approche présentée, nous décrirons un cursus d'anglais dans lequel nous demandons aux apprenants d'appliquer les techniques d'analyse de corpus à un corpus existant ou confectionné par leurs soins. Nous décrirons ensuite les productions de nos propres étudiants en utilisant les mêmes techniques et outils, disponibles gratuitement sur Internet, et qui ne nécessitent qu'un degré minimal de maîtrise de l'informatique.

INDEX

Keywords: corpus linguistics, data-driven learning, distance learning, incidental learning, learner corpora, non-linguistic application
Mots-clés: application non linguistique, apprentissage fortuit, apprentissage guidé, corpus d'apprenants, linguistique de corpus, télé-enseignement

AUTHORS

ALEX BOULTON

Alex Boulton est Maître de conférences à l'Université Nancy 2 où il enseigne au Centre de Télé-enseignement Universitaire. Il est membre du CRAPEL-ATILF/CNRS où il poursuit ses recherches sur les technologies de l'information de la communication, et plus précisément sur les applications didactiques de la linguistique de corpus. alex.boulton@univ-nancy2.fr

STEPHAN WILHELM

Stephan Wilhelm travaille en formation continue chez Force Langues à Maubeuge, principalement en anglais de spécialité auprès de cadres en entreprise. Outre son intérêt pour l'exploitation de la linguistique de corpus, ses travaux de recherche portent principalement sur la variation des accents et des dialectes en Grande Bretagne. sb.wilhelm@tele2.fr