# Corpus linguistics and second/foreign language learning: exploring multiple paths

## *Linguística de corpus e aprendizagem de segunda língua/língua estrangeira: explorando caminhos múltiplos*

Fanny Meunier*
Catholic University of Louvain
Louvain-la-Neuve / Belgium

ABSTRACT: The aim of this article is twofold: first, to briefly assess the influence that corpus linguistic research has had on second/foreign language learning so far, and second, to suggest future directions for a more coherent and well thought out integration of corpora in instructed settings. In section 1, the influence of *native* and *learner* corpus research on second/foreign language learning will be assessed in turn, and some reasons for the overall lack of uptake of corpora in educational contexts will be put forward. In section 2, I will argue that multiple paths will have to be explored for a better integration of corpora in instructed settings. The fact that various – and sometimes even radically opposite – directions will be proposed might appear conflicting at first sight, but it will be demonstrated that opting for a multiplicity of perspectives is the only way to lay the foundations of a healthy cross-fertilization between corpus linguistics and the current multi-faceted language learning and teaching cultures.

KEYWORDS: corpus linguistics; second/foreign language corpora; applications; second/foreign language teaching and learning.

RESUMO: O objetivo deste artigo é duplo: primeiramente, avaliar de forma sucinta a influência que a pesquisa da linguística de corpus tem tido sobre a área de aprendizagem de segunda língua/língua estrangeira; e, em segundo lugar, sugerir direções futuras para uma integração mais coerente e bem refletida sobre a integração de corpora aos ambientes de ensino. Na seção 1, a influência da pesquisa de corpora de língua nativa e corpora de aprendizes na aprendizagem de segunda língua/língua estrangeira será avaliada e algumas razões para a não-adoção dessas metodologias nos ambientes de ensino serão apontadas. Na seção 2, argumentarei que diferentes caminhos deverão ser adotados para que haja uma melhor integração entre corpora e ambientes de ensino. O fato de vários – e às vezes, até mesmo – caminhos opostos, serem propostos, pode parecer conflitante à primeira vista. Mas, será mostrado que, a opção por perspectivas múltiplas é o único caminho para que se estabeleçam bases saudáveis para a interação entre as culturas das áreas de estudos de corpora e aprendizagem e ensino de línguas.

PALAVRAS-CHAVE: linguística de corpus; corpora de segunda língua; aplicações; ensino/aprendizagem de segunda língua/língua estrangeira.

* fanny.meunier@uclouvain.be

## 1. Corpus linguistic research and second/foreign language learning: a brief state-of-the art

When trying to assess the influence of corpus linguistic research on second/foreign language learning (henceforth abbreviated as L2[1] language learning or L2L), it seems reasonable to address the influence of native and learner corpus research separately given that they have had rather different implications on instructional settings.

### 1.1 Native corpus research and L2 language learning

Widdowson (2004, p. 357), in an article addressing the recent trends in English language teaching, acknowledges the impact of technology on the current modes of language use and communication but also on "ways in which the language so used is recorded and analysed". He adds (2004, p. 357) that "the most striking development in linguistic description over the past twenty years has been the use of the computer to collect and analyse vast corpora of actually occurring language data" and speaks of an "abundance of dictionaries and grammatical **descriptions** which are corpus-based and which chart the patterns of the contemporary usage of English". In Meunier & Gouverneur (2009) we also argued that corpora have found their way to the offices of major ELT publishers and are increasingly used as a source of authentic data to inform new series of **reference and pedagogical materials** such as dictionaries, grammar or vocabulary books. Cambridge University Press offers a 'real English guarantee'[2] to the buyers and users of their material, Longman assures its readership that [they] 'only see real English, as it is really used'.[3] As for MacMillan, the use of their World English Corpus is described as 'a unique modern database of over 200 million words revealing fresh information on how words are used and natural examples of English as it is written and spoken now!'[4]

Römer (2006, p. 121) states however that "despite the progress that has been made in the field of corpus linguistics and language teaching, the **practice** of ELT has so far been largely unaffected by the advances of corpus research";

[1] In this article, no distinction will be made between second and foreign language learning, hence the general L2 abbreviation (which also encompasses the learning of possible third, fourth, etc. languages).

[2] See <http://www.cambridge.org/elt/corpus/corpus_based_books.htm>.

[3] See <http://www.longman.com/dictionaries/corpus/index.html>.

[4] See <http://www.macmillandictionary.com/aboutcorpus.htm>.

Breyer (2009) concurs by highlighting the opposition between the enthusiasm of the research community and the dearth of applications of corpus tools and resources in the classroom. There is thus a clear divide between the exponentially growing number of publications in applied native corpus research and the introduction of corpus data in reference books and teaching materials on the one hand and everyday teaching practices on the other. Furthermore, as English is the language that has been described most fully thanks to corpus methods, the gap or opposition mentioned by Römer and Breyer is likely to be even much wider for other languages.

Several reasons account for the lack of uptake of corpus-oriented tools and methods in the classroom, and I will expand on four of them. One rather fundamental reason is that the enthusiasm for corpus methods is mostly expressed by linguists and that the authority of the linguists does not always find an echo among teachers. As Widdowson (2004, p. 359) puts it whilst "the case for 'real' English" is in itself very appealing to teachers, it is often proposed "**on the authority of the linguists**". As a result, the so-called 'corpus revolution' (RUNDELL; STOCK, 1992) may either not have reached the teachers yet, or may be intentionally rejected by them. In the unintentional case, teachers are often not aware of the possibilities offered by corpora or not aware of the changes that corpus methods have brought to materials that they are sometimes actually using. This may be due to a lack of information at the pre- or in-service teacher training levels and/or to the sometimes too vague statements found in the introductions to teaching materials, which might refer to the corpus-informed nature of the materials but not explicitly list the pedagogical implications of this corpus-based nature. As for the intentional negation of the benefits of corpus research in L2L, it may be caused by the oft-cited 'ivory tower effect', i.e. the perception by teachers that linguists work in their offices at university and have no idea of what teaching is about - and this despite the fact that some of those linguists are also teachers. This feeling of distance is usually reinforced by the fact that the types of examples or applications provided in the literature are often meant for EAP/ESP audiences. The collection of corpora (be they native or learner corpora) is usually coordinated by university teams and a vast majority of applied uses of corpora are found at university level (see for instance FEAK; SWALES, 2010; JONES; SCHMITT, 2010). This focus on advanced and specialized levels of proficiency does not always facilitate a transposition to learners with less advanced, non-academic needs.

Another reason explaining the lack of uptake of corpus methods in instructional settings (and one that is especially worth taking into account by corpus linguists advocating the applied relevance of their research whilst not being directly involved in teaching) is that the importance of using **authentic, corpus-based descriptions of the target language** is **only one line of thinking among many other influential ones** in L2 language learning. Socio-politically oriented considerations have led some to reject the promotion of standard native speaker usage as the norm for L2L as a manifestation of linguistic imperialism (PHILLIPSON, 1992) that must be abolished. This corresponds to what Davies (1996) and Widdowson (2004) call the 'conspiracy' view.

More practical and methodologically oriented views have however also been put forward against the use of a native speaker model in L2L: it might set goals that are unachievable, unrealistic and unnecessary for the actual needs of the learners, and might not be appropriate to the socio-cultural conventions of the groups acquiring the L2.

Another key issue that has attracted controversy in L2 language learning and teaching circles is the **issue of frequency** which is unmistakably entwined with corpus linguistic research. As Gries (forthcoming) explains, many linguistic fields have witnessed "a development towards more rigorous data analysis: statistical analysis of various levels of complexity have become a mainstream component of linguistic analysis". This quantitative/statistical development can be considered as most welcome as it definitely promotes objectivity in research. Two main uses of frequency can be distinguished in L2L: a) the role of input frequency on L2 acquisition and b) the use of frequency information to help select which aspects of language (be they words, expressions, grammatical structures, errors, etc.) deserve more attention form the part of the learners and/or teachers. Whilst the role of frequency effects in SLA has clearly been demonstrated (see for instance ELLIS, N., 2002a; 2002b; SIYANOVA; SCHMITT, 2008; COLLINS; ELLIS N., 2009; ELLIS, N.; LARSEN-FREEMAN, 2009), the picture is perhaps less clear when it comes to potential applications to the teaching of foreign languages. Leech (forthcoming, 2011) argues that when applied to teaching, the frequency principle is often interpreted as 'more frequent = more useful to teach'. Frequency lists are certainly not unknown to teachers and many are familiar with the notion of threshold levels (see van EK; ALEXANDER, 1980), that of vocabulary size needed to read and understand unsimplified texts (see HIRSH; NATION, 1992), or also the existence of an academic word list (COXHEAD, 2000).

Some teachers also use existing web tools that include vocabulary frequency lists for teaching and learning purposes, together with a variety of corpus tools (see for instance the Lexical Tutor, <http://www.lextutor.ca/>, maintained by Cobb at the Université du Québec à Montréal). The existence of frequency lists and corpus tools that can help access frequency information should however not be considered as an end in itself, and, whilst stressing the important role of frequency in L2L, Leech (forthcoming, 2011) nonetheless warns against a naïve interpretation of the frequency principle when it comes to teaching (see section 2.3 for further comments).

A last factor accounting for the lukewarm reception of corpora in the classrooms is the **lack of empirical studies exploring the actual impact of corpus methods on the learning outcomes**. The results of the few studies available (see for instance YOON, 2005, VANNESTAL; LINDQUIST, 2007; BELZ; VYATKINA, 2008; BOULTON, 2009; BREYER, 2009) present a contrasted picture and show that using corpora with students may require substantial support in some cases, that it takes time and practice to help students become independent users, that it does not appeal to all the students, and that it may prove beneficial for some skills and tasks but not for others. Yoon's study (2005) shows for instance that the use of corpora in writing classes provides students with common usage and collocation patterns that can be recycled immediately in their own writing, helps them develop longer-term cognitive skills (such as a greater awareness for lexico-grammatical aspects), and promotes independent learning. Vannestal and Lindquist (2007) find similar results but add that weak students find corpus consultation difficult or boring, and that some students do not find corpora very useful to help them improve their grammatical knowledge of the target language. All the studies listed above also stress the fact that teachers who want to use corpora with their students need to have a good understanding of the multi-faceted aspects of corpus literacy if they want the experiment to be successful. More research on the impact and learning outcomes of corpus methods is definitely in order to provide clearer evidence on the types of tasks and skills that would benefit most from a corpus approach.

Whilst far from being exhaustive,[5] the list of arguments provided in the preceding paragraphs nevertheless sheds some light on the reasons that have

[5] The access to well-equipped computer rooms with up-to-date software and hardware has for instance not been mentioned.

put a break on the expansion, integration, acceptance and understanding of native corpus research in educational settings.

### 1.2 Learner corpus research and L2 language learning

The fact that L2L has remained largely unaffected by the advances of native corpus research is even truer when one looks at the L2 teaching/learning applications of learner corpus research. Learner corpora are sometimes described as the 'missing link' in (EAP) pedagogy (see GILQUIN *et al.*, 2007) and they provide one ideal type of data to help linguists and teachers attest actual learners' needs on the basis of a careful analysis of their productions. Yet, despite the potential of learner corpora, Granger (2009, p. 24) provides a critical evaluation of their actual contribution to SLA and foreign language teaching and writes that "there is undeniably very little evidence of fully-fledged up-and-running applications".

It would be wrong to state that learner corpora have not been used by ELT publishers, but it must be acknowledged that they have been used to a much smaller extent than native corpora. When publishers refer to learner corpora, they seem to privilege in-house learner corpora.[6] Another problem is that, as was the case for the use of native corpora (see section 1.1), the exact use that is being made of the learner corpus is not always clearly documented.[7] This leads to the sometimes surprising inclusions of what Granger (2010, p. 32) describes as error notes "apparently based on learner corpora". She gives the examples of the use of *attend* instead of *wait for* (*Her mother was attending her outside the car) or of *piece* instead of *room* (*There are en suite bathrooms in every piece), examples not even attested once in the International Corpus of Learner English (GRANGER *et al.*, 2009), a learner corpus containing 3.5 million words produced by over 6,000 learners from 16 different mother tongue backgrounds.

---

[6] See for instance CUP and the Cambridge Learner Corpus at <http://ww.cambridge.org/elt/corpus/learner_corpus2.htm> or Longman and the Longman Learner Corpus at <http://www.longman.com/dictionaries/corpus/learners.html>.

[7] De Cock et al. (2007)'s section *Improving your writing skills*, which is included in the CD-ROM version of the Macmillan English Dictionary for Advanced Learners, is well-documented and constitutes a welcome exception to the publishers' use of in-house learner corpora.

Another reason accounting for the limited impact of learner corpora in instructional settings is put forward by Flowerdew, who states that in most studies of learner corpora "the implications for pedagogy are not developed in any great detail with the consequences that the findings have had little influence on [...] syllabus and materials design" (1998, p. 550). There is an urgent need to go beyond the usual last paragraph of articles (or last slide of conference presentations) stating that 'foreign language instruction could profit from this kind of investigation' and efforts should be made towards providing teachers with ready-to-use teaching materials, or at least free access, user-friendly and ready-to-use platforms which they could use to collect, analyse and exploit learner corpora on a more regular basis.

I have also recently argued (MEUNIER, 2010) that an additional reason accounting for the lack of direct influence of learner corpus studies on L2 syllabuses and materials[8] is that the topics covered in most existing learner corpora are often miles away from the everyday needs of a vast majority of L2 school teachers who target the L2 for general purposes, often for a teenage audience. Finding a learner corpus that meets their needs comes close to looking for a needle in a haystack. The Common European Framework of Reference for Languages (CEFR, Council of Europe 2001, p. 52) suggests that the following thematic categories should be addressed in L2 for General Purposes (EGP): personal identification, house and home, environment, daily life, free time, entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, and weather. Few, if any, native or learner corpus studies provide easily transferrable research results which could be integrated in a syllabus addressing these themes. Corpus compilers will urgently have to address the learner's needs for what Braun (2005) calls 'pedagogical relevance' or what Belz and Viyatkina (2008) call 'authentication'.

## 2. Corpus linguistic research meets second/foreign language learning: exploring multiple paths for a balanced integration

A multiplicity of paths will have to be explored if a fuller integration between the two domains is to be achieved. Various – and sometimes even

[8] This does not apply to language for academic/specific purposes – domains where teachers do actually use native and advanced learner corpora (see for instance Flowerdew, 2003; Gilquin *et al.*, 2007, or PAQUOT, 2008).

radically opposite – directions are proposed in the coming sections. I believe that this multi-directionality is necessary to promote a healthy cross-fertilization between corpus linguistics and the current multi-faceted language learning cultures. Multi-directionality is also the only way to help teachers and material designers cater for the particular needs of specific learning populations in no less specific socio-cultural contexts.

### 2.1 Go global: expand your horizons

The call for more **cross-disciplinarity** between the various research paradigms involved in L2L has repeatedly been made. Recent calls include Sorace (2010), who states that a lot of relevant SLA research is done in other fields and ignored by SLA researchers. Similarly, Norris (2010), a famous proponent of task-based language learning,[9] stresses the importance of understanding instructed SLA (ISLA) and of taking into account the needs of the teachers and learners in classrooms. In Granger and Meunier (2010), we plead for a closer integration between SLA and learner corpus research (LCR) and show that SLA studies can greatly benefit from the solid empirical base provided by learner corpora research tools and methods, whilst LCR needs a more solid grounding in SLA theory.

Whilst it is obviously impossible to acquire unlimited expertise (foreign/second language teacher, SLA researcher, corpus linguist, sociolinguist, computer programmer, statistician, etc.) it is nevertheless essential to make a conscious effort to be open to other academic cultures and working environments and hence, to leave one's comfort zone. An implicit corollary of cross-fertilization is that each field should make yet another conscious effort to highlight the convergences between its own domain and other related domains. This, in turn, implies a certain degree of elaboration, specification and sometimes even simplification. Some researchers set the example and provide clear introductions to their research area. Römer and Wulff (2010), in a paper entitled *Applying corpus methods to written academic texts: Explorations of MICUSP*, provide a most welcome step-by-step introduction to the central techniques in corpus analysis intended for students and/or corpus

[9] Task-based learning involves goal-oriented communicative activities, with a specific outcome, where the emphasis is on exchanging meanings and not on producing specific language forms (see WILLIS, 1996).

novices. This type of publication can only be encouraged as it corresponds to what Römer calls the missionary work of corpus linguists, i.e. the need to convince teachers, students, materials writers, and syllabus designers that corpora can be of great use in their everyday work (2006: 128). Similar missionary work from other fields must also be encouraged.

One welcome development illustrating the benefits of cross-fertilization is the increasing use of **triangulation methods** in L2L studies. Triangulation can for instance be obtained by combining corpus and experimental data (SIYANOVA; SCHMITT, 2008), or by revisiting/replicating earlier SLA studies on (more) learner corpus data (as exemplified by HOUSEN, 2002, who revisited the verb morphemes study initially carried by DULAY *et al.*, 1982, or by WULFF *et al.*, 2009, who reinterpreted tense-aspect studies conducted by BARDOVI-HARLIG, 1998; 2002). Also important among triangulation methods is what Ellis, Simpson-Vlach & Maynard (2008) call the validation of the instructional value, i.e. the assessment by experienced language instructors and testers of the teaching-worthiness of the linguistic output obtained thanks to corpus metrics.

The challenge of going global and expanding the horizons can also be taken up within the corpus linguistics field. **More (learner) corpora** should be collected to represent:

- **more languages**, to counterbalance the predominance of anglo-saxon native and learner corpora and to foster the computer-aided analysis of different languages and language families,
- **more communicative modes**: spoken corpora, interactional corpora (classroom interactions, authentic interactions representing what Wagner (2010) calls 'language learning in the wild', multimodal corpora, corpora of textbook materials, etc.,
- **more text types and genres**, to cover text types which are less represented in corpora to date (letters, emails, twits, leaflets, TV programmes, book synopses, recipes, short notes, chat room logs, etc.),
- **more longitudinal language data**, from beginners to advanced levels, from children to adults, from L1 to L2s, but also attrited language and language impaired data,
- **more variables**: more language learning variables should be collected and encoded at the time of corpus collection (proficiency, language aptitude, motivation, more precise description of the task, of temporal, social or situational settings, etc.).

Adopting a cross-disciplinary perspective, and making a disciplinary effort to expand data types and quality within corpus linguistics, will lead to a better understanding of the processes at play in L2 learning and acquisition. It will also make it possible for more social, cultural and situational variables to be taken into account in instructed environments.

### 2.2 Go local: create a sense of community

The call for going global is mainly meant for researchers and teachers. If one adopts a more **learner-centred** view, a reverse call - viz. going local - is probably more appropriate.

One way of encouraging learners to use corpora is to enhance the **pedagogical relevance** (BRAUN, 2005) and **authentication** (BELZ; VIYATKINA, 2008) of corpus use (see section 1.2). In *From Corpus to Classroom*, O'Keeffe *et al*. (2007, p. xi) mention the "frequent mismatch between corpus linguistic research and what goes on into materials and resources, and what goes on in the language classroom". It is actually fair to argue that corpora will only be used by language learners if they can interpret, analyse and understand them in a personally meaningful way (BELZ; VIYATKINA, 2008). A direct involvement of learners in corpus collection and use corresponds to what Granger (2009, p. 25) calls corpora for immediate pedagogical use (IPU), i.e. data "collected by teachers as part of their normal classroom activities [...] [and where] the learners are at the same time producers and users of the corpus data". Examples of such IPU include telecollaborative interactions (oral or written) during which learners build personal relationships with other speakers. Those other speakers can be native speakers (as in BELZ; VIYATKINA's, 2008 study) but they can also be other non-native speakers. Once the oral and written interactions are archived they can be accessed and explored to serve as a basis for pedagogical interventions. Teachers can focus on specific linguistic forms produced by the learners themselves, in the context of **meaningful interactions** (see KASPER; ROSE, 2002) in **communicative tasks**. Learners are then more likely to feel a sense of authentication and pedagogical relevance.

Braun (2006)'s project provides another good illustration of pedagogical relevance; she uses a small English Interview corpus (ELISA) containing 26 interviews of approximately 10 minutes each, for a total of about 60,000 words. Despite its limited size in words, the corpus covers a variety of communicatively relevant topics from the broad area of professional, social and cultural life. Braun (2006) also argues that homogeneity and **topical**

**relevance** are more important than representativeness in the traditional sense, and that learners and teachers are more likely to adopt a more **qualitative approach** to corpus analysis as it is more appropriate and manageable for them.

Teachers who decide to join the corpus bandwagon will definitely have to go '**beyond the pen and paper**' (WIBLE, 2008) but this does not necessarily imply a technological big bang. Some teachers will no doubt be better served than others (well-equipped computer labs, school technicians, projects carried out on a large scale) but it has been shown that corpus collection and use can be started on a smaller scale. Illustrations of larger-scale projects, which nonetheless promote a sense of community, are presented in Wible *et al.* (2001) and Simpson *et al.* (2002). Wible *et al.* explain how learner corpora can be collected and annotated by teachers, and subsequently be used for further pedagogical exploitation. They have set up an interactive online environment in which essays written by learners, together with the comments provided by teachers are archived in a searchable online database. Corpus collection and error annotation are **integrated in the normal teaching activities**: the student composes an essay offline and hands it in to the teacher over the Internet; the teacher marks the essay and sends it back to the student who revises his/her essay. During the correction phase, teachers insert comments in the student's essay by keying in a comment or by choosing a comment from an already existing 'Comment Bank' (e.g. '*wrong verb*', '*wrong tense*', etc.). Learners can get a cumulative comment list which will give them an idea of their most error-prone patterns; teachers can use these lists to design exercises that target learners' most frequent errors. In addition, the online corpus can be searched to get more instances of error-prone patterns. The human and computing resources required in that project are impressive (large number of teachers and learners taking part in the project all over Taiwan). As for the MICASE project (Michigan Corpus of Spoken Academic English), presented in Simpson *et al.* (2002), it consists of a collection of nearly 1.8 million words of spoken academic English recorded on the University of Michigan campus, and transcribed into searchable documents. MICASE can freely be browsed and searched online, notably to find recurrent grammatical and phraseological patterns or track generalized changes in speech patterns as people gain experience of university culture and academic speech.

Regardless of the size of the corpus collected, such projects subscribe to the learning-driven data methodology advocated by Seidlhofer (2002) by promoting a **learner-centred**, **context-dependent and culture-bound approach**. The

fact that learners analyse their own productions favours the individualization of learning (and teaching), and helps learners monitor their own production and the effects of their production on others (MEUNIER, 2010).

To achieve the sense of community mentioned in the heading of the present section it is also essential to promote **care and attentiveness** in the use of corpora, both from a teacher- and learner- perspective. This care and attentiveness can be offered to teachers thanks to projects such as the Web 2.0 ERC (see <http://www.web20erc.eu/>), a new European Union funded education project to help educators who find ICT confusing and difficult to access. As for learners, Vannestal and Lindquist (2007) insist that using corpora with students requires time and a large amount of introduction and support, an issue that should certainly not be neglected when teachers opt for a corpus approach.

Going local will undoubtedly help promote corpus literacy as a useful tool for the empowerment of learning and teaching communities alike.

### 2.3. Let computer technology, frequencies and figures help and inform you; do not let them dictate

It would be most unreasonable to minimize the impact of technology without which frequency lists, specifications of textual features across languages, text types and genres, pattern grammar (the corpus-driven approach to the lexical grammar of English), collostructions (degree of attraction or repulsion between words and constructions), word sketches[10] *(*summaries of a word's lexical and grammatical collocational behaviour), or data-driven learning activities would still be unknown to date. The corpus revolution would have been impossible without the exponential increase in computing power, storage capacities, and programming and analysis skills of competent language experts. Jarvis (forthcoming) states that an important characteristic of (learner) corpus analysis is its heavy reliance on computer automation for purposes of discovering patterns in the data. Because of the size and complexity of most language corpora, it would be infeasible to perform comprehensive analyses of the data without computer automation. Gries *et al.* (2010, p. 4) also convincingly argue that "maybe the most dramatic changes that the field of corpus linguistics is witnessing these days concerns its methodologies… [and that] the field of corpus linguistics is rapidly being enriched with methodological

[10] For illustrations of word sketches, see <http://www.webdante.net/the_project.html>

expertise borrowed from other fields such as statistics, computational linguistics, and even artificial intelligence."

This said, the fact that sophisticated multi-factorial analyses are needed to refine the results of corpus studies should not divert our attention away from other key issues in language learning. The validation of the instructional value of the results of corpus studies, together with the pedagogical relevance of tools and methods used, have already been expanded on in sections 2.1 and 2.2 respectively. In the present section, I would like to come back to the frequency issue mentioned in section 1.1., and particularly to the 'more frequent = more useful to teach' approach.

A first warning must be made against an exclusive instructional focus on the more frequent lexical items in vocabulary lists, be they single words or multiword units. Vocabulary acquisition studies have demonstrated that higher **proficiency levels** correlate with the knowledge of less frequent words together with the knowledge of phraseological (and less common) uses of frequent words. It is therefore essential to gradually cover the whole frequency spectrum and even to come back to very frequent items in more advanced stages of acquisition in order to cover their phraseological and less common uses.

A slightly different perspective could probably be adopted for grammatical and syntactic patterns. While presenting highly infrequent structures to learners for **receptive** purposes makes sense, it is much less sensible to prompt learners to use these structures in **productive** tasks.

Whilst access to frequencies (in all its guises) is per se a very good thing, Leech (forthcoming, 2011) also warns that frequency counts are *least* useful when they are based on a general corpus covering the range of the language and are *more* useful if they are more specific, i.e. differentiated for mode, register, text type or region. A desirable evolution in corpus linguistics would then be to provide teachers and learners with more specific lists in line with teachers' and learners' communicative needs. On a more negative note, however, teachers and learners alike might wonder where to draw the line. The tensions between the precision and accuracy of the descriptions provided by corpus specialists can be perceived by teachers (and learners) as 'too much of a good thing', as shown by Coxhead (2008) who analysed the learners' negative perceptions of the importance of learning multi-word units when one word does the 'communication' trick.

In sum, I would recommend a flexible, teacher-validated and informed use of frequency lists. Teachers and learners would be wrong to do without or

to ignore the information contained in frequency lists, but they would be equally wrong to abide by them dogmatically.

## 3. Concluding remarks

The influence of native and learner corpus research on second/foreign language learning has been discussed in section 1. There is no denying that a perceptible divide exists between the numerous publications in applied corpus research and the actual use of corpus data in instructional settings. Rather than sticking to that rather pessimistic conclusion, I have expanded on four possible reasons which may explain why instructional settings have tended to shy away from corpus use. Acknowledging these issues and actually addressing them is a vital step in promoting corpus uptake.

In section 2, I have put forward some suggestions to foster a healthy cross-fertilization between corpus linguistics and the current multi-faceted language learning and teaching cultures. I mentioned the importance of going global and expanding our horizons by encouraging cross-disciplinarity, promoting the use of triangulation methods in L2 studies, and further refining the learner, task and situational variables in the compilation of new types of corpora. I also suggested an opposite trend, which consists in going local and creating a sense of community. Taking a successful digital turn requires pedagogical relevance and authentication. If corpus methods are to be integrated in normal teaching activities they must be learner-centred, context-dependent and culture-bound. Time, care and attentiveness are also essential when promoting corpus literacy as empowerment tools for learners and teachers. The third line of discussion was devoted to frequencies. I have highlighted their overall importance in corpus studies but have nevertheless suggested the need for a flexible, teacher-validated and informed use of frequencies for pedagogical purposes.

## References


BARDOVI-HARLIG, K. Narrative structure and lexical aspect: Conspiring factors in second language acquisition of tense-aspect morphology. *Studies in Second Language Acquisition*, 20, p. 471-508, 1998.

BARDOVI-HARLIG, K. Analyzing aspect. In: SALABERRY, R.; SHIRAI, Y. (Ed.). *Tense-aspect morphology in L2 acquisition.* Amsterdam: John Benjamins, 2002.

BELZ, J.; VYATKINA, N. The Pedagogical Mediation of a Developmental Learner Corpus for Classroom-Based Language Instruction. *Language Learning & Technology*, v. 12, n. 3, p. 33-52, 2008.

BOULTON, A. Testing the Limits of Data-Driven Learning: Language Proficiency and Training. *ReCALL*, v. 21, n. 1, p. 37-54, 2009.

BRAUN, S. From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, v. 17, n. 1, p. 47-64, 2005.

BRAUN, S. ELISA - a Pedagogically Enriched Corpus for Language Learning Purposes. In: BRAUN, S.; KOHN, K.; MUKHERJEE, J. (Ed.). *Corpus Technology and Language Pedagogy*: New Resources, New Tools, New Methods. Frankfurt: Peter Lang, 2006.

BREYER, Y. Learning and Teaching with Corpora: Reflections by Student Teachers. *Computer Assisted Language Learning*, v. 22, n. 2, p. 153-172, 2009.

COLLINS, L. ; ELLIS, N. C. (Ed.). Input and second language construction learning: frequency, form, and function. Special issue. *Modern Language Journal*, v. 93, n. 3, 2009.

COXHEAD, A. A New Academic Word List. *TESOL Quarterly*, 34, v. 2, p. 213-238, 2000.

COXHEAD, A. Phraseology and English for academic purposes: Challenges and opportunities. In: MEUNIER, F.; GRANGER, S. (Ed.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: Benjamins, 2008.

DAVIES, A. Ironising the Myth of Linguicism. Review of Linguistic Imperialism, by Robert L.H. Phillipson. Oxford: Oxford University Press, 1992. *Journal of Multilingual and Multicultural Development*, v. 17, n. 6, p. 485-496, 1996.

DULAY, H.C.; BURT, M.; KRASHEN, S. *Language Two*. Rowley: Newbury House, 1982.

ELLIS, N. C. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, v. 24, n. 2, p. 249-60, 2002a.

ELLIS, N. C. Reflections on frequency effects in language processing. *Studies in Second Language Acquisition*, v. 24, p. 297-339, 2002b.

ELLIS, N. C.; LARSEN-FREEMAN, D. Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning*, v. 59, 2009. Supplement 1, p. 93-128.

ELLIS, N.C.; SIMPSON-VLACH, R.; MAYNARD, C. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, v. 42, p. 375-396, 2008.

ELLIS, R. *The Study of Second Language Acquisition*. Oxford: Oxford University Press, 1994.

ELLIS, R. *The Study of Second Language Acquisition.* 2. ed. Oxford: Oxford University Press, 2008.

FEAK, C.B.; SWALES, J. M. Writing for publication: Corpus-informed materials for post-doctoral fellows in perinatology. In: HARWOOD, N. (Ed.). *English Language Teaching Materials*. Theory and Practice. Cambridge: Cambridge University Press, 2010.

FLOWERDEW, L.J. Corpus Linguistic Techniques Applied to Textinguistic. *System*, v. 26, n. 1, p. 545-556, 1998.

GILQUIN, G.; GRANGER, G.; PAQUOT, M. Learner corpora: the missing link in EAP pedagogy. THOMPSON, P. (Ed.). *Corpus-based EAP Pedagogy*. *Special issue of the Journal of English for Academic Purposes*, v. 6, n. 4, p. 319-335. 2007.

GRANGER, S.; MEUNIER, F. SLA research and learner corpus research: Friend or foe? Paper presented at the *20th Annual Conference of the European Second Language Association*. Reggio Emilia, Italy, 1-4 September, 2010.

GRANGER, S. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In: AIJMER, K. (Ed.). *Corpora and Language Teaching.* Studies in Corpus Linguistics 33. Amsterdam: John Benjamins, 2009.

GRANGER, S. Vingt ans d'analyse de corpus d'apprenants : leçons apprises et perspectives. In : CAPPEAU, P. ; CHUQUET, H. ; VALETOPOULOS, F. (Ed.). *L'exemple et le corpus*. Quel statut? Travaux linguistiques du CerLiCo. Numéro 23. Presses Universitaires de Rennes, 2010.

GRANGER, S.; DAGNEAUX, E.; MEUNIER, F.; PAQUOT, M. *The International Corpus of Learner English*. Version 2. Handbook and CD-Rom, Louvain-la-Neuve: Presses Universitaires de Louvain, 2009.

GRIES, St. Th. Statistical tests for the analysis of learner corpus data. In : DIAZ-NEGRILLO, A.; THOMPSON, P.; BALLIER, N. (Ed.). *Multidisciplinary perspectives to learner corpora*. Amsterdam & Philadelphia: John Benjamins. Forthcoming.

GRIES, St. Th.; WULFF, S.; DAVIES, M. (Ed.) *Corpus-linguistic applications*. Current studies, new directions. Amsterdam/New York: Rodopi, 2010.

HIRSH, D.; NATION, P. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language,* v. 8, n. 2, p. 689-696, 1992.

HOUSEN, A. A corpus-based study of the L2-acquisition of the English verb system. In: GRANGER, S.; HUNG, J.; PETCH-TYSON, S. (Ed.). *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam: John Benjamins, 2002.

JARVIS, S. Data Mining with Learner Corpora: Choosing Classifiers for L1 Detection. In: MEUNIER, F.; DE COCK, S.; GILQUIN, G.; PAQUOT, M. (Ed.). *A Taste for Corpora*. In honour of Sylviane Granger. Amsterdam/ Philadelphia: Benjamins. Forthcoming 2011.

JONES, M.; SCHMITT, N. Developing materials for discipline-specific vocabulary and phrases in academic seminars. In: HARWOOD, N. (Ed.). *English Language Teaching Materials*. Theory and Practice. Cambridge: Cambridge University Press, 2010.

LEECH, G. Frequency, corpora and language learning. In: MEUNIER, F.; DE COCK, S.; GILQUIN, G.; PAQUOT, M. (Ed.). *A Taste for Corpora*. In honour of Sylviane Granger. Amsterdam/Philadelphia: Benjamins, forthcoming 2011.

MEUNIER, F.; GOUVERNEUR, C. New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material. In: AIJMER, K. (Ed.). *Corpora and Language Teaching*. Amsterdam & Philadelphia: Benjamins, 2009.

MEUNIER, F. Learner Corpora and English Language Teaching: Checkup Time. *Anglistik: International Journal of English Studies*, v. 21, n. 1, p. 209-220, 2010.

NORRIS, J. Understanding instructed SLA: Constructs, contexts, and consequences. Plenary address at the *20th A*nnual Conference of the European Second Language Association. Reggio Emilia, Italy, 1-4 September, 2010.

O'SULLIVAN, I. Enhancing a Process-Oriented Approach to Literacy and Language Learning: The Role of Corpus Consultation Literacy. *ReCALL*, v. 19, n. 3, p. 269-286, 2007.

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. *From Corpus to Classroom*. Language use and language teaching. Cambridge: Cambridge University Press, 2007.

PHILLIPSON, R. *Linguistic Imperialism*. Oxford: Oxford University Press, 1992.

RÖMER, U.; WULFF, S. Applying corpus methods to writing research: Explorations of MICUSP. *Journal of Writing Research*, v. 2, n. 2, p. 99-127, 2010.

RÖMER, U. Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments. GAST, V. (Ed.). The Scope and Limits of Corpus Linguistics – Empiricism in the Description and Analysis of English. Special Issue: *Zeitschrift für Anglistik und Amerikanistik*, v. 54, n. 2, p. 121-134, 2006.

RUNDELL, M.; STOCK, P. The Corpus Revolution. *English Today,* v. 8, p. 9-14, 1992.

SEIDLHOFER, B. Pedagogy and local learner corpora: Working with learning-driven data. In: GRANGER, S.; HUNG, J.; PETCH-TYSON, S. (Ed.). *Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching*. Amsterdam: John Benjamins, 2002.

SIMPSON, R. C.; BRIGGS, S. L.; OVENS, J.; SWALES, J. M. *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan, 2002.

SIYANOVA, A.; SCHMITT, N. L2 Learner Production and Processing of Collocation: A Multi-study Perspective. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, v. 64, n. 3, p. 429-458, 2008.

SORACE, A. SLA as bilingualism: Or, it's time to see the forest for the trees. Plenary address at the *20th Annual Conference of the European Second Language Association.* Reggio Emilia, Italy, 1-4 September, 2010.

VAN EK, J.A.; ALEXANDER, L.G. *Threshold Level English*. Oxford: Pergamon, 1980.

VANNESTAL, M.E.; LINDQUIST, H. *Learning English Grammar with a Corpus*: Experimenting with Concordancing in a University Grammar Course. ReCALL, v. 19, n. 3, p. 329-350, 2007.

WIBLE, D. Multiword expressions and the digital turn. In: MEUNIER, F.; GRANGER, S. (Ed.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: Benjamins, 2008.

WIBLE, D.; KUO, C. W.; CHIEN, F., LIU, A.; TSAO, N. L. A webbased EFL writing environment: integrating information for learners, teachers, and researchers. *Computers and Education*, v. 37, p. 297-315, 2001.

WIDDOWSON, H. A perspective on recent trends. In: HOWATT, A. P. R.; WIDDOWSON, H. (Ed.). *A History of English Language Teaching*. Second edition. Oxford: Oxford University Press, 2004.

WILLIS, J. *A Framework for Task-Based Learning*. Longman. 1996.

WULFF, S.; ELLIS, N.C.; RÖMER, U.; BARDOVI-HARLIG, K.; LEBLANC, C.J. The Acquisition of Tense–Aspect: Converging Evidence From Corpora and Telicity Ratings. *The Modern Language Journal*, v. 93, n. 3, p. 354-369, 2009.

YOON, H. *An investigation of students' experiences with corpus technology in second language academic writing dissertation*. 2005. PhD (Dissertation, Degree Doctor of Philosophy) – Graduate School of The Ohio State University, 2005. Available at: <http://etd.ohiolink.edu/send-pdf.cgi/Yoon%20Hyunsook.pdf? osu1109806353>

Recebido em 22/09/2010. Aprovado em 08/05/2011.