

Language for Specific Purposes Learner Corpora

Sylviane Granger & Magali Paquot

[DRAFT VERSION]

Learner corpora are electronic collections of (near-)natural foreign or second language learner (L2) language assembled according to explicit design criteria (Granger, 2008). So far, they have predominantly been collected in the context of foreign/second language courses for general purposes (e.g. the *International Corpus of Learner English*). A very promising but as yet less widespread data type is the Language for Specific Purposes (LSP) learner corpus, which contains discipline and genre-specific texts written by learners within the framework of LSP or content courses. These texts are exemplars of the different genres (e.g. papers, lab reports, case studies, project reports or letters) that are most typical of specific disciplinary communities (e.g. physics, linguistics, history, engineering, medicine or business). The aim of LSP learner corpus research is to uncover L2 learners' difficulties in these specific domains and design tailor-made learning and teaching methods and tools. This type of corpus has connections with English as a Lingua Franca (ELF) corpora but differs from them in that the subjects are language learners of varying degrees of proficiency in educational settings, while the ELF subjects are characterized primarily as users of the language involved in real life communication events. It also has similarities with corpora of proficient student writing such as the *Michigan Corpus of Upper-Level Student Papers* (MICUSP) or the *British Academic Written English* (BAWE) corpus, which contain data from native and international students in mixed naturalistic and instructional environments.

In view of the status of English as the dominant language for knowledge dissemination and international communication, LSP learner corpus research has mainly centred on *English* for Specific Purposes (ESP), with a large number of ESP learner corpus compilations, especially in Asia. In spite of this vitality, it is not easy to survey the field as research is scattered and tends to be rather confidential, a situation recognized by Alsop and Nesi (2009, p. 75), who state that discipline-specific student writing “has tended to be collected for individual scholarly purposes rather than as part of formal corpus-building projects”. They tend to be “small private collections of student assignments” (ibid.) that are not widely available. One of the consequences of the localized nature of LSP learner corpora is that they are usually L1-specific, i.e. contain data from learners sharing a single mother tongue background (e.g. Chinese for Lee & Chen, 2009). A few ESP learner corpora include texts produced by international students from several mother tongue backgrounds collected in a specific institution but as the samples are aggregated, it is not possible to search the L1 populations (e.g. Hewings & Hewings, 2002). A notable exception is the *Indiana Business Learner Corpus* which is made up of application letters from native and non-native speakers of English studying in three different undergraduate business classes in Belgium, Finland and the United States (Connor et al., 2002).

There is clear scope for the collection and dissemination of multi-L1, multi-discipline and multi-genre corpora of LSP learner language. A step in this direction is the *Varieties of English for Specific Purposes dAtabase* (VESPA) learner corpus which contains ESP texts produced by English as a Foreign Language (EFL) writers from various mother tongue

backgrounds in a wide range of disciplines (linguistics, business, engineering, experimental sciences, environmental science, etc), genres (research papers, project reports, MA dissertations) and degrees of writer expertise in academic settings (from first-year students to PhD students). The database is searchable via a range of variables relating to learner (e.g. age, gender, educational background, country) and task (e.g. genre, type of reference tools used). Two online LSP learner corpora are also worth mentioning here: the *Learner Corpus of Essays and Reports* and the *Learner Corpus of English for Business Communication*. These two corpora are searchable online via the PolyU Language Bank platform (<http://langbank.engl.polyu.edu.hk/index1.html>), but their usefulness for research is somewhat limited by the fact that they do not allow for genre- or discipline-specific searches.

Another type of corpus relevant for LSP research is the translation learner corpus, which contains translations that have been produced by trainee translators. While translation learner corpora such as the *Student Translation Archive* (Bowker, 2002) and the *MeLLANGE Learner Translator Corpus* (Castagnoli et al., in press) mainly consist of translations into the first language, the *PELCRA* learner translation corpus (Uzar & Waliński, 2001) also includes trainee translations from Polish to L2 English.

Unsurprisingly perhaps, there are many more written than spoken LSP learner corpora. The potential of the latter is however demonstrated by the *International Teaching Assistants corpus* (ITAcorp) which contains samples of the language production of international graduate students in ITA language classes as they engage in a number of role-play situations designed to mirror university instructional contexts, e.g. office hour role plays, brief lecture presentations and discussion leadings (Thorne et al., 2008). It was built to identify aspects of teachers' professional discourse that are problematic for ITAs and help them negotiate the many spoken genres that are typical of instructional activity. The design of spoken and multimodal LSP corpora is clearly one of the key desiderata for the future.

Many studies of LSP learner corpora rely on quantitative and qualitative comparisons of learner and expert language to uncover the distinctive features of learner language. For example, Flowerdew (1998) compared the use of cause and effect markers in a corpus of 80 student assignments which discussed various environmental problems and an expert corpus of edited papers, viz. *Global Warming: The Greenpeace Report*. Among other things, she found that prepositions were commonly used to mark causality in the expert corpus, but occurred far less frequently in the learner corpus and that, unlike expert writers, Hong Kong EFL learners rarely used modifiers with causative verbs and complex prepositions to attenuate the proposition. Comparisons of learner and native or expert language constitute one branch of Contrastive Interlanguage Analysis (CIA) (Granger, 2008), a methodology that has been widely used to analyze learner corpora (e.g. Nesselhauf, 2005; Paquot, 2010). The other branch of CIA, which consists in comparing different varieties of interlanguage, has been less popular in LSP learner corpus research and has tended to focus on cross-discipline rather than cross-L1 comparisons (see Upton & Connor, 2001 and Connor et al., 2002 for exceptions). Hyland (2004), for example, compared the use of metadiscourse markers in dissertations by L2 postgraduate writers in six academic disciplines (electronic engineering, computer science, business studies, biology, applied linguistics and public administration) and found substantial variations in the ways L2 writers present their research and engage with their readers across disciplinary communities. Hyland interpreted these differences in terms of the norms and expectations of particular cultural and professional communities and showed, for example, that the high density of citations found in biology dissertations is also a feature of research articles in biology. Cross-L1 comparisons, however, constitute a major avenue for future

research as they make it possible to distinguish the difficulties that are proper to one L1 population from those that cut across L1 backgrounds.

The linguistic analysis of LSP learner corpora relies on genre analysis and/or discourse analysis, often used in combination with corpus-linguistic techniques. Upton & Connor (2001), for example, have adopted the Swalesian tradition of genre (Swales, 1990) and combined it with a computerized analysis of lexico-grammatical features of texts to analyze politeness strategies in application letters written by Belgian, Finnish and American students. They carried out a hand-tagged moves analysis, classified segments of text according to their communicative purpose and examined the linguistic features related to the concept of politeness (e.g. modals and formulaic expressions). Differences in the use of politeness strategies across countries were partly attributed to differences in audience expectations and writer concepts of how politeness is expressed. Other researchers have adopted a more discourse or functional perspective. Hewings & Hewings (2002) compared the use of clauses with an anticipatory *it* and extraposed subject in an expert corpus of research articles and a corpus of dissertations written by non-native speakers as the last component of their MBA programmes. Student writers appeared to use more *it*-clause emphatics and attitude markers to persuade readers of the validity of claims. Similarly, Flowerdew (2008) investigated the lexico-grammatical patterning of key words for the problem-solution pattern in a corpus of group project reports written by EFL undergraduate Science and Engineering students on a Technical Communication Skills course in the language centre at the Hong Kong University of Science and Technology. She found that students used a very narrow range of verbs as well as register-inappropriate verbs (e.g. *get rid of*) and confused cause/effect with result/effect verbs. She attributed these difficulties to “the lack of an appropriate grammar system and vocabulary range which play a role in the co-construction of meaning through the blending of collocational and colligational features of language” (ibid., p. 133) in academic discourse. The influence of a Sinclairian approach to corpus-driven lexico-grammatical patterns (Sinclair, 2004) is apparent in a large number of studies. Lee & Chen (2009), for example, relied on corpus linguistic tools and techniques to identify keywords in the *Chinese Academic Written English* (CAWE) corpus, a corpus of dissertations written by Chinese undergraduates majoring in English linguistics or applied linguistics. The keyword analysis made it possible to identify salient features in the Chinese learner corpus - the high frequency of the verb *make* and the function words *can*, *besides* and *according to* - which were then submitted to a careful lexico-grammatical analysis. The high frequency of the modal *can*, for example, appeared to be largely due to Chinese learners’ use of phrases such as *we can see*, *we can observe*, and *we can find* to refer to tables and figures, a pattern that is quite rare in expert and native student writing corpora where phrases such as ‘*Table 4 displays*’, ‘*are illustrated in Figure 1*’ or ‘*As shown in Table 2*’ are preferred. A similar approach, adopted by Gilquin et al (2007), helped uncover a wide variety of lexico-grammatical difficulties in English for Academic Purposes (EAP) writing, some proper to foreign language learners, others shared by novice native writers.

Although research in the field is both active and varied, it has not yet resulted in off-the-shelf LSP learner-corpus-informed pedagogical materials. This should not be taken to mean that the research has remained purely theoretical or descriptive. As Thompson (2006, p. 7) aptly reminds us, “A feature of ESP work is that the needs of learners are often highly specific and teachers may have to develop materials and resources for the specific context”. In many cases, the LSP learner corpus analyst is also the corpus compiler, a situation which offers the significant advantage of enabling the teacher to authenticate the data for classroom use to fit the students’ reality (Flowerdew, 2008, p. 134). Hewings (2000) illustrates how teachers can

use learner texts collected in their ESP courses to identify common areas of difficulty for a particular group of students and inform classroom practice. In the same vein, Lee & Swales (2006) report on an experimental corpus-based course for international doctoral students in which they made students compile corpora of their own writing for comparison with expert writing in the discipline. For less localized pedagogical resources, the future clearly lies in electronic resources which can be customized to learners' L1s and disciplines. The Louvain EAP dictionary-cum-writing aid, a web-based tool, demonstrates how insights gained from the analysis of both learner and expert corpora can help L2 writers from several L1s express key academic functions such as contrasting, introducing a topic, quoting, etc. in a range of disciplines (Granger & Paquot, 2010).

SEE ALSO: Learner corpora; Corpus Analysis of Written English for Academic Purposes; Applied corpus linguistics; Corpora in the language teaching classroom; Genre and Discourse Analysis in Language for Specific Purposes

References

- Alsop, S. & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1), 71-83.
- Bowker, L. (2002). Corpus-based applications for translator training: Exploring the possibilities. In S. Granger, J. Lerot & S. Petch-Tyson (Eds.) *Corpus-based Approaches to Contrastive Linguistics and Translation Studies* (pp. 169-183). Amsterdam / New-York: Rodopi.
- Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N. & Volanschi, A. (in press). Designing a learner translator corpus for training purposes. In N. Kübler (Ed.) *Corpora, Language, Teaching and Resources: From Theory to Practice*. Peter Lang.
- Connor, U., Petch, K. & Upton, T. (2002). Business English: Learner data from Belgium, Finland and the U.S. In S. Granger, J. Hung & S. Petch-Tyson (Eds.) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. (pp. 175-194). Amsterdam/Philadelphia: John Benjamins.
- Flowerdew, L. (1998). Integrating 'expert' and 'interlanguage' computer corpora findings on causality: discoveries for teachers and students. *English for Specific Purposes*, 17(4), 329-345.
- Flowerdew, L. (2008). *Corpus-based Analyses of the Problem-Solution Pattern*. Amsterdam / Philadelphia: John Benjamins.
- Gilquin G., Granger S. & Paquot M. (2007) Learner corpora: the missing link in EAP pedagogy. In Thompson, P. (ed.) *Corpus-based EAP Pedagogy*. Special issue of *Journal of English for Academic Purposes* 6(4), 319-335.
- Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook. Volume 1* (pp. 259-275). Berlin & New York: Walter de Gruyter.
- Granger, S. & Paquot, M. (2010). The Louvain English for Academic Purposes dictionary. In S. Granger & M. Paquot (Eds.) *eLexicography in the 21st century: New applications, new challenges. Proceedings of eLEX2009*. Cahiers du Cental 7. Louvain-la-Neuve: Presses universitaires de Louvain, 87-96.
- Hewings, M. (2002) Using computer-based corpora as a teaching resource.
<http://artsweb.bham.ac.uk/MHewings/Corpora%20and%20teaching%20paper.doc>

- Hewings, M. & Hewings, A. (2002). "It is interesting to note that ...": a comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes*, 21(4): 367-383.
- Hyland, K. (2004). Disciplinary interactions: metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing*, 13, 133-151.
- Lee, D & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25, 56-75.
- Lee, D. Y. W. & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, 18, 281-296.
- Nesselhauf, N. (2005) *Collocations in a learner corpus*. Amsterdam / Philadelphia: John Benjamins.
- Paquot, M. (2010) *Academic vocabulary in learner writing. From extraction to analysis*. Continuum.
- Sinclair, J. (ed.) (2004). *Trust the Text: language, corpus and discourse*. New York: Routledge.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Thompson, P. (2006). Assessing the contribution of corpora to EAP practice. In Z. Kantaridou, I. Papadopoulou & I. Mahili (Eds.), *Motivation in learning language for specific and academic purposes*. Macedonia: University of Macedonia.
- Thorne, S., Reinhardt, J., & Golombek, P. (2008). Mediation as objectification in the development of professional academic discourse: A corpus-informed curricular innovation. In J. Lantolf & M. Poehner (Eds.), *Sociocultural Theory and the Teaching of Second Languages* (pp. 256-284). London: Equinox.
- Upton, T.A. & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, 20, 313-329.
- Uzar, R. & Waliński, J. (2001). Analysing the fluency of translators. *International Journal of Corpus Linguistics* 6, 155-166.

Suggested readings

- Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Amsterdam and Philadelphia: Benjamins.
- Flowerdew, J. (Ed.) (2002). *Academic Discourse*. London: Longman.
- Flowerdew, L. (2009). Applying corpus linguistics to pedagogy. A critical evaluation. *International Journal of Corpus Linguistics*, 14(3), 393-417.
- Granger, S., Hung, J. & Petch-Tyson, S. (Eds.) (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam / Philadelphia: Benjamins.