# Learner corpora

Diaz-Negrillo, Ana; Thompson, Paul

*DOI:*
10.1075/scl.59

# Learner corpora

## Looking towards the future*

Ana Díaz-Negrillo and Paul Thompson

Although still a relatively young field, learner corpus research is showing a remarkable rate of development that extends beyond corpus linguistics to other areas such as FLT, SLA research and, more recently, computational linguistics. This paper presents a state of the art overview of learner corpus research which illustrates the wide range of uses, users and activities that surround learner corpora. It provides a critical appraisal of what has been done so far and points to future lines of ~~development.~~

### 1. Introduction

Learner corpus research is a relatively young field of research, dating back to the late eighties (Granger 2004: 123). It is also vibrant: the last decade has seen a rapid expansion of activity in this field, resulting in more and more corpus resources,[1] in a broadening of the range of uses that learner corpora are put to, and in a diversity of the types of user. This expansion is due to a great extent to the pioneering work of Sylviane Granger and her team at the *Université Catholique de Louvain*, but is due also to a widespread embracing of mainstream corpus linguistics across many research and teaching communities, with a growth particularly in the number of Second Language Acquisition (SLA) researchers taking an interest in learner corpus research and development.

Our aim in this chapter is to provide an overview of recent work in learner corpus research and development that will show the multifaceted nature of work in this area. We will argue that there is a need for greater dialogue between: the

* Ana Díaz-Negrillo's contribution to this paper has been written within the research project with ref. FFI2012-30755 by the Spanish Ministy of Economy and Competitiveness.

1. A regularly updated listing of learner corpora around the world can be found at <https://www.uclouvain.be/en-cecl-lcworld.html> (Learner corpora around the world, CECL).

compilers and users of learner corpus data; between teachers, researchers and learners; and between corpus linguists and computational linguists. In the next section, we discuss what learner corpora are, and the issues involved in processing them. In Section 3 of the chapter, we examine the range of users of learner corpora, the uses the corpora are put to and the growing multidisciplinarity of the field. We then conclude with a discussion of the directions that learner corpus studies will take in the coming years.

## 2. Corpora types, processing and annotation

### 2.1 Types of learner corpora

Granger (2002: 5) defines learner corpora as "electronic collections of authentic FL/SL textual data according to explicit design criteria for a particular SLA/FLT purpose". As these collections are computerized, they can easily be searched and manipulated, and, because of their size, they provide a reliable basis on which to describe and model learner language use.

As shown in Figure 1, learner corpora can be placed at different points on at least six gradients, or axes. These gradients are: **mode** (spoken/written), **annotation** (unannotated/annotated), **language** (multilingual/monolingual), **data collection conditions** (± control), **time** (longitudinal/cross-sectional), and **breadth** (general/specialised).

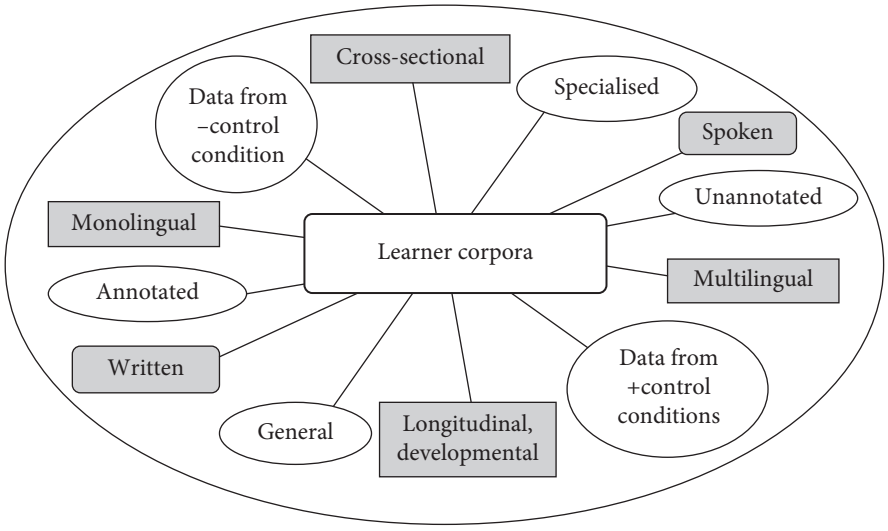


**Figure 1.** The gradients that learner corpora can be placed on (represented here, for convenience's sake, in two dimensions but intended to be multidimensional)

Learner corpora typically have been composed of written language data (Granger 2002). Nesselhauf (2004) observed that the majority of learner corpora (at that time of writing) were made up of academic essays, for the reason that they could easily be acquired by university researchers and in many cases they were already digitised. There is a steady growth, however, in the number of learner oral language corpora being produced, such as the LeaP (Gut 2012) and ISLE (<http://nats-www.informatik.uni-hamburg.de/~isle/speech.html>) corpus ~~projects~~ and also of multimodal corpora (see, for example, Reder et al. 2003). Furthermore, while the majority of learner corpora consist of English language learner data, there are other now reasonably developed corpora of other languages (Falko – German, CEDEL2 – Spanish, FLLOC – French).

Sinclair (1996), in writing of language corpus typologies, proposed that the default setting for *quality* of corpus data should be that the data is *authentic*. Granger (2002) rightly points out that the concept of *authentic* language learner production is problematic as much language learner output is produced in structured learning environments but she overstates the case by claiming that data obtained under experimental conditions do not qualify as learner corpus data on the grounds that such production does not equate to *authentic* language use. Granger herself reports in a footnote that Sinclair recommended that the term *experimental corpus* should be used of data obtained under experimental conditions and he observed that "it is important that serious intervention by the linguist, or the creation of special scenarios, is recorded in the name of the corpus". Following Sinclair's argument, then, one can label learner corpora that contain data obtained under +control conditions as *experimental learner corpora*. Experimental learner corpora are of value in certain lines of research, and in particular for phonetic analysis of learner oral output (cf. LeaP corpus; see also Ballier & Martin, Ferragne, Méli and Tortel this volume). In such research, it is essential to have high quality recordings of data and it is often important to control output so that a defined set of words are produced, allowing precise comparisons of how certain phonemes are realised.

Differing degrees of control can also be imposed in task specification. Where learner corpora tended in the past to be quite permissive in the specification of the types of performance admissible for inclusion in the corpus (often this was a pragmatic decision based on practicality), there is now a trend towards clearer task specification with an emphasis on the importance of comparable performances, that can be attributed at least in part to the growing participation of SLA researchers and language testing specialists in learner corpus work. The French Learner Language Oral Corpora (FLLOC) project (<http://www.flloc.soton.ac.uk/>), for example, has collected recordings of children at different stages of French language learning performing a range of tasks, for assessment purposes, all of which are

clearly specified on the project website. The scores on the tasks establish the proficiency level of the learner as well, which makes it possible to examine the features of learner performances not only at different ages but also at different proficiency levels. Similarly, large language test performance corpora, such as the Cambridge Learner Corpus (CLC), which contain examples of texts in different score bands, provide data for investigation of the linguistic and rhetorical features of learner performances in those bands.

Granger (2002) has also observed that learner corpora tend to be synchronic, and this remains the case. There is however an increase in diachronic corpora, such as the SILS (School of International Liberal Studies at Waseda University) corpus of undergraduate EFL writers, which allows both developmental and longitudinal studies, as it contains learner texts from different years of study and it also has a number of texts for each year by the same individual writers (Muehleisen 2006). Another example is the Japanese EFL Learner (JEFLL) corpus which is a collection of free writings by Japanese EFL learners in the six years of Junior and Senior School study, on any of six specified topics, that supports studies of writing development over time (<http://jefll.corpuscobo.net/index.htm>). A further example is the LONGDALE (Longitudinal Database of Learner English) corpus, compiled at Louvain-la-neuve (<https://www.uclouvain.be/en-cecl-longdale.html>). It is to be expected that the number of diachronic learner corpora will increase in coming years.

Corpus annotation will be discussed in detail below but it is worth noting here that the trend seems to be towards more annotation rather than less. Similarly, there appears to be a tendency towards greater specificity, both in the types of language learning covered and in the learners profiled. The MeLLANGE corpus (http://mellange.eila.jussieu.fr/index.en.shtml), for example, is a collection of learner translator written texts, which is a rich resource to be used by translator trainers, trainees and professional translators in the study of translation alternatives and of translator errors. The English Speech Corpus of Chinese Learners (ESCCL), developed for phonetic analysis of Chinese learner English, contains ~~samples speech~~from speakers of ten different regional dialects in China, and thus goes beyond the broad characterisation of a national grouping (Chinese) of language learners to the exploration of regional variation (Chen et al. 2008).

By definition, a learner corpus is a collection of learner output, of language produced by learners. A recent development, however, is the creation of complementary corpora of input, such as the textbooks that the learners are using in their instructed learning environment. A survey of textbooks can show how linguistic features are, or are not, treated in teaching materials. McEnery & Kifle (2002), for example, observe that Eritrean ELT textbooks do not cover the use of strong modality, and they link this to the underuse of strong modality markers in their

learner corpus data. Meunier & Gouverneur (2009) argue the case cogently for creating textbook corpora as an important resource in learner corpus studies; such corpora, which they term *pedagogical corpora*, make possible rapid and thorough analyses of textbook coverage. Meunier & Gouverneur also present the annotation scheme used to mark up the data in their textbook material (TeMa) corpus, a scheme which distinguishes between textbook rubric and language presentation, and also classifies activities into types, such as matching or completion activities, as well as sub-types.

Finally, a further important variable category is the *learner* in *learner corpora*. Variables that have been controlled for include age, gender, L1, L2 exposure, region, motivation, proficiency level, but these variables have not been controlled consistently, across corpora and are seldom incorporated in metadata and query options. As Granger (2004: 126) observes, "one must admit that ... there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one may want to control". Typically, the learners represented in learner corpora are school or university students, but an interesting exception is the multimedia adult ESL learner corpus (MAELC, Reder et al. 2003) which, as its name indicates, contains data obtained from adult ESL learners.

## 2.2 Annotation

Many studies have shown that it is feasible to do research on raw learner corpus data (cf. for example, Aijmer 2002; Nesselhauf 2004). This type of study typically focuses on a limited range of items or addresses questions in which the relevant linguistic features can be formally identified, and therefore also easily retrieved. Still, and just as is the case with L1 corpora, learner corpora have a much greater potential if specific language properties have been previously identified and signalled in the corpus, that is, if the corpora have been annotated.

Since the emergence of learner corpus work, the form of annotation that has been most often associated with learner corpora has been *error annotation*. Interest in this form of annotation can be seen in the number of attempts to design a gold standard error-tagging scheme (for an overview of error tagsets, ~~see Diaz-Negrillo and Fernandez-Dominguez~~). In addition, the study of error-annotated learner corpora has also been established as the one methodological approach that is specific to learner corpora, that of *computer-aided error analysis* (Granger 2002). This approach has had an influence in the three major areas associated with learner corpus research: SLA, FLT and computational linguistics. Some SLA researchers have questioned the adequacy of error analysis as a method for building a complete picture of the properties that can explain language acquisition; however,

error annotation can provide insights into proficiency stages, as shown in Abe and Tono (2005) and also Tono in this volume, and can be combined with other methods for the identification of properties that govern SLA. In the field of FLT, a thoroughly error-analysed corpus can be an invaluable resource, in that it can inform and constitute in itself a pedagogical tool (Granger 2009: 24). One clear example is provided by the error annotations in the Cambridge Learner Corpus, which inform the development of Cambridge University Press course and remedial materials. On a smaller scale, Mukherjee and Rohrbach (2006) and Mendikoetxea et al. (2010) report on the use of error-tagged local learner corpora for in-house pedagogical applications. Finally, error annotation is also relevant to computational linguists, as will be discussed in Section 3.3 below, since computational linguists are, for example, interested in the design of automatic annotation of learners' errors by using previously annotated learner data (cf. for example, Lee et al. 2009 on error annotation of Korean particles).

In terms of implementation, error-tagging practices have developed a degree of sophistication over the past few years. Earlier attempts consisted of pasting or typing in error codes in the learners' texts, or of coders relied on basic editors with menus that enhanced the tag insertion process. Tags were inserted directly in the learners' texts and queries were made on inline tags. While this approach is perfectly adequate for practices like data-driven learning using local learner corpora (cf. Mukherjee & Rohrbach 2006) or in small-scale studies (cf. Tono 2000), the format may impose a number of limitations on the research. More recently, annotations have been stored separately from the texts, in XML and in a multi-layered fashion. This new way of implementing annotations has a number of advantages over previous practices, among others, that various types of annotation can be added in various separate layers without interfering in the learner text, and this in turn strengthens the research potential of the learner data (see Reznicek et al. in this volume for further discussion). Error-tagging has also benefited from initiatives which aim to develop annotation tools that support manual annotation. These tools can be used for multi-layered annotation and sometimes also provide other functionalities such as searching and performing statistical tests. One example is the UAM CorpusTool (O'Donnell 2009), which has the advantage that it does not require programming on the part of the user.[2] Similar tools are MMAX2

2. The user designs the tagging scheme in the UAM CorpusTool (<http://www.wagsoft.com/CorpusTool/>) graphically, according to a hierarchical organisation of disjunctive and conjunctive options. This means that the user does not need to deal directly with XML. Another feature that makes the tool user-friendly is that glosses can be added to each feature in the tagset in order to facilitate the selection of tags during the tagging process. The coding can be modified during the annotation process and changes will as a result also be incorporated in the annotation carried out so far.

(Müller & Strube 2006), Dexter (Garretson 2006) or the SACODEYL Annotator (Pérez-Paredes & Alcaraz-Calero 2009). ExMERaLDA (http://www.exmaralda.org/) can also be used to handle multimodal corpora (cf. Sarré 2011). A visualization and query tool for multi-layered annotated corpora is ANNIS (Zeldes et al. 2009), which is currently used for Falko (see Reznicek et al. in this volume). Another is IMS Corpus Workbench (<http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>) used for instance for the ASK corpus.

Despite the degree of sophistication and the number of applications that have been developed recently, there are several issues associated with error-tagging that hinder its development and, consequently, large fully error-tagged corpora remain the exception. The first and most obvious issue has to do with its implementation. Even though there have been initiatives to automate the process, large-scale error-tagging still remains essentially a manual enterprise, which is naturally rather costly (see, however, Gamon et al. in this volume on crowdsourcing, which provides an alternative that can reduce costs). Subjectivity is a further issue. As a consequence, detailed documentation about the tagging scheme and the tagging guidelines has been considered essential to achieve systematicity in error-tagging. However, provision of documentation about error tagsets or error-tagged corpora still remains the exception, which, in turn, may be one of the reasons why a gold standard for error annotation is still to be arrived at. Reznicek et al. (in this volume), on the other hand, stress the importance of supporting error annotations with more than one *target hypothesis* in order to cover several tagging possibilities. While Falko includes this feature, this does not seem to be common practice in error-tagged corpora.

Similarly, inter-coder agreement, which is crucial to annotation reliability and validity of results, still remains a challenge for error-tagged learner corpora (Meurers 2009), first, because very few studies report on inter-coder reliability figures (cf. however, Fitzpatrick & Seegmiller 2004 or MacDonald et al. 2011) and, second, because of the challenge of attaining high inter-coder agreement figures. Difficulty in achieving high inter-coder agreement kappa figures (Carletta 1996; Artstein & Poesio 2008) has been reported for other types of manual annotation, in particular discourse annotation (Spooren & Degand 2010). It may be the case that inter-coder agreement in manual annotations, including error annotations, requires special treatment. This is a question that needs to be reflected upon if progress in error annotation is to be made.

A final issue is the validity of error annotation, that is, whether the tagset adopted is actually valid for learner corpus research. Error-tagging imposes an error categorization on learner data which may not always be adequate for the end user's research, because it may not cover the categories the researcher is interested in, or simply because the error categorization may be unsuitable for the actual target

research (Tono 2003: 804; Gamon et al. 2009). This seems an inescapable issue since tagging always implies the imposition of a given set of categories. One possible solution is problem-oriented annotation (de Haan 1984), that is, a form of annotation that suits the requirements of a particular research topic (see, for example, Tono 2000). This may provide a way of undertaking error-tagging in learner corpora, until large-scale error-tagging based on robust schemes that may support a variety of research agendas reveals itself as a feasible enterprise.

More recently, the annotation of linguistic properties of learner language, in the form of POS tagging, has drawn increasing interest. A number of learner corpora incorporate POS tagging (cf. ICLE v.2 Granger et al. 2009 or the ASK corpus of learner Norwegian, Tenfjord et al. 2006). However, POS tagging of learner corpora seems to have been treated as an instance of domain transfer just as when automatic taggers trained on particular text genres are run on corpora of texts from a different genre. The performance of the tagger on the new genre is usually lower and therefore post-editing techniques need to be applied to improve the quality of the tagging (cf. for example, van Rooy & Schäfer 2002; Thouësny 2009).[3]

As shown in the latest POS annotated version of the ICLE, the use of post-editing techniques can result in high-quality POS annotation. However, POS annotation of learner language continues to be debated in the literature. It has been argued that in learner language, stem, distribution and morphology do not always match as they do in native language, and therefore learner language provides linguistic properties which diverge from those defined in native language grammatical categorizations.[4] Díaz-Negrillo et al. (2010) discuss this issue and explain that in transferring native categories to learner language POS tagging, the actual learner grammatical categories become concealed behind native POS categorizations. Rastelli (2009) and Ragheb & Dickinson (2010) in the same vein, argue that learner language should be described in its own right and also advocate an *ad hoc* POS tagging of learner language. Finally, Díaz-Negrillo et al. (2010) argue that instead of associating learner language with native POS tags, stem, morphological marking and syntactic distribution should be individually described in a multi-level annotation fashion.

There have also been attempts to annotate more complex linguistic units in learner corpora. Syntactic annotation has been recently been looked at in Dickinson & Ragheb (2009) and Rosén & De Smedt (2010). Functional annotation has also been attempted, for example, in Schiftner in this volume, which tackles annotation

**3.** According to the literature, spelling errors, incorrectly inflected words and syntactic problems in learner data seem to pose the majority of problems to native POS taggers (van Rooy & Schäfer 2002).

**4.** This can be seen for example in *one of the favourite places to visit for many foreigns*, where *foreigns* exhibits an adjectival stem but behaves morphologically and syntactically as a noun.

of text features within the framework of rhetorical structure theory. Just as in POS annotation of learner data, in these types of annotation there is also the issue of whether the categorization of learner language is to be made only on the basis of native categories or whether *ad hoc* annotation of learner features should instead be pursued. In the latter case the great internal variation in learner language is one of the most problematic issues to tackle.

## 3. Uses and users of learner corpus data

### 3.1 Overview

As learner corpora have grown in number, the range of creators and users of learner corpora has also expanded and the number of uses of learner corpora has also amplified. Figure 2 shows in diagrammatic form the range of users (outer circle) and the types of activities that users are involved in (inner circle).

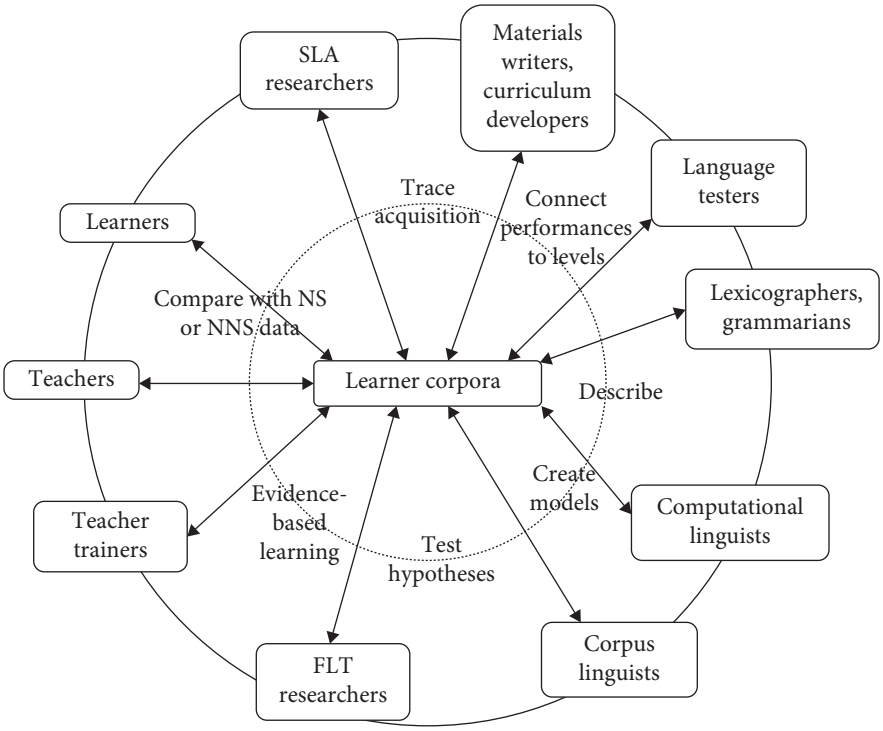


**Figure 2.** Users and activities surrounding learner corpora (users are depicted on the outer circle and activities are shown on the inner circle). Activities are shown near to the typical users but can be associated with other user groups as well.

The two main research user groups are those of FLT and SLA researchers, along with a further researcher group, that of corpus and computational linguists. The other user groups are to differing degrees involved in foreign language teaching, either in practice or in providing reference, testing and pedagogical materials for use in teaching. The various user groups are not intended to be represented here as separate entities; work on learner corpora is often multidisciplinary with, for example, computational linguists working with FLT researchers, or corpus linguistics working with language testers. Moreover, the work of researchers or practitioners may influence or be influenced by those in other areas.

### 3.2 Foreign language teaching

Learner corpus data can be used for pedagogical purposes by incorporating the findings of SLA and FLT research into the language classroom or into teaching materials, or after undertaking surface research into the language learners' production. Following this distinction, the terms *delayed pedagogical use* and *immediate pedagogical use* are often used (Granger 2009: 20–22).

Delayed pedagogical use of learner corpus data involves a two-stage process. First, the learner corpus researcher compiles the corpus data which can at a later stage be used by publishers to inform course books and dictionaries designed for similar language populations to those represented in the corpus. O'Dell (2005), for example, describes the value of having access to learner corpus data while she writes FLT materials for publication; Gilquin et al. (2007) argue the case persuasively for the use of learner corpora in FLT materials and dictionary development. Instances of teaching materials and dictionaries based on learner corpus data are, for example, *Learning from Common Mistakes* (Brook-Hart 2009), published by Cambridge University Press and based on the Cambridge Learner Corpus, or the *Macmillan English Dictionary for Advanced Learners* (Rundell 2007) which is partly based on the ICLE.[5] Delayed pedagogical use of learner corpus data can also be made by foreign language teachers in order to design classroom materials for areas where course books fail to pay sufficient attention or where students experience high levels of difficulty. These materials often include error correction or error identification activities and can be complemented with native corpus data.

Learner corpus data become more relevant to language learners if the data they are presented with are data they have produced. This is what Granger (2009: 20–22) calls *immediate pedagogical use* of learner corpora, which is reported in, for example, Seidlhofer (2002) and, more recently, Mukherjee & Rohrbach

5. Actual examples of how learner corpus data and findings have informed this dictionary are described in detail in De Cock & Paquot (2009).

(2006). Immediate pedagogical use of learner corpora can be effective where the data are perceived as relevant to the learner, because this can increase levels of motivation (Mukherjee & Rohrbach 2006: 228). A corpus of student writings by writers of the same L1 as the learners can, for example, be investigated to explore variation between the language of that group and the target language user group, or it can be used by the learner to set realistic goals for language attainment by looking at examples of production by learners at a higher proficiency level (cf. Franca 1999: 116). Learner corpora composed of texts similar to the texts that the learners are preparing to produce can also be used, and in some cases these may be post-edited by the teacher for correction of errors (cf. Al-Lawati 2011); the rationale for using such data is that the texts in the corpus are culturally familiar to the learners and they represent performances that are attainable by the learners.

Personal experience tells us that the use of learners' language for pedagogical treatment with the same set of language learners is something that teachers were doing long before learner corpora came onto the scene. The difference now is that this can be done with corpus linguistic techniques (such as using annotations for later retrieval of relevant examples, sorting, counting, etc.), and consequently teachers can have more objective information about their students' difficulties, on the one hand, and more powerful tools with which to work on their students' data, on the other.

McCarthy (2008) has made a strong case for the introduction of corpus training into language teacher training. He argues that the role of the teacher should be shifted from that of the consumer of corpora to that of a "researcher ... someone more actively involved in their own professional development and in what happens in their classrooms" (McCarthy 2008: 564) and that teachers should be given training in corpus evaluation and exploitation. To date, however, large-scale integration of corpus training into teacher training courses is limited. Allan (2002) describes the use of the TeleNex network in Hong Kong to allow trainee teachers the opportunity to work with the TELEC learner corpus, but this is a relatively rare example.

Much language teaching is concerned with preparing learners for language tests and these tests are also increasingly influenced by research done on learner corpora. Learner corpora can be used for example in the compilation and grading of wordlists into different Common European Framework of Reference (CEFR) levels (Capel 2010) or can be used to create profiles of learner performances at different levels (Hawkins & Buttery 2009). These profiles can then be exploited in exam preparation materials and they can be used to construct computer programmes that can give preliminary gradings of exam performances.

### 3.3 Second language acquisition research

It is usually assumed that a "learner's performance is indicative of what learners know of the L2" and, as a result, that "learner language should constitute the primary data for the study of L2 acquisition" (Ellis & Barkhuizen 2005: 359). Learner corpora, being large and carefully designed electronic collections of learner data, constitute invaluable sources of evidence for the study of L2 acquisition.

Among the various types of learner language, learner corpus data typically comply with what Ellis & Barkhuizen (2005: 30–31) call *clinically elicited samples*, that is, samples of language collected for research purposes, in classroom settings and as part of tasks in which learners are required to use a foreign language to achieve a particular purpose. As observed in Section 2 above, this type of linguistic learner data takes an intermediate position between experimental data (+control) and naturally occurring data (–control). In contrast with experimental data, first, learner corpus data (as distinct from experimental corpus data) is usually structured in full texts. This means that any language instance under investigation is contextualised and therefore can be analysed within a wider picture of the learner's performance. In addition, since most learner corpus elicitation tasks do not aim to retrieve samples representative of a very specific research question, learner data can be used for a variety of research topics.[6] Finally, although learner corpora are still relatively small compared to native corpora,[7] they are intended to be large data collections, which makes it possible to observe the occurrence of a wide variety of language uses.

At the extreme of –control are naturally occurring data. Typically, learner corpus data do not comply with the features associated with naturally occurring data, that is, language produced in real-life situations for communicative purposes and subjected to no elicitation. Naturally occurring data may be easier to retrieve in SL contexts in which students can *naturally* use a SL in real-life communicative contexts. However, it seems less straightforward in FL contexts, where the language classroom is probably the only setting where the FL is used by the learners, and therefore where the FL will be more naturally used (Granger 2002: 8). As a result, particularly in the case of FL contexts, clinically elicited data seem not only more readily available to the FL learner corpus researcher but also closer to what naturally occurring data are understood to be in the case of FL learners.

---

**6.** Cf. however, SPLLOC corpus (<http://www.splloc.soton.ac.uk/>), which includes the use of controlled tasks to test the order and gender of clitics (<http://www.splloc.soton.ac.uk/splloc1/cpt.html>)

**7.** See, however, Section 4 for information on some initiatives that foster corpus data sharing and large-scale corpus collection. These and other similar initiatives might well begin to level out the difference in size between native and learner corpora in no more than a decade.

All this said, the combination of learner corpus data with other more controlled language data types, which has already been exhibited in, for example, Gilquin (2007), and recently argued for in Granger (2012), represents a new avenue to be explored. In what follows we explain the reasons for such a match. First, as is often argued, experimental data sometimes offer the only way to have access to infrequent features, which may be harder to explore with corpus data. In addition, experimental data can also be used, not just to have access to more infrequent language uses but also to simply provide more fine-grained insights into the acquisition of a particular aspect in the study of, for example, avoidance strategies or degrees of acquisition. In addition, a criticism often levelled at learner corpus data is that language producers are no longer accessible to the researcher and therefore their language cannot be further examined. Triangulation, the use of various sources of data, may help overcome this limitation by providing further sources of information about the research question under study. Finally, combining learner corpus data with other data types, such as experimental learner corpus data or non-corpus experimental data, may be beneficial not just for a better understanding of research questions, but also for the further development of learner corpus studies. An assessment of what SLA experimental data and learner corpus data can offer to each other may encourage the use of learner corpus data by SLA researchers and, by and large, result in better communication between SLA researchers and learner corpus users (see also Section 2 in Lozano & Mendikoetxea in this volume for more details of learner corpora and SLA research).

The exploitation of learner corpus data critically depends on the design of the learner corpus. Aspects that have often been explored in the history of SLA research, like L1 transfer or the relevance of input, can now be more systematically investigated and replicated using learner corpus data, especially if learner corpora contain a component containing L1 data, or of the input such language learners are exposed to, for example in the form of the textbook data (see Section 2 above). Interlanguage development, which is one the main concerns of SLA research, can be now explored in longitudinal corpora (see 2.1 for examples of longitudinal corpora). Other factors such as age in interlanguage development can be explored with corpora containing data from language learners of different ages, from young learners to adults. Finally, the influence of genre and register in learner language use can be studied more systematically if corpora contain data produced according to varied tasks types in varied contexts. All in all, it seems reasonable to encourage the collection of learner corpus data designed with a clear SLA research agenda in mind. This, in turn, may lead to the development of a substantial body of SLA interpretative studies in learner corpus research which, as is often suggested, are still very much in need in the field (Granger 2012: 21).

### 3.4 Corpus and computational linguistics

As learner corpora are in essence a language corpus type, they share with other language corpora basic corpus linguistic principles that relate to corpus design, data processing, data analysis and corpus tools design, albeit with an obvious degree of specialisation. There has been a great effort to define learner corpus data gathering, processing and analysis techniques over the past years especially by the Louvain group, as mentioned in the introduction to this chapter. At the same time, corpus and computational linguists have also worked on the adaptation of corpus tools that can be used in the description of learner language-specific features and that can cope with this type of language-specific need. The design of error-tagging systems, as described above, as well as of techniques that may foster automatic error-tagging is an example (cf. Rayson & Baron 2011; Tono in this volume). Recent approaches to grammatical and syntactic annotation of learner corpora mentioned earlier (Dickinson & Ragheb 2009; Rastelli 2009; Díaz-Negrillo et al. 2010; Ragheb & Dickinson 2010) also give evidence of work in this direction. To these should also be added automatic analytic tools to measure learner language features, such as the L2 Syntactic Complexity Analyser described in Lu (2010) and also in Ai & Lu in this volume.

In terms of data analysis, the development of learner corpus research was to a large extent due to the amount of descriptive research carried out by linguists, especially at the inception of learner corpus work. Although the applications of findings within FLT and SLA research are greatly emphasized nowadays, pure corpus descriptive studies are still in progress. Considering that learner corpora are collected to be representative of the language use of a particular population, learner corpus language can be viewed as representing another language variety that shows specific linguistic features and which may be described using corpus linguistics techniques. In this sense, learner corpora share the research agenda with other language varieties, such as second and native language. This has already been evidenced in studies such as Nesselhauf's (2009) work on collocations, or by some studies collected in the volume edited by Mukherjee & Hundt (2011), which comprises corpus-based empirical research on computerised corpora of learner Englishes and second-language varieties of English, and examines the presence of common features across the different English varieties.

Learner corpora constitute large sets of naturalistic data, which sometimes have the added value that they are also error-tagged, so they can be used for NLP intrinsic purposes as shown in Gamon et al. and suggested by Tono (both in this volume). In their paper, Gamon et al. show the potential of learner corpora for the development, training and evaluation of error detection and correction systems. In a related manner, Tono suggests that learner corpora can be used to identify

criterial features that at a later stage can be used by automatic performance analysis systems to decide on the proficiency level of a given learner text. This may have a direct application in the grading of placement tests. Finally, some computational linguists work on aspects of learner language that may have a direct application in language teaching and learning, for example to design language learning applications, such as ESL Tutor (Cowan et al. 2003, in Granger et al. 2007: 256; see also Granger et al. 2007 for an overview of learner corpus-informed CALL systems).

The increasing interest of computational linguists in learner corpora is further confirmed in the organisation of two preconference workshops at the CALICO conferences in 2008 and 2009 (*Automatic analysis of learner language, AALL 2008, AALL 2009*). The two events brought computational linguists together to discuss issues in learner language modelling and came to a general conclusion that there was a clear need for greater collaboration with other areas of learner corpus research (Meurers 2009: 469–470). Similarly, the recent '*HOO Challenges*' (<http://clt.mq.edu.au/research/projects/hoo/>), a shared task intended for NLP specialists and concerned with automated correction of learners' errors, stands as evidence of the computational interest in learner language-related topics.

## 4. Looking forwards

In this chapter, we have provided an overview of recent developments and issues in the field of learner corpus research, and indicated some of the directions in which the field is developing. We have argued that the field is expanding as the range of users and uses of learner corpora has widened, and we predict that this trend (of expansion) will continue.

It is likely that more publishers and testing organisations will develop their own in-house learner corpus collections in the way that Cambridge University Press and Cambridge ESOL have done, and also that larger numbers of SLA researchers will engage in corpus-based studies of language acquisition. The latter will necessarily require the development of new corpus resources, addressing a wider range of languages, and with a greater emphasis on task specification and on capturing more fine-grained learner and context variables that relate to particular SLA research agendas.

The collection of larger and more diversified corpora may be easier if portals for data collection were established, as is the case of English Profile (Cambridge University Press, (<https://epp.ilexir.co.uk/>) or CHILDES (<http://childes.psy.cmu.edu/>), and also, at a smaller scale of CEDEL2 (<http://www.uam.es/proyectosinv/woslac/start.htm>), which collects its own data using the same means. We expect also that there will be an increase in the number of learner corpora that allow for

longitudinal/development studies. All this will be motivated by the desire to learn more about language acquisition sequences, to profile learner language performances at different proficiency levels and to identify the lexis and grammatical structures for inclusion in language learning syllabi.

Another area of expansion in learner corpus research is that of spoken language, as shown in the presence of four chapters in this volume that discuss learner oral data (Ballier & Martin, Ferragne, Méli and Tortel). Some of the difficulties of dealing with oral data are data collection and processing. Nowadays, oral data can be more easily collected and handled with management systems like IPS Wikispeech (Institute of Phonetics and Speech Processing of Munich, (<https://webapp.phonetik.uni-muenchen.de/wikispeech/>). The development of resources to deal with oral data along with greater collaboration with computational scientists and phoneticians will greatly benefit research on learner spoken data (see Ballier & Martin in this volume). Concomitant with the growth of learner corpora of oral data will come a move towards multimodal corpora where the text files will be supplemented by corresponding sound files or filmed material (for the FLLOC project, the audio files can be downloaded from the website, for example) or integrated so that links are inserted into the text files to specific points in the audio/video files. The alignment of the transcript to audio file is one further layer of mark-up of the data.

Finally, as we have argued above, annotation of the data is also likely to become more sophisticated with annotation layers at different levels and with more research into automatic annotation. Automation of learner corpus annotation is clearly an improvement as it saves time and money and cuts out the stages of inter-rater agreement testing, even though some stages of manual post-editing may be necessary. However, automation seems to have also introduced a shift towards more coarse-grained annotation, losing therefore the detail that manual annotation could provide (see, for example, Díaz Negrillo 2009), which is something we should not lose sight of. Different groups are developing their own approaches to error annotation, for example, or to mark-up of syntactic or functional features of the data, which is indicative of the broadening range of activity in this area of learner corpus research. Still, there is a need for discussion between these groups so that researchers can work towards some degree of standardisation of approach to annotation, where possible, and so that learner corpus resources can achieve a reasonably high level of interoperability.

## References

Abe, M. & Tono, Y. 2005. Variations in L2 spoken and written English: Investigating patterns of grammatical errors across proficiency levels. Paper presented at the *Corpus Linguistics 2005*

*Conference*, Birmingham, UK. http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/LanguageLearningandError/variationsinL2.doc [Accessed 31.7.2012]

Ai, H. & Lu, X. this volume. A corpus-based comparison of syntactic complexity in NNS and NS university students' writing.

Aijmer, K. 2002. Modality in advanced Swedish learners' written interlanguage. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung & S. Petch-Tyson (eds), 55–76. Amsterdam: John Benjamins.

Al-Lawati, N. 2011. Learning strategies used and observations made by EFL Arab students while working on concordance-based grammar activities. *AWEJ* 2(4): 302–322. http://www.awej.org/awejfiles/_80_6_11.pdf [Accessed 31.7.2012]

Allan, Q. 2002. The TELEC secondary corpus: A resource for teacher development. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. S. Granger, J. Hung & S. Petch-Tyson (eds), 195–212. Amsterdam: John Benjamins.

Artstein, R. & Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4): 555–596.

Ballier, N. & Martin, P. this volume. Developing corpus interoperability for phonetic investigation of learner corpora.

Brook-Hart, G. 2009. *Learning from Common Mistakes. Intermediate.* Cambridge: Cambridge University Press.

Capel, A. 2010. Insights and issues arising from the English Profile Wordlists project. *Cambridge ESOL Research Notes* 41: 2.

Carletta, J. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* 22(2): 249–254.

CEDEL2. http://www.uam.es/proyectosinv/woslac/start.htm [Accessed 12.8.2012]

Chen, H., Wen, Q. & Li, A. 2008. A Learner Corpus – ESCCL. In *Proceedings of the Fourth Conference on Speech Prosody*. Campinas, Brazil. http://sprosig.isle.illinois.edu/sp2008/papers/id187.pdf [Accessed 31.7.2012]

CHILDES. http://childes.psy.cmu.edu/ [Accessed 12.8.2012]

Cowan, R., Choi, H.E. & Kim, D.H. 2003. Four questions for error diagnosis and correction in CALL. *CALICO Journal* 20(3): 451–463.

De Cock, S. & Paquot, M. 2009. The monolingual learners' dictionary as a productive tool: The contribution of learner corpora. In *Corpus-Based Approaches to English Language Teaching*, M.C. Campoy, B. Bellés-Fortuno & M.L. Gea-Valor (eds), 195–204. London: Continuum.

de Haan, P. 1984. Problem-oriented tagging of English corpus data. In *Corpus Linguistics: Recent Developments in the Use of Computer Corpora*, J. Aarts & W. Meijs (eds), 123–139. London: Addison Wesley Longman.

Díaz Negrillo, A. 2009. *EARS: A User's Manual*. Munich: LINCOM Academic Reference Books.

Díaz-Negrillo, A. & Fernández-Domínguez, J. 2006. Error tagging systems on learner corpora. *RESLA* 19: 83–102.

Díaz-Negrillo, A., Meurers, D., Valera, S. & Wunsch, H. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum* 36(1–2): 139–154.

Dickinson, M. & Ragheb, M. 2009. Dependency annotation for learner corpora. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories*, M. Passarotti, A. Przepiórkowski, S. Raynaud & F. Van Eynde (eds), 59–70. Milan: EDUCatt.

Ellis, R. & Barkhuizen, G. 2005. *Analysing Learner Language*. Oxford: Oxford University Press.

English Profile. Data Collection portal. https://epp.ilexir.co.uk/ [Accessed 13.8.2012]

ExMERaLDA. http://www.exmaralda.org/ [Accessed 13.8.2012]
Ferragne, E. this volume. Automatic suprasegmental parameter extraction in learner corpora.
Fitzpatrick, E. & Seegmiller, M.S. 2004. The Montclair electronic language database project. In *Applied Corpus Linguistics: A Multidimensional Perspective*, U. Connor & T.A. Upton (eds), 223–237. Amsterdam: Rodopi.
FLLOC project. http://www.flloc.soton.ac.uk [Accessed 2.8.2
Franca, V.B. 1999. Using student-produced corpora in the L2 classroom. In *IATEFL 1999 Edinburgh Conference Selections*, P. Grundy (ed.), 116–117. Whitstable: IATEFL.
Gamon, M., Leacock, C., Brockett, C., Dolan, W., Gao, J., Belenko, D. & Klementiev, A. 2009. Using statistical techniques and web search to correct ESL errors. *CALICO Journal* 26(3): 491–511.
Gamon, M., Chodorow, M., Leacock, C. & Tetreault, J. this volume. Using learner corpora for automatic error detection and correction.
Garretson, G. 2006. Dexter: Free tools for analyzing texts. In *Academic and Professional Communication in the 21st Century: Genres, Rhetoric and the Construction of Disciplinary Knowledge. Proceedings of the 5th International AELFE Conference*, M.C. Pérez-Llantada Auría, R. Pló Alastrué & C.P. Neumann (eds), 659–665. Zaragoza: Prensas Universitarias de Zaragoza.
Gilquin, G. 2007. To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik* 55(3): 273–291.
Gilquin, G., Granger, S. & Paquot, M. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6: 319–335.
Granger, S. 2002. A bird's-eye view of learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. S. Granger, J. Hung & S. Petch-Tyson (eds), 3–33. Amsterdam: John Benjamins.
Granger, S. 2004. Computer learner corpus research: Current status and future prospects. In *Applied Corpus Linguistics: A Multidimensional Perspective*, U. Connor & T. Upton (eds), 123–145. Amsterdam: Rodopi.
Granger, S. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In *Corpora and Language Teaching*, K. Aijmer (ed.), 13–32. Amsterdam: John Benjamins.
Granger, S. 2012. How to use second and foreign language learner corpora. In *Research Methods in Second Language Acquisition: A Practical Guide*, A. Mackey & S.M. Gass (eds), 7–29. London: Wiley-Blackwell.
Granger, S., Kraif, O., De Ponton, C., Antoniadis, G. & Zampa, V. 2007. Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL* 19(3): 252–268.
Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English. Handbook and CD-ROM. Version 2.* Louvain-la-Neuve: Presses universitaires de Louvain.
Gut, U. 2012. The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In *Multilingual Corpora and Multilingual Corpus Analysis*, T. Schmidt & K. Wörner (eds), 3–23. Amsterdam: John Benjamins.
Hawkins, J. & Buttery, P. 2009. Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment*, L. Taylor & C. Weir (eds), 158–175. Cambridge: Cambridge University Press.
HOO Challenges. http://clt.mq.edu.au/research/projects/hoo/ [Accessed 13.8.2012]

IPS Wikispeech. https://webapp.phonetik.uni-muenchen.de/wikispeech/ [Accessed 13.8.2012]
ISLE corpus. http://nats-www.informatik.uni-hamburg.de/~isle/speech.html [Accessed 2.8.2012]
JEFLL corpus project. http://jefll.corpuscobo.net/index.htm. [Accessed 13.8.2012]
Learner corpora around the world. CECL. https://www.uclouvain.be/en-cecl-lcworld.html [Accessed 13.8.2012]
Learner corpus bibliography CECL. http://www.uclouvain.be/en-cecl-lcbiblio.html [Accessed 13.8.2012]
Lee, S.-H., Jang, S.-B. & Seo, S.-K. 2009. Annotation of Korean learner corpora for particle error detection. *CALICO Journal* 26(3): 529–544.
LONGDALE. https://www.uclouvain.be/en-cecl-longdale.html [Accessed 13.08.2012]
Lozano, C. & Mendikoetxea, A. this volume. Learner corpora and Second Language Acquisition: The design and collection of CEDEL2.
Lu, X. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4): 474–496.
MacDonald, P., Murcia, S. Boquera, M., Botella, A., Cardona, L., García, R., Mediero, E., O'Donnell, M., Robles, A. & Stuart, K. 2011. Error Coding in the TREACLE project. In *Las Tecnologías de la Información y las Comunicaciones: Presente y Futuro en el Análisis de Córpora. Actas del III Congreso Internacional de Lingüística de Corpus*, M.L. Carrió Pastor & M.A. Candel Mora (eds), 725–740. Valencia: Universitat Politècnica de València.
McCarthy, M. 2008. Accessing and interpreting corpus information in the teacher education context. *Language Teaching* 41(4): 563–574.
McEnery, T. & Kifle, N. 2002. Epistemic modality in argumentative essays of second-language writers. In *Academic Discourse,* J. Flowerdew (ed.), 182–195. Harlow: Longman.
Méli, A. this volume. Phonological acquisition in the French-English interlanguage: Rising above the phoneme?
MeLLANGE project. http://mellange.eila.jussieu.fr/index.en.shtml [Accessed 2.8.2012]
Mendikoetxea, A., Murcia Bielsa, S. & Rollinson, P. 2010. Focus on errors: Learner corpora as pedagogical tools. In *Corpus-Based Approaches to English Language Teaching*, M.C. Campoy, B. Bellés-Fortuño & M.Ll. Gea-Valor (eds), 180–194. London: Continuum.
Meunier, F. & Gouverneur, C. 2009. New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In *Corpora and Language Teaching*, K. Aijmer, (ed.), 179–201. Amsterdam: Benjamins.
Meurers, D. 2009. On the automatic analysis of learner language. Introduction to the special issue. *CALICO Journal* 26(3): 469–473.
Muehleisen, V. 2006. Introducing the SILS Learners' Corpus: A tool for writing curriculum development. *Waseda Global Forum* 3: 119–125.
Mukherjee, J. & Hundt, M. (eds). 2011. *Exploring Second-Language Varieties of English and Learner Englishes*. Amsterdam: John Benjamins.
Mukherjee, J. & Rohrbach, J.-M. 2006. Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In *Planning, Painting and Gluing Corpora. Inside the Applied Corpus Linguist's Workshop*, B. Kettmann & G. Marko (eds), 205–232. Frankfurt: Peter Lang.
Müller, C. & Strube, M. 2006. Multi-level annotation of linguistic data with MMAX2. In *Corpus Technology and Language Pedagogy*, S. Braun, K. Kohn & J. Mukherjee (eds), 197–214. Frankfurt: Peter Lang.
Nesselhauf, N. 2004. Learner corpora and their potential for language teaching. In *How to Use Corpora in Language Teaching,* J. Sinclair (ed.), 125–152. Amsterdam: John Benjamins.

Nesselhauf, N. 2009. Co-selection phenomena across new Englishes: Parallels (and differences) to foreign learner varieties. *English Word-Wide* 30(1): 1–26.

O'Dell, F. 2005. How the Cambridge Learner Corpus helps with materials writing. *Humanising Language Teaching* 7(1) http://www.hltmag.co.uk/jan05/idea01.htm [Accessed 13.8.2012]

O'Donnell, M. 2009. The UAM CorpusTool: Software for corpus annotation and exploration. In *Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente*, C.M. Bretones et al. (eds), 1433–1447. Almería: Universidad de Almería.

Pérez-Paredes, P. & Alcaraz-Calero, J.M. 2009. Developing annotation solutions for online Data Driven Learning. *ReCALL* 21(1): 55–75.

Ragheb, M. & Dickinson, M. 2010. Avoiding the comparative fallacy in the annotation of learner corpora. In *Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions*, G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. Prieto Botana & E. Rhoades (eds), 114–124. Somerville MA: Cascadilla Proceedings Project.

Rastelli, S. 2009. Learner corpora without error tagging. *Linguistik Online* 38: 57–66.

Rayson, P. & Baron, A. 2011. Automatic error tagging of spelling mistakes in learner corpora. In *A Taste for Corpora. In Honour of Sylviane Granger*, F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds), 109–126. Amsterdam: John Benjamins.

Reder, S., Harris, K. & Setzler, K. 2003. The multimedia adult ESL learner corpus. *TESOL Quarterly* 37(3): 546–558.

Reznicek, M., Lüdeling, A. & Hirschmann, H. this volume. Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture.

Rosén, V. & De Smedt, K. 2010. Syntactic annotation of learner corpora. In *Systematisk, variert, men ikke tilfeldig*, H. Johansen, A. Golden, J.E. Hagen & A.-K. Helland (eds), 120–132. Oslo: Novus forlag.

Rundell, M. (ed.). 2007. *Macmillan English Dictionary for Advanced Learners. Second Edition*. Oxford: Macmillan Education.

Sarré, C. 2011. Computer-mediated negotiated interactions: How is meaning negotiated in discussion boards, text-chat and videoconferencing? In *Second Language Teaching and Learning with Technology*, S. Thouësny & L. Bradley (eds), 189–210. Dublin: Research Publishing.

Schiftner, B. this volume. Analysing coherence in upper-intermediate learner writing.

Seidlhofer, B. 2002. Pedagogy and local learner corpora: Working with learning-driven data. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, S. Granger, J. Hung & S. Petch-Tyson (eds), 213–234. Amsterdam: John Benjamins.

Sinclair, J. 1996. *EAGLES: Preliminary Recommendations on Corpus Typology* http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html [Accessed 31.7.2012]

SPLLOC project. http://www.splloc.soton.ac.uk/ [Accessed 13.08.2012]

Spooren, W. & Degand, L. 2010. Coding coherence relations: Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2): 241–266.

Tenfjord, K., Meurer, P. & Hofland, K. 2006. The ASK corpus – a language learner corpus of Norwegian as a second language. *Proceedings from 5th International Conference of Language Resources and Evaluation (LREC)*, 1821–1824. http://hnk.ffzg.hr/bibl/lrec2006/pdf/573_pdf.pdf [Accessed 13.08.2012]

Thouësny, S. 2009. Increasing the reliability of a part-of-speech tagging tool for use with learner language. Paper presented at *Automatic Analysis of Learner Language (AALL'09): From a*

*Better Understanding of Annotation Needs to the Development and Standardization of Annotation Schemes*. Arizona State University, Tempe.

Tono, Y. this volume. Criterial feature extraction using parallel learner corpora and machine learning.

Tono, Y. 2000. A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In *Rethinking Language Pedagogy from a Corpus Perspective*, L. Burnard & T. McEnery (eds), 123–133. Frankfurt: Peter Lang.

Tono, Y. 2003. Learner corpora: Design, development and applications. In *Proceedings of the 2003 Corpus Linguistics Conference*, D. Archer, P. Rayson, A. Wilson & T. McEnery (eds), 800–809. UCREL: Lancaster University.

Tortel, A. this volume. Prosody in a contrastive learner corpus.

UAM CorpusTool. http://www.wagsoft.com/CorpusTool/ [Accessed ~~13.08~~]

van Rooy, B. & Schäfer, L. 2002. The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies* 20: 325–335.

Zeldes, A., Ritz, J., Lüdeling, A. & Chiarcos, C. 2009. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics 2009*, M. Mahlberg, V. González-Díaz & C. Smith (eds), 20–23. University of Liverpool, UK.