14 Learner corpora

Introduction

In this chapter we explore the insights to be gained into the nature of second language acquisition from the analysis of *learner corpora*, which are digital representations of the performance or output, typically written, of language learners. The use of learner corpora in SLA studies thus falls within the tradition of treating learner language as expression, as described in earlier chapters.

The way in which the emerging field of learner corpus analysis differs from other approaches to SLA is detailed in the following sections. First we examine the design and compilation of learner corpora, which is a crucial component of data gathering, since learner corpora are in some ways more complex as data sets than native speaker corpora and the appropriate encoding of information about the learners and the setting is a key prerequisite for ensuing analyses. Next we look at research on learner corpora, which as Granger (forthcoming) notes is a relatively new development and consequently it is not possible to provide a comprehensive assessment of the place of learner corpora research in SLA studies at this time. Nevertheless, what we can do is review some existing research studies in order to present some of the findings coming out of learner corpus research and to illustrate different aspects of the analysis of learner corpus data.

One striking characteristic of learner corpus research is the prevalent use of frequency data, which might include the learners' over use or under use of lexical or grammatical forms or an analysis of the frequency of error forms. The emphasis on frequency, combined with the fact that learner corpora are quite large, means that software tools are an essential part of learner corpus research, and in the third section in this chapter we illustrate the application of corpus analysis software.

One impetus to compile learner corpora follows from the Error Analysis tradition of identifying, describing and explaining errors, and many of the issues related to error analysis and linguistic analyses based on categorization of errors described in Chapter 3 apply to the analysis of learner corpora. There are, however, some important differences in approach and also in procedure, since the analysis of learner corpora encompasses the techniques and many of the assumptions of *corpus linguistics*. Applying the methodology of corpus analysis (Sinclair 1991) to the investigation of learner corpora entails the collection of fairly large samples, typically hundreds of thousands of words or more of learner language. The general technique consists of trawling through learner corpora using searching software to reveal and quantify recurrent patterns, typically lexico-grammatical patterns, that characterize the learner language associated with different learners and different settings.

The links to corpus linguistics have influenced the models of grammar and acquisition processes typically used in learner corpus studies. This influence has been manifested in an emphasis both on lexis in grammatical descriptions and on frequency of occurrence of language structures as an important factor in research studies. (See Barlow 1996 and Kemmer and Barlow 2000 for a description of grammatical models influenced by the analysis of corpus data.)

Since the use of learner corpora is a new development, many of the results must be regarded as preliminary until a wider range of learner corpora are available for analysis, covering a range of proficiency levels and a number of L1-L2 combinations. The existing learner corpora tend to contain little in the way of analytical markup, i.e., annotations of the raw corpus data that code grammatical information, which means that the form-function or form-meaning part of the analysis must be completed manually, with all the practical and theoretical problems typically encountered when assigning language forms to abstract categories. Still more problematic is the uncertainty about the exact nature of the relationship between particular learner corpora and a more general characterization of interlanguage. A learner corpus often represents just a single genre, such as argumentative essays, and so some features of the learner's production may be closely associated with that genre rather than being more generally representative of interlanguage. (See Dagneaux 1995.) Thus while we might expect some aspects of the learner's production to be relatively invariant over a range of modalities and genres, other aspects of production are likely to be highly modality-specific or genre-specific. But, at present, the required range of learner corpus types is not available to provide the information necessary to assess the variability of different aspects of language production.

Design and compilation of learner corpora

Whether a learner corpus is to be used to help identify those aspects of the student's language that are due to L1 influence or those aspects that are due to developmental processes, then clearly the variables associated with the learners' production must be systematically encoded. While all corpora must be well-designed and well-documented, the recording of data concerning the individual learners and the tasks and settings associated with the learners' language production is all the more important for learner corpora because of the central importance assigned to knowledge of the characteristics of the language learners. In short, we may not care about the background of writers employed by the *New York Times*, but we need to know some basic information about learners and about the conditions under which their language is produced, if we are to draw any useful generalizations related to second language acquisition.

The collection of data for a learner corpus typically involves the sampling of language production (i.e., speech or writing) along with descriptions of the setting and a description of the variables for each learner (Granger 2002:5), as shown in Table 14.1.

Setting	**Task:** A description of the nature of the task that provides the language sample. It could be a written prompt for an argumentative essay, a picture, or cartoon. Additional details may be furnished, depending on the particular nature of the task.
	Audience/Interlocutor: Identification of the person(s) interacting with the student, along with their role (teacher, tester, etc.).
	Time Limit: If the task is timed, what is the time allowed?
	Use of reference materials: Are dictionaries and other reference materials allowed?
Learner	**Mother tongue:** The primary language of the student.
	Other languages: Languages that the student knows with an assessment of

	competence with respect to speaking/writing/listening/reading
	L2 level of proficiency: An assessment of the level of the student. Such assessments are sometimes difficult to equate across institutions and across countries.
	Location: The country or region that the students come from.
	Education: This variable may include general information about education as well as an indication of the nature of language classes.
	Age:/Sex:/… and other attributes of the learner

Table 14.1: Examples of task and learner variables

These variables may be stored in external files or database with an ID number linking the information in the database with the language production data. Alternatively, information concerning variables can be encoded as tags or annotations within the corpus itself (Figure 14.1). The form in which this information is stored is not important in itself. The key point lies in structuring the data to allow retrieval software to reveal the links between a set of lexical or grammatical units in the learner corpus and the values of variables such as proficiency, age, L1, etc.

Database format

ID	Sex	Proficiency	Mother tongue	Occupation	Age	Learning Context
028	male	intermediate	Dutch	Student	16	school
029	female	intermediate	French	Student	17	school
030	male	advanced	French	Student	18	school
031	female	intermediate	Dutch	Student	17	school

Tag format

```
<student id="028">
<sex>male</sex>
<proficiency>intermediate</proficiency>
<language>Dutch</language>
<age>16</age>
<occupation>student</occupation>
<learning_context>school</learning_context>
</student>
```

Figure 14.1: Alternative representations of learner variables

The most extensive and best-known collection of learner corpora is the International Corpus of Learner English (ICLE), a project started by Sylviane Granger at the University of Louvain in 1990. The learner corpus which consists of essays of about 700 words produced by Advanced EFL learners in a variety of countries. The aim is to compile a corpus of about 200,000 words per language or country (Granger 1998:10–11); currently there are seventeen international partners participating in the project (Granger, Dagneaux, and Meunier 2002:11). The first release of the corpus on CD-ROM contains samples of English from Bulgarian, Czech, Dutch, French, Finnish, German, Italian, Polish, Russian, Spanish and Swedish learners of English, with data sets from Brazilian, Chinese, Japanese, Norwegian, Portuguese and South African learners to follow in subsequent

releases. The Polish component of ICLE, called PICLE, can be accessed through a web-based search engine described in more detail below.

The ICLE subcorpora are typically made up of academic essays produced within a classroom-based EFL learning context. A learner profile questionnaire is used to collect information on a range of variables, both a core set of variables and a set of variables relevant for some subcorpora only. Granger, Dagneaux, and Meunier (2002:13) list the variables shared by all the subcorpora in ICLE as follows:

Learner variables
- age
- learning context
- proficiency level

Task variables
- medium (for example, writing)
- field (for example, general)
- genre
- length

The information on the relevant variables is stored in separate files, leaving the texts themselves with very little mark-up. The ICLE texts are not annotated with part-of-speech (POS) tags, but some preliminary investigations on the feasibility of tagging of learner corpora have been carried out (de Haan 2000; Meunier and de Mönnink 2001). The difficulty in applying standard taggers arises from misspellings and unusual word sequences in learner corpora, but with further work on the development of appropriate taggers, we can expect to see more learner corpora annotated with POS tags in the future.

While most learner corpora are based on writing, typically essay writing, there are some spoken learner corpora. While not generally referred to as a learner corpus, the ESF (European Science Foundation) Second Language Database consists of spoken data collected in France, Germany, Great Britain, The Netherlands and Sweden (Feldweg 1991). For the project, the spontaneous productions of forty adult immigrant workers living in Western Europe were sampled and transcribed. There are five target languages: Dutch, English, French, German and Swedish, and for each target language, there are two source languages.

A more conventional spoken learner corpus is the LINDSEI Project (Louvain International Database of Spoken English Interlanguage) (De Cock, Granger, and Petch-Tyson 1995). The first component of the learner corpus contains spoken transcripts of fifty French learners of English, yielding a corpus of around 100,000 words.(1)

The Standard Speaking Test (SST) Corpus was started in Japan in 1999 with the aim of creating a one million word spoken corpus of Japanese learners of English (Tono, Kaneko, Isahara, Saiga, Izumi, Narita, and Kaneko 2001), which is due to be released in 2004. The corpus data are collected from students taking an Oral Proficiency Interview and thus is based on the use of interviews and picture prompts to elicit speech in a situation that approximates a natural dialogue. A smaller test-based spoken corpus collected at the University of Bergen contains the output of 62 Norwegian pupils aged 14-15 who were asked to complete a variety of tasks such as role-play and describing

pictures (Hasselgren 2002). These tasks were also carried out by 26 British pupils, providing an NS benchmark for the language of the Norwegian learners of English.

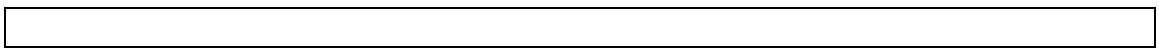
Commercially-based learner corpora, which are proprietary and thus generally unavailable to researchers, include the Longman Learner's Corpus and the Cambridge Learner Corpus. These corpora are large, about 10 million words each, and consist of the writings of a wide variety of students learning English around the world. The data in these corpora are analysed by lexicographers and materials developers in order to improve the usefulness of dictionaries and coursebooks for language learners.

There has been a considerable amount of work related to learner corpora in Hong Kong, such as the 25 million-word HKUST Learner corpus compiled by John Milton, the TeleNex Student Corpus, and the Chinese Learner English Corpus (CLEC). Other learner corpora described in Pravec (2002) and Nesselhauf (In press) include the Uppsala Student English project, which contains writings of Swedish undergraduates; the Learner Business Letters Corpus, consisting of business letters written by Japanese business people; and the Montclair Electronic Database (MELD), which is unusual in that it contains the language production of ESL learners rather than EFL learners.

Following in the EA tradition, one approach to the analysis of learner corpora is to annotate the learner language with error tags (Milton and Chowdhury 1994; Dagneaux, Denness and Granger 1998). An error tag is an annotation added to the corpus to explicitly mark an error, as in 'The main feature of campus is (GA) the its conviviality' (Dagneaux, Denness, Granger, and Meunier 1996). In this example, the error tag GA is used to indicate a problem with article use and the target form 'its', signalled by $, has also been added by the annotator.

The identification of errors is, however, not at all straightforward since a sentence with errors can often be corrected in multiple ways, making the labelling of individual errors within the sentence quite difficult, as discussed in some detail in Chapter 3. For an error-tagged corpus to be useful for research purposes, clearly the assignment of error tags has to be consistent. The difficulties of tagging errors is compounded by the fact that following the corpus tradition of more is better, learner corpora tend to be quite large, perhaps containing millions of words. Such large data sets make error identification a time-consuming process, even with the aid of error taggers, since each error must be individually located and classified by a researcher.

An error-tagged learner corpus enables researchers to search for different types of error and may also allow searches of correct target forms in addition to the actual, possibly erroneous forms. As noted above, the marking of errors is not at all straightforward and involves some interpretation. A hierarchical error tagging system, developed at Louvain (Dagneaux, Denness, Granger, and Meunier 1996) assigns a major category type to each error: grammatical (G), lexical (L), lexico-grammatical (X), formal (F), register (R), syntax (W) and style (S). Additional specifications include GV (grammatical verb error), GVAUX (grammatical auxiliary verb error), and GVT (tense error), among others. Documentation is available that defines and illustrates all the error codes so that the coding performed by different researchers is consistent. The extract from the Error Tagging Manual (Version 1.1) shown in Figure 14.2 illustrates the entry for the tagging of missing words.



5.3 WORD MISSING: (WM)

This subcategory is for errors involving the omission of words
e.g. *We have a meeting* (WM*) 0 on Monday*.
I do not (WM*) 0 see why I should go there.*

The following types of errors should not be classified in this category:
(LP): where the missing word forms part of a set phrase:
e.g. (LP) *on equal footing $on an equal footing$*
(GA): for a missing article or misapplication of the zero article:
e.g. (GA) *He works in* (GA) *0 an office.*
(LC): missing connective or a word missing from a connective:
e.g. (LCLC) *on other hand $on the other hand$*

Figure 14.2: An extract from the Error Tagging Manual (Dagneaux, Denness, Granger, and Meunier 1996)

The work on error-coding on the ICLE Project informed the development of markup software, called *TagEditor*, for use with the SST Corpus (Tono, Kaneko, Isahara, Saiga, Izumi, Narita, and Kaneko 2001). The software is used to facilitate the marking of errors, using XMLtags (2) and to search the corpus in different ways.

Milton and Chowdhury (1994) describe the error-tagging of a corpus of Chinese learners of English. They make the interesting suggestion that the problem of multiple analyses of an error should, in some cases, be dealt with by encoding multiple alternative corrections within the tagging scheme (1994:129).

Error tags are the most common type of annotation used to mark up learner corpora, but any aspect of linguistic structure can be coded explicitly using an appropriate annotation scheme. Burdine (2002), for example, annotated a corpus comprising interviews with French immersion students in order to show the occurrence of different communication strategies. In the following exchange, for example, a foreignization strategy <FOR> is followed by an unsuccessful attempt at correction <OC>: *c'est c'est facile à* <FOR> *marquer* </FOR> <OC> *à correcter* </OC>. Having identified the different communication strategies used by the students, Burdine is able to assess the relative frequency of different strategies as well as show the relationship between preferences for particular communication strategies and other parameters such as proficiency.

In this section we have focussed on the structure of learner corpora, but as Altenberg (2002:38) points out, in order to investigate and understand the nature of interlanguage, information must be gathered on not only the form and functioning of the interlanguage, but also the features of the learner's mother tongue and of the target language. Thus learner corpora, monolingual corpora and bilingual or parallel (translation) corpora all have a part to play. Parallel corpora can be used to establish equivalences or non-equivalences in lexis or grammar holding between two languages. For example, the word *line* in English and *ligne* in French take part in a variety of metaphorical extensions and occur in different phrases. Some of these uses are congruent in the two languages, while others are not. Using a parallel corpus, we can find, for instance, those uses of *line* in English which typically do not correspond to *ligne* in French. These include: *in line with, bring/fall into line*, *poverty line*, *down the line, the bottom line*, and *line of*

reasoning/*argument/thought.* Data such as these provide background information on equivalences and non-equivalences between two languages, which can be used as part of the evaluation of the language of French learners of English and that of English learners of French.

Research on learner corpora

Research on learner corpora is often inherently contrastive and to some extent it follows some of the general concepts and aims associated with **c***ontrastive analysis*. Granger (1998:12) refers to a new research paradigm of *contrastive interlanguage analysis*, which covers both NS/NNS and NNS/NNS comparisons. Studies which use native speaker corpora as a benchmark for the analysis of learner corpora (i.e. NS/NNS comparisons) provide evidence for the nature of interlanguage, focusing on the non-native aspects of learners' speech or writing. Alternatively, a comparison of different NNS corpora can be used to highlight aspects of language use and development shared by learners with different language backgrounds. In cases where differences emerge among learners with different language backgrounds, the analyst will explore the likelihood that the variation is due to L1 influence (Granger 2002:13). In practice, many studies of learner corpora are designed to include both NS/NNS and NNS/NNS comparisons.

A comparison of learner corpora with NS corpora provides data on the properties of interlanguage, covering features which are typically overused or underused, in addition to those which are misused by language learners (Leech 1998:20). Thus learner corpora studies often involve the counting of particular words or grammatical categories, a process which is not as simple as it sounds because of the ill-formed or variable nature of L2 production data. A further source of complexity is that automated counting routines are based on formal identity of linguistic items, not form-function identity. For instance, a count of the word *can* in an untagged corpus does not discriminate between *can* as a noun and *can* as a modal auxiliary. Furthermore, in some studies an even finer-grained functional categorization may be necessary to distinguish, for example, the ability uses of the modal *can* from the permission uses. The tabulation of form-function linguistic items in learner corpora can be time-consuming in cases where the corpora are not already annotated for the categories of interest. Yet such fine-grained analyses are often needed to give an accurate picture of the nature of interlanguage. For example, an investigation of essay-writing by Lin (2002) showed that the word *it* was used less frequently by Chinese learners of English than by native speakers. But when the different functions of *it* were examined, it turned out that some functions of *it* were underused by language learners (for example, *tough*-constructions such as *It is easy to please John*), while some were overused (for example, dummy *it*).

Once corpus-based data on particular characteristics of interlanguage have been analysed, it is possible to look for explanations for these features, which typically involve factors such as:

L1 transfer
Some forms or grammatical patterns found in the learner's language production may result from the intrusion of L1.
general learner strategies

To help deal with the complex task of speaking or writing in a second language, the learner may adopt some coping strategies such as the use of L1 forms, circumlocution, avoidance strategies, etc.

paths of interlanguage development

Some aspects of interlanguage, such as the development of negation or the development of tense/aspect marking proceed in a series of stages which may be tracked using longitudinal studies of learner output.

intralingual overgeneralization

Some features of the learner's language may be due to overgeneralisation of an aspect of L2 grammar such as the use of *-ed* to mark past tense.

input bias

The form of the learner's production may reflect the particular input received, such as the language used in coursebooks. (See Römer forthcoming.)

genre/register influences

Researchers working with learner corpora have suggested that the writing of L2 learners contains a variety of informal patterns that are characteristic of spoken discourse.

We can distinguish two principal methodologies associated with learner corpus investigations. One is to use learner corpus data to test specific hypotheses about the nature of interlanguage generated through introspection, SLA theories, or as a result of the analysis of experimental or other non-corpus-based sources of data. In this case, text analysis software is used to extract data in a way that specifically relates to the hypothesis; the retrieved data are viewed as valuable only in so far as they confirm or disconfirm the hypothesis. One such study is exemplified in Tono (2000). His hypothesis is that the overall morpheme ordering data revealed by Dulay and Burt (1974) will be confirmed by an analysis of a written corpus consisting of English essays written by Japanese speakers. In his study, Tono found some differences in the morpheme acquisition order for Japanese learners in that the possessive *–s* morpheme was acquired relatively early and article usage emerged relatively late. Another instance of hypothesis-testing is Housen (2002) which aims to investigate the Aspect Hypothesis put forth in Andersen and Shirai (1996) and Bardovi-Harlig (1999). See below for further discussion of Housen's study.

An alternative is to investigate learner corpora data in a more exploratory manner and initiate analyses that yield patterns of data, which can then be inspected for unusual features. Such features may then be used to generate hypotheses about learner language. This general approach is illustrated in the section below on Corpus Analysis Software and in work such as Aijmer (2002). In her study, Aijmer starts out from the general observation that non-native speakers find it difficult to use English modal verbs appropriately. She then exploits different corpora to compare modal use by Swedish learners of English and by native-speakers and finds that there is a general overuse of modals by Swedish learners and a particular overuse of the modals *will, must, have (got) to, should* and *might*.

These two methodologies correspond loosely to the contrast between *hypothesis-driven* and *hypothesis-finding* approaches (Granger 1998:15), and to the general *corpus-based* versus *corpus-driven* distinction (Tognini-Bonelli 2001). In a corpus-based approach, a search is selected to find data that are relevant to a particular hypothesis. On

the other hand, in a corpus-driven approach large amounts of data derived from corpus analysis are used in the formulation of grammatical descriptions.

In practice, researchers may well make use of a combination of approaches, but there are biases in practice such that broadly speaking the experimental/generative tradition favours hypothesis-driven, corpus-based approaches, while corpus linguists have a preference for a hypothesis-finding, corpus-driven methodology.

Table 14.2 provides a schematic overview of how a hypothesis-driven analysis of learner corpora might proceed; it is based on a study in Tono (2000).

1 Initial hypothesis	An analysis of a learner corpus will support the morpheme order results of Dulay and Burt.
2 Corpus selection/ compilation	Selection or compilation of learner corpora to use.
3 Preliminary data analysis	Tagging of relevant morphemes in the learner corpus. (e.g. *I have hardly had <ART> a </ART> bad dream.*)
4 Further data analysis	Searching for all instances of each morpheme tag and marking errors manually. (e.g. *Do I see <ER_ART> the </ ER_ART> movies too much?*) Assessment and computation of frequency of errors, following Dulay and Burt's methodology.

Table 14.2: A hypothesis-driven learner corpus study

Table 14.3 provides a schematic overview of the stages in a topic-driven analysis of learner corpora.

1 Initial topic	Example: The use of hedges by language learners. (The identification of items of potential interest may be based on a preliminary learner corpus analysis as described below in Preliminary data analysis.)
2 Corpus selection/ compilation	Selection or compilation of learner corpora to use. Depending on the nature of the investigation, other types of corpora may be used: an NS corpus, a bilingual (translation) corpus, or a textbook corpus. See Contrastive studies below.
3 Preliminary data analysis	Analysis or identification of the item in a learner corpus: using word frequency lists, n-grams (e.g. 3-word sequences ordered by frequency), collocation lists, etc.
4 Further data analysis	Use of concordance searches on target forms to show the forms in their linguistic context. Sorting concordance lines may serve to group similar uses together. This format facilitates the identification of the range and frequency of form-function mappings. Coding of target forms.

5 Contrastive studies	Evaluation of the patterning revealed by learner corpus data analysis, typically based on one or more of the following: – Comparison of learners with different L1 – Comparison of learners in different instructional situations – Contrast of learner language with an L2 reference corpus – Comparison of different levels of proficiency – Longitudinal (or cross-sectional) analysis based on learner corpora – Evaluation of learner language based on L1-L2 equivalences revealed by analysis of a L1-L2 bilingual corpus and/or by analysis of L1 and L2 monolingual corpora – Comparison of different modalities or genres – Comparison with corpora representing input to the learners (textbooks, classroom talk, etc.)

Table 14.3: A hypothesis-finding, corpus-driven learner corpus study

The most common form of learner corpora research involves contrasting an NNS with an NS corpus or with a corpus-based reference such as the Longman Grammar of Written and Spoken English (Biber, Johansson, Leech, Conrad, and Finnegan 1999). If a learner corpus is to be contrasted with an NS corpus, then a variety of issues arise, as they always do when corpora are compared. It is always possible to find fault with research relying on the comparison of corpora that have been compiled by different groups for different purposes. In such situations the corpora necessarily differ along several dimensions, making comparability of linguistic features open to question. Given that there is no perfect benchmark, an appropriate benchmark must be found for each study. One issue concerns the variety of NS English to be used. Is the reference corpus based on British, American or Australian English, for example? More important perhaps is the question of text type. A corpus such as the Brown corpus (Francis and Kucera 1967) or the American National Corpus (Ide, Reppen, and Suderman 2002) contain such a wide variety of different types of language that they are probably unsuitable for many studies (Granger and Tyson 1996). The problem here is that the combination of genres in the general corpus does not provide a good reference point for the learner corpus, which invariably consists of a single genre. Thus, using a general corpus introduces an additional variable, as any comparisons made would be based on one genre versus multiple genres as well as on learner language versus non-learner language.

In cases where the reference corpus is based on a single text type, such as a newspaper, variability in language use still occurs due to the styles of different writers and differences associated with different sections of the newspaper. Nevertheless, the variability is much reduced and the writing in a newspaper may be taken as a target that

student writers may aim for. An alternative option is to use a reference corpus that is as close as possible in genre and other dimensions to the learner corpus. In fact, most studies of ICLE subcorpora use the LOCNESS corpus as a benchmark. This corpus consists of essays written by British and American undergraduates.

Comparisons of learner corpora with a reference corpus have been carried out on a variety of lexical and grammatical topics: complement clauses (Biber and Reppen 1998), direct questions (Virtanen 1998), causatives (Altenberg 2002), tenses (Granger 1999; Housen 2002), modals (Aijmer 2002; McEnery and Kifle 2002), hedges/certainty markers (Flowerdew 2000, Milton and Hyland 1997), adjective intensifiers (Lorenz 1998), formulae (de Cock 1998) and connectors (Altenberg and Tapper 1998).

A learner corpus study

An example of learner corpus analysis is Housen's 2002 study of the development of the English verbal system. This particular study is a quite complex analysis of the formal and functional development of the verbal system in learners. One goal of the investigation is to determine the applicability of the Aspect Hypothesis (Andersen and Shirai 1996; Bardovi-Harlig 1999), which suggests that the initial uses of verb morphology are constrained by the inherent semantics of the verbs used. Thus following this hypothesis, we might, for instance, expect the learners to initially use the morpheme *–ing* solely with activity verbs.

The patterns of language emerging from the study of actual production data reflect a variety of influences, including language processing, L1 influence, conceptual predisposition and frequency of forms in the input (Housen 2002:108). The study is summarized in Table 14.4.

Research question	Housen (2002) investigates how second language learners of English acquire the forms and functions of the English verbal system.
Participants	23 Dutch-speaking and 23 French-speaking students , 9-17 years old. Also 8 native speakers, aged 11-13.
Data collection	The data were collected by interview and semi-guided speech tasks. The recorded data were transcribed, segmented, and annotated, forming the Corpus of Young Learner Interlanguage.
Analysis	1 The verbs were coded for morphosyntactic form (e.g. *-ing, -s, -ed, en*), agreement values, tense (past, present, etc.), aspect (imperfect, progressive, etc.), and inherent aspect (state, activity,etc.) 2 The learner transcripts were grouped according to proficiency level: Low, Lower Intermediate, Higher Intermediate and High. 3 The clauses in the transcripts were analysed for the underuse/overuse of inflectional verb categories (*V, Ving, Ved, Vs*). Underuse in this study measures the instances in which a particular form is omitted from an obligatory context. Overuse refers to instances of use of a form in inappropriate contexts.
Results	Based on the data, Housen described three formal stages. Stage 1. Invariant default forms. Verbs appear as invariant forms, typically the unmarked base form, but high frequency irregular *Ven* forms (e.g. *got*) also occur. Stage 2. Non-functional variation. The order of emergence of forms

	is *V0* > *Ving*; *was* > *Ven* > *Ved*; *going+Vinf* > *have+V*; *Vs*; *will+V.* Stage 3. More target-like use of verb morphology to encode tense, aspect and agreement. The patterns of underuse and overuse decrease with increasing proficiency, although there is still variation among different verb forms. The results support the predictions of the Aspect Hypothesis for development of the *Ving* form, but not for the *Vs* form.
Discussion	The results reveal general patterns in the development of the English verbal system and also the variability in development.

Table 14.4: A study of the development of the English verbal system

Housen's study can be seen as being in the tradition of research on the order of acquisition of morphemes, but in a corpus-based study such as this one considerable coding is required to identify the different verb forms embedded in the mass of corpus data. It is worth pointing out that in this particular study, the development of the verb by individual learners is not tracked and hence it is not possible to assess the variability in the developmental sequences followed by individuals.

Explaining patterns in learner corpora

Determining the source of the patterns detected in corpus analyses presents considerable difficulties. In some cases, researchers feel confident in attributing the patterns of interlanguage to transfer from L1, especially in cases of lexical patterns. For instance, Lu (2002: 51) found in a comparison of noun compounds used in a Chinese learner corpus (CLEC) that the phrase *we college students,* a translation of *wo men da xue sheng,* was used quite frequently.

The assessment of the source of grammatical patterns is much more difficult. In her investigation of modals in a learner corpus, Aijmer (2002: 60) points out that the overuse or underuse of particular modals may be due to L1 influences or to general learner strategies, but may also relate to the different distribution of modals in spoken and written modalities. Several researchers have found that the writing of English learners had the characteristics of more informal, spoken usage. For instance, Altenberg and Tapper (1998) found that formal connectors such as *therefore* and *thus* were underused, while more informal markers, *but* and *still,* were overused.

Altenberg (2002) undertakes an interesting contrastive study using parallel English-Swedish texts to analyse the range of causative constructions (for example, *make* causative, synthetic causatives, and other construction types) and to assess how the causative constructions correspond to each other in the two languages. As a result of his investigations on translated texts, Altenberg found that the English *make* causative is generally equivalent to the Swedish *göra* construction. Importantly, however, English *make* is in competition with a variety of other causative constructions, and many of the *göra* uses in Swedish are not translated using causative *make* in English. Thus, while there is a general equivalence between the two analytic causatives in English and Swedish, there are also some important differences which become apparent only after a detailed analysis of parallel texts.

What do Swedish learners of English do with respect to the production of English causative constructions? Not surprisingly, given the facts discussed above, Swedish learners of English tend to overuse the *make* causative, which Altenberg (2002:52) suggests is due to transfer propelled by cross-linguistic similarity. He proposes that Swedish learners see the similarity between the prototypical causative construction in Swedish and the English construction, and tend to use *make* causatives in a way that mimics the wide functionality of the *göra* causative. This contrasts with the behaviour of French learners of English, for example, who do not overuse the English *make* causative.

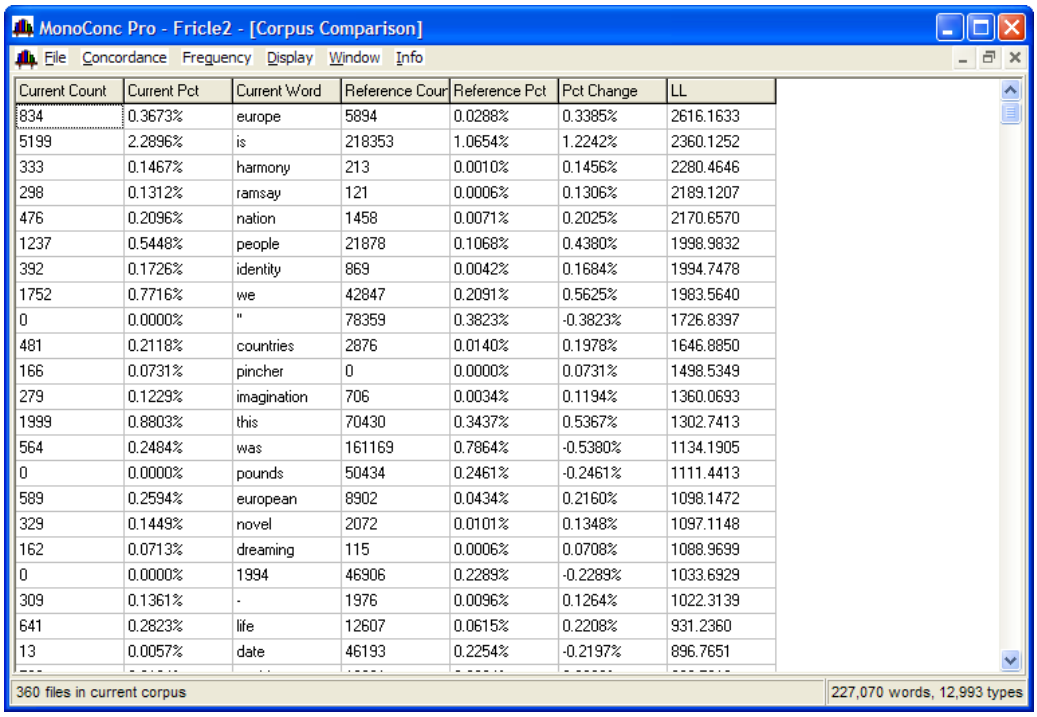
Like NS–NNS studies, the analysis of NNS–NNS contrasts provides evidence of L1 influence on learner output, but such studies also provide evidence of general learner strategies, such as simplification, and other general aspects of L2 development that are unrelated to L1. Thus NNS–NNS comparisons can be seen as a way to increase our understanding of the characteristics of interlanguage and assess the influence of particular variables on the form of interlanguage. Aijmer (2002: 57) cautions that the results from a single study which point to the existence of learner strategies need to be followed up with more extensive investigations on a range of learners and on different types of data.

In order to fully understand the process of second language acquisition, it is necessary to trace the interlanguage of individual learners over time. In other words, it would be beneficial to carry out longitudinal studies to complement the kinds of investigations described above. Many learner corpora contain data from students at different proficiency levels, which can be used to suggest hypotheses about the paths of language development. Such 'quasi-longitudinal' data studies (Granger 2002: 11) can be checked using longitudinal corpora in which the progress of individual students can be tracked. Research by Housen (2002: 95), described above, suggests that the aggregate data on the development of the third-person singular morpheme *-s* broadly follows the patterns emerging from a longitudinal study (Housen 1995; 1998). It is also possible, however, that the aggregate view offered by the corpus as a whole will mask changes in the language of individuals or the differences in individual development paths. These longitudinal or quasi-longitudinal studies (Housen 2002; Tono 2000) have the potential to provide corpus evidence relating to the morpheme acquisition study of Dulay and Burt (1974). (See Chapter 4.)

Corpus analysis software

An example of a common type of data used in corpus linguistics is the word frequency list, which can easily be generated for any learner corpus and the results inspected for unusual patterns. But rather than simply examine a word frequency list of a learner corpus, we can compare the frequency of words in a learner corpus with a reference corpus, which might be another learner corpus or a native speaker corpus.

The screen shot in Figure 14.3 illustrates the data produced by the 'corpus comparison' command in the software program *MonoConc Pro* (Barlow 2002). In this example, the frequency of the words in a corpus, the French learners component of ICLE, are compared with the frequency of words in a reference corpus, in this case a corpus based on *The Times* newspaper.



Current Count	Current Pct	Current Word	Reference Cour	Reference Pct	Pct Change	LL
834	0.3673%	europe	5894	0.0288%	0.3385%	2616.1633
5199	2.2896%	is	218353	1.0654%	1.2242%	2360.1252
333	0.1467%	harmony	213	0.0010%	0.1456%	2280.4646
298	0.1312%	ramsay	121	0.0006%	0.1306%	2189.1207
476	0.2096%	nation	1458	0.0071%	0.2025%	2170.6570
1237	0.5448%	people	21878	0.1068%	0.4380%	1998.9832
392	0.1726%	identity	869	0.0042%	0.1684%	1994.7478
1752	0.7716%	we	42847	0.2091%	0.5625%	1983.5640
0	0.0000%	"	78359	0.3823%	-0.3823%	1726.8397
481	0.2118%	countries	2876	0.0140%	0.1978%	1646.8850
166	0.0731%	pincher	0	0.0000%	0.0731%	1498.5349
279	0.1229%	imagination	706	0.0034%	0.1194%	1360.0693
1999	0.8803%	this	70430	0.3437%	0.5367%	1302.7413
564	0.2484%	was	161169	0.7864%	-0.5380%	1134.1905
0	0.0000%	pounds	50434	0.2461%	-0.2461%	1111.4413
589	0.2594%	european	8902	0.0434%	0.2160%	1098.1472
329	0.1449%	novel	2072	0.0101%	0.1348%	1097.1148
162	0.0713%	dreaming	115	0.0006%	0.0708%	1088.9699
0	0.0000%	1994	46906	0.2289%	-0.2289%	1033.6929
309	0.1361%	-	1976	0.0096%	0.1264%	1022.3139
641	0.2823%	life	12607	0.0615%	0.2208%	931.2360
13	0.0057%	date	46193	0.2254%	-0.2197%	896.7651

Figure 14.3: The comparison of words in a learner corpus and reference corpus

The columns in the table in Figure 14.3 provide different kinds of data related to the words in the learner corpus. The leftmost column gives the count or frequency of each word in the learner corpus. In the next column these data are expressed as a percentage, i.e., the frequency divided by the total number of words in the corpus multiplied by 100. The third column lists all the words occurring in the learner corpus or reference corpus. The next two columns give the count and percentage for the word in the reference corpus.

The two rightmost columns provide comparative information for each word: Percentage Change and the Log Likelihood (LL) value, a statistical measure of difference (Rayson and Garside 2000) of words in two corpora. (3) Percentage change is simply (Percentage in current corpus) – (Percentage in reference corpus). A positive value for Percentage change indicates that the word is over-represented in the learner corpus; a negative value indicates under-representation.

The words shown in the screen shot in Figure 14.3 do not constitute a random selection; they represent those words which, according to the Log Likelihood value, are most distinct, based on a comparison of their occurrence in the two corpora. The next step is to examine the list more closely in order to eliminate those words that are not associated with learner language. The word *pounds*, for instance, does not occur in the learner corpus, but this under-representation occurs because the French learners of English happen not to be writing about British currency or weight. The words, *europe, harmony*, and *nation*, for example, are over-represented in the learner corpus, but this is due to the essay topics assigned to the students and again we can ignore such words.

After removing words which are not of interest, we are left with the selected data in Table 14.5. All the words listed in this table are significantly over-represented in the learner corpus and are candidates for follow-up investigations. We find various forms of

the copula, as well as connectors such as *moreover, nowadays, because*, and *indeed*, and the modals *will* and *can*.

Word	Percentage Difference	Log Likelihood
is	1.22%	2360.12
people	0.43%	1998.98
we	0.56%	1983.56
this	0.53%	1302.74
will	0.40%	838.57
our	0.21%	811.08
they	0.37%	693.85
are	0.41%	677.17
moreover	0.05%	587.99
be	0.42%	577.98
not	0.37%	574.91
also	0.21%	572.66
can	0.23%	560.46
nowadays	0.04%	471.06
because	0.16%	452.60
think	0.11%	442.59
of	0.77%	419.51
to	0.70%	397.25
their	0.25%	378.56
it	0.36%	365.15
these	0.12%	328.59
indeed	0.06%	325.67
does	0.09%	319.61
that	0.38%	316.85

Table 14.5: Words over-represented in a learner corpus

It goes without saying that results obtained in this manner are merely suggestive. The *Times* corpus is being used here as a representative of a target language, but clearly there are genre differences between a newspaper corpus and academic essays, and it may be these genre differences which are behind the difference in relative frequencies of these words. This may be the case for *is*, for example, but it may also be a fact that the French learners do overuse copula sentences or phrases such as *it is* or *there is* in their writing. Thus for each of these words, further investigations are needed in order to determine whether the overuse is a reflection of frequent syntactic or discourse patterns or whether it is simply a matter of lexical choice. Once this is determined, it is possible to take the next step and investigate possible reasons for the patterns found.

Many variations on word counts are possible. For example, it may be profitable to count n-grams, for example, word pairs or word triples, etc., to see whether certain sequences appear to be overly frequent. (4) To briefly illustrate these kinds of data, Table 6 shows the rank order of the ten most frequent word pairs from the French, German, and Czech components of ICLE, and again these results are compared with data from the *Times* newspaper corpus.

French	German	Czech	Times

of the	*of the*	*of the*	*of the*
in the	*in the*	*in the*	*in the*
to the	*to be*	*it is*	*to the*
to be	*to the*	*to be*	*on the*
it is	*it is*	*do not*	*for the*
of a	*on the*	*to the*	*to be*
is a	*have to*	*It is*	*and the*
on the	*of a*	*is the*	*at the*
will be	*in a*	*is a*	*that the*
is the	*and the*	*is not*	*of a*

Table 14.6: Frequent bigrams in learner corpora and a newspaper corpus

Note that the same two most frequent bigrams occur in all the subcorpora and it is only in the third row that some differences start to appear. Looking at the table as a whole, we can see that further analysis of the bigram *will be* from the French subcorpus, the bigram *have to* from the German subcorpus, and the use of the copula in the Czech subcorpus might be fruitful as these results suggest an overuse of these forms in the learners' English. This methodology can be extended to trigrams, tetragrams, and so on. It should be noted, however, that the larger the n-gram, the more idiosyncrasies appear, due to the particular content being described. (5) It is also possible to examine POS tag sequences rather than word sequences (Aarts and Granger 1998).

Another variant on word counts is shown in Table 14.7. This table illustrates the most common sentence-initial words in three learner corpora and in *The Times* newspaper corpus. We find that the same set of words tend to be used sentence-initially in all four subcorpora, but the use of *as* in the French subcorpus and the use of *if* in the German subcorpus stand out.

French		**German**		**Czech**		**Times**	
The	10.2%	*The*	7.5%	*The*	7.5%	*The*	11.4%
In	5.0%	*I*	5.9%	*It*	7.2%	*It*	3.3%
This	4.9%	*But*	5.0%	*They*	5.8%	*He*	3.1%
But	4.2%	*It*	4.0%	*I*	4.8%	*But*	2.8%
It	4.0%	*In*	3.2%	*But*	3.5%	*In*	2.5%
He	3.4%	*They*	2.8%	*We*	3.5%	*I*	1.8%
They	3.1%	*And*	2.1%	*In*	2.8%	*A*	1.6%
We	2.8%	*This*	2.0%	*And*	2.6%	*They*	1.3%
As	2.1%	*There*	1.6%	*There*	2.6%	*This*	1.3%
She	1.6%	*If*	1.5%	*He*	2.2%	*In*	1.0%

Table 14.7: Common sentence-initial words in learner corpora and a newspaper corpus

Following up on the data in Table 14.7, we can look at the most common sentence-initial word pairs and we find *as a* in the French subcorpus and *if you* in the German subcorpus (Table 14.8).

French	**German**	**Czech**	**Times**
It is	*It is*	*It is*	*It is*
On the	*In the*	*They are*	*In the*
This is	*There is*	*There is*	*It was*

As a	*There are*	*There are*	*He was*
In the	*On the*	*I think*	*There is*
He is	*When I*	*On the*	*But the*
In this	*This is*	*In the*	*This is*
There is	*It was*	*We can*	*It is*
Let us	*If you*	*This is*	*He is*
They are	*They are*	*It was*	*Yours faithfully*

Table 14.8: Common sentence-initial bigrams in learner corpora and a newspaper corpus

The advantage of performing these preliminary kinds of text analysis based on word frequency measures is that few, if any, assumptions are made about the nature of learner language. An initial examination of the results may reveal interesting data and can suggest hypotheses, which can be pursued by follow-up studies.

Different kinds of text analysis provide different views of language data (Barlow In press) and one obvious problem with wordlists is that all the context of the words has been removed. Once a word or phrase has been selected for further study on the basis of wordlist data, the next step is typically to perform a KWIC (keyword in context) concordance search. As noted above, any of the words in Table 14.5 might be selected for a follow-up study. In this case, we can choose *indeed* and perform a concordance search for the word, as illustrated in Figure 14.4. Such a search will reveal all the instances, 81 in this case, of the keyword (or phrase) within its linguistic context, allowing for further classification on the basis of meaning, function, error type, etc. In other words, the text on either side of the keyword is used to classify each instance based on an error analysis or based on form-function mappings or whatever other level of analysis is deemed necessary. Such an analysis is necessary to determine which of the uses of *indeed* are being overused by the French learners of English. If the narrow context of a few words either side of the keyword is not enough for these classificatory purposes, then a wider context can be obtained by clicking on a particular line, as shown in Figure 14.4.

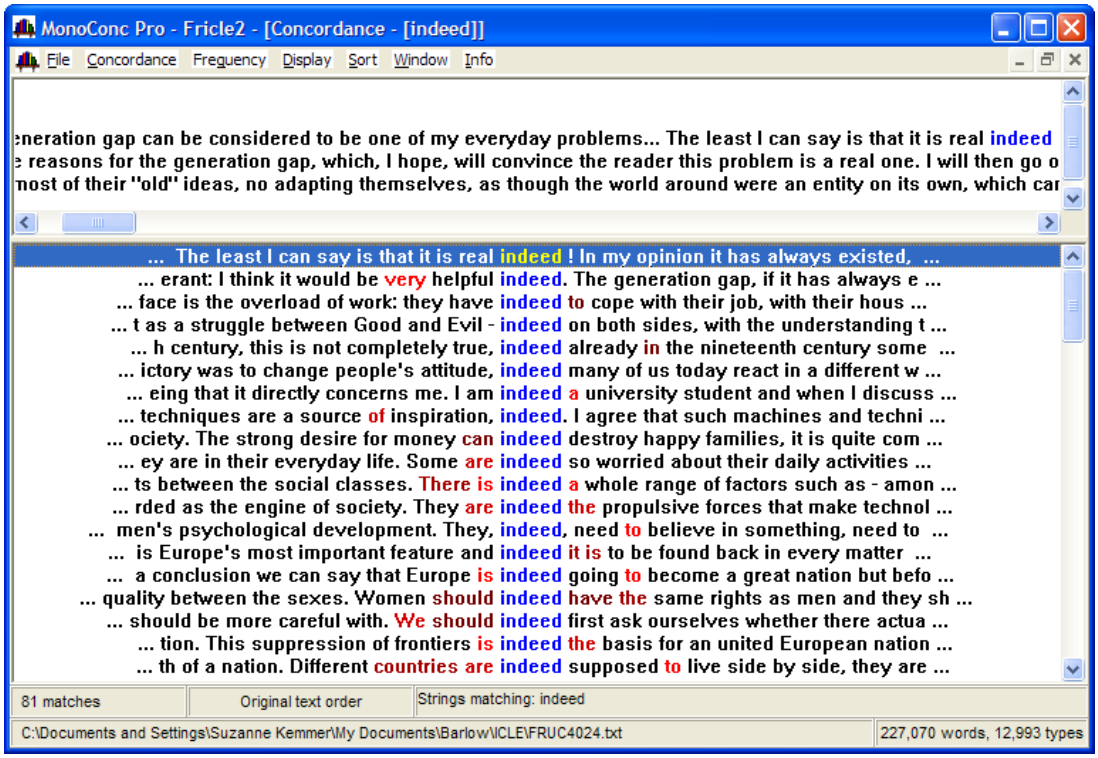


Figure 14.4: A concordance search for indeed *in a learner corpus*

The analysis of concordance results is often aided by the ability to re-sort the concordance lines based, for instance, on the word preceding or the word following the keyword. Rearranging the lines in this way often highlights regularities in the data because repeated patterns stand out visually and this may make the classification of the concordance lines easier. The screen shot in Figure 14.5 shows the results of a search for sentence-initial *It is*. The concordance lines have been sorted 1st right, 2nd right, which alphabetizes the lines based on the word following *It is* and in those cases where the word is repeated, the lines are ordered secondarily according to the alphabetical order of the second word following *It is*.

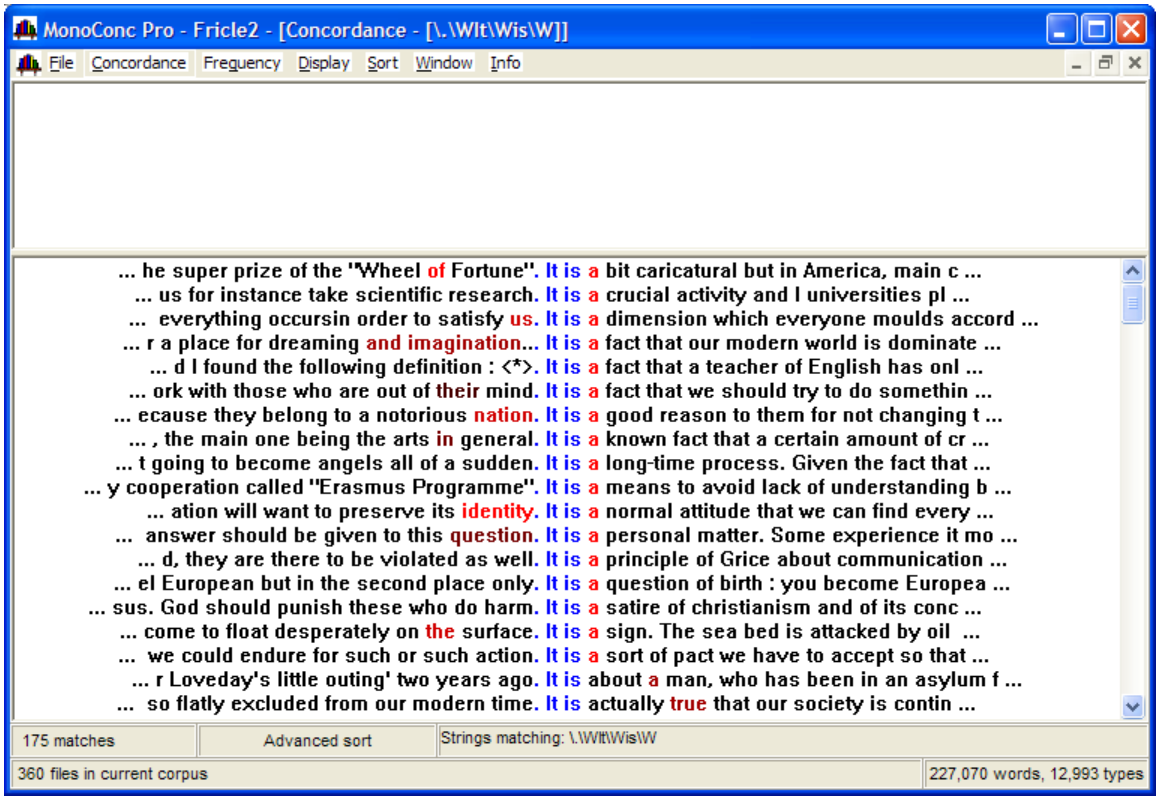


Figure 14.5: A concordance search for It is *sorted 1st right 2nd right.*

In this section, we have discussed two uses of text analysis software. One is to generate wordlists and similar data structures that may reveal interesting patterns in the language used by learners. The second is to use concordance searches and sorting as an aid for in-depth analyses or classifications of errors, form-function mappings, or other linguistic functions. The fact that samples of learner performance data are stored in computer-based digital format means that it is simple to search for and extract particular lexical items (for example, *nevertheless, on the one hand*) and grammatical forms (for example, *have* + participle), along with as much linguistic context as is desired. Searches can be performed fairly easily and quickly, making it feasible to carry out multiple exploratory investigations.

Task: Analysing a learner corpus

You can examine the Polish component (350,000 words) of ICLE at the website http://elex.amu.edu.pl/~przemka/concord2/search.html

A preliminary analysis of Polish learners of English shows that the sentence-initial phrase A*s a* is used quite frequently, as it is in French. Analyse the use of the sentence initial phrase *As a* in this Polish learner corpus to find out if the patterns of use are similar to those found in the French data where three main phrases used were *as a conclusion* (41%), *as a matter of fact* (29%) and *as a result* (10%).

To perform the analysis, carry out the follow steps.

1 Perform a KWIC search for *As a (*i.e. *As a* followed by space*),* remembering to search for *As a* , not *as a* .
2 How many instances occur in the corpus?
3 List the main sentence-initial phrases based on *As a* (*As a result*, etc.). Give the percentages for these phrases.
4 How do these percentages compare with the results for French, given above?
5 Describe how you would construct a follow-up investigation to test for overuse or underuse of these phrases.

Final comment

Learner corpora potentially provide a very rich source of data, which may be used to overcome some of the problems with EA noted in Chapter 3, such as the inability of the method to account for learners' avoidance of L2 forms. On the other hand, the size and complexity of learner corpus data means that in the absence of automatic analysis software, researchers must perform a considerable amount of manual coding of errors or form-function categories.

Most of the existing learner corpora are based on the writing of fairly advanced language learners. In order to play a central role in understanding SLA a wider range of learner corpora, including spoken learner corpora, will have to be created. As Granger (forthcoming) notes, 'learner corpora should not be seen as a panacea, but rather as one highly versatile resource which SLA/FLT researchers can usefully add to their battery of data types'.

Notes

1 Currently four subcorpora are complete (French, Chinese, Italian, Japanese) and a CD-ROM release of LINDSEI (with audio-synchronization) is planned for 2005. See http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl–Projects/Lindsei/lindsei.htm
2 XML stands for Extensible Markup Language and is similar in many ways to the html tags used in webpages. See www.w3.org/ XML for details. An example of a sentence from a learner corpus with an error tagged using XML is given below.
The main feature of campus is <GA> the "Corr=it"</GA> conviviality.
3 This statistic is similar to the better-known chi-squared test, but is more reliable for low scores. The test compares the frequency of word A in corpus 1 with the frequency of word B in corpus 2, based on the assumption that the word occurrences follow a binomial distribution.
4 See De Cock, Granger, Leech, and McEnery (1998) for an analysis which takes distribution as well as frequency into account.
5 The n-gram data were produced by the software program *Collocate* (Barlow 2003), but it is also possible to write simple Perl scripts or use UNIX commands such as *sort*, *uniq*, *tail*, and so on, to produce such lists.

Further reading

Granger, S. (ed.) (1998) *Learner English on Computer*. Addison Wesley Longman, London and New York.
Granger, S., Hung, J. and Petch-Tyson, S. (eds) (2002) *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.