

**DRAFT****Computer Learner Corpora and Monolingual Learners' Dictionaries:  
the Perfect Match**

Sylvie De Cock & Sylviane Granger  
*Lexicographica* 20, 2005, 72-86.

## 1. Introduction

Monolingual learners' dictionaries (MLDs) are dictionaries that are specially designed to cater for the needs of foreign language learners and provide all the information in the learners' target language. They differ from general-purpose dictionaries (GPDs) in several respects. One major difference is that MLDs contain a more detailed description of the grammatical and lexical behaviour of words, which is meant to help users in their encoding activities, while GPDs are primarily decoding tools and hence contain less detailed information on a larger number of words, including highly specialised and infrequent ones. The encoding priorities of MLDs go back to the earliest dictionary of this type, HORNBY/GATENBY/WAKEFIELD's (1942) *IDIOMATIC AND SYNTACTIC ENGLISH DICTIONARY*, whose "design specification included the distinction between countable and uncountable nouns, the presentations of as many collocations as possible and the introduction of a system of construction-patterns" (COWIE 1999, 42). Over the years, as a result of the increased use of computer corpora in lexicographical practice, productively oriented information has both expanded and diversified. RUNDELL's (1999) review of these improvements includes the following six areas:

- *meanings and explanations*: more fine-grained semantic analysis; more user-friendly definition style;
- *syntactic behaviour*: more transparent coding system;
- *phraseology and collocation*: more phrasally oriented approach;
- *frequency*: frequency data on headwords; frequency-based ordering of meanings;
- *lexical sets*: information on synonyms, antonyms and superordinate terms via cross-references and usage notes;
- *prevention of errors*: focus on learners' problem areas.

All but the last of these have benefited from the use of large computerised databases of *native* English. For the last area, however, a totally different type of corpus is necessary, a corpus containing foreign/second language learner productions, commonly referred to as a 'computer learner corpus' (CLC) or 'learner corpus' for short. The addition of this new resource to the lexicographer's workstation constitutes a new departure in pedagogical lexicography, which RUNDELL (1999, 47) describes as follows:

As well as providing acceptable models of performance, MLDs are now taking more proactive steps to help learners negotiate known areas of difficulty. (...) But just as lexicographers' intuitions are now powerfully complemented by corpus data, so too the instincts of experienced language teachers regarding likely problem areas for learners are backed up by empirical data in the form of learner corpora. (...) Corpora of this type enable lexicographers to identify recurrent sources of difficulty, and to use this information to anticipate learners' errors.

For nearly a decade Longman was the only ELT publisher to show an interest in learner corpora. The collection of the *Longman Learners' Corpus* started in the late 1980s and the first learner-corpus-informed dictionary, the *LONGMAN LANGUAGE ACTIVATOR*, was

published in 1993, soon to be followed by two other CLC-informed dictionaries, the 3rd edition of the LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH (LDOCE 3) (1995) and the LONGMAN ESSENTIAL ACTIVATOR (1997). Ten years later, the idea has clearly caught on. Two of the current MLDs explicitly claim to have incorporated data from large learner corpora: the 4th edition of LDOCE (2003) and the first edition of a newcomer to the MLD landscape, the CAMBRIDGE ADVANCED LEARNER'S DICTIONARY (CALD) (2003). A third dictionary, the MACMILLAN ENGLISH DICTIONARY FOR ADVANCED LEARNERS (MED) (2002), also refers to learner corpus data but does not seem to have exploited the corpus as systematically as the other two, as evidenced by the fact that the dictionary does not contain explicit error notes.<sup>1</sup>

In view of these developments it seems timely to reflect on the exact role that learner corpora are playing and could come to play in EFL dictionaries. In the following, we first describe learner corpora and highlight the features that are relevant for lexicographical use (section 2). We then examine the types of information that can be extracted from learner corpora to inform MLDs, explore how and to what extent they are integrated into current learner-corpus-informed MLDs and suggest possible avenues for improvement (section 3).

## 2. Key features of learner corpora for lexicographical use

### 2.1. Learner corpora: a versatile resource

Computer learner corpora are electronic collections of authentic foreign/second language data. Like native corpora, they come in many shapes and sizes: written vs. spoken, general vs. technical, synchronic vs. longitudinal, monolingual vs. multilingual (for a more detailed survey, see PRAVEC 2002 and GRANGER forthcoming). Although English is clearly the language that can boast the highest number of learner corpora, this may be about to change as several new learner corpus projects concern languages other than English.

Based on the status of the corpus building entities, learner corpora can be grouped into two major categories: academic and commercial. Most academic learner corpora are mono-L1 corpora, i.e. they contain data from only one national learner group; multi-L1 academic corpora like the *International Corpus of Learner English (ICLE)* (GRANGER/DAGNEAUX/MEUNIER 2002), which covers 11 different mother tongue backgrounds, are much rarer. They vary a great deal in size, ranging from c.100,000 words to 15 million. The objective of academic learner corpus collection can be theoretical, i.e. to provide a better description of interlanguage and a better understanding of the factors that influence it, and/or applied, i.e. to contribute to the development of pedagogical tools and classroom practices that target more accurately the needs of the learner. Commercial learner corpora, on the other hand, are multi-L1, i.e. they cover more than one mother tongue background, and tend to be very big. The two major commercial CLCs for English are the *Longman Learners' Corpus* and the *Cambridge Learner Corpus*, which respectively contain 10 and 16 million words. In collecting learner corpus data, publishers aim to inform their EFL reference tools, in particular monolingual learners' dictionaries.

Some learner corpora are better than others for lexicographical purposes. In the following sections we highlight two key features of learner corpora for MLDs: size and representativeness, and error tagging.

### 2.2. Size and representativeness

---

<sup>1</sup> See also G. FOX's comment on the MED website: "Looking to the future, Fox says she and her team at Macmillan are eager to develop and refine the learners' corpus they have been working on" <http://www.macmillandictionary.com/aboutreviews1.htm>.

Relatively small learner corpora (c.100,000 words) can be extremely valuable for the analysis of high-frequency linguistic items. To investigate article usage, for example, there is no need to have a 10 million word corpus. Indeed, in view of the very high frequency of articles, an investigation based on a corpus of this size would be completely unmanageable. For lexical analysis, however, things are very different as “investigating lexical behaviour in general, and combinatorial behaviour in particular, requires very large volumes of text” (KILGARRIFF/RUNDELL 2002, 807). But size is not the only factor to consider.

Another essential factor is that of representativeness. FRANCIS (1982, 7) defines a corpus as “a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis”. A learner corpus should therefore be representative of learner language. The problem is that learner language is a very heterogeneous variety: it can cover a wide range of L2 data that differ in terms of learner characteristics (such as age, gender, level of proficiency, mother tongue background) and task characteristics (such as medium, text type, task setting). All these variables influence learner language and are therefore critical to data interpretation and subsequent integration into teaching materials. If a learner corpus is to be a useful resource for lexicographical purposes, one clearly needs to ban what RUNDELL (1996) calls the ‘opportunistic’ data collection approach, which essentially amounts to “collecting whatever data one can easily get hold of”, and opt for a ‘principled’ approach.

For lexicographical purposes, two variables in particular are of crucial importance: the learners’ level of proficiency and mother tongue background. MLDs have traditionally appeared in two versions: one designed for intermediate learners and another for advanced learners. A learner corpus stratified in terms of *proficiency level* will ensure that the attested difficulties extracted from it and focused on in the dictionary match the difficulties that the dictionary’s intended users are likely to be faced with. Another important distinction among MLDs, i.e. between generic and bilingualised, requires control of the learners’ *mother tongue background*. For generic dictionaries, the learner corpus used needs to cover a wide range of language backgrounds, ideally represented in similar proportions, while for bilingualised dictionaries, only data from one specific national learner group can be used (e.g. Chinese learner data for an MLD intended for Chinese learners).

As data collection (and annotation, see section 2.3) is a highly time-consuming and hence costly business, it seems advisable to build one large learner corpus that can be used for many applications rather than to collect different learner corpora for different applications. However, for this large resource to be useful, it needs to fulfil two main conditions. Not only does it need to be well documented for learner and task variables, but in addition lexicographers need to be able to use these variables as search criteria to parametrise their analysis. In other words, what is needed is not simply an L2 corpus, but an *L2 database and workbench* designed to help lexicographers and other potential users of the system to correlate the particular word they are describing with the profile information of each text. The software tools developed by Cambridge to exploit the *Cambridge Learner Corpus* do precisely that. Using these tools, lexicographers can easily find out, for example, which national learner groups are most likely to use a particular word or phrase or make a particular error. As the data come from the Cambridge ESOL exams, it is also possible to correlate this information with the examination levels (NICHOLLS 2003).

### 2.3. Error tagging

Uncovering errors in learner corpora is not a straightforward task. In this section we describe four ways of accessing learner errors and expand on one extremely demanding but highly promising method, that of error annotation.

### 2.3.1. Accessing learner errors

The first method is a computerised version of the traditional intuition-based method. It consists in identifying error-prone items on the basis of intuition and/or published lists of common errors and using the learner corpus to check whether the errors actually occur and, if so, in what contexts and with what frequency. It is an extremely worthwhile method, but, as in so many other fields, intuition proves to have its limits. A quick look at usage guides shows that it is always the same mistakes that are listed. Practically all of them, for example, will contain a note on the possible confusion between *as* and *like* (*she sang \*as a bird*), but the equally frequent confusion between *as* and *such as* (*a drug \*as opium, problems \*as poverty*) is very rarely mentioned. In addition, published lists of common errors regularly contain errors that hardly ever occur. Highly revealing in this connection is HERBST's (1996, 338) comment on the list of false friends in 14 languages included in the CAMBRIDGE INTERNATIONAL DICTIONARY OF ENGLISH (CIDE) (1995):

Although German learners may have difficulty using words such as *actual-aktuell*, the idea of confusing *afterthought* and *Nachdenken*, *fail* and *Fehler*, *famous* and *famos*, *floor* and *Flor*, *rusty* and *rüstig* is more than far-fetched – at least if it is not absolute beginners (who would probably not be familiar with a word such as *Flor*) but the target group of a dictionary such as CIDE that one has in mind.

Similarly, SWAN/SMITH's (1987) *Learner English*, which lists characteristic difficulties of learners from 19 different mother tongue backgrounds, gives examples of words that French learners are claimed to misuse because of similarity with English words. Among these words one finds cases such as the English noun *essence* used instead of the noun *petrol* (= French *essence*) or the verb *command* used instead of *order* (= French *commander*), which in our experience are highly infrequent.

Whereas the first method is corpus-based, i.e. it uses the corpus as a repository of examples to test a hypothesis, the second is corpus-driven, i.e. it uses the corpus heuristically with no a priori hypothesis. Frequency lists in a learner corpus and a comparable native speaker corpus are compared automatically and lists of words that display significantly different frequencies are drawn up.<sup>2</sup> Such instances of over- and underuse signal two types of difficulties that MLDs may want to address: downright errors (e.g. the overuse of *indeed* by French learners is due to an erroneous perceived equivalence between *indeed* and the French connector *en effet*) and lexical poverty (e.g. the overuse of the most common words in the language in lieu of more specific words, see GILLARD/GADSBY 1998, 160-162).

A third option consists in looking up every word in the learner corpus. The advantage of this method is that, unlike the first method, it is exhaustive: all words are looked up, not only those that are presupposed to be problematic. The drawback is that it is very time-consuming, especially in the case of closed-class items or high-frequency words (such as the heavy-duty verbs *make*, *take*, *put*, *get*) that count a high number of occurrences, many of which are correct. Lexicographers therefore have to scan through a large number of concordance lines to spot the mistakes. When a corpus has over 10 million words, this is a bit like trying to find a needle in a haystack.

The fourth and last method relies on error annotation, i.e. all errors in the corpus are tagged with an appropriate error tag. This is a very arduous method, both complex and time-

---

<sup>2</sup> The Wordlist tool in *WordSmith Tools* (SCOTT 1996) allows for this type of comparison.

consuming, but Cambridge, which unlike Longman, has opted for error annotation, notes that “its benefits for our purposes (...) have proved to far outweigh the difficulties” (NICHOLLS 2003, 572). One obvious advantage of the method is that the corpus, once annotated, can be used for a variety of applications: not only dictionaries but also grammars, textbooks, CALL programs, testing, etc.

### 2.3.2. Error coding systems

DULAY/BURT/KRASHEN (1982, Ch. 7) suggest two major types of error taxonomy: one based on linguistic categories (general ones, such as morphology, lexis or grammar, and more specific ones, such as auxiliaries, passives or prepositions), the other focusing on the way surface structures have been altered by learners (omission, addition, misformation, misordering). The error annotation system devised by Cambridge uses a combination of surface structure alteration and word category: “The majority of the error codes are based on a two-letter coding system in which the first letter represents the general type of error (e.g. wrong form, omission), while the second identifies the word class of the required word” (NICHOLLS 2003, 573). In the two error coding systems developed at Louvain – one for English, the other for French – the first level of annotation classifies errors in terms of error domain, i.e. as being errors of form, morphology, grammar, lexis, etc. Subsequent levels indicate the error category (number, tense, etc.) and word category of the misused item (verb, noun, etc.).<sup>3</sup> Figure 1 illustrates the three different error coding systems: a missing preposition error annotated using the Cambridge system (1), a verb-dependent preposition error annotated using the Louvain error coding system for English (2) and an adjective gender agreement error annotated using the Louvain error coding system for French (3).

- (1) He said <#MT|to</#MT> me that
- (2) The citizens (XVPR) participate to \$participate in\$ those important decisions.
- (3) L'héritage du passé est très <G><GEN><ADJ> #fort\$ forte </ADJ></GEN></G>

Figure 1: Three error annotation systems

In each of the three systems the error annotation includes the correction of the error. This makes it possible to look up a particular correction and find all the types of error it gives rise to. For the preposition *for*, this procedure gives access not only to the oft-cited misuses of *since* (*it has settled deeply in our roots \*since centuries*) but also to many erroneous uses of *during* (*you can walk \*during hours*). It is also possible to retrieve all the possible corrections of a particular erroneous item. Depending on the context, the adjective *great* is corrected to *big*, *large*, *important*, *considerable* or *high*.

While undoubtedly very useful, error annotation is also beset with difficulties. All three stages of the error coding procedure – error detection, correction and annotation – involve a high degree of subjectivity. As pointed out by MILTON/CHOWDHURY (1994, 129) “Tagging a learner corpus allows us, at least and at most, to systematize our intuitions”. To minimise subjectivity and increase inter-coder reliability, it is essential to have a coherent error coding system and a detailed error tagging manual, where all categories are defined and illustrated and error coding principles are noted.

<sup>3</sup> For more information on the error tagging systems developed at Louvain, see DAGNEAUX/DENNESS/GRANGER (1998) for the English system and GRANGER (2003) for the French system.

### 3. Integration of learner corpus information into MLDs

Before exploring how and to what extent information extracted from learner corpora has been integrated into the latest editions of the two MLDs that claim to have made extensive use of such corpora, i.e. LDOCE 4 and CALD, let us first turn our attention to the types of information learner corpus-based research can offer lexicographers setting out to compile more learner-aware MLDs.

#### 3.1 Types of CLC-based information

There are two major types of CLC-based information that can usefully be extracted to enhance MLDs: on the one hand, information relating to learners' misuse of target language (TL) words or phrases (i.e. downright errors), and, on the other, information relating to learners' overuse and underuse of TL words or phrases. The focus is on information that can be obtained from *written* learner corpora, as the bulk of CLC-based research has so far been conducted on written rather than spoken corpora<sup>4</sup> and both the *Longman Learners' Corpus* and the *Cambridge Learner Corpus* contain written learner productions only.

##### 3.1.1 Learners' misuse of TL words or phrases

The major types of learners' misuse of TL words or phrases that are of potential interest to lexicographers are orthographic errors, lexical errors, lexico-grammatical errors and register errors<sup>5</sup>. Research has shown that these errors can be cross-linguistic, i.e. shared by learners from a variety of mother tongue backgrounds, or L1-specific, i.e. shared by learners from one specific mother tongue background. Lack of space compels us to confine our discussion to a few representative examples of each category of errors. Additional examples can be found in the ample learner corpus research literature<sup>6</sup>.

As word processors all include automatic spell-checkers, orthographic errors could be regarded as of only marginal relevance. There are, however, a whole series of orthographic errors that learners tend to make that are not flagged by spell-checkers and that are therefore worthy of attention. Errors that stem from the confusion of homonyms like *loose/lose*, *practice/practise*, *advice/advise* or *price/prize* are a case in point. An investigation of items of this type in the *ICLE* corpus reveals that such errors tend to be cross-linguistic and frequent even at an advanced level. Another example of a frequently recurring cross-linguistic orthographic error not flagged by spell-checkers is *can not*.

Research has shown that lexical errors involving not only single words but also multi-word units (e.g. *\*do an error*) tend to be quite numerous. A large proportion of these can be labelled L1-specific (e.g. French learners' use of *on the contrary* instead of *on the other hand*). Deceptive cognates pose a major problem for learners (e.g. French learners' use of *society* when they mean *company*).

Lexico-grammatical errors are errors in the area of word grammar. These typically include errors relating to verb complementation, dependent prepositions, the countable or uncountable nature of nouns (*\*advices*, *\*informations*) and the attributive or predicative-only use of some

<sup>4</sup> The *Louvain International Database of Spoken English Interlanguage (LINDSEI)* is the only large-scale spoken learner corpus project of its kind.

<sup>5</sup> Spoken learner corpora would give researchers access to other types of errors as well, e.g. pronunciation errors and pragmatic misuse.

<sup>6</sup> A comprehensive learner corpus bibliography is available from the Centre for English Corpus Linguistics website: <http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/cecl.html>.

adjectives. While countable/uncountable errors tend to be largely cross-linguistic, other lexico-grammatical errors such as dependent prepositions errors can be more L1-specific.

Register errors typically stem from learners' lack of awareness of speech vs. writing and formal vs. informal distinctions. Learners have, for example, been shown to display a tendency to include a whole series of speech-like features such as *a lot of* or *lots of* (DE COCK 2003) or the pragmatic prefab *you see* in their formal argumentative essays (WIKTROSSON 2001).

### 3.1.2 Learners' overuse and underuse of TL words or phrases

As mentioned above (section 2.3.1), learners' overuse and underuse of words or phrases is brought to light by contrasting the frequency counts of words or phrases in a learner corpus and in a comparable native speaker corpus. French learners' overuse of *indeed* mentioned above is an illustration of L1-specific learner overuse. Groups of learners with related mother tongues have also been shown to underuse and overuse some words and phrases along similar lines. For example, NEFF/DAFOUZ/HERRERA/MARTINEZ/RICA's (2003) study of reporting verbs in learner argumentative essay writing shows that Italian, French and Spanish learners can all be seen to overuse *say* and underuse the more specific reporting verbs *state* and *argue*. A number of studies reveal that learners from a wide variety of (unrelated) mother tongue backgrounds display a common tendency to overuse common, non-specific words such as *important* (GRANGER/TRIBBLE 1998) or *big* or *nice* (GILLARD/GADSBY 1998).

It is important to point out that instances of learner overuse and underuse are, on the whole, not recoverable from error-tagged corpora. The words and phrases that learners overuse or underuse tend to be overlooked by annotators, who focus on downright errors, unless they are also erroneously used. Error-tagged learner corpora should therefore not be seen as the be-all and end-all of learner corpus-informed pedagogical lexicography. Both error-tagged and unannotated corpora have a valuable contribution to make to the field.

## 3.2 Integration into MLDs

### 3.2.1 General issues

Learner corpus research provides lexicographers with a wealth of information about learners' problems. One of the major issues involved in the integration of this information is undoubtedly that of *selection*. According to STEPHEN BULLON (personal communication), 250 notes in LDOCE 4 were specifically based on the analysis of the *LLC*. CALD contains c. 200 Common Learner Error notes (PATRICK GILLARD, personal communication). The preliminary results from an error-tagging project involving eight partners in the *International Corpus of Learner English* project reveal that there are on average 3,500 errors per 50,000-word sample. This figure, which does not include cases of learners' 'non-erroneous' overuse and underuse of words and phrases, gives a rough idea of the mammoth and challenging task facing lexicographers working with learner corpora containing several million words. Assuming that a similar ratio of errors per number of words applies to the 6 million word error-tagged *Cambridge Learner Corpus*, lexicographers are faced with over 400,000 errors, well over a third of which are potential candidates for integration into MLDs (i.e. belong to the category of orthographic, lexical, lexico-grammatical or register errors).

Two key factors relating to the MLDs' targeted users ought to guide lexicographers' choice. The first factor concerns whether the MLD is aimed at learners of English regardless of their mother tongue (generic MLD) or whether it is intended for a specific group of learners who

share the same mother tongue (bilingualised or L1-specific MLD). The second key factor concerns the targeted learners' level of proficiency: is the MLD aimed at intermediate or advanced learners of English? Including notes dealing with errors that learners at an advanced level are no longer seen to make would be a waste of precious space in an MLD designed for this type of learner. In view of these factors, it is essential for lexicographers to be able to have easy access to this type of information.

The frequency with which errors recur in the corpus can also usefully be brought into the equation. Errors that tend to recur frequently in the corpus are more likely candidates for inclusion than errors that occur only once or twice.<sup>7</sup>

When it comes to the *format* in which learner corpus-based prevention of errors is presented, RUNDELL (1999, 47) makes a useful distinction between implicit and explicit use of learner corpus-based information. Implicit learner corpus-based prevention of errors involves, for example, “highlighting and clearly exemplifying a ‘correct collocation’”, presenting learners with more specific alternatives to frequently-used core-vocabulary items such as *important*, *big*, *happy*, *nice*<sup>8</sup> or providing learners with notes that point out the differences between closely related (yet non-interchangeable), easily confused words. Explicit learner corpus-based prevention of errors, on the other hand, involves issuing specific warnings to alert learners to typical erroneous usage.

The *place* where the information is included also requires careful consideration. Besides attempting to make the information as prominent as possible, lexicographers need to address the issue of optimal placement when the errors involve the confusion between two or more different words recorded in separate dictionary entries. For example, should the typical learner error *\*the leader of a great amount of* (instead of *number of*) *people* or *\*money works as* (instead of *like*) *a drug* be dealt with under the incorrect or the correct form? Although inclusion under the incorrect form may intuitively strike one as the preferred option, optimal placement is yet to be empirically tested. Optimal placement is arguably less of an issue in electronic versions of MLDs thanks to the hyperlinks system, which makes it possible for a note to appear both under the incorrect and correct forms without taking up any ‘valuable’ space.

### 3.2.2 Learner corpus-based information in LDOCE 4 and CALD

Let us now explore and compare the integration of learner corpus-based information into the latest edition of the forerunner in the field, LDOCE 4, and into the newcomer, CALD, which has devoted considerable time and energy to the collection and error-tagging of a large learner corpus.

Very little is mentioned in the dictionaries themselves and on the publishers' websites about the exact contribution of learner corpus-based information to LDOCE 4 and CALD.<sup>9</sup> From the available description, it seems to be the case that both dictionaries have opted for explicit rather than implicit learner corpus-based prevention of errors. While the ‘How to use the dictionary’ section in LDOCE 4 mentions that “Warning notes tell you about common mistakes that people make when using a word, based on research from the Longman Learners' Corpus (a database of over 10 million words of English written by students from around the world)”, in his introduction to CALD, Gillard informs users that “You will find ‘Common Learner

<sup>7</sup> The speech vs. writing dimension could also be taken into consideration depending on whether or not – and on the extent to which – the MLD is writing- and/or speech-oriented and provided lexicographers also have a spoken learner corpus at their disposal.

<sup>8</sup> The ‘words you can use instead of’ notes used in MED appear to have specifically been designed to address learners' overuse of core-vocabulary items.

<sup>9</sup> Detailed descriptions of the learner corpora used are regrettably also absent.

Error' notes spread throughout the dictionary, to make sure that you don't make the mistakes that many learners make. The notes are based on the Cambridge Learner Corpus." It is regrettably unclear whether or not – and the extent to which – CLC-based information has been used to compile the wide variety of other notes and extra information (CALD: Usage notes, Collocation boxes; LDOCE 4: Grammar, Word Focus, Word choice notes, Collocation box) included in the two dictionaries.

Our analysis is essentially based on an examination of the various notes provided for the entries under the letters 'A' and 'P'. We have decided to focus our attention on Warning and Common Learner Error notes as these are clearly presented by the editors as resulting from learner corpus investigation. The information provided is scrutinised from both a quantitative (i.e. in terms of the number of notes) and a qualitative point of view (i.e. in terms of presentation, selection and content). A distinction is made between the paper and the electronic versions of the two dictionaries where relevant. The analysis is supplemented by limited spot checks carried out for a few learner corpus-attested errors<sup>10</sup> which, as experienced EFL teachers, we have frequently come across.

|                                                                                   | <b>CALD<br/>(paper)</b> | <b>LDOCE 4<br/>(paper)</b> | <b>CALD<br/>(electronic)</b> | <b>LDOCE 4<br/>(electronic)</b> |
|-----------------------------------------------------------------------------------|-------------------------|----------------------------|------------------------------|---------------------------------|
| <b><i>Letter A</i></b>                                                            |                         |                            |                              |                                 |
| Explicit notes (Common Learner Error/Warning notes)                               | 24                      | 16                         | 24                           | 16                              |
| Implicit notes (Usage notes, Grammar, Word Focus, Word Choice, Collocation notes) | 2                       | 25                         | 4                            | 25                              |
| <b>Total</b>                                                                      | <b>26</b>               | <b>41</b>                  | <b>28</b>                    | <b>41</b>                       |
| <b><i>Letter P</i></b>                                                            |                         |                            |                              |                                 |
| Explicit notes                                                                    | 14                      | 40                         | 16                           | 40                              |
| Implicit notes                                                                    | 2                       | 38                         | 17                           | 38                              |
| <b>Total</b>                                                                      | <b>16</b>               | <b>78</b>                  | <b>33</b>                    | <b>78</b>                       |

Table 1: Number of explicit and implicit notes in the paper and electronic versions of CALD and LDOCE 4.

The results in Table 1 seem to point to a greater number of notes in LDOCE 4. This is mainly due to the large number and the wide variety of implicit notes. When it comes to explicit notes, however, no conclusion can be drawn as to which dictionary contains the most notes of this type: whereas there are more explicit notes in CALD for the entries included under the letter 'A', the situation is reversed for the entries under the letter 'P'. It is worth mentioning that there are far fewer differences between the number of notes in the paper and electronic versions than the increased availability of space offered in electronic versions might have led one to expect. This is especially striking for LDOCE 4 (no differences).<sup>11</sup> The differences between the CALD paper and electronic versions are essentially due to the inclusion on the CD-ROM of (implicit) collocation notes, which are not included in the paper version.

<sup>10</sup> For this purpose we have made use of the *International Corpus of Learner English* (see GRANGER/DAGNEAUX/MEUNIER 2002).

<sup>11</sup> That said, the excellent additional features 'Phrase bank' and 'Examples bank' in the electronic version of LDOCE 4 are well worth highlighting.

Beyond these purely quantitative considerations closer inspection reveals that a large number of the explicit notes provided in the dictionaries do actually address corpus-attested frequently recurring advanced learner errors. Good illustrations include the notes for *accept* (confusion *accept/agree*), *actual* (when erroneously used to mean *current*), *amount of* (confusion *amount of/number of*), *parking* (erroneous countable use of the word), *police* (when erroneously used as a singular noun), *percent* (confusion *percent/percentage*) and *progress* (erroneous countable use of the noun). Good-quality explicit notes of this type have enormous potential as they can genuinely help raise learners' awareness of common (persistent) errors - the first step in their eradication.

What is striking and perhaps somewhat unexpected, however, is that there is no consensus about the actual entries ('flagged entries') that include explicit and/or implicit notes in the two dictionaries. As far as explicit notes are concerned, LDOCE 4 and CALD share only four flagged entries under the letter 'A' in both the paper and electronic versions (*a*, *actual*, *advice*, *afford*) and four for the letter 'P' in the paper version (*pay*, *possibility*, *propose* and *problem*) or six in the electronic version (*pay*, *possibility*, *propose* and *problem* + *phone* and *part*). Taking both explicit and implicit notes into account does not lead to any significant increase in the number of common flagged entries as a mere four extra entries would be added to the list (*afternoon*, *ask*, *provide* and *place*). The majority of errors are thus only dealt with either in CALD or in LDOCE 4. The absence of the frequent and typical error *\*persons* (instead of *people*, ICLE) in CALD or of the recurring error *\*to avoid to do* (instead of *avoid doing something*, ICLE) in LDOCE4 is rather puzzling. It is important to note that the explicit notes contained in the entries that are flagged in the two dictionaries do not always target the same learner errors. For example, while LDOCE 4 and CALD both warn learners against confusing the verb *advise* and the noun *advice*, CALD includes an additional note to explicitly alert learners to the uncountable nature of the noun *advice*. Under the entry for the noun *part*, LDOCE 4 draws users' attention to the erroneous *most part of* (instead of *most of*, as in *\*we spent most part of the afternoon shopping*) and *take a part of* (instead of *take part in*). The CALD CD-ROM, on the other hand, warns its users against confusing *take part in something* and *take place*. These observations highlight the central role played by selection criteria in the integration of learner corpus-based information into MLDs.

The problematic selection of learner errors is also reflected in a number of rather questionable notes. In most cases, although the words flagged can be shown to pose problems to learners, the actual errors the notes focus on are either inappropriate to the learners' (advanced) level of proficiency or they are not among the most frequently recurring errors that can be seen to affect the word. Examples include the common learner notes for *piece* and for *attend* in CALD. These notes warn advanced learners against using *piece* instead of *room* and *attend* instead of *wait*. There are no instances of these errors in the ICLE corpus and in our experience advanced learners are very unlikely to make such errors. It therefore strikes us as rather dubious that these notes are actually based on evidence from advanced learner productions in the *Cambridge Learner Corpus*. That said, *piece* and *attend* are in fact error-prone items for advanced learners. A study of *piece* in ICLE and in the comparable *Louvain Corpus of Native English Essays* (LOCNESS) reveals that learners tend to significantly overuse the word in phrases like *piece of advice* or *piece of information*. Typical attested erroneous uses of the verb *attend* tend to be of the type *\*a shop assistant refused to attend her* (instead of *attend to her*, ICLE corpus) or *\*attend to classes* (instead of *attend classes*). Similarly, a warning note in LDOCE4 draws learners' attention to the fact that the verb *participate* is never followed directly by a noun (as in *\*Everyone can participate an election*). In the ICLE corpus dependent preposition errors such as *to participate to* or *at an election* far outnumber the error mentioned in the warning note. In the same vein, an error note in CALD warns learners against using the verb *prevent* in the meaning of *protect*. However, the most frequent error affecting

this verb is the use of the infinitive instead of the gerund (e.g. *\*the financial aspect might prevent some to acquire a higher education*). It is rather surprising that this typical learner error should have been overlooked both in CALD and LDOCE 4.

Another important issue that needs to be taken on board, but one that lies beyond the scope of this article, is the visual presentation of the error sections in MLDs. In this connection, a prominent eye-catching presentation similar to that in CALD ought to be favoured to ensure that learners have rapid and easy access to the information provided.

#### 4. Conclusion

While monolingual learners' dictionaries aim to help foreign language learners in their encoding activities, learner corpora contain evidence of what it is that learners have difficulty with when carrying out these activities. In principle, therefore, the two resources make a perfect match. However, as our analysis of two CLC-informed monolingual learners' dictionaries reveals, even if a step has been taken in the right direction, the match is not perfect yet. This is hardly surprising, as learner corpus investigation is still a relatively new field of research. What is more, extracting relevant information from learner corpora tends to be fraught with difficulties. The frequency, level and L1-specificity of errors all need to be carefully assessed. The lack of overlap between the errors targeted in the two dictionaries shows that this task is far from straightforward. One of the main difficulties lexicographers have to contend with is that learner corpora, even advanced ones, contain a high number of errors and space limitations only allow for the inclusion of a very small proportion in the dictionary. Fortunately, one can reasonably expect future developments in electronic lexicography to remedy this problem. As space is much less of an issue in electronic MLDs, many more error sections can be included, targeting not only cross-linguistic difficulties but also L1-specific ones like false friends. This development is perfectly consistent with RUNDELL's (1996) vision of the electronic ELT dictionary of the future in which "many of the old distinctions – eg dictionary vs grammar, (...) global-monolingual vs bilingual dictionary – become much less relevant, and it is possible, for example, to envisage an all-purpose learners' dictionary that could be customized according to the specific first language, level of competence, and field of interest (eg Business English) of an individual user".

One thing is certain, however: for maximum effectiveness, the analysis of learner corpora will always require not only expert lexicographical skills but also solid teaching experience. As it is no easy matter to find people who have both of these skills, lexicographical exploitation of learner corpora will generally need to rely on close collaboration between trained lexicographers and experienced ELT teachers.

#### 5. Bibliography

##### 5.1 Dictionaries

CALD = CAMBRIDGE ADVANCED LEARNER'S DICTIONARY. PATRICK GILLARD (ed.) Cambridge 2003.

CIDE = CAMBRIDGE INTERNATIONAL DICTIONARY OF ENGLISH. PAUL PROCTER (ed.). Cambridge 1995.

HORNBY/GATENBY/WAKEFIELD 1942 = ALBERT SYDNEY HORNBY, EDWARD. V GATENBY, HAROLD A WAKEFIELD: IDIOMATIC AND SYNTACTIC ENGLISH DICTIONARY. Tokyo 1942.

LDOCE 3 = LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH. DELLA SUMMERS (ed.). Harlow 1995.

LDOCE 4 = LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH. DELLA SUMMERS (ed.). Harlow 2003.

LONGMAN LANGUAGE ACTIVATOR. DELLA SUMMERS (ed.). Harlow 1993.

LONGMAN ESSENTIAL ACTIVATOR. DELLA SUMMERS (ed.). Harlow 1997.

MED = MACMILLAN ENGLISH DICTIONARY FOR ADVANCED LEARNERS. MICHAEL RUNDELL (ed.). Oxford 2002.

## 5.2 Other literature

COWIE 1999 = ANTHONY PAUL COWIE: English Dictionaries for Foreign Learners. A History. Oxford 1999.

DULAY/BURT/KRASHEN 1982 = HEIDI DULAY/ MARINA K. BURT/ STEPHEN KRASHEN: Language Two. New York 1982.

DAGNEAUX/DENNESS/GRANGER 1998 = ESTELLE DAGNEAUX/SHARON DENNESS/SYLVIANE GRANGER : Computer-aided Error Analysis. In: System. An International Journal of Educational Technology and Applied Linguistics 26/2. 1998, 163-174.

DE COCK 2003 = SYLVIE DE COCK : Recurrent Sequences of Words in Native Speaker and Advanced Learner Spoken and Written English. A Corpus-driven Approach. Unpublished PhD thesis, Université catholique de Louvain, Louvain-la-Neuve 2003.

FRANCIS 1982 = W. NELSON FRANCIS: Problems of assembling and computerizing large corpora. In: STIG JOHANSSON (ed.): Computer corpora in English language research. Bergen 1982, 7-24.

GILLARD/GADSBY 1998 = PATRICK GILLARD/ADAM GADSBY: Using a learners' corpus in compiling ELT dictionaries. In: SYLVIANE GRANGER (ed.): Learner English on Computer. London and New York 1998, 159-171.

GRANGER 2003 = SYLVIANE GRANGER: Error-tagged learner corpora and CALL: a promising synergy. In: CALICO 20/3. 2003, 465-480.

GRANGER forthcoming = SYLVIANE GRANGER: Computer learner corpus research: current status and future prospects. In: ULLA CONNOR, THOMAS UPTON (eds.): Applied Corpus Linguistics: A Multidimensional Perspective. Amsterdam and Atlanta forthcoming.

GRANGER/DAGNEAUX/MEUNIER 2002 = SYLVIANE GRANGER/ESTELLE DAGNEAUX/FANNY MEUNIER (eds.): The International Corpus of Learner English. Handbook and CD-ROM. Version 1.1. Louvain-la-Neuve 2002.

GRANGER/TRIBBLE 1998 = SYLVIANE GRANGER/CHRISTOPHER TRIBBLE: Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In: SYLVIANE GRANGER (ed.): *Learner English on Computer*. London and New York 1998, 199-209.

HERBST 1996 = THOMAS HERBST: On the way to the perfect learners' dictionary: a first comparison of OALD5, LDOCE3, COBUILD2 and CIDE. In: *International Journal of Lexicography* 9/4. 1996, 321-357.

KILGARIFF/RUNDELL 2002 = ADAM KILGARIFF/MICHAEL RUNDELL: Lexical profiling and its lexicographic applications – a case study. In: ANNA BRAASCH, CLAUS POVLSEN (eds.): *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002*. Copenhagen, Denmark, August 13-17, 2002. Copenhagen 2002, 807-818.

MILTON/CHOWDHURY 1994 = JOHN MILTON/NANDINI CHOWDHURY: Tagging the interlanguage of Chinese learners of English. In: LYNNE. FLOWERDEW, ANTHONY.K.K. TONG (eds.): *Entering Text*. Hong Kong 1994, 127-143.

NEFF/DAFOUZ/HERRERA/MARTINEZ/RICA 2003 = JOANNE NEFF/EMMA DAFOUZ/HONESTO HERRERA/FRANCISCO MARTINEZ/ JUAN PEDRO RICA: Contrasting learner corpora: the use of modal and reporting verbs in the expression of writer stance. In: SYLVIANE GRANGER, STEPHANIE PETCH-TYSON (eds.): *Extending the Scope of Corpus-based Research. New Applications, New Challenges*. Amsterdam and New York 2003, 211-230.

NICHOLLS 2003 = DIANE NICHOLLS: The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In: DAWN ARCHER, PAUL RAUSON, ANDREW WILSON, TONY MCENERY (eds.): *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL technical paper number 16. Lancaster 2003, 572-581.

PALMER 1931 = HAROLD .E. PALMER: *Second Interim Report on Vocabulary Selection*. Tokyo 1931.

PRAVEC 2002 = NORMA A. PRAVEC: Survey of learner corpora. In: *ICAME Journal* 26. 2002, 81-114.

RUNDELL 1996 = MICHAEL RUNDELL: The corpus of the future, and the future of the corpus. Available from <http://www.ruf.rice.edu/~barlow/futcrp.html>.

RUNDELL 1999 = RUNDELL, MICHAEL: Dictionary use in Production. In: *International Journal of Lexicography* 12/1. 1999, 35-53.

SCOTT 1996 = MIKE SCOTT: *WordSmith Tools*. Oxford 1996.

SWAN/SMITH 1987 = MICHAEL SWAN/ BERNARD SMITH: *Learner English. A teacher's guide to interference and other problems*. Cambridge 1987.

WIKTORSSON 2001 = MARIA WIKTORSSON: Register Differences between Prefabs in Native and EFL English. In: SATU MANNINEN, CARITA PARADIS (eds.): *Working Papers in Linguistics* (vol 1). Lund 2001, 85-94.