

Sylvie De Cock and Magali Paquot
Centre for English Corpus Linguistics
Université catholique de Louvain

The monolingual learners' dictionary as a productive tool: the contribution of learner corpora

[DRAFT]

1. Introduction

This paper reports on the outcomes of a large-scale collaborative corpus-based project between the Centre for English Corpus Linguistics and Macmillan Education. The main aim of this project was to produce learner corpus informed materials for inclusion in the second edition of the *Macmillan English Dictionary for Advanced Learners* (Rundell 2007; hereafter *MED2*).

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 learner's target language' (De Cock and Granger 2004). As highlighted by Rundell (1999), MLDs have reaped enormous benefits from the increased use of computer corpora in dictionary making. Corpus-related improvements (e.g. enhanced descriptions of word meanings and phraseological patterns) have so far mainly involved the use of corpora of native speaker speech and writing. The use of learner corpora (i.e. see De Cock, this volume) when compiling MLDs is comparatively relatively recent: the first learner corpus informed dictionary, the *Longman Language Activator*, was published in 1993. *The Longman Dictionary of Contemporary English* (1995), the *Longman Essential Activator* (1997) and the *Cambridge Advanced Learners' Dictionary* (2003) followed suit.

Findings from learner corpus-based research, which shed light on learners' difficulties and deficiencies, can be used by lexicographers to anticipate learners' errors (Rundell 1999). De Cock and Granger (2004) identify two types of learner corpus-based information that can help enhance MLDs: information relating to learners' misuse of target language words or phrases and information relating to learners' overuse and underuse of target language words or phrases. While the extraction of downright errors from learner corpora has greatly been facilitated by error annotation (De Cock and Granger 2004), cases of overuse and underuse can be uncovered by contrasting the frequency counts of words and phrases in a learner corpus and in a comparable native speaker corpus.

Information from learner corpora has overwhelmingly been included in the form of explicit 'warning' or 'common learner error' notes in MLDs. An investigation of these error notes in two learner corpus informed advanced MLDs (De Cock and Granger 2004) revealed that there was still room for improvement. It showed that, although some of the notes in the MLDs addressed corpus-attested frequently recurring advanced learner errors, a number of them were misguided as they focused on errors that were clearly inappropriate to the learners' advanced level of proficiency or on errors that were either rare or non-existent in a corpus of advanced learner writing like the *International Corpus le Learner English*. The study also highlighted a distinct lack of overlap between the errors targeted in the two dictionaries, which points to the problematic selection of the learner errors that should be focused on.

2. *The Macmillan English Dictionary for Advanced Learners* (second edition, *MED2*)

2.1 Objective of the project

The main objective of the collaborative corpus-based project between the Centre for English Corpus Linguistics and Macmillan Education reported on in this paper was to produce learner corpus informed materials that would help advanced learners cope with attested areas of difficulty in their writing. Three learner corpus-based components were developed with this objective in mind: 100 'Get it right' (GIR) boxes, six Grammar Sections and twelve EAP Writing Sections.

The three components were compiled on the basis of detailed analysis of (1) learner corpus data from the 3.5-million-word *International Corpus of Learner English (ICLE)*, which contains essay writing by EFL learners from 16 different mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish) and (2) native speaker data from a 15-million-word corpus of academic writing.

2.2 'Get it right' boxes

GIR boxes are located at individual dictionary entries. They are designed to alert learners to typical erroneous usage associated with a particular word or phrase. *MED2* is aimed at all advanced learners regardless of their mother tongue backgrounds. It was therefore decided that, to make the shortlist of flagged words or phrases, the errors associated with these items had to be shown to be both frequent and widespread (i.e. attested in data from learners from at least five different mother tongue backgrounds) in the learner corpus. Possible candidates for the shortlist were identified on the basis of, on the one hand, careful scrutiny of the 680,000-word error-tagged component of *ICLE* available at the time and, on the other, systematic comparisons of word frequency lists from the advanced learner corpus and the native speaker corpus used. Once identified, the possible candidates were subjected to rigorous analysis in the whole 3.5-word *ICLE* corpus. The errors in the GIR boxes cover a number of categories including 'countability', 'register', 'verb patterns' or 'spelling'.

As Figures 1, 2, 3 and 4 illustrate, GIR boxes typically contain authentic erroneous learner examples clearly marked as incorrect (the examples are preceded by a cross and the erroneous items have been crossed out), clear explanations of the source of the problem and clear and practical advice on how the error can be corrected and avoided.

It is worth noting that emphasis was also laid on prominent eye-catching presentation (with the use of colour, bold type and underlining) to ensure that the information in GIR boxes is not overlooked by learners.

Get it right: damage

When **damage** means 'harm or injury' it is an uncountable noun, and so:

- it is never used in the plural
- it never comes after **a** or a number

✗ These toxins can cause **damages** to the lungs and brains.

✓ These toxins can cause damage to the lungs and brains.

✗ They should consider the serious **damages** that their decisions may cause.

✓ They should consider the serious damage that their decisions may cause.

✗ ~~A great damage~~ has been done to agriculture, forests, and people's health.

✓ Great damage has been done to agriculture, forests, and people's health.

The plural form **damages** is a specialized legal term meaning 'money that a court orders you to pay someone because you have harmed them or their property'.

Mr Galloway was awarded substantial **damages**.

Figure 1: GIR box at 'damage' in *MED2*

Get it right: more

The expression **more and more** is used mainly in speech and informal writing. In academic and professional writing, the adverb **increasingly** is much more common:

✗ Europe is becoming ~~more and more unified~~ and therefore people are afraid of losing their own identity.

✓ Europe is becoming increasingly unified and therefore people are afraid of losing their own identity.

✗ Problems include the loss of national identity, ~~more and more competitive~~ lifestyles, and declining moral values.

✓ Problems include the loss of national identity, increasingly competitive lifestyles, and declining moral values.

Figure 2: GIR box at 'more' in *MED2*

Get it right: research

Research is an uncountable noun, and so:

- it is hardly ever used in the plural
- it never comes after **a** or a number

✗ Her latest work confirms the findings of earlier **researches**.

✓ Her latest work confirms the findings of earlier **research**.

✗ According to **one recent research**, women's earnings are still 27% lower than men's.

✓ According to **recent research**, women's earnings are still 27% lower than men's.

Q: What should I say if I want to refer to one particular study or to several studies of this type?

A: You can say: a **study**, several **studies**, **some research**, a **piece of research**, or a **programme of research**:

Her latest work confirms the findings of earlier **studies**.

According to **one study**, women's earnings are still 27% lower than men's.

a detailed **programme of research** on the economics of nuclear energy

Figure 3: GIR box at 'research' in *MED2*

Get it right: pay

The verb **pay** is never followed by a direct object that refers to the thing you are buying. We **pay for** a product or service:

✗ Credit cards are used ~~to pay the product~~ you purchased without using cash.

✓ Credit cards are used to pay for the product you purchased without using cash.

✗ At that time, very few people could ~~pay a university education~~.

✓ At that time, very few people could pay for a university education.

You can also use **pay** in these patterns:

- pay someone for something
- pay an amount of money for something
- pay someone an amount of money for something

It was rumoured that Texaco had **paid** the government over \$800 million for drilling rights.

Have you **paid** your brother for the cinema tickets?

However, **pay** can be used with a direct object which refers to money that is paid for a specific purpose. The nouns most frequently used in this pattern are:

bill, charge, compensation, debt, fine, price, fee, rent, salary, tax, wage

Around one third of schoolchildren failed to enrol this year because their parents could not **pay the school fees**.

Married couples are taxed independently, and each spouse is responsible for **paying tax** on his/her own income.

Figure 4: GIR box at 'pay' in *MED2*

2.3 Grammar Sections

Six so-called 'Grammar Sections' are included as part of the 'Improve your Writing Skills' component (De Cock et al. 2007) located in the 'Study Section' in the middle of the dictionary. The sections address key areas of difficulty relating to either word grammar (namely Articles, Complementation: Patterns used with nouns, verbs and adjectives, Countable and Uncountable Nouns, and Quantifiers) or specific aspects of written discourse (i.e. Punctuation and Spelling). The focus on word grammar and punctuation and spelling was deemed particularly appropriate in a dictionary (words, words, words!) that seeks to help learners with their writing.

Research based on the error-tagged component of *ICLE* (cf. above) served as the starting point for the selection and content of the sections. The larger unannotated *ICLE* corpus and the native speaker corpus (cf. above) were also used in the compilation of the sections.

The Grammar Sections are typically structured around the following key elements:

- (1) an introductory presentation and description of the phenomenon in the spotlight
- (2) learners' problems and errors relating to it (with authentic learner examples)
- (3) clear explanations of why the authentic learner examples included are problematic
- (4) suggested corrections of the authentic learner examples
- (5) general advice on how to avoid making such errors

Frequency data is also provided where relevant to raise learners' awareness of inappropriately overused items in academic writing, as illustrated by Figures 5 and 6 from the sections Punctuation and Quantifiers respectively.

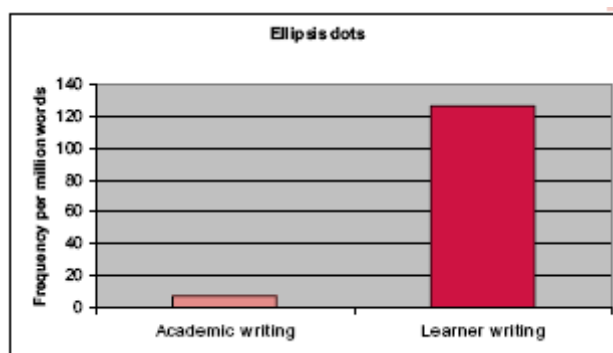
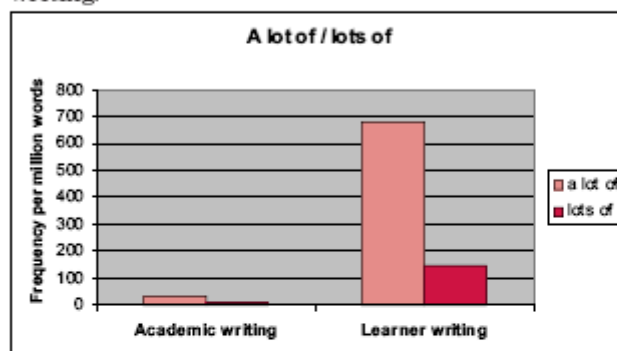


Figure 5: Frequency of ellipsis dots in learner and native speaker writing (section about Punctuation, *MED2*)

4. Quantifiers and register: *lots* and *a lot*

The quantifiers *a lot of* and *lots of* are common in informal conversation and in informal writing (for example, in a letter to a friend). However, they are rarely used in academic writing or professional reports. As the graph below shows, learners tend to massively overuse these expressions in academic writing.



? Some would argue that banning smoking in restaurants will bring ~~lots of advantages~~.

✓ Some would argue that banning smoking in restaurants will bring many advantages.

✓ Some would argue that banning smoking in restaurants will bring a number of advantages.

? During the last few decades, there has been ~~lots of discussion~~ about the possibility of machine translation.

✓ During the last few decades, there has been a great deal of discussion about the possibility of machine translation.

? ~~A lot of effort~~ has been made to persuade these people of the immorality of their behaviour.

✓ Much effort has been made to persuade these people of the immorality of their behaviour.

Figure 6: Frequency of the quantifiers *a lot of* and *lots of* in learner and native speaker writing (section about Quantifiers, MED2)

The final paragraph of each section is entitled 'Advice on avoiding errors'. In addition to providing learners with general advice on how to avoid making the errors discussed in the sections, the paragraph also systematically highlights how the dictionary can help them deal with these problem areas. For example, the section Complementation is rounded off with clear explanations and illustrations of where and how information relating to complementation can be found at dictionary entries (cf. Figure 7). The sections also include systematic cross-references to related 'Get it right' boxes and to related Grammar Sections and/or Writing Sections. These 'Advice on avoiding errors' paragraphs can thus also be seen to provide learners with some form of dictionary training, which has been shown to be extremely valuable if learners are to make the most of the riches contained in MLDs.

4.2 Using the *Macmillan English Dictionary*

Your dictionary is a reliable source of information on complementation. In the *Macmillan English Dictionary*, information on the patterns that can be used with a word is shown after the definition and before the example sentences:

enjoy /m'dʒɔɪ/ verb ★★★

1 [T] to get pleasure from something: *Do you enjoy cooking or do you just see it as a chore?* ♦ **enjoy doing sth** *I don't enjoy going on holiday as much as I used to.*

Here, the grammar code [T] indicates that **enjoy** is a transitive verb, and in the first example sentence the verb is followed by a simple direct object ('your time at university'). The code **enjoy doing sth** indicates that **enjoy** can also be used with a verb in the -ing form ('enjoy going on holiday').

participate /pɑː(r)'tɪsɪpeɪt/ verb [I] ★★ to take part in something: *Members are eligible for a 50% saving on room rates at all hotels that are participating in the scheme.* ♦ **+in** *The rebels have agreed to participate in the peace talks.*

Here, the grammar code [I] shows that **participate** is intransitive. Before the second example sentence, the code **+in** shows that **participate** can be used with a preposition, and the correct preposition is *in* ('participate in the peace talks').

Make sure that the pattern you choose corresponds to the particular meaning of the verb you are using. As explained above (see section 3.1), some verbs have different patterns for different meanings. For example:

remember /rɪ'membə(r)/ verb ★★★

1 [I/T] to have an image in your mind of a person, a place, or something that happened or was said in the past: *I can still remember every word of our conversation.* ♦ *That was a beautiful summer, as I remember.* ♦ **remember doing sth** *She remembers seeing him leave an hour ago.*

...

3 [T] to do something that you promised to do or that you have to do, and not forget about it: *I hope she remembers my book when she comes* (=remembers to bring it.) ♦ **remember to do sth** *He never remembered to lock the door when he went out.*

Notice that in meaning 1, **remember** can be used with a verb in the -ing form ('remembers seeing him'), while in meaning 3 the correct pattern is an infinitive ('remembered to lock'). So when you are using a word that has more than one meaning, check that you have chosen the right pattern.

Remember also that closely related words (for example, a related verb and noun) sometimes have different patterns (see section 3.4), so always check the dictionary to make sure you have the right pattern for the right word.

4.3 Using the 'Get it right' boxes

The *Macmillan English Dictionary* includes over 100 'Get it right' boxes at individual dictionary entries. These boxes deal with many different issues that cause difficulties for learners. About 40 of these boxes give advice about complementation.

The following 'Get it right' boxes deal with complementation problems:

Source of error
verb patterns: <i>accept, access, afford, agree, approve, arrive, ask, attend, avoid, contribute, discuss, enter, help, make, marry, mean, pay, prevent, provide, risk, say, spend, stop, suggest, tell, think</i>
noun patterns: <i>ability, decrease, desire, difference, difficulty, increase, influence, interest, knowledge, need, possibility, reason, right, risk, solution, tendency</i>
adjective patterns: <i>capable, dependent, guilty, harmful, independent, responsible, worth</i>

Figure 7: Advice on avoiding errors using the *Macmillan English Dictionary* (section about Complementation, MED2)

2.4 EAP writing sections

Beside the Grammar Sections described above, the 'Improve your Writing Skills' component in the middle of the dictionary also includes twelve 'Writing Sections'. These Writing Sections focus on rhetorical or organisation functions that are particularly prominent in academic writing: (1) Adding information, (2) Comparing and Contrasting: Describing similarities and differences, (3) Exemplification: Introducing examples, (4) Expressing Cause and Effect, (5) Expressing Personal Opinions, (6) Expressing Possibility and Certainty, (7) Introducing a Concession, (8) Introducing Topics and Related Ideas, (9) Listing Items, (10) Reformulation: Paraphrasing or clarifying, (11) Quoting and reporting, and (12) Summarizing and Drawing conclusions. The Writing Sections were compiled on the basis of findings from a detailed analysis of 350 EAP markers (see Paquot 2007) in the *ICLE* corpus and in the native speaker corpus used in the project. As well as explaining the main strategies that writers can use to perform the twelve functions in formal writing, the sections also specifically address the wide range of problems advanced learners can be shown to experience when performing these functions. Special 'Be careful' notes, 'Get it right' boxes and Collocation boxes are included in the Writing Sections to help learners deal with problems of frequency, register, positioning, semantics and phraseology. For a detailed description of the Writing Sections and the method used to compile them, see Gilquin et al. (2007), Gilquin and Paquot (2008), and Granger and Paquot (forthcoming).

4. Conclusion

The outcomes of the collaborative project reported on in this paper are a good example of the invaluable contribution learner corpus-based research can make to MLDs. In this project trained lexicographers have worked in very close collaboration with learner corpus-based researchers (who are also experienced ELT teachers) to produce materials that can be seen to help boost the confidence of the MLD as a truly productive tool.

References

- De Cock, S., Gilquin, G., Granger, S., Lefer, M-A., Paquot, M., Ricketts, S. (2007) Improve your writing skills. In Rundell, M. (editor in chief) *Macmillan English Dictionary for Advanced Learners* (second edition). Oxford: Macmillan Education, IW1-IW50.
- De Cock, S. and Granger, S. (2004) Computer Learner Corpora and Monolingual Learners' Dictionaries: the Perfect Match. In Teubert W. and Mahlberg M. (eds.) *The Corpus Approach to Lexicography*. Special issue of *Lexicographica* 20: 72-86.
- Gillard, P. (ed.) (2003) *Cambridge Advanced Learner's Dictionary*. Cambridge: Cambridge University Press.
- Gilquin, G., Granger, S. and Paquot, M. (2007) Learner corpora: the missing link in EAP pedagogy. In Thompson, P. (ed.) *Corpus-based EAP Pedagogy* . Special issue of *Journal of English for Academic Purposes* 6(4): 319-335.
- Gilquin, G. and Paquot, M. (2008) Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1): 41-61.
- Granger, S. and Paquot, M.(forthcoming) From dictionary to phrasebook? In Proceedings of the 2008 EURALEX Conference..
- Paquot, M. (2007) Towards a productively-oriented academic word list. In Walinski J., Kredens K., Gozdz-Roszkowski S. (eds) *Corpora and ICT in language studies. PALC 2005*. Lodz studies in LANGUAGE 13. Frankfurt am Main: Peter Lang: 127-140.
- Rundell, M. (1999). "Dictionary use in production". *International Journal of Lexicography* 12(1): 35-53.
- Rundell, M. (ed.) (2007). *Macmillan English Dictionary for Advanced Learners* (second edition). Oxford: Macmillan Education.
- Summers, D. (ed.) (1995) *Longman Dictionary of Contemporary English*. Harlow: Longman.
- Summers, D. (ed.) (1993) *Longman Language Activator*. Harlow: Longman.
- Summers, D. (ed.) (1997) *Longman Essential Activator*. Harlow: Longman.