

# The English Grammar Profile of learner competence

## Methodology and key findings

Anne O’Keeffe and Geraldine Mark

Mary Immaculate College / University of Limerick

*English Profile* (EP) is an ongoing empirical exploration of learner English initiated by Cambridge University Press and Cambridge English, among others. EP aims to create a set of empirically-based descriptions of language competencies for English. ‘Reference Level Descriptors’ already exist as part of the Common European Framework of Reference (CEFR) but are intuitively derived and not designed for one specific language. The *English Grammar Profile* (EGP, [www.english-profile.org/english-grammar-profile](http://www.english-profile.org/english-grammar-profile)) is a sub-project of EP which aims to profile learner competence in grammar. This paper details the rationale for the study and the methodology that was developed to investigate the Cambridge Learner Corpus to arrive at over 1,200 grammatical competence statements. Key findings which link to existing corpus-based second language acquisition work are also presented.

**Keywords:** learner corpus, learner competence, learner grammar, CEFR, English Grammar Profile

### 1. Introduction

The Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001a) is an established benchmark for language competence (Jones & Saville 2009). It comprises six levels of competence from A1 (lowest) to C2 (highest). Anderson (2007: 660) notes that these levels have become a common currency in language education, prevalent in curricula, syllabuses, textbooks, teacher training courses. In the early 1990s, as part of a Council of Europe project, the CEFR established a set of statements, illustrated in Figures 1 and 2, defining what is minimally required for each stage within the framework in terms of grammar, vocabulary and skills development as well as functional and notional objectives. These

performance-based “can-do statements”, or ‘Reference Level Descriptors’, evolved from the collective judgements of a body of experts (Van Ek & Trim 1991a, 1991b).

- 
- C2 *Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.*
- C1 *Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.*
- B2 *Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.*
- B1 *Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.*
- A2 *Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language.  
Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.*
- A1 *Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.*
- 

**Figure 1.** Example of Descriptors within the CEFR for Overall Reading Comprehension, from levels C2 (highest) to A1 (lowest) (Council of Europe 2001b: 10)

### 3.3.4 asking someone for something

- **I’d (very much) like + NP (please)**  
I’d like a gin and tonic, please.
- **Please may I have + NP**  
Please may I have a piece of cake?
- **Can/could I have + NP (please)**  
Could I have a smaller piece, please?
- **Do you think I could have + NP please**  
Do you think I could have my tea without sugar, please?

**Figure 2.** Extract from the CEFR Vantage (B2) functional framework (Van Ek & Trim 2001: 47)

The CEFR is not without its critics (see, *inter alia*, Fulcher 2004, Weir 2005, O'Sullivan 2011), especially because it concerns itself primarily with the *validity* of performance-based descriptors of competence. Apart from reliability concerns (Osbourne 2014, Figueras et al. 2005), it is criticised for its eurocentric nature and for obliging teachers to work prescriptively (Anderson 2007). Carlsen (2012) notes the arbitrariness of such competence frameworks and points out that levels are rarely supported by empirical evidence. This is particularly pertinent when commercially competing language test developers link and calibrate their tests to these levels of proficiency. O'Sullivan (2011) starkly warns that different developers interpret the CEFR levels differently and therefore the resultant tests cannot be shown to be at the level they purport to be.

The language-neutral nature of the CEFR results in statements of competence that are quite generic and underspecified (see Milanovic 2009, Hawkins & Filipović 2012, Callies & Zaytseva 2013). Encouraged by the Council of Europe to (i) adapt the generic CEFR to local contexts and purposes, and (ii) address the need for a more empirical investigation, Cambridge University Press and Cambridge English Language Assessment, along with other partner institutions and individual researchers, established the English Profile (EP) research network in 2006. EP was founded as the English language arm alongside parallel programmes in other European countries (Harrison & Barker 2015: 3).

The EP is therefore an attempt to provide empirical detail about levels of competence in learner English, to complement and supplement the intuitively-derived language-neutral CEFR for languages. To date, a substantial amount of work has been done under the EP umbrella (see Harrison & Barker 2015; Capel 2010, 2015; Green 2012 on English functions). Central to this paper, the *English Grammar Profile* (EGP) resource, a database of over 1,200 empirically-derived statements, provides greater detail on learner grammar competence. The current authors were commissioned, as independent researchers, to use the Cambridge Learner Corpus (CLC) to investigate what learners at each of the six CEFR levels "can do" with grammar in English. The benefits and rationale for undertaking the EGP are clear but our challenge was to build a methodology for the large-scale task at hand. Many learner corpus studies have looked at individual grammatical items in learner English from one L1 background. Our brief was to look at all learner grammar, across all the CEFR levels. The aim of this paper is to detail the methodology of the EGP and to point generally to its key findings.

## 2. Corpus studies of learner language

Corpora and corpus linguistics, as Gablasova et al. (2017: 2) note, are becoming increasingly more central to the investigation of second language acquisition (SLA). In its early stages, learner corpus research was mostly concerned with describing learner language rather than interpreting it within the context of SLA theory (Myles 2015). From the early 1990s, the scale and impact of the pioneering International Corpus of Learner English (ICLE) project (Granger 1994) facilitated comparative observations of interlanguage, learner variation and levels of proficiency. A growing body of learner corpora has added much to our ongoing quest to understand and describe the factors at play in learning language (Gablasova et al. 2017, Ädel 2015, Callies 2015, Granger 2012). In recent years, with a greater SLA focus, learner corpus research has enhanced our understanding of the language learning process and has changed some long-held paradigms (e.g. Murakami 2013 on the influence of the learner's L1 on the universal order of acquisition theory).

Additionally, there is a growing body of work where learner corpus research is being used to triangulate experimental methods in psycholinguistics. Specifically, the use of corpus data has proved useful in evidencing usage-based theories of language which, in an SLA context, hold that usage leads to commonly occurring form-meaning pairings becoming entrenched as grammatical knowledge in the learner's mind, and that the degree of entrenchment is proportional to the frequency of usage (Ellis & Ferreira-Junior 2009: 188). Ellis & Ferreira-Junior (2009) propose that creative linguistic competence comes from the aggregation of the memories of all of the utterances in a learner's entire history of language use and experience. Learner corpora have proved essential in showing the evidence of learner formulaic use and have allowed the charting of the growth of learner usage (Ellis et al. 2015).

SLA researchers have also used corpora to address longitudinal research questions by using data that is divided according to cross-sectional groupings using variables such as age, year of study or proficiency level as a proxy for the variable of "time" (see also Section 3). As the present work hopes to show, the CEFR offers a useful structure for investigating proficiency.

Of relevance to this study, Thewissen (2013) investigates learner development (with the CEFR) by looking at error data. Thewissen (2013) finds lack of significant progress in accuracy between CEFR levels B2 and C2 (e.g. lexical phrases). She cites the accuracy-complexity trade-off effect whereby higher level learners' attempts to use more complex language increase their risk of error, hindering an improvement in accuracy. Thewissen (2013: 87) notes that this B2-C2 stabilization could be related to the 'ceiling effect' (after Milton & Meara 1995). She advises that although

errors still remain, a significant amount of learning has already taken place and she stresses that the stabilization tendency should not be narrowly interpreted as an absence of any development whatsoever. Drawing on Larsen-Freeman's (2006: 597) words, it is not "linguistic rigor mortis". The present study looks at patterns of grammatical development across all of the levels of the CEFR but, unlike Thewissen (2013), it investigates growth in learner competence rather than patterns of error. Essentially, this study focuses on the flip-side of errors and this brings some complementary insight to the ceiling effect which Thewissen (2013) has discussed (see Section 6).

### 3. Methodological issues in LCR

With the "remarkable developments" in LCR in a short period of time comes an increasing understanding of the need for careful learner corpus design (Granger et al. 2015: 2). To this end, there is growing debate and discourse about some key issues which we summarise here, as these are salient to our work.

#### 3.1 Contrastiveness and representativeness

Many LCR studies have compared the use of a linguistic feature in one or more L2 corpora with its use in L1 data. In doing so, they have focused on the frequency of a given linguistic pattern in the interlanguage often in relation to L1 transfer effects (Ädel 2015, Granger et al. 2015, Meunier 2015). Despite the obvious value of frequency analysis, many of these researchers have critically evaluated contrastive approaches, pointing to an inability to control variables. Gablasova et al. (2017) encourage critical interpretation of comparisons of frequency, particularly in the context of representativeness and comparability, and variation between language users and proficiency levels. They point out that most LCR studies link patterns in L2 production to the effects of L1 transfer, and emphasize the need to explore other effects such as task effect and L2 proficiency.

Issues of representativeness and comparability go hand in hand. As Gablasova et al. (2017) emphasize, if two corpora differ in too many respects, it may be difficult to ascertain the source of any difference observed. They argue that hitherto corpus-based SLA studies have largely ignored representativeness and comparability despite their major implications for the validity of corpus findings. While this study does not have a contrastive focus at its heart, there is an inherent contrastive element in that all evaluation takes place with reference to where L2 use converges with L1 use. The Cambridge Learner Corpus is a collection of exam data and there

is no comparable L1 corpus with respect to representativeness and task. We are not able to directly compare groups of language users with comparable backgrounds in comparable contexts and yet we are in need of a baseline comparison. We opted to use the BNC written corpus of c50 million words (excluding BNC fiction and BNC academic) for several reasons: firstly, because we were able to ascertain with a reasonable degree of certainty that the target linguistic features would be present in the data, secondly, to give us a baseline indication for typical frequencies of these structures, and, thirdly, because of the comparability of size and content of the two corpora. We are aware of the limitations of comparability with any other corpus and therefore we use the BNC written corpus solely as a baseline for frequency per million as described in the methodology in Section 5.

In making generalisations about a structure, Ädel (2015) discusses the issue of task effect and the need for a range (of contexts). Studies such as Myles (2004) and Tavakoli (2009) show that learners tend to use language that they are comfortable with. This can lead to an over-representation of some features over others. The current study also indicates over-representation in the form of exam ‘display’ where the learners use a language feature in order to display knowledge, often with a structure which is perceived to be structurally complex. The alternative issue of zero occurrence of a feature, as Granger (2012) notes, may be considered a matter of ability rather than choice in an L2 user whereas absence in an L1 speaker may be interpreted as choice rather than ability. We also note in this study where low occurrence of a feature may be due to a lack of opportunity (Buttery & Caines 2012). As regards overall representativeness in the Cambridge Learner Corpus, each exam script is tagged for candidate information and task details allowing the researchers to normalise frequencies across any one of these variables (see Section 4). However, we are aware that the Cambridge Learner Corpus solely represents the written output of students undertaking Cambridge English examinations and acknowledge the issues concerning generalisations about implicit internalised knowledge of the language from written production (Myles 2015).

### 3.2 Interspeaker variation and proficiency

Another area receiving attention in LCR is the issue of speaker variation and proficiency (Myles 2015). Variation has often been central to SLA studies, focusing on the effect of both speaker-internal (e.g. proficiency, age, educational history) and speaker-external (e.g. task, education history) variables. Gablasova et al. (2017) note that there has been a recent gradual shift towards consideration of speaker-external related variables (Ädel, 2008, Gablasova et al. 2015, Tracy-Ventura & Myles 2015) but there is a dearth of corpus-based studies on speaker-internal variables, such

as L2 proficiency, despite its use as a control and outcome variable in SLA studies not using corpora. Despite its significance, standards for establishing and reporting learner proficiency in L2 corpora lack uniformity, making comparability of results difficult. Estimates of proficiency are often based on year of study, which can realistically mean a span of actual proficiency levels within one grouping, as discussed by Díez-Bedmar (2012), Tono & Díez-Bedmar (2014) and Myles (2015). The CEFR, as observed by Tono & Díez-Bedmar (2014: 165), is increasingly being used as a standardising measure for comparing learner corpus data (cf. Díez-Bedmar 2012, Harrison & Barker 2015, Hawkins & Buttery 2009, Hawkins & Buttery 2010, Hawkins & Filipović 2012, Negishi et al. 2013, Thewissen 2013).

Central to the discussion of proficiency is the assumption that L2 speakers are working towards an ‘L1-norm’ and therefore that all L1 speakers are a homogenous group, producing similar language in similar contexts. As Larsen-Freeman (2014) points out, however, a homogenous native speaker speech community does not exist and yet, as many studies illustrate, the L1 speaker is depicted as a goal and L1 language as a kind of endpoint for L2 users. However, the notion of an L1 target is an idealized competence dependent on a consensus of “success” (Larsen-Freeman 2014).

Assignment of proficiency levels is central to this study. Each candidate is tagged for the exam taken as well as the level of proficiency achieved, benchmarked to the standardized scale of the CEFR. While there continues to be a debate about the reliability and validity relating to subjectivity in assigning levels, proficiency tagging is one of the main strengths of this study.

This research acknowledges the issues surrounding comparability, variability and proficiency. It takes a fine-grained descriptive approach to documenting the competencies that are evident at each CEFR level. The scale of the data allows us to see systematic aggregated use across a range of internal and external variables, provided by the metadata (see Section 4). They allow us to move from analysis of a single value (frequency of forms) to analysis against a range of values (including correct usage, spread of users, L2 background, context of use and task). We note some important departures in our study:

- i. Using the CEFR as the calibration for proficiency, we are looking at learner language across *all* six proficiency points;
- ii. We are investigating a summative aggregation of grammatical competences of learner English (from those who have taken Cambridge English exams). Hitherto studies usually examine one or a set of grammatical features;
- iii. We do not take a systematic contrastive focus between learners’ L1 and the target language;
- iv. We are not setting out to examine the learner English of one L1 cohort;

- v. We are not tracking the evolution of any one interlanguage feature;
- vi. We are investigating the quasi-longitudinal development;
- vii. We are not setting out to chart learner development in relation to errors. We are tracking and describing learner grammar competence towards an “idealized L1 norm”.

In summary, through the EGP, it is aimed to arrive at a corpus-based description of what learners can do with grammar at each level of the CEFR based on what they have written in Cambridge exams. The present study therefore represents a large-scale attempt to look at learner language at all levels across the CEFR (see Section 3.2). It is a quasi-longitudinal study tracking the evolution of a form and its use, as consistently used by different learners, in different contexts, with different L2 backgrounds, at given levels of proficiency.

#### 4. The data

According to Cambridge English Language Assessment figures, over 5 million Cambridge English exams are taken each year in 130 countries, across approximately 2,800 authorised exam centres ([www.cambridgeenglish.org/](http://www.cambridgeenglish.org/)). The Cambridge Learner Corpus (CLC) is an ever-growing collection of some of these exam data. At the time of writing, the CLC stands at over 55.5 million words (63,759,660 tokens). It comprises 266,600 exam documents, spanning 143 different first languages backgrounds, from a 17-year period (1993–2012). Only open-ended student writing is included in the CLC (i.e. it does not include gap-filling tasks). For this study we used pass scripts only.

Each document is tagged with metadata which includes candidate information: first language, nationality, level of education, age, gender; general exam information: exam taken, CEFR level, year of exam, exam performance; task specific information: question number, task style (e.g. advice, argument/opinion, complaint, criticism, descriptive, news), task format (e.g. article, essay, letter, email, report, proposal, speech, story), task register (e.g. formal, informal, neutral).

The data is hosted on the *Sketch Engine* (Kilgariff et al. 2014) online platform, using a bespoke instance of the corpus-analysis suite of tools customised for use on the CLC across the metadata (see Section 5). As well as giving us access to learner grammar use on a large and CEFR-calibrated scale, the CLC allowed us to break the data down in fine detail according to the metadata so as to refine our results. As exemplified below, this functionality was crucial to our methodology. Within the CLC, the exam data is spread across the CEFR levels as detailed in Table 1.



**Table 1.** Breakdown of exam level calibration in the CLC

| CEFR proficiency level | Number of words   | %     |
|------------------------|-------------------|-------|
| A1                     | 159,427           | 0.29  |
| A2                     | 3,614,974         | 6.56  |
| B1                     | 12,885,966        | 23.40 |
| B2                     | 15,710,082        | 28.52 |
| C1                     | 12,556,502        | 22.80 |
| C2                     | 10,148,733        | 18.43 |
| <b>Total</b>           | <b>55,075,684</b> |       |

In the open writing tasks that comprise the data, students have to complete writing tasks across informal, neutral and formal registers, over a range of task formats (letter/reference, essay, article, report, email, review, speech, etc.) and styles (critical, argument/opinion, complaint/apology/response, business, descriptive/creative, autobiographical, informative/news, application/response); see Appendix 1 for a profile of task formats and styles across levels. The number and range of these tasks is relative to the level of the exam. Examples of typical tasks are:

i. Informal/note/informative/news:

Your friend Sam is coming to your house tomorrow evening. Write a note to Sam. Tell Sam: what time to come; what to bring; how to get to your house. Write 25–35 words. (KET, A2)

ii. Neutral/letter/descriptive/informative:

Last week you went to a music concert. Now you are writing a letter to an English-speaking friend. Explain why you decided to go to the concert, give some information about who was performing and describe what happened at the concert. Finish the letter on your answer sheet, using about 100 words. (PET, B1)

iii. Formal/essay/argument/opinion:

Experts are constantly telling us what is good or bad for our health. Outline two or three of their concerns and say how valid you think their advice is. (About 350 words) (CPE, C2)

We also had access to the 1.9 billion word Cambridge English Corpus, which includes some existing spoken and written corpora such as the BNC (Burnard 1995), MICASE (Simpson et al. 2002), CANCODE (McCarthy 1998), CANBEC (Handford 2010), among others. As discussed, we elected to use the BNC written (excluding fiction and academic) as a comparison for frequencies per million (see Section 5).

## 5. Methodology

In this section, we detail our methodology, including how we established a search inventory for our analysis based on what we term the ‘ELT canon of grammatical structures’. As we outline, we began with a linear approach, which was frequency-driven, but this proved too simplistic. Ultimately, we evolved an iterative approach based on a range of key criteria. We overview the steps in this model for analysis and finish this section with a case study of the application of the model using one grammatical item.

### 5.1 Establishing a grammar search inventory

Within the ELT community there is widespread agreement on what grammar structures are taught and when (apart from advanced levels, see McCarthy 2013). We refer to this convergent understanding as the ELT canon of grammatical structures. This is perceived as a “must-teach” list of items in an English grammar syllabus which has evolved over many years, within a Community of Practice (Lave & Wenger 1991) and is perpetuated and sustained through materials and examinations, supported by a publishing industry that is seen as inherently conservative (Burton 2012, Littlejohn 1992).

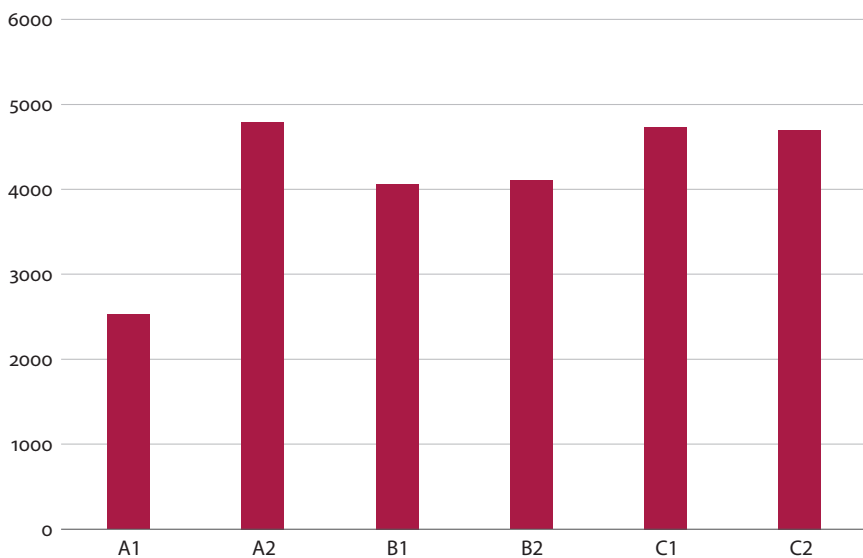
The ELT canon is a smörgåsbord of items, departing from a traditional linguist’s taxonomy, and typified by a mix of syntactic items and more thematic and functional clusterings such as “talking about the future”, “conditional sentences” and “reported and direct speech”, and so on. The terminology carries currency within the ELT community and it was important for EGP (as a pedagogical resource) to work within this understanding. To this end, we began by collating a list of all of the grammar items in the ELT canon (see Appendix 2 for the complete list). To arrive at this list, we collated the grammar syllabi of ELT pedagogical grammars and course books (see Appendix 3 for list of sources). We did not collate any information about the level at which these are typically presented in syllabi. This inventory of grammatical items was then categorised into 19 grammatical superordinate categories, each with multiple subordinate categories:

adjectives; adverbs; clauses; conjunctions; determiners; discourse markers; future time; modality; negation; nouns; passives; past time; prepositions; present time; pronouns; questions; reported speech verbs; focus.

## 5.2 Developing an iterative process

We first anticipated taking a linear approach to our analysis, searching for the canon of grammatical items in the data, from pass-only scripts, and pinpointing where items were being used competently by learners. The process of analysis which we ultimately evolved (as described in Section 5) was far more iterative. Here we illustrate why a linear process did not work.

Pilot examination of any grammatical item from the canon using a corpus query language (CQL) search across the CEFR levels gave us quantitative results for a given form, irrespective of its use. It quickly became apparent that this neat correlation between a grammatical item and a single level at which it is “competently” used was too simplistic. By way of example, a simple search for modal verb *would* (occurrences as *would* or modal *’d*), reveal the results shown in Figure 3:



**Figure 3.** Normalised frequencies per million words (PMW) of modal verb *would* across pass only data

Figure 3 shows a spread of forms and frequencies across the CEFR levels but represents none of the complexity of patterns for the modal verb either in a syntactic or functional sense. Quantitatively, we see that there are proportionally more instances at A1 than at A2. However, when we examine these instances qualitatively through concordance lines, we begin to see the texture and complexity of grammatical patterns that are at play. At A1, 80% of occurrences collocate only with the verb *like*,

typically in the string *I + would + like* as in: *I would like to eat some Spanish food* (A1, Italian, 2006). Apart from one instance of *you*, the pronoun *I* is the only pronoun used in the subject position at A1.

Further concordance analysis shows that by B1 level, learners have moved from using one typical pattern, *I + would + like*, to a range of patterns. Table 2 summarises the patterns that learners can do with the modal verb *would* at B1 which they cannot do at the previous levels, A1 and A2.

**Table 2.** Range of “can-dos” with *would* from Cambridge Learner Corpus at B1 level (from EGP)

| B1 “Can-do” statement  | Learner example   |
|--|---|
| Can use <i>would have + -ed</i> .  | That answer <i>would have changed</i> my life.<br>(B1, Italian, 2008)   |
| Can use <i>would not have + -ed</i> or <i>wouldn’t have + -ed</i> .  | Without your help, I <i>wouldn’t have got</i> this result. (B1, Spanish – European, 2010)   |
| Can use <i>would</i> in the main clause of a conditional sentence to talk about an imagined situation, often in the context of advice or opinion giving. | It’s your decision but if I had the chance, I <i>would go</i> with my friends. (B1, German 2008)  |
| Can use a limited range of adverbs with <i>would</i> , including <i>really</i> , <i>probably</i> , <i>certainly</i> , <i>definitely</i> .                | I <i>would definitely choose</i> to live by the seaside!<br>(B1, Chinese, 2010)   |
| Can use <i>would</i> to talk about the future in the past.   | [...] we thought that the film <i>would be</i> horrible because of the title but when the film started, the story was beautiful [...]<br>(B1, Portuguese, 2002) |
| Can use <i>would</i> to talk about imagined situations in the past.  | If I were in your shoes I <i>would have chosen</i> the school which is in the centre of town [...]<br>(B1, Farsi, 2008)   |
| Can use <i>would</i> as the reported form of <i>will</i> in reported clauses.  | I felt better when he said that he <i>would be</i> in Fenerbahce. (B1, Turkish, 2009)   |
| Can use the negative forms of <i>would</i> to talk about willingness in the past.  | We had to leave without you because the bus driver <i>wouldn’t wait</i> .<br>(B1, Spanish – Latin American, 2004)   |

The quantitative data hides a leap in the use of more complex grammatical patterns, as exemplified in the case of *would*. This acutely brought to light the need for a more refined methodology that took into consideration the developmental and evolving nature of competence, and complexity, across levels. What became clear was that there was unlikely to be a one-to-one correspondence between grammatical form (in its range of use) and level(s). Pilot investigations showed us that learners seemed

to learn a “form” and, at subsequent levels, there appeared to be growth in how and when they deploy it.

To this end, the methodology which we devised is best described as iterative in nature (see Figure 4). Within this process, there was need for qualitative as well as quantitative consideration. Due to the multiplicity of variables involved, we needed to set criteria to guide the process so as to take account of more variables beyond frequency of form.

### 5.2.1 *Establishing criteria*

Our methodology involved analysing the data with reference to six criteria. These are summarised and explained here. In Section 5.2.2, we take one structure through these steps by way of example.

- i. Criterion 1 – Frequency of use: Is there sufficient evidence of a structure at this level to warrant investigation?

The first criterion is frequency and this essentially involves finding the CEFR level at which typical occurrences of the structure first appear. It is important to note that because we are dealing with exam data, where grammatical items are being tested, their PMW frequency may be higher than in a broader non-exam corpus. As a benchmark for typical frequency, we used the BNC written corpus (see Section 3). This gave us a threshold below which PMW frequencies were not considered.

- ii. Criterion 2 – Rate of correct uses: Is there an adequate rate of correct uses?

Brown (1973) establishes a 90% accurate suppliance rate as the criterion for mastery of a grammatical or morphological structure in the study of children’s first language acquisition. In SLA studies, Wang (2013) proposes an accuracy cut off rate of 50% in a study of Chinese learners’ interlanguage (after Pienemann’s (1998) study which shows that accuracy rates develop with highly variable gradients in relation to grammatical items and individual learners in SLA). Because of the breadth of the CLC data, and based on pilot investigations, it was decided to raise the accuracy cut off rate to 60% to counter the varying impact of L1 and CEFR level span. Critical to this study, correct use of a grammatical form refers to both syntactic and pragmatic “correctness”. For instance, while the following example of the modal verb *must* in the pattern “pronoun + modal verb + main verb” is syntactically correct, it is pragmatically incorrect in this context of an invitation. In our search for 60% correct uses, instances such as this do not meet Criterion 2: *Hi Mike. Next Saturday in the morning I go to the park. If you want go to the park with me, you must come my house at 8.00 am in the morning. Park in nearly my house [...]* (Polish, A1, 2011).

- iii. Criterion 3 – Range of users: Is the usage distributed across a range of individual users?

This criterion relates to the dispersion of a given form under investigation. If it is the case that only a narrow range of learners are repeatedly using this form, it will not meet this criterion (despite having met Criteria 1 and 2). Dispersion can be measured statistically, for example, using standard deviation (McEney et al. 2006).

- iv. Criterion 4 – Spread of first language families: Is the usage distributed across a range of first language families?

This criterion is very relevant in the case of the CLC. Since one of the central aims of the study is to aggregate the learner data from a large number of L1 backgrounds, it would be flawed to base a competence statement on results dominated by a small number of language families. The results had to be spread over multiple language families. If more than 30% of results are produced by one language family, Criterion 4 was not met.

- v. Criterion 5 – Spread of contexts of use: Is the usage distributed across a range of contexts?

This criterion addresses whether the occurrences of a given form, at a given level, are adequately spread across a number of contexts of use, relative to the form. Clearly not all forms can appear in all contexts but to meet Criterion 5 results must be spread over a range of “styles”, “formats” and “registers” (Appendix 1). These are pre-defined meta data categorisations within the CLC bespoke instance of *Sketch Engine*. In some instances, however the specific use of a feature is restricted by a particular style or register. In such cases, this is noted in the competence statement.

- vi. Criterion 6 – Avoiding the effect of a task: Is the usage affected by a task?

Task effect is where patterns recur because of a task. This can often be as a result of a task rubric where learners see a form and then reproduce it, either as a display tactic or out of necessity. For example, a letter requesting a change to the time of an appointment would yield a high frequency of the pattern *Would you mind [...]*. This can be quickly checked automatically using the metadata in *Sketch Engine*.

### 5.2.2 Steps in the iterative process

The iterative framework for using this criteria-based approach is summarised in Figure 4. The framework charts the process of establishing what learners can do with a given grammatical item at a given level. Each grammatical item under scrutiny at any given level must begin at Step 1 and go through four steps of the process, and meet all of Criteria 1 to 6 (Step 2) to be considered competently used at this level.

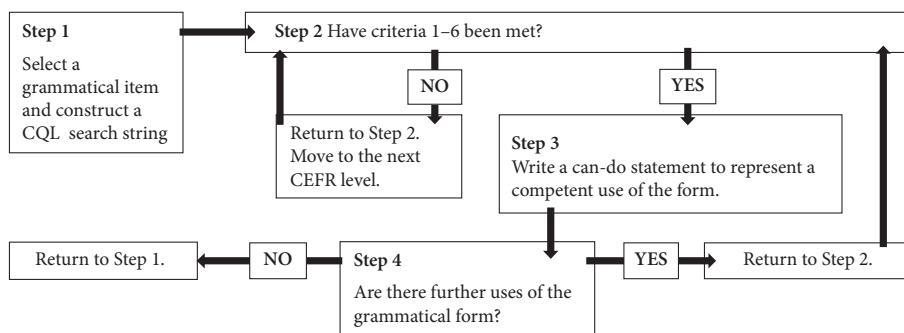


Figure 4. Iterative process of establishing grammatical can-do statements

The four steps in our process of analysis are summarised in Table 3:

Table 3. Steps in the iterative process

| Steps |   |
|-------|---|
| 1     | Select a grammatical item. Identify the search terms/exponents of a grammatical item and construct an appropriate CQL search string.  |
| 2     | Scrutinising the results, establish whether the grammatical item at that level meets the Criteria 1 to 6. If so, move to Step 3. If not, move to the next CEFR level of data and repeat Step 2 until data at a given level meets Criteria 1 to 6. |
| 3     | Establish a “can-do” statement to represent the use of a grammatical item with a particular form and/or use, at a given level.  |
| 4     | Identify whether or not there are other uses of the grammatical form defined by the same search string. If there are, return to Step 2 and begin the process again. If not return to Step 1 and select a new grammatical item.                    |

### 5.3 Exemplifying the iterative process using a criteria-based methodology

In this section, we take one grammatical item through the steps of the criteria. We will use what is referred to in ELT as ‘the past perfect simple’ in its affirmative form (e.g. *I had eaten*) as our example.

In Step 1, we construct an appropriate CQL search string for the search item, in this case: [word="had'd"][tag="RB"]{0,2}[tag="V.N"]. This string is intended to capture affirmative instances of the past perfect simple, with an optional adverb [RB] in mid-position, before the past participle [V.N].

In Step 2, we establish whether evidence of use of the grammatical item at that level meets the competence Criteria 1 to 6:

## i. Criterion 1 – Frequency of use:

Using the search string from Step 1, generate the frequencies across CEFR levels (see Table 4).

**Table 4.** Raw and normalised frequencies of past perfect simple forms across CEFR levels

| Level | Raw frequencies | PMW  |
|-------|-----------------|------|
| A1    | 1               | 11   |
| A2    | 194             | 188  |
| B1    | 2,730           | 712  |
| B2    | 8,093           | 1333 |
| C1    | 2,756           | 460  |
| C2    | 13,213          | 1999 |

As our guide, the BNC written corpus shows a frequency of 61 PMW for the search item. This provides our threshold frequency from a non-exam context and guides us to start our search at A2 level, which is the nearest PMW frequency above the BNC written threshold.

## ii. Criterion 2 – Rate of correct uses:

By looking qualitatively at the concordances of the 194 instances of the past perfect simple affirmative at A2, we find that there is not an adequate number of correct uses. To calculate the error rate within the CLC, we use the error-coded component of the corpus. We note that in some cases the form is syntactically correct but used incorrectly, often instead of a past simple form, for example: *I <#TV> had seen | saw </#TV><#MD> | the </#MD> Akropolis yesterday* (A2, Dutch, 1999) (within CLC error codes, TV = tense error; MD = missing determiner). In total at A2, there are 96 tense errors out of 194 occurrences of the form which means an error rate of 49.5% (i.e. 51.5% correct uses). Our threshold for correct uses is set at 60% so this does not meet the criterion. Since Criterion 2 has not been met, we investigate the data at the next CEFR level (B1) and begin again with Criterion 1. At B1 level, we can say that the data meets Criterion 1, in that there is above the threshold frequency of occurrence at B1 (see Table 4). Under Criterion 2 (rate of correct uses), there was a correct use rate of 81.5% (i.e. an 18.5% error rate). This meets the criterion for rate of correct uses and so we move to Criterion 3.

## iii. Criterion 3 – Spread of users:

As mentioned, dispersion can be calculated across the tasks at this level using standard deviation. However, while our *Sketch Engine* interface showed us all individual learners document at B1 level and the frequency of the pattern of



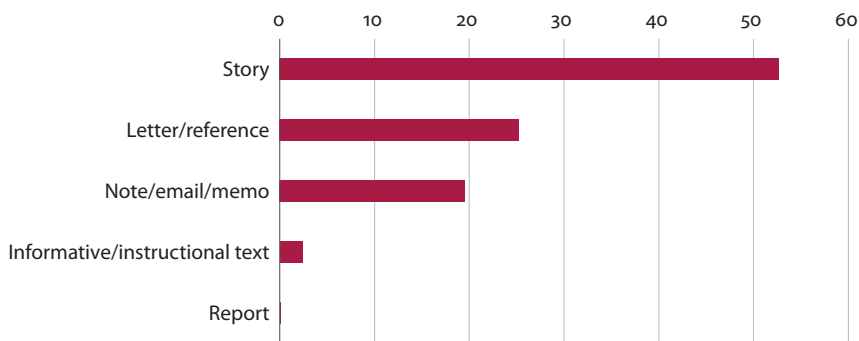
occurrences in each document, it did not allow us to automatically calculate the standard deviation (it could be done manually by copying and pasting almost 40 pages of results but this was not plausible in the scale of our overall task). Our means of checking dispersion relied on (i) manually examining the frequencies across the documents, screen by screen; (ii) calculating the mean; and (iii) calculating the range (i.e. the difference between the highest and the lowest frequencies). On calculating the mean, we could see a result of 1.5 uses per document. The range was from a max of 10 uses (in one document) to a minimum of 1 use, which gave a range of 9. Based on these qualitative and quantitative explorations, we are happy that the form meets the criterion of spread of users.

iv. Criterion 4 – Spread of first language families:

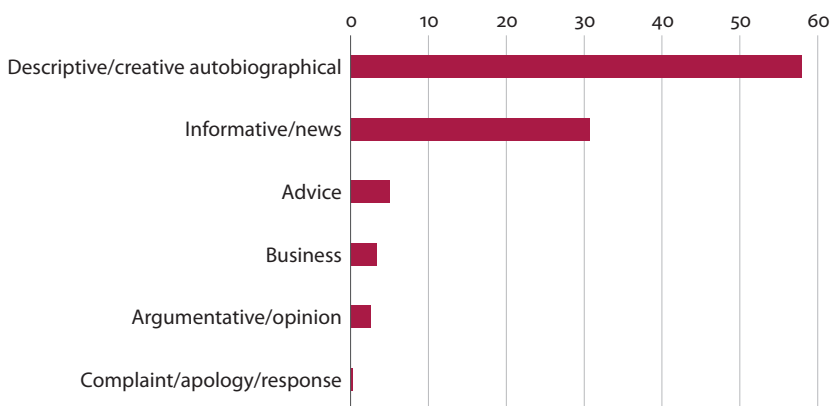
The form is used across 50 first languages, at B1 level, and these represent many language families including Indo-European, Afro-Asiatic and Austronesian.

v. Criterion 5 – Spread of contexts of use:

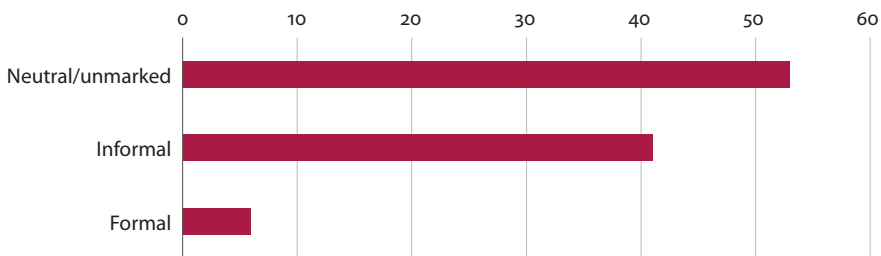
Based on the BNC written data, we expect the past perfect simple to span a range of contexts of use. The metadata show us a spread of formats, stylistic contexts (tagged in CLC as “style”) and registers within the CLC bespoke instance of *Sketch Engine* (Figures 5–7) and so we can say that the form meets this criterion.



**Figure 5.** Spread of past perfect simple affirmative occurrences across “formats”, at B1 (percentage)



**Figure 6.** Spread of the past perfect simple affirmative occurrences across “style”, at B1 level (percentage)



**Figure 7.** Spread of the past perfect simple affirmative occurrences across “register” at B1 level (percentage)

vi. Criterion 6 – Avoiding the effect of a task:

*Sketch Engine*’s visualization function allows us look at where the form under investigation occurred in the corpus across ten time intervals from 1999 to 2012. This is a useful way of spotting the possibility of task effect across this large dataset. As Figure 8 shows, there is abnormally high frequency of this form around the 60% mark in the corpus position (which equated to 2008). We can instantly check whether this was generated by one task through concordancing these two peak frequencies.

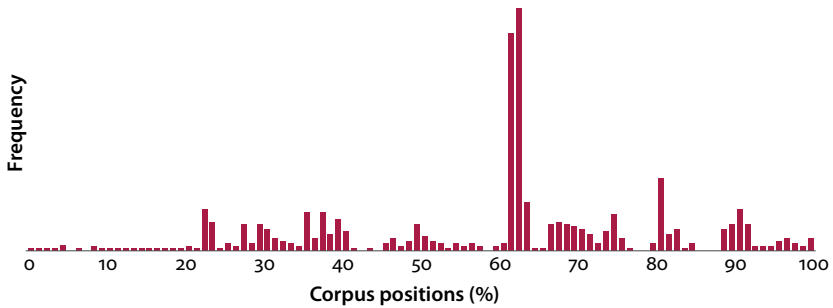


Figure 8. *Sketch Engine* visualization of frequency over corpus position (1999–2012) at B1 pass only scripts

All of the instances at the two peak points come from 2008 but the concordances do not show one task having a bearing on these uses. The patterns generated show a range rather than a convergence of verb forms (Figure 9). Based on this, we can say that the frequencies are not skewed by one task in the data.

At this point, we can say that the past perfect simple affirmative at B1 satisfies Criteria 1 to 6.

|  |               |  |
|--|---------------|--|
| envelope I smiled. I smiled because I knew that nobody else could have written that, yes, he and only he, smiled. I had been waiting for this letter for more than two weeks. Indeed Delphine, my twin sister, As soon as I saw the handwriting on the envelope I smiled. The letter I was waiting for many time.  | had written   | me that letter. I smiled completely in love. I had waited for that letter many time and right now he for Australia on June 6 and I was worrying about her. Australia is far away from France, it is a long |
| <p> Bye. </p><p> As soon as I saw the handwriting on the envelope I smiled. Finally my friend Alexandra the money she had, but a green light took his wand from his hand. It was Sarah my cousin that already to Egypt she found her lost sister. She had been looking for her for a long time. </p><p> Laura also Tom and we slowly became more and more friends. With him I could speak about anything and he always               | had left      | had come   |
| <original_answer></p> As soon as I saw the handwriting on the envelope I smiled. My penpal Aly from America nice and they could pay for everything but on another hand I would be so bored! I spoke with them and  | had sent      | had left   |
| </p><p> Hi Alice, yesterday there was a TV company at my school. They chose our school because our band envelope I smiled: "My mother still reminds me", I said. That envelope was surely a gift for me, because soon as see the handwriting on the envelope I smiled. I couldn't believe my eyes, my old boyfriends handwriting on the envelope I smiled. </p><p> I had been waiting for that letter for weeks. My friend Charlotte | had left      | had put  |
| <original_answer></p><p> Dear Alice </p><p> Yesterday, there was a TV company at my school. My class danger of drugs, the reason is that many students had even been faced to drugs. Seven of my classmates  | had helped    | had returned   |
|  | had explained | had won  |
|  | had managed   | had written  |
|  | had moved     | had made   |
|  | had been      |  |

Figure 9. Sample of concordance of past perfect from the 60% peak point (2008) in the visualization (see Figure 8)

In Step 3, we establish a “can-do” statement to represent the use of a grammatical item with a particular form and/or use, at a particular level. Based on the evidence in the CLC in relation to the past perfect simple affirmative, at B1, the following “can-do” statements (competency descriptors) are established (Table 5).

At B1, we find the affirmative form is competently used. We also find the pattern is competently used with a limited range of adverbs. In addition to these two forms, we then describe the use which meets the six criteria, illustrated in Table 5. ‘Form’ refers to the syntactic structure of the grammatical item whereas ‘use’ describes how

**Table 5.** Range of “can-do” statements for the past perfect simple form at B1 level

| “Can-do” statement   | Examples from the CLC  |
|--|--|
| FORM: Can use the affirmative form of the past perfect simple.   | <i>I’d forgotten</i> that I have an appointment with the doctor [...]<br>(B1, Spanish – Latin American, 2008)                          |
| FORM: Can use the past perfect simple with a limited range of adverbs (including <i>never, ever, just, always, already</i> ) in the normal mid-position. | <i>I had always wanted</i> to visit European cities and Paris was one of the easiest cities to get to from England. (B1, Korean, 1999) |
| USE: Can use the past perfect simple to talk about a time before another time in the past.   | Actually <i>I’d arranged</i> an appointment with my doctor before calling you.<br>(B1, Arabic – Meghreb, 2008)                         |

a syntactic structure functions. This distinction between form and use is common place in pedagogical grammars in ELT (see for Swan 2005, Carter et al. 2011) and, for this project, it was important that the EGP resource also made this distinction.

In Step 4, we identify whether or not there are other uses of the grammatical form defined by the same search string. If there are, we return to Step 2 and begin the process again. If not, we return to Step 1 and select a new grammatical item. As there are no further uses of the past perfect simple affirmative form at B1 level, we then move back to Step 2 and begin the iterative process again with B2 data, going through the six criteria, so as to look at the same form at the subsequent levels to appraise any new uses that emerge (for description of new uses at B2, and beyond, in the EGP online).


As part of our analysis at B2, we also note the growth in lexical range in relation to the use of adverbs. While at B1, we noted that learners can use the past perfect simple with a limited range of adverbs (including *never, ever, just, always, already*) in the normal mid-position (Table 5); we note at B2 that learners can use the past perfect simple with a wide range of adverbs (including *finally, recently, simply*) in the normal mid-position (see Section 6.1).




## 6. Discussion of initial findings

In the EGP, our focus was on the development of grammar competence rather than on patterns of error decline, plateau or regression. We see our work on grammar competence as a useful complement to studies on errors. In combination, looking at error and competence gives a fuller picture of learner grammar. Here we spotlight and discuss briefly some of the indicative insights from our research which

parallel or complement previous corpus and non-corpus SLA work on learner grammar. These correlations will require further investigation in our ongoing work with the CLC.

## 6.1 Lexico-grammatical development and the ceiling effect

Our findings offer insight into lexico-grammatical development which complements corpus-based SLA work discussed in Section 2. In summary, we observed that knowing more lexis appears to go hand-in-hand with an expanding repertoire of grammatical uses of a form, level by level. In differing ways, this complements the work of Thewissen (2013) and work by Ellis and his associates (e.g. Ellis et al. 2015). For instance, in the case of the past simple affirmative form, the syntactic pattern can be used competently at A1 but it can only be used with a limited range of verbs. This is because a learner at A1 has not yet acquired many verbs (this is marked in EGP resource with the lexical range symbol , see Figure 10):

|                             |  |
|-----------------------------|--|
| Element                     |  FORM: AFFIRMATIVE  |
| SuperCat                    | PAST   |
| SubCat                      | past simple  |
| Lexical Range               |   |
| Level                       |  A1   |
| Cando                       | Can use the affirmative form with a limited range of regular and irregular verbs.  |
| Corrected Learner Example   | Every day at college I went to class at 9 am. (A1 BREAKTHROUGH; 2007; Portuguese; Pass)<br>I met her about ten years ago. (A1 BREAKTHROUGH; 2006; Others; Pass)<br>My grandmother lived in a village and I often went there. (A1 BREAKTHROUGH; 2007; Polish; Pass) |
| Uncorrected Learner Example |  |
| Comments                    | N/A  |

**Figure 10.** Extract from the EGP showing lexical range of the past simple affirmative at A1 (Cambridge University Press 2015)

By A2, learners can use the same form with an increasing range of verbs and, by B1, the learner can use it with a wide range of verbs echoing the morphosyntactic pattern of acquisition discussed in Ellis et al. (2015) (cf. Section 2).

Adverbs, as discussed in Section 5, offer another example. At A1 level, learners can use the “adverb + adjective” pattern, as in *He’s a very good man* (A1, Mandarin, 2006). Again, while the syntactic form stabilises at A1, it is constrained by the lexical repertoire of this level, in this case: “very + adjective”. By C1 level, we see

the same basic pattern being used with an advanced lexical repertoire to bring pragmatic force to a sentence as part of a focusing device: *It is painfully obvious that the Internet is crucial thing nowadays and it is helpful in improving language skills* (C1, Polish, 2006).

One could say that the pattern has stabilised by A1 level but this hides the sophisticated growth in use of this form relative to lexical and pragmatic development. Borrowing the notion of polysemy in vocabulary development, we see a type of “grammatical polysemy”. In vocabulary development, lower level learners might know one meaning of a polysemic form, such as *bark* (a dog barking); as their lexical repertoire grows, so does their polysemic range so that by C1 a learner knows the metaphoric use of *bark* (shouting) and *bark* in the sense of the outer layer of a tree (as detailed in the *English Vocabulary Profile*). Similarly, learners acquire a grammatical pattern at a lower level and, as they learn more vocabulary, not only are they able to deploy the same grammatical pattern to a wider range of meanings and for greater pragmatic effect, they also become aware of the collocational and colligational limitations of the pattern. Take, for example, a wider syntactic context of the “adverb + adjective” combination (“pronoun + linking verb (+ adverb) + adjective + (*that*)-clause”) and an exponent at each of the three levels of the CEFR (Table 6).

**Table 6.** Examples of “pronoun + linking verb (+ adverb) + adjective + (*that*)-clause” across levels

| Level | “pronoun + linking verb (+ adverb) + adjective + ( <i>that</i> )-clause”                          |                       |
|-------|---|-----------------------|
| A2    | <i>I am sure</i> we will find something to do.  | (A2, Norwegian, 2003) |
| B2    | <i>It seems obvious that</i> this oil comes from the gas station.                                 | (B2, French, 2008)    |
| C1    | <i>It is highly unlikely that</i> the goods can vanish from your warehouse without being noticed. | (C1, Russian, 2008)   |

In these examples, the syntactic choices remain constant but at the paradigmatic level (Halliday 1985), the learner becomes increasingly aware of the range of lexis they can use in each “slot” and how one word might be lexically primed to follow another (Hoey 2005).

As outlined in Section 2, Thewissen (2013) discusses a ceiling effect where learning appears to stabilise. In line with this, we too see the development pattern where learners acquire a form and this stabilises at a lower level. At this point, a form seems to have reached its syntactic ‘developmental endpoint’ (after Thewissen 2013) at A and B levels but as learners go beyond these levels, they deploy the stabilised forms with a greater complexity of meaning and with greater dexterity of use. These findings echo Larsen-Freeman’s (2015: 496) encouragement to consider learner grammar as a dynamic system which rather than being complete and acquired is “never complete” and is in constant development.

## 6.2 Pragmatic development

As discussed above (Section 5 and elsewhere), we see pragmatic competence as part of the bigger picture of competence. This complements work such as Bardovi-Harlig (2013) and Bardovi-Harlig & Bastos (2011). In line with such studies, our research shows many instances where, a form is acquired and learners then develop more pragmatic competence in its use. At higher levels, this complexity and dexterity, described in Section 6.1 is constantly borne out in how a learner can use a syntactic form pragmatically and how they are able to skillfully play with a form for greater subtlety of meaning or focus. The past simple tense offers one of many examples of this developing pragmatic competence where learners acquire a form at A1 and use it in more and more complex ways as they progress (Table 7). At B2, for instance, the form is deployed for pragmatic effect where learners use the past simple affirmative form of the verbs *wonder* and *want* as a politeness structure in thanking and requesting.

**Table 7.** A sample of the “can-do” statements and examples, from the EGP, illustrating the use of the affirmative form of the past simple (affirmative form), at A1 and B2 level

| Level | “Can-do” statement   | Learner example  |
|-------|--|--|
| A1    | Can use the affirmative form with a limited range of regular and irregular verbs.  | My grandmother <i>lived</i> in a village and I often <i>went</i> there.<br>(A1, Polish, 2007)  |
| B2    | Can use the past simple with <i>I wondered</i> and <i>I wanted</i> as politeness structures, when making polite requests and thanking. | I <i>wanted</i> to know if the rooms are single or double, if they have showers and if there is room service? Secondly, I also <i>wanted</i> to know if there is a gym in the hotel?<br>(B2, Spanish – European, 1997) |

## 7. Conclusion

The English Grammar Profile project set out to characterise learner grammar competence across the CEFR levels, using the quasi-longitudinal CLC. The original goal was to create an open online resource to enhance our understanding of what learners can do (rather than cannot do) with English grammar. In doing so, a criteria-based approach was taken in order to investigate competence. Focussing on competence rather than errors allowed for the description of an emerging picture of syntactic, functional and pragmatic competence beyond the point of stabilisation of form alone.

By way of concluding remarks, it is hoped that the methodology and the grammar profile results (see EGP online) will lend to furthering the empirical basis for and understanding of what it means to say a learner is at a particular level of the CEFR, in terms of their use of grammar. However, we note limitations to our study. It is limited to written exam data, from Cambridge exams. The calibration of the data to the CEFR is solely within the assessment criteria of this examination board (as discussed in Section 1). However, we note that, through the EP project, the CLC has been opened up for large-scale corpus linguistic analyses and so the results of EP research have an important role to play in investigating what language a learner can use proficiently at any given level of the Cambridge exams. It will be important to have access to more learner exam data, from other major exam boards, to compare their calibrations of competence at any given CEFR level. The competence statements derived from our study are openly available and can be compared against any other examination dataset to expose any anomalies and issues alluded to in O’Sullivan (2011) and Gablasova et al. (2017). We also note the opening up of learner exam corpus data by other major examination boards, such as the Spoken Learner Corpus Project, a collaboration between Trinity College London and the Centre for Corpus Approaches to Social Science. This corpus, to date, comprises 300 hours of transcribed speech from Trinity College London oral exams forming a corpus of 3.5 million words, across 1,500 L2 English speakers, at B1 to C2 level CEFR proficiency levels, across 9 L1 backgrounds.

A replication of this study using CEFR-calibrated oral exam data would be insightful, as would parallel work using non-Cambridge exam data. Another crucial study would be to look at grammar competence within non-examination CEFR-calibrated learner data, both spoken and written, so as to examine competence in a broader range of use, beyond the constraints of an exam. We hope that our work will form a comparative baseline for other research into CEFR-calibrated learner language and that our methodology can be refined within that endeavour.

## References

- Ädel, A. (2015). Variability in learner corpora. In Granger, S., Gilquin, G., & Meunier, F. (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 401–421). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.018
- Anderson, C. (2007). The CEFR and the need for more research. *Modern Language Journal*, 91(4), 659–663. doi:10.1111/j.1540-4781.2007.00627\_4.x
- Bardovi-Harlig, K. (2013). Developing L2 pragmatics. *Language Learning*, 63(s1), 68–86. doi:10.1111/j.1467-9922.2012.00738.x



- Bardovi-Harlig, K., & Bastos, M. T. (2011). Proficiency, length of stay, and intensity of interaction, and the acquisition of conventional expressions in L2 pragmatics. *Intercultural Pragmatics*, 8(3), 347–384. doi:10.1515/ipro.2011.017
- Brown, R. (1973). *A First Language: The Early Stages*. London: George Allen & Unwin. doi:10.4159/harvard.9780674732469
- Burnard, L. (1995). *Users' Reference Guide to British National Corpus*. Oxford: Oxford University Computing Services.
- Burton, G. (2012). Corpora and coursebooks: Destined to be strangers forever? *Corpora*, 7(1), 91–108. doi:10.3366/cor.2012.0019
- Buttery, P., & Caines, A. (2012). Normalising frequency counts to account for 'opportunity of use' in learner corpora. In Y. Tono, Y. Kawaguchi & M. Minegishi (Eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* (pp. 187–204). Amsterdam/Philadelphia: John Benjamins. doi:10.1075/tufs.4.16but
- Callies, M. (2015). Learner corpus methodology. In Granger, S., Gilquin, G., & Meunier, F. (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 35–55). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.003
- Callies, M., & Zaytseva, E. (2013). The Corpus of Academic Learner English (CALE): A new resource for the study and assessment of advanced language proficiency. In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Corpora and Language in Use – Proceedings 1* (pp. 49–59). Louvain-la-Neuve: Presses Universitaires de Louvain.
- Cambridge University Press (2015). *English Grammar Profile*. Cambridge: Cambridge University Press. Retrieved from <http://www.englishprofile.org/english-grammar-profile/egp-online> (last accessed March 2016).
- Capel, A. (2010). A1 – B2 Vocabulary: Insights and issues arising from the English Profile Wordlists project. *English Grammar Profile Journal*, 1(1), 1–11. Retrieved from [http://journals.cambridge.org/download.php?file=%2F1715\\_9743C07C92A1D32D9C5923E4578849FE\\_journals\\_\\_EPJ\\_EPJ1\\_01\\_S204153621000048a.pdf&cover=Y&code=b5d9787b-d17ae6a4a22b52fd9ff5084a](http://journals.cambridge.org/download.php?file=%2F1715_9743C07C92A1D32D9C5923E4578849FE_journals__EPJ_EPJ1_01_S204153621000048a.pdf&cover=Y&code=b5d9787b-d17ae6a4a22b52fd9ff5084a) (last accessed November 2014).
- Capel, A. (2015). The English Vocabulary Profile. In J. Harrison & F. Barker (Eds.), *English Profile in Practice, English Profile Studies*, Vol. 5 (pp. 9–27). Cambridge: Cambridge University Press.
- Carlsen, C. (2012). Proficiency level: A fuzzy variable in computer learner corpora. *Applied Linguistics*, 3(2), 161–183. doi:10.1093/applin/amr047
- Carter, R., McCarthy, M., Mark, G., & O'Keeffe, A. (2011). *English Grammar Today*. Cambridge: Cambridge University Press.
- Council of Europe (2001a). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Council of Europe (2001b). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment Structured Overview of all CEFR Scales*. Strasbourg: Council of Europe. Retrieved from [http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key\\_reference/Overview\\_CEFRscales\\_EN.pdf](http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/Overview_CEFRscales_EN.pdf) (last accessed November 2014).
- Díez-Bedmar, M. B. (2012). The use of the Common European Framework of Reference for Languages to evaluate compositions in the English exam section of the University admission examination. *Revista de Educación*, 357, 55–79.

- Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 7(1), 188–221. doi:10.1075/arcl.7.o8ell
- Ellis, N. C., O'Donnell, M., & Römer, U. (2015). Usage-based language learning. In B. MacWhinney & W. O'Grady (Eds.), *The Handbook of Language Emergence* (pp. 163–180). Hoboken, NJ: John Wiley & Sons, Inc.
- Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing*, 22(3), 262–279. doi:10.1191/0265532205lt3080a
- Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, 1(4), 253–266. doi:10.1207/s15434311laq0104\_4
- Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, Advance online publication. doi:10.1093/applin/amv055
- Gablasova, D., Brezina, V., & McEnery, T. (2017). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 67(s1), 130–154.
- Granger, S. (1994). The learner corpus: A revolution in applied linguistics. *English Today*, 10(3), 25–33. doi:10.1017/S0266078400007665
- Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 7–29). Malden: Blackwell. doi:10.1002/9781444347340.ch2
- Granger, S., Gilquin, G., & Meunier, F. (Eds.) (2015). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414
- Green, A. (2012). *English Profile Studies. Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition across the Ability Range*. Cambridge: UCLES/Cambridge University Press.
- Halliday, M. A. K. (1985). *Introduction to Functional Grammar*. London: Edward Arnold.
- Handford, M. (2010). *The Language of Business Meetings*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139525329
- Harrison, J., & Barker, F. (Eds.) (2015). *English Profile in Practice. English Profile Studies*, Vol. 5. Cambridge: Cambridge University Press.
- Hawkins, J. A., & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency: Insights from the English Profile Programme. In L. Taylor & C. J. Weir (Eds.), *Language Testing Matters: Investigating the Wider Social and Educational Impact of Assessment* (pp. 158–175). Cambridge: Cambridge University Press.
- Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(1). doi:10.1017/S2041536210000103
- Hawkins, J., & Filipović, L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. Cambridge: Cambridge University Press.
- Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Jones, N., & Saville, N. (2009). European language policy: Assessment, learning, and the CEFR. *Annual Review of Applied Linguistics*, 29, 51–63. doi:10.1017/S0267190509090059
- Kilgarrriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36. doi:10.1007/s40607-014-0009-9

- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4), 590–619. doi:10.1093/applin/aml029
- Larsen-Freeman, D. (2014). Another step to be taken: Rethinking the end point of the interlanguage continuum. In Z. -H. Han & E. Tarone (Eds.), *Interlanguage: Forty Years Later* (pp. 203–220). Amsterdam/Philadelphia: John Benjamins. doi:10.1075/llt.39.11ch9
- Larsen-Freeman, D. (2015). Saying what we mean: Making a case for 'language acquisition' to become 'language development'. *Language Teaching*, 48(4), 491–505. doi:10.1017/S0261444814000019
- Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511815355
- Littlejohn, A. (1992). *Why are ELT coursebooks the way they are?* (Unpublished doctoral dissertation). Lancaster University, Lancaster, UK.
- McCarthy, M. J. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.
- McCarthy, M. J. (2013). Corpora and the advanced level: Problems and prospects. *English Australia Journal*, 29(1), 39–49.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge
- Meunier, F. (2015). Developmental patterns in learner corpora. In Granger, S., Gilquin, G., & Meunier, F. (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 379–400). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.017
- Milanovic, M. (2009). *Cambridge ESOL and the CEFR*. University of Cambridge ESOL Examinations Research Notes, 37, 2–5. Retrieved from <http://www.cambridgeenglish.org/images/23156-research-notes-37.pdf> (last accessed November 2014).
- Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL Review of Applied Linguistics*, (107–08), 17–34.
- Murakami, A. (2013). Cross-linguistic influence on the accuracy order of L2 English grammatical morphemes. In S. Granger, S. Gaëtanelle & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research. Looking back, Moving ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)* (Vol. 1) (pp. 325–334). Louvain: Presses universitaires de Louvain.
- Myles, F. (2004). From data to theory: The over-representation of linguistic knowledge in SLA. *Transactions of the Philological Society*, 102(2), 139–168 doi:10.1111/j.0079-1636.2004.00133.x
- Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 309–331). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139649414.014
- Negishi, M., Tono, Y., & Fujita, Y. (2013). A validation study of the CEFR levels of phrasal verbs in the English Vocabulary Profile. *English Profile Journal*, 3, 1–16. Retrieved from <http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8663926> (last accessed November 2014).
- Osbourne, J. (2014). Multiple assessments of oral proficiency: Evidence from a collaborative platform. In P. Leclercq, A. Edmonds & H. Hilton (Eds.), *Measuring L2 Proficiency: Perspectives from SLA* (pp. 54–70). Bristol: Multilingual Matters.
- O'Sullivan, B. (2011). Language testing. In J. Simpson (Ed.), *The Routledge Handbook of Applied Linguistics* (pp. 259–273). London: Routledge.
- Pienemann, M. (1998). Developmental dynamics in L1 and L2 acquisition: Processability Theory and generative entrenchment. *Bilingualism*, 1(1), 1–20. doi:10.1017/S1366728998000017

- Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.
- Swan, M. (2005). *Practical English Usage* (3rd ed.). Oxford: Oxford University Press.
- Tavakoli, P. (2009). Assessing L2 task performance: Understanding effects of task design. *System*, 37(3), 482–495. doi:10.1016/j.system.2009.02.013
- Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error tagged learner corpus. *The Modern Language Journal*, 97(S1), 77–101. doi:10.1111/j.1540-4781.2012.01422.x
- Tono, Y., & Díez-Bedmar, M. B. (2014). Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics*, 19(2), 163–177. doi:10.1075/ijcl.19.2.01ton
- Tracy-Ventura, N., & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, 1(1), 58–95.
- Van Ek, J. A., & Trim, J. L. M. (1991a). *Waystage 1990*. Cambridge: Council of Europe/Cambridge University Press.
- Van Ek, J. A., & Trim, J. L. M. (1991b). *Threshold 1990*. Cambridge: Council of Europe/Cambridge University Press.
- Van Ek, J. A., & Trim, J. L. M. (2001). *Vantage*. Cambridge: Council of Europe/Cambridge University Press.
- Wang, X. (2013). *Grammatical Development of Chinese among Non-native Speakers: From a Processability Account*. Cambridge: Cambridge Scholars. Retrieved from <http://www.cambridgescholars.com/download/sample/58986> (last accessed April 2015).
- Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave MacMillan.

## Appendices

**Appendix 1.** Overview of “styles” and “formats” by level, as defined by the CLC customised version of *Sketch Engine*

| Level                                 | A2  | B1                           | B2                           | C1                           | C2  |
|---------------------------------------|-----|------------------------------|------------------------------|------------------------------|-----|
| Exams                                 | KET | PET<br>BECP<br>BEC1<br>CELSP | FCE<br>BECV<br>BEC2<br>CELSV | CAE<br>BECH<br>BEC3<br>CELSH | CPE |
| Style                                 |     |                              |                              |                              |     |
| Informative/news                      | ✓   | ✓                            | ✓                            | ✓                            | ✓   |
| Complaint/apology/response            | ✓   | ✓                            | ✓                            | ✓                            | ✓   |
| Business                              |     | ✓                            | ✓                            | ✓                            |     |
| Descriptive/creative autobiographical |     | ✓                            | ✓                            | ✓                            | ✓   |
| Advice                                |     | ✓                            | ✓                            | ✓                            | ✓   |
| Argumentative/opinion                 |     | ✓                            | ✓                            | ✓                            | ✓   |
| Critical                              |     |                              | ✓                            | ✓                            | ✓   |
| Application/response                  |     |                              | ✓                            | ✓                            |     |

| Level                          | A2  | B1                           | B2                           | C1                           | C2  |
|--------------------------------|-----|------------------------------|------------------------------|------------------------------|-----|
| Exams                          | KET | PET<br>BECP<br>BEC1<br>CELSP | FCE<br>BECV<br>BEC2<br>CELSV | CAE<br>BECH<br>BEC3<br>CELSH | CPE |
| <b>Format</b>                  |     |                              |                              |                              |     |
| Note/email/memo                | ✓   | ✓                            | ✓                            | ✓                            |     |
| Informative/instructional text | ✓   | ✓                            | ✓                            | ✓                            | ✓   |
| Letter/reference               | ✓   | ✓                            | ✓                            | ✓                            | ✓   |
| Story                          |     | ✓                            | ✓                            |                              | ✓   |
| Survey/questionnaire/form      |     | ✓                            | ✓                            | ✓                            |     |
| Composition/essay              |     |                              | ✓                            | ✓                            | ✓   |
| Article                        |     |                              | ✓                            | ✓                            | ✓   |
| Report                         |     |                              | ✓                            | ✓                            | ✓   |
| Proposal                       |     |                              | ✓                            | ✓                            | ✓   |
| Review                         |     |                              | ✓                            | ✓                            | ✓   |
| Speeches/spoken response       |     |                              | ✓                            |                              |     |

## Appendix 2. Grammar items typically taught and tested in ELT syllabi

| Superordinate categories | Subordinate categories  |
|--------------------------|---|
| adjectives               | position<br>modifying<br>combining<br>superlatives<br>comparatives  |
| adverbs                  | adverbs and adverb phrases: types and meaning<br>adverbs as modifiers<br>adverbs: position<br>adverb phrases: form                  |
| clauses                  | declarative<br>interrogatives<br>imperative<br>exclamative<br>coordinated<br>subordinated<br>comparative<br>relative<br>conditional |

| Superordinate categories | Subordinate categories  |
|--------------------------|---|
| conjunctions             | coordinating<br>subordinating   |
| determiners              | demonstratives<br>articles<br>possessive<br>quantity  |
| discourse markers        | written discourse markers: text organisation, word order, chunks ( <i>to begin with, in my opinion, etc.</i> )  |
| future                   | future <i>be going to</i><br>future <i>will</i> and <i>shall</i><br>future continuous<br>future perfect simple<br>future perfect continuous<br>present simple for future<br>present continuous for future<br>future expressions with <i>be</i><br>future in the past  |
| modality                 | modals: <i>can</i><br>modals: <i>could</i><br>modals: <i>have (got) to</i><br>modals: <i>may</i><br>modals: <i>might</i><br>modals: <i>must</i><br>modals: <i>shall</i><br>modals: <i>should</i><br>modals: <i>will</i><br>modals: <i>would</i><br>semi-modal verbs: <i>dare</i><br>semi-modal verbs: <i>need</i><br>semi-modal verbs: <i>ought to</i><br>semi-modal verbs: <i>used to</i><br>modality: expressions with <i>be</i> ( <i>be bound to, be to, be likely to, be supposed to, etc.</i> )<br>modality: adverbs, adjectives and nouns |
| negation                 | negation  |

| Superordinate categories | Subordinate categories  |
|--------------------------|---|
| nouns                    | nouns types<br>noun phrases<br>plural uncountable<br>noun phrases: grammatical functions  |
| passives                 | passives: form<br>passives with <i>get</i> and <i>have</i>  |
| past                     | past simple<br>past continuous ( <i>We were waiting</i> )<br>present perfect simple ( <i>I've finished</i> )<br>present perfect continuous ( <i>We've been waiting</i> )<br>past perfect simple ( <i>We'd waited</i> )<br>past perfect continuous ( <i>She's been talking about him</i> ) |
| prepositions             | prepositions: simple and complex  |
| present                  | present simple<br>present continuous  |
| pronouns                 | subject, object<br>possessive<br>reflexive<br>demonstrative ( <i>this/that/these/those</i> )<br>quantity<br>substitution ( <i>one/ones/none</i> )<br>generic use<br>reciprocal<br>indefinite ( <i>someone, anyone, everyone, -thing, -where, etc.</i> )                                   |
| questions                | questions: form, affirmative questions, negative questions<br><i>wh-</i> questions<br><i>yes-no</i> questions<br>alternative questions ( <i>x or y?</i> )<br>tags   |
| reported speech          | reported speech   |
| verbs                    | types<br>phrasal verbs<br>prepositional verbs<br>phrasal-prepositional verbs<br>linking<br>patterns   |
| focus                    | focus   |

**Appendix 3.** ELT pedagogical grammars and courses surveyed as the basis for the list of “superordinate” categories and “subordinate” categories for investigation

*ELT Pedagogical grammar books*

- Beaumont, D. (1989). *The Heinemann English Grammar*. Oxford: Heinemann Educational.
- Beaumont, D. (1993). *The Heinemann Elementary English Grammar* (2nd ed.). Oxford: Heinemann Educational.
- Biber, D., Conrad, S., & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Carter, R. A., & McCarthy, M. J. (2006). *The Cambridge Grammar of English*. Cambridge: Cambridge University Press.
- Carter, R. A., McCarthy, M. J., Mark, G., & O’Keeffe, A. (2011). *English Grammar Today*. Cambridge: Cambridge University Press.
- Conrad, S., Biber, D., & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English Workbook*. Harlow: Pearson Education.
- Dean, M. (1993). *English Grammar Lessons: Upper-Intermediate*. Oxford: Oxford University Press.
- Eastwood, J. (2005). *Oxford Learner Grammar: Grammar Finder*. Oxford: Oxford University Press.
- Hewings, M. (2013). *Advanced English Grammar in Use* (3rd ed.). Cambridge: Cambridge University Press.
- Murphy, R. (2012). *English Grammar in Use* (4th ed.). Cambridge: Cambridge University Press.
- Murphy, R. (2007). *Essential Grammar in Use* (3rd ed.). Cambridge: Cambridge University Press.
- Parrott, M. (2010). *Grammar for English Language Teachers*. Cambridge: Cambridge University Press.
- Swan, M. (2005). *Practical English Usage* (3rd ed.). Oxford: Oxford University Press.
- Thomson, A. J. (1980). *A Practical English Grammar*. Oxford: Oxford University Press.

*ELT coursebooks*

| Coursebook             | Levels | Publisher                  |
|------------------------|--------|----------------------------|
| <i>Inside Out</i>      | All    | Macmillan                  |
| <i>English in Mind</i> | All    | Cambridge University Press |
| <i>Cutting edge</i>    | All    | Longman (Pearson)          |
| <i>face2face</i>       | All    | Cambridge University Press |
| <i>English file</i>    | All    | Oxford University Press    |
| <i>Matrix</i>          | All    | Oxford University Press    |



*Authors' addresses*

Anne O'Keeffe

Department of English Language & Literature, Mary Immaculate College

University of Limerick

South Circular Road

Limerick V94 VN26

Ireland

anne.okeeffe@mic.ul.ie

Geraldine Mark

Department of English Language & Literature, Mary Immaculate College

University of Limerick

South Circular Road

Limerick V94 VN26

Ireland

gmark@germark.org