CrossMark

Lang. Teach. (2016), **49.1**, 99–115 © Cambridge University Press 2013
doi:10.1017/S0261444813000189 First published online 19 June 2013

Plenary Speech

Putting the CEFR to good use: Designing grammars based on learner-corpus evidence

Michael McCarthy University of Nottingham, UK
mactoft@gmail.com

This lecture considers what reference and pedagogical grammars and grammar teaching materials for L2 learners should ideally include, based on corpus evidence from both native-speaker and learner corpora. I demonstrate how learner corpora can be used to track the emergence of grammatical features, from the elementary level to advanced, how learners get to grips with new grammar and what we can learn from the statistical output of error-coded corpora. Additionally, we look at how the divide between lexis and grammar becomes progressively blurred and how corpus information can best be used to produce useful grammars and teaching materials for students at different levels. The advanced level in particular is focused on. The lecture is presented within the framework of the English Profile Project (EPP), a large, international, inter-disciplinary project which uses corpora to investigate learner competence at different levels of the Common European Framework of Reference (CEFR).

1. Introduction

Let me begin this lecture by looking at some of the main factors that underlie its theme and contributed to its present form. We are concerned here with the creation of good, useful reference and pedagogical grammars for learners of English, though the principles and practices that I shall refer to could be put to use for any language for which learner corpora are available, and for any context in which the learners have been pre-assessed with reference to accepted systems and standards. There is undoubtedly growing pressure around the world for standardisation and common understandings with regard to what levels of language proficiency actually entail. No country seems to be immune from these pressures. The clamour for standardisation manifests itself at all levels, from primary to tertiary education and in professional and vocational contexts. In the present case, of particular relevance is the growth of the Common European Framework of Reference (CEFR) (Council of Europe 2001), which in recent years has influenced assessment systems way beyond the borders of the

Revised version of a plenary address given at the Languages in the Globalised World conference, Leeds Metropolitan University, 6 September 2012.

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

European nations and is now widely referred to in countries around the world. The refinement and updating of the descriptors of the CEFR for English is the principal motivation behind the English Profile Project (EPP), a massive research undertaking which I shall refer to in greater detail later. A further motivation for this lecture is the growth in the availability of spoken corpora, and sophisticated software that enables complex searches for grammatical and discoursal patterns to be carried out with relative ease. Allied to the development of corpora in general and the growing sophistication of software is the steady expansion of learner corpora and in particular, in recent years, the compilation of spoken learner corpora.

2. The CEFR

The CEFR consists of three primary levels: A, the basic user level, B, the intermediate level and C, the advanced level. Each of these primary levels is subdivided into two parts. Each sub-part has a metaphorical name that attempts to capture the learner's experience at that particular level; for example, A1 is *Breakthrough*, A2, *Waystage*: metaphors that are, if anything, military, but that aptly and colourfully describe the initial struggle with a new language which, hopefully, brings the learner to a satisfactory staging post on the long trek to advanced proficiency. The B level – *Independent User* – continues the metaphorical portrayals of the learner's experience; B1 is referred to as the *Threshold*, with the B2 level returning to the military metaphor of *Vantage*, both of which suggest a transition into new and rewarding zones of development. What I would like to note in particular here is the term *Independent User*, for it implies a fundamental change in the learner's experience with the language, suggesting that he or she is entering a new territory with less dependence on the teacher, a territory in which, ideally, the learner is set to become an autonomous language animal, a point we shall return to later. Finally, the C level is that of the proficient learner who ultimately arrives at a high level of control of the target language. This system has served us and continues to serve us well, providing both a common standard for assessment across different countries, languages and contexts and a framework for the creation of materials and resources to support language learning. The major ELT publishers now routinely give information on their language teaching products in terms of the CEFR level at which the materials and resources are aimed. The framework has, in its turn, been cross-referenced to other systems of assessment and descriptions of proficiency such as Cambridge English Language Assessment and other systems around the world (Milanovic 2009; Khalifa & French 2009).

However, it should be recalled that the CEFR was devised before the widespread use of and easy access to language corpora. This is not to suggest that the framework itself is deficient, but that it is a manifestation of the vast pool of expertise representing the experience of decades of examinations, curriculum design and the almost limitless input of generations of teachers. As such, it is a powerful instrument resting on a solid base. What we have in the framework is a robust, tried and tested set of labels with which we assign proficiency levels to our learners. This pre-existing population has been an invaluable resource for the research I shall talk about in the course of this lecture.

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

3. The English Profile Project

Recent years have allowed us to take a new and fresh look at the CEFR and work towards an empirical underpinning for any language of the detailed descriptions that accompany each level and sub-level in the framework's documentation, outlined above. Empirical evidence of learner performance at the different levels is much easier to obtain because of developments in corpus linguistics, and this is what the EPP has set out to provide with respect to proficiency in English. The project is an inter-disciplinary research project which brings together several major bodies concerned with language education, including Cambridge English Language Assessment, the British Council, several universities and language institutions in the UK and abroad and major ELT publishers. The project's website, www.englishprofile.org, gives full details of its scope and aims but suffice it to say that the goals it has set itself are ambitious but essential if we are to provide the objective evidence that will take the CEFR to a new level of rigour and reliability.

The two main goals of the EPP are (1) to find out what learner English is really like at the different levels of the framework and (2) to provide empirical evidence for the detailed descriptors for each level based on the investigation of corpora. Sub-projects within the project include the English Vocabulary Profile, already available as a superb online resource providing a profile of typical lexical proficiencies at each level of the CEFR (see www.englishprofile.org/index.php?option=com_content&view=article&id=4&Itemid=5). Other current projects include the English Grammar Profile, as well as projects investigating communicative-functional aspects of language proficiency (Green 2008, 2012), writing, phonology and spoken fluency (McCarthy 2010). In this lecture, I will focus on the challenges presented by creating grammar profiles, with particular reference to the work of the two principal researchers, Dr Anne O'Keeffe of Mary Immaculate College, University of Limerick, Ireland, and Geraldine Mark, an ELT author and freelance researcher (Mark & O'Keeffe 2012), along with that of Professor John Hawkins of Cambridge University, who contributed greatly to the original theoretical grounding of the grammar project. In addition, I shall report some of my own research in connection with the preparation of advanced level grammatical materials (Carter et al. 2011; McCarthy, McCarten & Sandiford 2012). The EPP grammatical research is work in progress and will be disseminated over the next couple of years. Here I want to lay out some of the principles involved in such research and some of the interesting insights which learner corpora can provide in connexion with grammatical proficiency.

4. Questions for the design of learner grammars

Let us now consider some of the issues that need to be addressed in the design of grammars for learners at different levels, not least the basic question of how well the items and structures chosen for inclusion in a learner grammar align with and match what we know about learners at each level of the CEFR – or any other similar framework, for that matter. Secondly, it is important to recognise that the three primary levels of the CEFR are not simply

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

lock-step levels. Transition from one level to another does not happen overnight, while, as we have already noted, the transition from the A level to the B level involves not just learning more language but actually becoming a different kind of learner, an Independent User. This means that a grammar for the B-level learner will ideally be not just quantitatively but also qualitatively different from that for a beginner. As we shall see later, the B-level learner has to cope not only with an exponential increase in language input, but the allied challenge of matching new input to the ever-increasing stock of already processed input. Equally, it is, I hope, uncontroversial to say that the advanced level is largely unexplored territory into which language teaching materials characteristically inject the leftovers of grammatical description, often rather unsystematically, relying on notions of difficulty and the feeling that particular items are complex and sophisticated (an exception is Hewings 1999/2005).

Thus the challenge of designing a learner grammar has, at the very least, two major sticking points: the explosion of input and the qualitative leap to independence that takes place at the B level, and the boundless and unexplored territory of the advanced level (not least the intriguing question of where the C level ends and whether there are D or even E levels!). A final, major preoccupation that will affect the B and C levels is the need for recycling, given what we know from our common experience about the fossilisation of learner errors and other problems and, as I shall attempt to demonstrate, the evidence of such fossilisation as presented in learner corpora. And error is not the only motivation for recycling: we cannot be certain that anything will be learned and retained after only one or a couple of exposures. For some items and patterns, multiple exposures in different contexts may be necessary (Nunan 1988: 34), alongside the careful pairing and combination of items to enable comparative and contrastive descriptions and pedagogy which may assist the learning process. Additionally, there has been debate over the role of repeated exposure for the purposes of consciousness-raising and/or incidental learning of grammatical items (Ellis 2002).

5. Learner corpora

Researching learner corpora enables us to find out what learners who have been pre-assigned to CEFR levels can do with the language, having achieved certain levels in existing examination systems. Among the questions of interest are vocabulary size, or breadth, and the depth of the learner lexicon at any given level, functional and strategic expertise, level of fluency and, in the present case, ability to control and use appropriately the resources of the grammar. Conversely, corpus evidence enables us to see what learners are NOT doing at a given level, either because of lack of knowledge of the repertoire required to operate at that level, or because of lack of opportunity of use to display competences which they might possess but which have not come to light in the available data (Buttery & Caines 2012).

Learner corpora linked to database information in which learners are tagged according to such factors as first language background, age, gender, proficiency level and length of time learning the target language also enable us to make statements about developmental patterns. Truly longitudinal studies are not easy to carry out, but quasi-longitudinal studies, looking at how learners typically expand their knowledge as they move from one level of the CEFR to another, offer a great deal of insight into how grammatical knowledge grows

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

and develops and also how it relates to and intertwines with other types of knowledge such as lexical or communicative-functional knowledge. Learner corpus evidence is clearly not enough to enable us fully to understand the process of grammatical development; the EPP also addresses itself to syllabuses and teaching materials on the clear understanding that development is influenced by numerous different factors, including not least the input that learners receive. Overarching these preoccupations are more general questions connected with second language acquisition (SLA) and whether the posited orders of acquisition are borne out in the corpus evidence or whether other factors come into play (e.g. L1 influence): factors that may affect the order in which grammatical development takes place (Murakami 2013).

6. Data for the present study

Research within the EPP is based both upon native-speaker corpora and the multi-million word Cambridge Learner Corpus. This corpus consists primarily of Cambridge examination scripts representing 150 different first language backgrounds, with 200,000 examination scripts providing more than 55 million words of data for corpus research. More than half of this total has been error-coded, a painstaking process in which raters assign codes to perceived errors, which can then be searched for and the relevant error information retrieved automatically. This is an invaluable resource for the investigation of learner problems at any level, as much to see where errors are NOT occurring or where they are declining as where they are present and persistent. As well as the examination scripts, an increasing amount of spoken corpus data is progressively becoming available, including both oral examination data from the Cambridge examinations suite and non-examination data collected in learning settings, for example the one-million word CLAS spoken corpus, collected in a vocational training context (Healy & Onderdonk Horan 2012). The data collection process is ongoing and includes plans to collect further spoken data in both classroom and non-classroom settings.

7. Examples of grammatical analysis

Equipped with a large corpus, the researcher can either take a broad, bird's-eye view of the data in the hope of observing general developmental patterns, or home in on specific items and trace their trajectory across the different levels of proficiency. We will look at examples of both in this lecture. Figure 1, based on the EPP research of Mark & O'Keeffe (2012), gives us a very broad picture for two kinds of error as they occur across the levels of the CEFR. Both are concerned with errors within the verb phrase and the use of past simple tense form. The darker line shows errors in tense choice, while the lighter line traces the occurrence of errors in verb choice. Immediately noticeable is the similarity in the shape of the trajectory. A kind of witch's hat indicates a peak of error in both categories around the B1 level.

What does this mean? Clearly, we can relate this to our earlier remarks about the transition from the A level to the B level. After the A2 level, the learner is exposed to massive amounts of new input, particularly in relation to categories such as tense and aspect within the verb

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

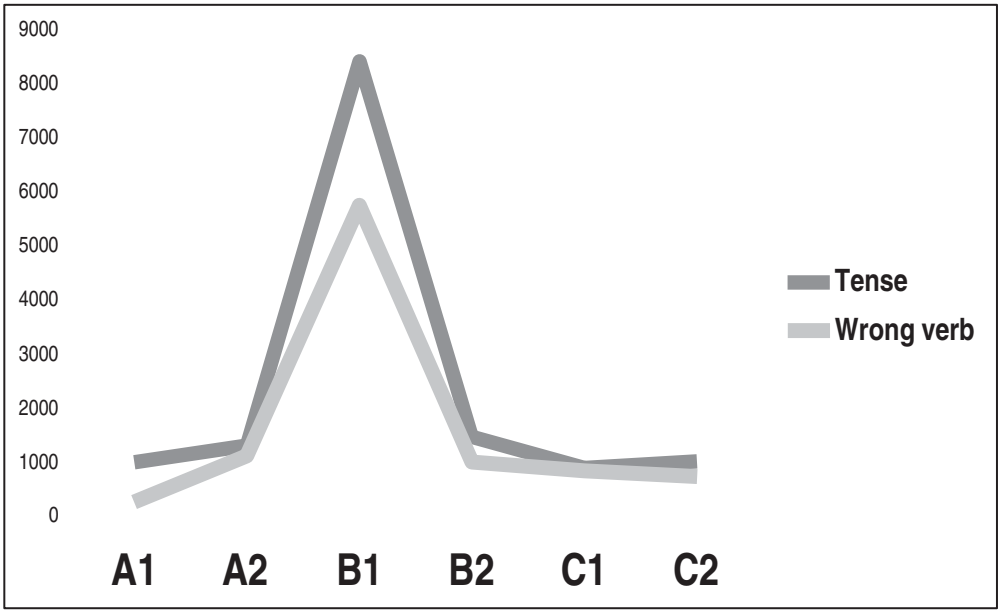


Figure 1 Tense errors and wrong verb choice: Cambridge Learner Corpus (Mark & O'Keeffe 2012)

phrase, in tandem with a massive expansion of the vocabulary base of lexical verbs. So, on the one hand, learners have to cope with a greatly increased amount of new knowledge that needs to be matched with and assimilated into the existing knowledge base, which at this level was rather low at the outset. On the other hand, as we have also noted, the learner is entering a new territory, qualitatively different from that of the A level, where he or she is struggling to achieve some kind of independence or autonomy in their learning. Achieving these comes not least from experimentation, trial and error and the inevitable pitfalls presented by the expanded opportunities of use offered by the materials and activities in the classroom. Typically, at this level, the length of writing tasks increases and greater opportunities are provided for oral production. It is not surprising, therefore, that in this context of experimentation and expanded use, error should greatly increase. We need not see this as a problem or as a failure in the curriculum, the materials, or the teaching. Just as in L1 acquisition, where over-extension and greatly increased use result in infelicitous performances (Clark 1987), so too may we expect the same in SLA (e.g. Shirai & Kurono 2002).

The pattern of the witch's hat repeats itself, not always in the same place, depending on the point at which input occurs, the type and extent of input and the difficulty for the learner in the new material encountered. Error levels for the present perfect continuous, for example, drop after the B1 level, indicative of the early stage of the syllabus at which present perfect continuous forms are typically privileged in teaching, regardless of frequency of occurrence or difficulty in use. Notwithstanding such environmental circumstances, searches of this type in the learner corpus regularly yield some high point of error-occurrence which then tails

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

off as target items are consolidated and controlled more effectively. Carrying out this kind of investigation for particular grammatical items or patterns provides us with invaluable information when designing learner grammars, information which may influence our choice of ideal points at which to introduce particular items, the extent to which particular items may need repetition, recycling and reinforcement, the typical level of difficulty presented by items or patterns, and so on. In the case of tense choice, Mark & O'Keeffe (2012) report the widespread use of past simple forms instead of present simple forms as one of the most common errors persisting up to the B2 level. They also present evidence which shows that the addition of new tense and aspect forms (e.g. the teaching of new forms such as perfect and continuous aspects) predictably introduces new types of confusion between forms, with some evidence of fossilisation – probably caused by the increasing complexity that students have to mentally disentangle when confronted with forms such as past continuous and past perfect.

The broad, bird's-eye view of grammatical development enables us to assign to any particular item or pattern a typical stage when it emerges, when it is marked by extensive experimental use and recurrent error, when it becomes more consolidated and finally when it is adequately controlled. This broad picture is based on the evidence of thousands of examples and the performance of many, many individuals. In this sense, the evidence gives us a profile of a typical learning trajectory. Since we could carry out this operation for any of the hundreds of items which make up the grammar of English, we seem to be landing ourselves with a task of gargantuan proportions and more than a lifetime's work! What is more, the whole picture is complicated by the fact that some learners seem to know particular items or patterns and are able to use them way in advance of the general mass of the student population. This led to the need for a conceptual framework that adequately grasps what the generality of learners are able to do at any given level, which was elaborated by Hawkins & Buttery (2009) and Hawkins & Filipović (2012) in the notion of CRITERIAL FEATURES, that is to say, major features of the grammar which cluster in terms of emergence, use and consolidation at particular points along the levels of the CEFR. An analogy here would be the police identikit picture, where the facial profile of a suspect can be accurately portrayed based on a relatively small number of key features which identify that individual, excluding myriad other features which may or may not be the unique possession of that individual (the odd facial freckle here and there, for instance). In this sense, the EPP is indeed concerned with creating profiles for each level, based upon the points at which grammatical categories are criterial, in that they are possessed and used successfully by the large majority of learners who have been pre-assigned to that level. Put simply, if you do not possess the criterial features for any given level, you may not label yourself as a member of that level. The added benefit of criterial features is that, because they accumulate individually, we are able to observe relatively seamless progression from one level towards, and ultimately into, the next level. Mark & O'Keeffe (2012) give a very clear example of criteriality in the case of simple and correlative conjunctions. Table 1 shows the levels at which each of the listed conjunctions becomes criterial, that is to say, the levels at which the majority of learners characteristically use the items appropriately in the majority of cases. It can be seen for example that *and*, *but* and *or* are criterial at all levels, while general control of the pattern *either. . .or* does not emerge till B2 level and the *neither* + inversion construction does not emerge as criterial until C2 level.

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

Table 1 Criteriality of simple and correlative conjunctions (Mark & O'Keeffe 2012)

	and, but, or	*both . . . and*	*either . . . or*	*neither . . . nor*	*not only . . . but also*	*neither* + inversion
A1	√					
A2	√	√				
B1	√	√				
B2	√	√	√	√		
C1	√	√	√	√	√	√
C2	√	√	√	√	√	√

8. Grammar and lexis: countability

An example of a category where fossilisation is clearly displayed is that of noun countability. The notion of countability in itself is not an easy one and mention of it will chime with many teachers' experiences of trying to put the concept over in the classroom, whether successfully or unsuccessfully. The corpus enables us not only to see where errors and problems with countability occur but also which particular nouns are prone to difficulty. In this respect, countability provides us with a neat segue into the overlap between grammar and lexis, which increases and becomes more blurred as learning progresses towards the advanced level. Grammatically, all uncountable nouns are equal; lexically, to coin a phrase, some uncountable nouns are more equal than others. Mark & O'Keeffe (2012) researched the occurrence of uncountable nouns as teaching items in a range of English language course books and other teaching materials and found that countability was most commonly taught in conjunction with expressions such as *there is/are*, *have got* + noun, following determiners such as *some*, *any*, *much*, *many* and, for uncountable nouns, within the topic domain most commonly of 'food'. The level at which uncountable nouns were typically introduced was either A1 or A2. The most common errors with the use of countable nouns were forms such as *informations*, *advices*, *equipments* and *transports*.

The distribution of errors with uncountable nouns peaked at B2 level, declined slightly at C1 level and more sharply at C2 level. So although the developmental curve resembled a slightly lopsided witch's hat, it is a witch's hat nonetheless; the persistence of error is apparent even at B2 and C1 levels, despite the fact that much of the input on countable nouns takes place at the A level and is thereafter assumed to be 'done'. *Information* and *advice*, the front runners in error-prone uncountable nouns, are both given as A2 level words in the English Vocabulary Profile; this is understandable owing to their widespread occurrence in the kinds of topics students typically write about at that level. Despite this, problems with their grammatical form persist, with *information* remaining the number one error-prone noun right up to level C1. *Advices* remains fossilised right up to C2 level. A similar picture of fossilisation manifests itself with *equipments* and *transports*, with both remaining problematic up to C1 level. This kind of distribution presents researchers and language educators alike with a problem.

While it may be possible to make a rather general statement that countability, on the whole, seems to be effectively controlled by the majority of students after the error-peak at B2 level,

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

the description in a reference grammar or pedagogical grammar, to be of any practical use, needs to specify those lexical items which are most error-prone, and teaching materials based on the evidence of learner data clearly need to recycle particular nouns even if the general category of countability is deemed to have been satisfactorily dealt with in the materials and in the classroom at a much earlier stage.

9. Extended lexico-grammatical patterns

I mentioned the way in which, as learning progresses, lexis and grammar seem to become increasingly interrelated and I wish to look closer at that problem now. The example I take here comes from research in connection with materials I have recently co-authored (McCarthy, McCarten & Sandiford 2012). One of the features of English grammar is that di-transitive verbs may express their complements in two different ways, as exemplified here:

indirect object + direct object

She always gave ***the children presents***.

direct object + prepositional phrase

She always gave ***presents to the children***.

Both sentences are perfectly well formed even if their distribution might be uneven in particular contexts of use because of the different focus on participants which each version represents. Both patterns occur in the learner data from the CEFR A level onwards, though characteristically with a much lower overall occurrence at the A level. Errors with the indirect object followed by direct object form decline sharply at B2 level and continue to decline up to C2 level, suggesting at the very least that we know how to teach this form and after a period of initial experimentation and error, the problems are minimal.

The same cannot be said of the second pattern, involving the direct object followed by a prepositional phrase. Here, in a search involving three high-frequency di-transitive verbs (*give*, *send* and *buy*), the overall total of errors is greater than that of the indirect object + direct object pattern and, after an encouraging sharp drop in errors from A2 to B1, the decline in the error total slows down and the error remains persistent, falling only to a satisfactory level at C2. Based on the three chosen verbs, the evidence seems to suggest that the direct object + prepositional phrase construction may need more and continued attention in pedagogy than the indirect object construction. However, the picture is not that simple and we need to look at the distribution of errors for each verb individually. For example, when we do this with the verb *give*, we see a similar pattern in the steady decline of errors with the indirect object construction, such that by the C level, problems with this form are minimal. Likewise, a decline in the occurrence of error with the direct object and prepositional phrase construction is observed after A2, but the error totals flat-line across B1, B2 and C1, only falling at C2 (Figure 2).

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

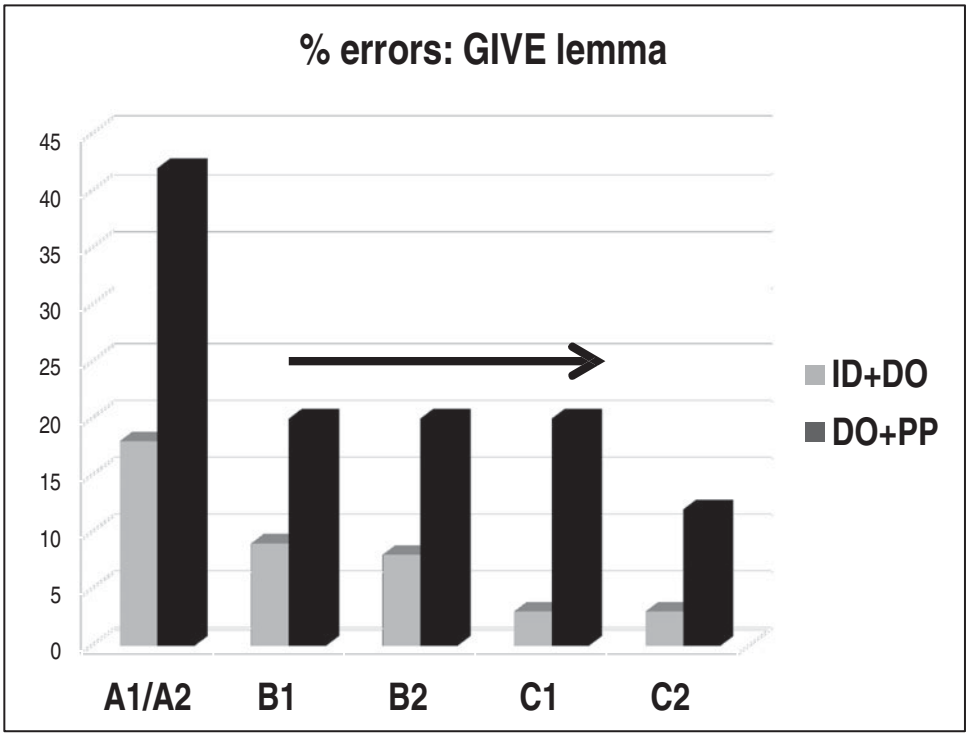


Figure 2 Errors with complements of verb *give* [ID+DO = indirect object + direct object; DO+PP = direct object + prepositional phrase], Cambridge Learner Corpus.

The verb *give*, therefore, seems to be a persistent trouble-maker. Cambridge Learner Corpus examples include the following:

*I will **give** a phone call **for you**.* (A2)
*It's important for me because my mother **gave to me**.* (A2)
*Maybe you can keep my work and you can **give to me** the next day.* (B1)
*I own a lovely old table which I've loved since I saw it three years ago, in my uncle's house. He **gave me** because he moved to Londrina.* (B1)
*People don't **give** too much importance **for** their health nowadays.* (B2)
*A friend of mine gave **to** me your advertisement about two nights for the price of one.*

Two patterns of error dominate: the wrong preposition (with *for* often being used instead of *to*) or the correct prepositional construction without the necessary direct object. Less seriously, errors also occur with inappropriate use of the preposition *to* before the indirect object (a feature which might possibly pass without notice in native speaker conversation). Errors with other di-transitive verbs often overlap with those found with the verb *give*, but also differ in individual verbs; for example, using the preposition *to* instead of *for* with the verb *make* is a common error, persisting up to C2 level. The verb *buy* displays similar difficulties with the

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press



Figure 3 Presentation of direct object + preposition patterns from McCarthy, McCarten & Sandiford (2012: 165)

use of the preposition *to* instead of *for* frequently presenting a problem, albeit in this case the error has virtually died out by C1, with only residual examples cropping up at C2. Such information should not only be of value to the researcher, not to mention the teacher, but can be presented directly in materials for students, to encourage the growth of language awareness and observation which, as we have noted, is important in the transition from dependent learning to more autonomous and independent learning when the student moves into the B level. Figure 3, an extract from recently published material (McCarthy, McCarten & Sandiford 2012), shows the presentation of the direct object + preposition pattern in relation to some frequent di-transitive verbs as attested in native speaker corpora. The focus here is not merely on the general pattern but on the specific sub-patterns relating to individual verbs.

10. Issues at the advanced level

I mentioned earlier that the advanced level has long been one at which we are never quite sure what we should be doing in terms of grammar, often settling for structures and patterns which are perceived to be complex, difficult, sophisticated or of low frequency. An example of this would be the English negative past subjunctive, in sentences such as *He insisted that they not be allowed to enter*. This structure is of very low frequency, is only found in formal writing and, anyway, can always be substituted by a more familiar construction with the modal verb *should* (*He insisted they should not be allowed to enter*). Its use, although frequently demanded by publishers and teachers alike, judging by descriptions in textbooks and accompanying practice, is far from essential for most students. On the other hand, the corpus reveals a number of features that textbooks may have overlooked for one reason or another but that are both relatively frequent and useful for learners either in their spoken or written production.

One problem is that, by the advanced level, most of the finite number of items and structures that constitute English grammar may well have already been covered to a greater or lesser extent at lower levels, and teachers and learners may be loathe to return to what they think is familiar territory at the expense of breaking new ground. However, there are very good reasons for looking again at some of the structures which may have been taught

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

and practised at lower levels, for what the corpus reveals is that structures may have meanings and uses which had not been covered in earlier materials. Some of these meanings, when exposed, feel quite familiar and one wonders why nobody bothered with them before. I have in previous works, such as McCarthy (1998: Chapter 8), pointed out that even major ELT grammatical resources have ignored the extremely common everyday pattern of introducing a newsworthy topic with the past continuous forms of reporting verbs such as *say* or *tell* in banal and familiar utterances such as *Mary was telling me you've changed your car*. The reason for the neglect of such a simple structure with a simple meaning and a simple function is probably, in this case, the preoccupation with the sequence of tenses in indirect reports, a theme much beloved of textbooks.

Equally, at the advanced level, we can take matters further, teaching new meanings and uses of structures that may in themselves be perceived as advanced and commonly taught at that level, but which have meanings and uses that simply get forgotten. An example here would be the use of the future perfect form to state assumptions about the present rather than predictions about the future. Participants in workshops would hardly be surprised to hear the workshop leader say *You will have been given a handout as you entered* and would not think of it as an unusual or particularly obscure use of the future perfect form. The problem is, as ever, the difficulty of being objective about one's own use of language. This is where the corpus comes in and I have long argued that the evidence from corpora levels the playing field between native users and expert non-native users of the language, with native users having no greater claim to objectivity in observation of how they use their language than the keen non-native observer. It requires long and careful study of the corpus to observe the various functions and uses of grammatical items but if we do have space at the advanced level to introduce new grammar, it is in my view a far better and far more efficient use of time and resources to present meanings and functions of grammatical elements which are attested in the corpus but which may simply have been neglected, rather than only seeking out the arcane.

What we learn from these statistics and examples is that although grammatical patterns may be abstracted for a particular set of lexical items, it may often be necessary to focus on the individual items within a given set that falls into the general category of that pattern. This blurring of the distinction between lexis and grammar increases as the learner's lexicon grows and as the focus moves from the grammatical properties of individual words towards more extended patterns, especially those of verb complementation, though noun complementation also produces similar problems in relation to prepositions (e.g. *the reason of* instead of *the reason for* is a common error).

11. Getting the most out of the grammar

A final dimension of the design of learner grammars, hinted at throughout this lecture, is the question of what we can do to best enhance learners' grammar so that what they learn is not only common and useful, as opposed to rare and of low value, but is also the kind of grammar that will lead to success in examinations or academic, professional and vocational contexts. I make this remark because research is steadily emerging which shows that some uses of grammar considerably impress readers and listeners, especially examiners

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

and teachers rating student work. What I am proposing here is that over and above the desire for correctness and accuracy in grammatical use, there are ways of using grammar which improve style, which show control and mastery of academic and professional conventions and which fit well with certain types of text genre. Research by Liu (2011) and Biber, Gray & Poonpon (2011), for example, suggests that the noun phrase may be a grammatical category that is often taken for granted or neglected in teaching, but which is a key component of success in academic writing.

The English noun phrase can be grammatically very complex, but the focus in teaching often goes rarely beyond the presentation and practice of the finite set of determiners, the so-called canonical order of adjectives, and post-modifying relative clauses. However, the noun phrase offers a number of other possibilities, including nouns pre-modifying other nouns, coordination of adjectives and a wide variety of types of post-modification of the head word. Over and above the structural configurations, there is the process of nominalisation, whereby meanings often encoded in other types of phrase, such as verb or adjective phrases, are instead packed into the noun phrase. Compare these pairs of sentences:

(a) *The president arrived but he was delayed by an hour.*
(a1) *The president's arrival was delayed by an hour.*
(b) *She will retire as Chief Executive in June.*
(b1) *Her retirement as Chief Executive will take place in June.*

Each is grammatically correct, but the second versions, in which the verbs have been transmuted into nouns, are more typical of academic and professional writing, so their use offers the learner opportunities for success in writing and improvement in their style. One of the reasons why nominalised structures of this type are so useful in academic and professional writing is that they can pack in a lot of information into a few words, taking advantage of the available slots both before and after the noun head. The relationship between grammar and information packaging is not often explored outside of academic contexts, but it is a good candidate for the grammatical syllabus at the advanced level for any learners concerned with writing or speaking in relatively formal contexts. Nominalisation is an example of what Halliday has referred to as grammatical metaphor (1985/1994), where elements normally associated with one of the principal grammatical roles (participants, processes and circumstances) are transferred to another (see also Halliday & Martin 1993). Thus processes such as *arriving* and *retiring* are normally associated with verbs; in the examples above they have become associated with nouns, which in their turn are normally associated with participants in the clause such as subjects and objects.

12. Conclusion

In this lecture I have sketched out some of the considerations that we need to take into account when designing learner grammars. There are undoubtedly many more which are beyond the scope of this talk, but I have focussed on what I regard as some of the most important. Most important, in my view, is that we use the powerful evidence of corpora, both native-speaker

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

and learner corpora, to achieve greater objectivity in our description of the grammar, both in terms of its forms and functions. Native-speaker corpora are valuable in that they can reveal to us how grammatical items or patterns are characteristically used by large numbers of users. Equally useful in this regard are corpora of expert non-native users. And in this talk I have attempted to show that learner corpora, especially error-coded learner corpora, are of immense value, in that, with enough data and in the present case, with learners pre-assigned to CEFR levels (or indeed any system describing levels of proficiency) we can build profiles for each level. In doing this, we can also observe typical trajectories from one level to another, the difficulties that may occur at one level more than another, the learning curves (which I have lightheartedly referred to as the witch's hat), the persistence and fossilisation of errors, all representing information of immense value to the materials writer and to the teacher. Additionally, a corpus offers us evidence of the ever-increasing inter-meshing of grammar and lexis that takes place as the learner's lexicon and grammar expand in tandem. Finally, I have attempted to show how we may solve some of the problems associated with selecting grammar for the advanced level, a territory which, until now, has remained confused, vague and unsystematic. By looking not only at the less frequent structures but also at new or previously neglected meanings and functions of structures already known, we can deliver genuine increments to the grammatical repertoire of learners which they will encounter in listening and reading and which may prove more useful to them than the more obscure and arcane structures. Alongside this, I mentioned an example of how the teaching of grammar at the advanced level can enhance performance in academic, vocational and professional contexts, with the example of nominalisation.

Much of the research I have reported here has taken place within the context of the EPP, whose website I recommend you to visit. Dissemination of the EPP's research, as it unfolds, will not only underpin the CEFR (as regards English) in terms of more objective, detailed and robust descriptions of each level, but will also provide an invaluable resource for the creation of materials tailored to different learning levels, and a bank of evidence of genuine and attested use. As a materials writer myself, I have found the resources of native-speaker and learner corpora to be irreplaceable. I do not wish to denigrate the past; we all, as teachers, have benefited from the outstanding reference grammars, pedagogical grammars and grammar practice materials that have been authored by people who are or were superb observers of the language as well as talented presenters of clear, practical and useful teaching resources, written in the era before corpora were available. Having said that, now that corpora are becoming increasingly available, it is my hope that they will become a tool routinely used in the creation of learner grammars and that some of the problems I have outlined in this talk will be at least reduced and made easier to handle, thanks to a better knowledge of how English grammar is used, how learners struggle to control it, and what problems they encounter along the way.

13. Postscript

The people who kindly gave up their time to review this lecture for inclusion in this journal quite properly made observations on its weaknesses and shortcomings. However, it is not an

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

easy task to rewrite a live performance as one would write a conventional article, nor would it be entirely honest; my lecture must stand as it was given, and its original audience will have made its own judgment on it. However, this printed version does give me the opportunity to clarify a few key points in the form of this epilogue.

Firstly, it does of course need to be pointed out that the CEFR is language-independent: it attempts to set standards for all and any of the Council of Europe's member-languages. That is not what this paper is about. As a teacher of English and materials writer for English-language learners, what I need to know is not just, for example, that a Vantage level (B2) learner can use a prepositional phrase post-modifier of a noun (van Ek & Trim 2001: 161). Ideally, I would like to know how rapidly or slowly, if introduced, say, at the B1 level, such a form is typically mastered by the majority of learners, whether there are particular nouns that habitually pose a problem for learners (I have already mentioned the stubborn infelicity of *reason of* instead of *reason for*), and whether I am in fact introducing the most useful and common nouns that typically take post-modifiers or complements, and so on. The EPP builds on the impressive groundwork of scholars such as Jan van Ek and the late John Trim (himself a steadfast and unwavering supporter of the project), adding, based on empirical evidence of native-, expert- and learner-user language, flesh to the bones of the original description. By definition, the more we try to flesh out the specifications for individual languages, the less we focus on the CEFR as language-independent.

One of the key aims of the EPP, as I have indicated, is a better understanding of the C level of the CEFR. The fact that the C level makes increasingly complex demands on the student in terms of language activity at all levels (grammatical, lexical, discoursal, sociolinguistic, etc.) is one very good reason for having the complex evidence of a corpus to bring to bear. Here I am only concerned with grammar, and the scope of my 45-minute lecture prevented the presentation of too many examples. I mentioned the future perfect aspect in English, where the functional focus in coursebooks and syllabuses is typically narrow, homing in on predictions about the future and ignoring its uncontroversial use to make assumptions about the present. One could add many other areas where the C level grammar could be improved: if one really wants to present the English subjunctive, corpus evidence shows a wide range of adjectives and nouns which are often followed by it, rather than the favourite pattern of verb + subjunctive *that*-clause. Beyond sentence-grammar, there are key aspects of the use of adjectives and adverbs as response items in conversation that contribute to what spoken language research tells us about the nature of fluency (McCarthy 2010). Examples where the C level can be vastly improved are myriad; the EPP, when disseminated, will offer a wealth of practical and usable grammatical information. Right now, the project is in its early stages as regards grammar specifications. The broad-brush view of learning curves may well offer us few surprises; that is precisely why the thrust of learner-grammar research needs to focus on lexico-grammatical issues, as I have attempted to argue. My position remains that learning trajectories are highly sensitive to the delicate relationship between the unfolding lexicon and the developing grammar. To this end, the project is not only exploring aspects of the grammar previously neglected or only partly understood, it is also constantly improving its computational tools for natural language analysis, especially in terms of the spoken grammar, moving more towards complex searches for extended lexico-grammatical patterns.

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

Another weakness of any individual project is that its data can hardly be expected to cover every eventuality and, as I have acknowledged, the pilot studies of the EPP have been conducted mainly on examination data. But this is rapidly changing and will have already changed greatly when this paper sees the light of day. One of the main thrusts of the project is to admit new, especially spoken, data, and already more than three million words of spoken data involving learners in both exam- and non-exam contexts has been collected, with more non-exam data in the pipeline. Few projects can claim to be as far-reaching in data collection. Equally, because the project surveys syllabuses and course books, it takes into account teaching contexts and teaching sequences that are, to a large extent, based on the intuition and experience of many practitioners over many years, what I referred to earlier as pooled expertise, a massive body of inter-subjective insight. It is worth recalling also that the Cambridge Assessment Examinations Suite celebrates its centenary in this year of 2013: a highly-respected institution with a century of reliability and excellence whose vast body of accumulated and validated learner data has only relatively recently been able to benefit from corpus research. It would be foolish to build the entire enterprise of grammar teaching on corpus evidence alone. Equally foolish would be to dismiss the powerful additional evidence that corpus analysis can provide where intuition fails. The EPP does neither.

Throughout this paper I have talked about 'teaching grammar'. One reviewer raised the issue of whether grammar can be 'taught'. There he/she and I must part company. Of course we can TEACH grammar: hundreds of thousands of teachers do it every day and any discussion that threatens to denigrate their efforts is impertinent. Learning and uptake are a different matter, and there most of us can only ever live in hope, and experience satisfaction if and when our learners show control and confident use of the grammar. I know of no research that convincingly shows me what, in the grammar of English, can and cannot be 'taught' to the generality of human beings. My own 47 years as a grammar teacher and materials writer does not chime with my reviewer's misgivings. Teaching is a process of intervention founded on what we deem to be best practice and develops with experience; learning is highly individual and extremely complex, certainly not reducible to observations of learners of the kind made on rats in the laboratory. The present lecture was about grammar TEACHING.

References

Biber, D., B. Gray & K. Poonpon (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45.1, 5–35.

Buttery, P. & A. Caines (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*. Amsterdam: John Benjamins, 187–204.

Carter, R. A., M. J. McCarthy, G. Mark & A. O'Keeffe (2011). *English grammar today*. Cambridge: Cambridge University Press.

Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum, 1–33.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Ellis, R. (2002). Grammar teaching: practice or consciousness-raising? In J. C. Richards & W. A. Renandya (eds.), *Methodology in language teaching: An anthology of current practice*. Cambridge: Cambridge University Press, 167–174.

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press

Green, A. (2008). English profile: Functional progression in materials for ELT. *Cambridge ESOL Research Notes* 33, 19–25.
Green, A. (2012). *Language functions revisited: Theoretical and empirical bases for language construct definition across the ability range*. English Profile Studies, vol. 2. Cambridge: UCLES/Cambridge University Press.
Halliday, M. A. K. (1985/2004). *An introduction to functional grammar*. London: Edward Arnold.
Halliday, M. A. K. & J. R. Martin (1993). *Writing science: Literacy and discursive power*. London: Falmer Press.
Hawkins, J. & P. Buttery (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal* 1.1, e5.
Hawkins, J. A. & L. Filipović (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. English Profile Studies, vol. 1. Cambridge: UCLES/Cambridge University Press.
Healy, M. & K. Onderdonk Horan (2012). Looking at language in hotel management education. In F. Farr & M. Moriarty (eds.), *Language, learning and teaching: Irish research perspective*. Berlin: Peter Lang, 141–165.
Hewings, M. (1999/2005). *Advanced grammar in use*. Cambridge: Cambridge University Press.
Khalifa, H. & A. French (2009). Aligning Cambridge ESOL examinations to the CEFR: Issues and practice. *Cambridge ESOL Research Notes* 37, 10–14.
Liu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly* 45.1, 36–62.
Mark, G. & A. O'Keeffe (2012). The English Grammar Profile. Paper presented at the English Profile Network Seminar. Hughes Hall, Cambridge, UK, 2–3 February 2012.
McCarthy, M. J. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
McCarthy, M. J. (2010). Spoken fluency revisited. *English Profile Journal* 1.1, e4.
McCarthy, M. J., J. McCarten & H. Sandiford (2012). *Viewpoint 1*. Cambridge: Cambridge University Press.
Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Cambridge ESOL Research Notes* 37, 2–5.
Murakami, A. (2013). Cross-linguistic influence on the accuracy order of L2 English grammatical morphemes. In S. Granger, S. Gaëtanelle & F. Meunier (eds.), *Proceedings of learner corpus research 2011*. Louvain-la-Neuve: Presses universitaires de Louvain.
Nunan, D. (1988). *Syllabus design*. Oxford: Oxford University Press.
Shirai, Y. & A. Kurono (1998). The acquisition of tense-aspect marking in Japanese as a second language. *Language Learning* 48. 2, 245–279.
Van Ek, J. A. & J. Trim (2001) *Vantage. Council of Europe*. Cambridge: Cambridge University Press.

MICHAEL MCCARTHY is Emeritus Professor of Applied Linguistics, University of Nottingham, UK, Adjunct Professor of Applied Linguistics, University of Limerick, Ireland and Visiting Professor in Applied Linguistics at Newcastle University, UK. He is author/co-author/editor of fifty books, including *Touchstone* 1–4 (Cambridge University Press, 2004–2006), *Viewpoint* 1 and 2 (Cambridge University Press, 2012 and 2013) the *Cambridge grammar of English* (Cambridge University Press, 2006), *English grammar today* (Cambridge University Press, 2011), *Grammar for business* (Cambridge University Press, 2009), *From corpus to classroom* (Cambridge University Press, 2007), *The Routledge handbook of corpus linguistics* (Routledge, 2010) and more than a dozen titles in the *English vocabulary in use* series (Cambridge University Press, 1994–2013), as well as more than 90 academic papers. His current research involves the creation and analysis of spoken learner corpora in connection with the EPP.

https://doi.org/10.1017/S0261444813000189 Published online by Cambridge University Press