Criterial Features in Learner Corpora: Theory and Illustrations

John A. Hawkins University of Cambridge, RCEAL, University of California, Davis, Department of Linguistics

Paula Buttery University of Cambridge, RCEAL, University of California, Davis, Department of Linguistics

Abstract

One of the major goals of the Cambridge English Profile Programme is to identify 'criterial features' for each of the Common European Framework of Reference (CEFR) proficiency levels as they apply to English, and to assess the impact of different first languages on these features (through 'transfer' effects). The present paper defines what is meant by criterial features and proposes an initial taxonomy of four types. Numerous illustrations are given from our collaborative research to date on the Cambridge Learner Corpus. The benefits and challenges posed by these features for corpus linguistics and for theories of second language acquisition are briefly outlined, as are the benefits and challenges for language assessment practices and for publishing ventures that make use of them as supplements to the current CEFR descriptors.

Keywords: criterial feature types, CEFR levels, CLC, learner corpora

1. Introduction: Criterial features

Several decades of practical work on language testing and teaching have led to the six proficiency levels of the Common European Framework of Reference (CEFR), as summarized in the Council of Europe's 2001 document *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Cambridge University Press). These levels are given in (1):

(1) The CEFR Levels: C2 Mastery
                           C1 Effective Operational Proficiency
                           B2 Vantage
                           B1 Threshold
                           A2 Waystage
                           A1 Breakthrough

The Council of Europe proposed a large number of 'illustrative descriptors' in chapter 3 of its document that were designed to distinguish between the levels. In Appendix D (pp. 244–57) the authors also summarize a set of 'can do' statements developed by the Association of Language Testers in Europe (ALTE) and anchored to the descriptors. For example, learners at B1 can 'express opinions on abstract/cultural matters in a limited way or offer advice

within a known area'. Learners at B2 'can follow or give a talk on a familiar topic'. And learners at C1 'can contribute effectively to meetings and seminars within own area of work or keep up a casual conversation with a good deal of fluency'.

These six CEFR levels and their illustrative descriptors have had a major impact on language assessment practices throughout Europe, and on the teaching and learning of foreign languages. The descriptors are given in functional terms, i.e. they describe the different uses to which language can be put and the various functions that learners can perform as they gradually master a second language (L2). The descriptors do not give language-specific details about the grammar and lexis that are characteristic of each of the proficiency levels for each L2. Even chapter 5 of the 2001 document on 'The user/learner's competences', which includes discussion of syntax, morpho-syntax and lexis, does not link particular grammatical and lexical properties to the CEFR levels with any degree of specificity. There was a reason for this: those who wrote the CEFR wanted it to be neutral with respect to the L2 being acquired and hence compatible with the different languages of Europe. In this way a given level of proficiency in L2 German, for example, could be compared with a corresponding level in L2 French or English.

The result is, however, that the CEFR levels are underspecified with respect to key properties that examiners look for when they assign candidates to a particular proficiency level and score in a particular L2 (cf. Milanovic 2009). Learners who perform each of the functions in the illustrative descriptors may be using a wide variety of grammatical constructions and words, and the ability to 'do' the task does not tell us with precision how the learner does it and with what grammatical and lexical properties of English (or of other target languages). It is this (deliberate) underspecification that has led to the research programme described in this paper.

This programme is embedded within a larger applied and theoretical research programme, the English Profile Programme, which was initiated by the Cambridge ESOL group of Cambridge Assessment in collaboration with Cambridge University Press and other stakeholders in 2005. One of the goals of this programme is to add specific grammatical and lexical details of English to the functional characterization of the CEFR levels based on the Cambridge Learner Corpus (CLC). The CLC is made up of approximately 45 million words of English taken from exam scripts written by learners around the world at all levels of proficiency. It has been tagged for parts of speech and parsed using a sophisticated automatic parser (Briscoe, Carroll & Watson 2006) permitting numerous grammatical as well as lexical searches to be conducted. Approximately half has been error-coded using grammatical and lexical codes devised by researchers at Cambridge University Press. English Profile builds on the pioneering work of van Ek and Trim, for example in their *Threshold* 1990 book, which added specific grammatical and lexical details of English to a functional and notional characterization of this level, and which linked grammatical and lexical exponents to a rich inventory of the functions in clear and practically useful ways. Van Ek and Trim did not have access to the rich database of the CLC for empirical testing of their claims however, and nor was their work guided by the search for criterial features at the different levels.

The basic intuition behind the criterial feature concept is that there are certain linguistic properties that are characteristic and indicative of L2 proficiency at each level, on the basis of which examiners make their practical assessments. Since there is a large measure of inter-examiner agreement, and since the illustrative descriptors are underspecified with respect to these L2 properties, we need to discover what it is exactly that the examiners

look for when they assign the scores they do. It is reasonable to assume that their collective experience over many years has led to an awareness of the kinds of properties that distinguish levels and scores from one another. The challenge is to discover what these properties are: this is the essence of the criterial feature concept. If we can make the distinguishing properties explicit at the level of grammar and lexis, and ultimately for phonology, semantics and form-function correspondences as well, then we will have identified a set of linguistic features which provide the necessary specificity to CEFR's functional descriptors for each of the proficiency levels. This will have considerable practical benefits for both examining and publishing. It can also contribute new patterns and insights to theories of second language acquisition.

As a background, notice that criterial features must be defined in terms of the linguistic properties of the L2 as used by native speakers, that have either been correctly, or incorrectly, attained at a given level. In the former case we can talk of 'positive grammatical properties' that are acquired at a certain L2 level, in the latter of 'negative grammatical properties of an L2 level', i.e. of errors and error frequencies characteristic of different levels. We can also define properties more gradiently as well as discretely, and subsume patterns of usage and frequency that are characteristic of L1 speakers of English. Again, learners may show a 'positive usage distribution for a correct property of L2' at a particular level that matches the distribution of native speakers (i.e. L1) of the L2, or they may show a 'negative usage distribution' that does not match that of native speakers.

Notice that defining what is criterial for each level involves multiple factors, since there are many properties of these different types that are potentially criterial. Also, criterial properties may not be unique to just one level. A given property may distinguish both of the C levels from the A and B levels. Or it may distinguish both B levels from the A and C levels, or it may be unique to B2, or to C2. The ultimate definition for what is criterial for a particular level, say B2, will be a cluster of criterial features, each of which distinguishes B2 either uniquely or non-uniquely from other levels. Obviously, the more unique a feature is to B2 alone, the more useful it will be as a diagnostic for that level. But non-uniqueness to B2 is not problematic. In fact, it will be an inevitable consequence of certain types of criteriality. If learners of English acquire a novel structure at B2, that structure will generally persist through the higher C levels and will be characteristic of B2–C2 inclusive, and will distinguish them from all lower levels.

One of the goals of English Profile from the outset has been to identify criterial features of these different types for each of the CEFR levels, and to assess the impact of different first languages on these features (through 'transfer' effects). The present paper defines the criterial feature concept and illustrates the different types. The order of presentation is as follows. Section 2 defines the four types of criterial features. Section 3 gives some brief initial illustrations of each. Section 4 defines L1-specific criterial features. Section 5 describes the Cambridge Learner Corpus, the error codes, and the tagging and parsing system that has been applied to it. Section 6 gives a more detailed illustration of different criterial features in the area of verb co-occurrence frames, i.e. for basic construction types of English defined in terms of the verb and its co-occurring phrases. Section 7 illustrates criterial features for relative clause constructions. Section 8 quantifies different error types and links them to the criterial feature concept. Section 9 illustrates L1-specific criterial features in the area of missing determiners. And section 10 concludes with a discussion of the benefits and challenges for this research programme. For a more detailed and much extended development of this research programme the reader is referred to Hawkins and Filipovic (2011).

2. Four types of criterial features

Based on the discussion above we can set up a preliminary four-way classification system. In this section we define the four classes briefly and in the next section we give illustrations of each.

In order to provide a framework that is useful to researchers within linguistics and applied linguistics we define two basic notions: linguistic properties can be **positive** or **negative**.

For a given language task we may ascertain how often a linguistic property is likely to be exemplified by a native speaker of English. We could, for instance, estimate this from examining native speaker corpora. When accurate models of the native speaker predict a non-zero probability of usage we say that the property in question is a **positive** linguistic property. In traditional linguistic terms positive properties (sentences, constructions, and their meanings, etc.) are those that are generated by the grammar of the relevant language – here English – and judged to be well-formed, and used, by native speakers. We can then ascertain how often a learner at a given proficiency level exemplifies the same property. Based on analysis of the learner data we might, for instance, discover that we do not expect any examples of the property to appear before the C1 proficiency level.

When the native speaker estimates tend towards a zero probability of usage but the learner estimates show a non-zero probability we say that the learners are exhibiting a **negative** linguistic property. Negative properties are those that fall outside the set generated and that are judged ill-formed by native speakers. It is generally desirable for learners to reduce their production of language containing these negative properties, or 'errors', in the interests of successful communication.

The number line in Figure 1 illustrates a positive linguistic property. For a given language task we can place an arrow along the line to indicate how often a linguistic property is likely to be exemplified by a native speaker of English. Other arrows may be placed to indicate how often learners at given proficiency levels are likely to exemplify the same property. Figure 2 illustrates a negative linguistic property.
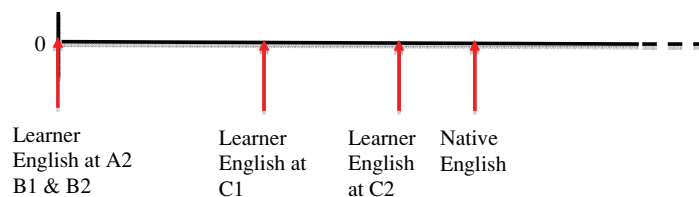

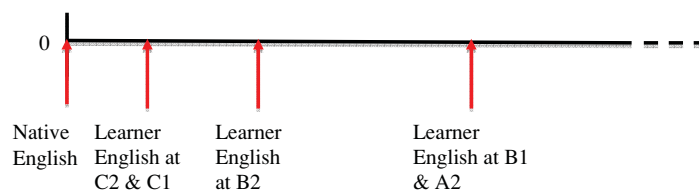**Figure 1** Example number line showing a positive linguistic property.

Figure 2 Example number line showing a negative linguistic property.

Additionally, we can refer to the extent to which the estimates for the learner usage approximate the native speaker usage (i.e. the distance between the learner arrows and the native speaker arrow in Figures 1 and 2). These estimates may be either positive or negative with respect to the native speaker benchmark.

From a practical perspective the classification we propose below is well-motivated and useful. (Here and throughout this paper the use of square brackets around a CEFR level or levels, e.g. [C1, C2], indicates that the levels have the criterial feature in question.)

2.1. Positive linguistic properties are correct properties of English that are acquired at a certain L2 level and that generally persist at all higher levels

For example, a property P acquired at B2 may differentiate [B2, C1, C2] from [A1, A2, B1] and will be criterial for the former. Criteriality characterizes a set of adjacent levels in this case. Or some property might be attained only at C2 and be unique to this highest level.

2.2. Negative linguistic properties of an L2 level are incorrect properties or errors that occur at that level, and with a characteristic frequency

Both the presence versus absence of errors, and the characteristic frequency of the error (the 'error bandwidth', cf. section 8 below), can be criterial for the given level or levels. For example, error property P with a characteristic frequency F may be criterial for [B1, B2]; error property P' with frequency F' may be criterial for [C1, C2].

2.3. Positive usage distributions for a correct property of L2 match the corresponding distribution for native speaking (i.e. L1) users of the L2

A positive usage distribution matching that of native speakers may be acquired at a certain level and will generally persist at all higher levels and be criterial for the relevant levels, e.g. [C1, C2].

2.4. Negative usage distributions for a correct property of L2 do not match the distribution of native speaking (i.e. L1) users of the L2

A negative usage distribution that does not match that of native speakers may occur at a certain level or levels with a characteristic frequency F and be criterial for the relevant level(s), e.g. [B2].

3. Initial illustrations of the criterial feature types

3.1. Positive linguistic properties of the L2 levels

Caroline Williams has examined the emergence of new basic construction types in the CLC, which she describes in terms of 'verbal subcategorisation frames' and which we shall describe

here as 'verb co-occurrence' patterns, see section 6 below (Williams 2007). For example, new verb co-occurrences that appear at B1, such as the 'ditransitive' NP-V-NP-NP structure (*she asked him his name*), are criterial for [B1, B2, C1, C2]; those appearing at B2, for example, the object control structure NP-V-NP-AdjP (*he painted the car red*), are criterial for [B2, C1, C2].

3.2. Negative grammatical properties of the L2 levels

Morpho-syntactic, syntactic and lexical error types frequently have significantly different 'error bandwidths' (cf. section 8) at different proficiency levels, and these bandwidths can be characteristic of, and criterial for, the different levels. For example, errors involving incorrect morphology for determiners, as in Derivation of Determiners (abbreviated DD) *Shes name was Anna* (instead of *Her name* . . .), show significant differences in error frequencies that decline from B1 > B2 > C1 > C2. The relevant bandwidths are criterial for B1 versus B2 versus C1 versus C2 here. Noun agreement errors, on the other hand (abbreviated AGN), such as *one of my friend*, show significant differences in frequency that increase from B1 to B2 and then decline again at each of the C levels, resulting in an 'inverted U' pattern of B2 > [B1, C1, C2]. The relevant error scores and bandwidths are criterial for B2 versus [B1, C1, C2] (B1 and C1 being non-adjacent levels grouped into the same criterial feature set on this occasion).

3.3. Positive and negative usage distributions for correct L2 properties (that match or do not match L1 frequencies respectively)

The distribution of relative clauses formed on indirect object/oblique positions (e.g. *the professor that I gave the book to*) to relativizations on other clausal positions (subjects and direct objects) appears to approximate to that of native speakers at the C levels, but not at earlier levels. Hence this is a positive usage distribution that is criterial for [C1, C2], while the negative usage distribution is criterial for the lower levels (see section 7 for details). There is a lexical frequency skewing for common verbs of English (e.g. *know, see, think*) compared with native speaking usage, whereby these verbs are over-represented in the learner corpora at lower proficiency levels, given the more general skewing at early levels of proficiency in favour of more frequent lexical items and constructions of the L2 (see Hawkins & Buttery 2009). These atypical frequency skewings are negative usage distributions that decline at higher levels of proficiency, as learners move to a more positive and native-like balance between high-, mid- and low-frequency items. The precise bandwidth of the skewing can be criterial for the relevant level(s).

4. L1-specific criterial features

The criterial feature types and illustrations given so far are based on English L2 data and on comparisons with native-speaking English data that hold regardless of the L1 of the learner. We can, in addition, define a set of 'L1-specific criterial features', consisting of the same four possible types, that hold only for a particular L1 or set of L1s, e.g. for Romance language

speakers learning English, or for speakers of languages without definite or indefinite articles learning English.

For example, learners of English whose native language belongs to the Germanic language family may acquire some positive property P of L2 English at an earlier level than Romance speakers, or than any other speakers; or speakers of Romance languages may acquire P earlier, etc.

Speakers of languages without definite and indefinite articles have error bandwidths for Missing Determiner errors in L2 English that are significantly higher, for all levels, than those for speakers of languages with articles. The characteristic error bandwidths for article-less languages generally follow a B1 > B2 > C1 > C2 pattern that makes the relevant bandwidths criterial for the relevant L1s at the respective levels. See section 9 below.

The positive and negative usage distributions for correct properties of English L2 may also vary by first language and first language type.

5. The Cambridge Learner Corpus (CLC)

Before providing a more detailed discussion of the criterial feature types that were defined in section 2 and illustrated in section 3, a brief summary of the CLC is in order.

The CLC currently comprises approximately 45 million words of written data from learners of English around the world who have taken Cambridge proficiency exams. As a result the CLC contains data from numerous (over 30) grammatically and typologically different first languages. Between a third and one half of the corpus is coded for errors based on a manual annotation conducted by our Cambridge University Press colleagues prior to the initiation of the interdepartmental and interdisciplinary English Profile Programme. The CLC's error codes classify over 70 error types involving lexical, syntactic and morphosyntactic properties of English which are judged by the Cambridge University Press coders to have been incorrectly used by non-native learners. A sample is given in (2), together with sentences exemplifying each:

(2) *Sample error codes in the CLC*

RN	**Replace noun**	*Have a good travel* (journey)
RV	**Replace verb**	*I existed last weekend in London* (spent)
MD	**Missing determiner**	*I spoke to President* (the)
		I have car (a)
AGN	**Noun agreement error**	*One of my friend* (friends)
AGV	**Verb agreement error**	*The three birds is singing* (are)

The CLC was originally searchable lexically, i.e. on the basis of individual words, and grammatically to the extent that a rule of English grammar was reflected in an error code. Within the EPP this search capability has been expanded and the CLC has been tagged for parts of speech and parsed using the Robust Accurate Statistical Parser (RASP) developed by Ted Briscoe and John Carroll of the Cambridge Computer Laboratory and the University of Sussex respectively, (see Briscoe, Carroll & Watson 2006). This is an automatic parsing system incorporating both grammatical information and statistical patterns, details of its operation

are summarized very briefly in this section. Applying the RASP toolkit to the CLC makes it possible to conduct more extensive grammatical, as well as lexical, analyses.

When the RASP system is run on raw text, such as the written sentences of the CLC, it first marks sentence boundaries and performs a basic 'tokenisation'. The text is then tagged with one of 150 part-of-speech and punctuation labels, on a probabilistic basis. It is 'lemmatized', i.e. analyzed morphologically, to produce dictionary citation forms (or lemmas) plus any inflectional affixes. A sentence like *This is an example sentence* would be represented as (3) at this point

(3) This_DD1 be+s_VBZ an_AT1 example_NN1 sentence_NN1 ._.
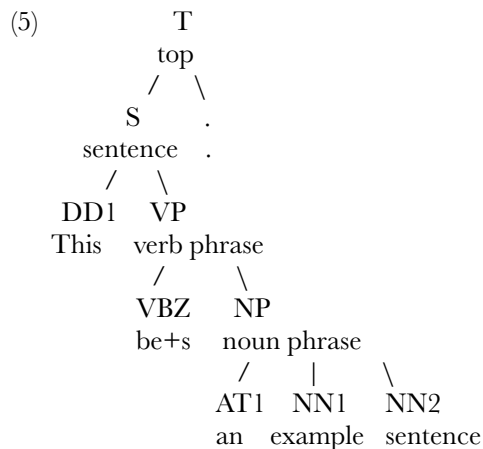
And with word numbering it would look like (3′):

(3′) This:1_DD1 be+s:2_VBZ an:3_AT1 example:4_NN1 sentence:5_NN1 .:6_.

5.1. Parsing using RASP

Parse trees are constructed from the sequence of tags using a set of around 700 probabilistic Context Free Grammar production rules (of the general form S $\Rightarrow$ VP NP). Features associated with the parse are passed up the tree during construction. A parse in its simplified form is shown in (4):

```
(4)  (TOP
       (S (DD1 This:1)
         (VP (VBZ be+s:2)
              (NP (AT1 an:3) (NN1 example:4) (NN2 sentence:5))))
        (. .:6))
```

which is the textual representation of the tree diagram in (5):

```
(5)              T
               top
              /   \
           S        .
        sentence    .
          /   \
       DD1    VP
      This   verb phrase
             /        \
          VBZ     NP
          be+s    noun phrase
                  /    |    \
               AT1   NN1    NN2
               an   example  sentence
```

All the possible parses for a sentence are referred to as a **parse forest** and RASP ranks these according to the probabilities associated with each of the production rules that are used to derive the tree. The parsing algorithm builds all the parses in the parse forest simultaneously, proceeding from left to right across the tag sequence. It activates all possible rules that are consistent with the partial parses already built and with the next tag in the sequence. This has the consequence that rules can be activated during processing that will not ultimately be used in any valid parse. The parse forest has a dynamic size, therefore, that alters as the parsing algorithm progresses from one tag to the next. The probabilities were originally derived from counting rule usage frequencies in a manually annotated corpus.

When a sentence is ambiguous there will be more than one valid parse tree. (6) below shows the parses for the sentence *She saw the girls with the telescope* which has an ambiguity over where to attach the prepositional phrase *with the telescope*. In Parse 1 (7) the girls were seen with the aid of a telescope, but in Parse 2 (8) the telescope is in the possession of the girls. The complete noun phrases containing *the girls* are highlighted in bold in each case for comparison.

(6) Parse: 1
  (TOP
    (S (PPHS1 She:1)
      (VP
        (VVD see+ed:2) **(NP (AT the:3) (NN2 girl+s:4))**
        (PP (IW with:5) (NP (AT the:6) (NN1 telescope:7)))))
  (. .:8))

  Parse: 2
  (TOP
    (S (PPHS1 She:1)
      (VP
        (VVD see+ed:2)
          **(NP (AT the:3) (NN2 girl+s:4) (PP (IW with:5) (NP (AT the:6) (NN1
                                                            telescope:7))))))**
  (. .:8))

5.2. Grammatical relations in RASP

Grammatical relations are assigned in RASP that capture reasonably theory-neutral binary relations between lexical items. They are expressed in the general format: (|relation-type| |head| |dependant|).

For example, RASP would annotate sentence (7) with the grammatical relations shown in (8):

(7) She was eating an apple.
(8) (|subject| |eating| |She| _)
   (|auxiliary| |eating| |was|)
   (|direct-object| |eating| |apple|)
   (|determiner| |apple| |an|)

Once such annotation has been provided, many detailed grammatical investigations become an efficient possibility. Expanding from the example above, imagine we now wish to extract all of the direct objects of 'eating' from the corpus. Traditionally we would have to do a string search for 'eating' and then manually check all of the returned concordances for the direct objects. This is tedious at best, and at worst we might find ourselves in a situation where the direct object lies outside the concordance window (e.g. in a sentence where the direct object is displaced such as in 'The boy is eating, or perhaps he would be better described as scoffing, a juicy red apple'). By collating over grammatical relations, a frequency list can be quickly constructed, as shown for instance in (9):

(9) Grammatical relation ➜ frequency count
    (|direct-object| |eating| |apple|) ➜ x
    (|direct-object| |eating| |pizza|) ➜ y
    . . . .
    (|direct-object| |eating| |words|) ➜ z

With respect to the English Profile analysis for this specific example, we should expect to find the more frequent literal uses of 'eating' to be present at all proficiency levels whereas the abstract usage ('eating words') would occur primarily at the higher C levels.

Beyond this simple word sense demonstration, the grammatical relations can also be used to identify particular grammatical constructions (for instance the verb subcategorisation constructions identified in section 6). To illustrate how this is done consider the following problem: how do we find all ditransitive verbs in the corpus? From experience we know that the ditransitive verb 'give' can occur in the structures shown in (10):

(10) Simon gives the book to her.
      Toby gives the flowers to his mum.

Using these forms as a template we could search in a suitably annotated corpus for the general pattern given in (11):

(11) SUBJECT gave DIRECTOBJECT to INDIRECTOBJECT

However, this search would fail to return the sentence 'Francis gave Jodie the job' and also any sentence containing a ditransitive with a lexical item that we haven't previously identified, e.g. the one in (12):

(12) Lucy passed her sister the butter.

The solution is to search within the grammatical relations annotation for the underspecified pattern that expresses the ditransitive verb frame, as illustrated in (13):
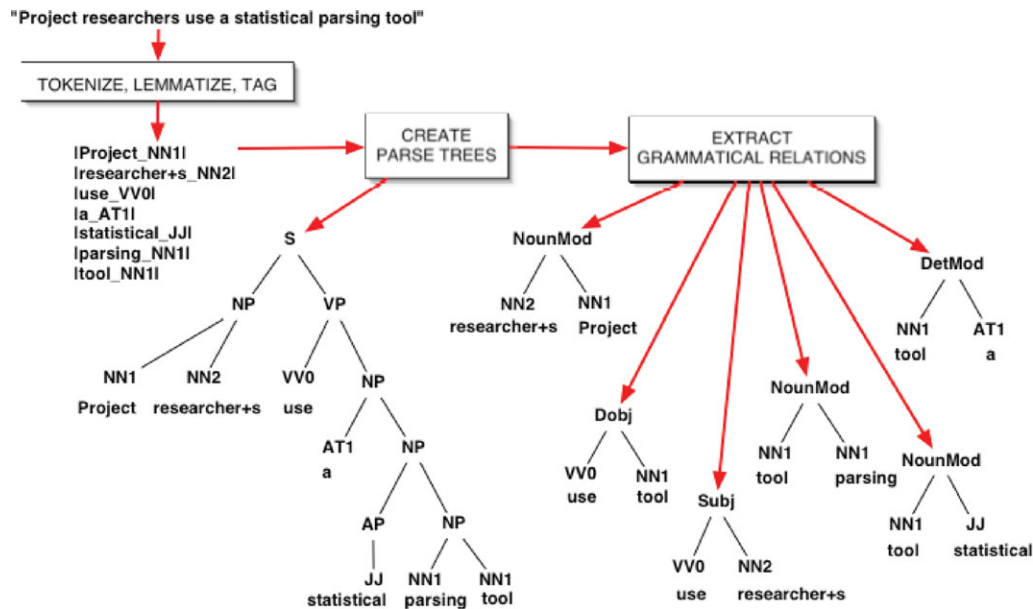
Figure 3 'Project researchers use a statistical parsing tool'.

(13) IF the grammatical relations for a sentence match:
      (|subject| ?x ?a _)
      (|direct-object| ?x ?b)
      (|indirect-object| ?x ?c)
      where ?x, ?a, ?b, ?c are variables
      AND ?x is a verb
      THEN the verb frame for ?x is DITRANSITIVE

Finally, RASP performs a semantic analysis that captures information about the argument-predicate structure of each sentence.

The full lexical and syntactic operations performed by RASP are summarized for a sample sentence in Figure 3, using just one illustrative parse tree and its associated grammatical relations.

Each (error-coded) sentence in the CLC has been automatically annotated with part-of-speech tags, word lemmas, phrasal parse groupings and grammatical relations along these lines, making a number of searches possible that help us to identify criterial features of the CEF proficiency levels.

6. Verb co-occurrence frames

Table 1 shows the verb co-occurrence frames identified by Williams (2007) that are found at A2. This table includes some of the most basic, most frequent and simplest construction types of English, such as intransitive sentences (*he went*), basic transitives (*he loved her*), intransitives

Table 1   *A2 Verb Co-occurrence Frames.*

NP-V	*He went*
NP-V (reciprocal Subj)	*They met*
NP-V-PP	*They apologized [to him]*
NP-V-NP	*He loved her*
NP-V-Part-NP	*She looked up [the number]*
NP-V-NP-Part	*She looked [the number] up*
NP-V-NP-PP	*She added [the flowers] [to the bouquet]*
NP-V-NP-PP (P = for)	*She bought [a book] [for him]*
NP-V-V(+ing)	*His hair needs combing*
NP-V-VPinfinitival (Subj Control)	*I wanted to play*
NP-V-S	*They thought [that he was always late]*

Table 2   *New B1 Verb Co-occurrence Frames.*

NP-V-NP-NP	*She asked him [his name]*
NP-V-VPinfin (Wh-move)	*He explained [how to do it]*
NP-V-NP-V(+ing) (Obj Control)	*I caught him stealing*
NP-V-NP-PP (P = to) (Subtype: Dative Movement)	*He gave [a big kiss] [to his mother]*
NP-V-NP-(to be)-NP (Subj to Obj Raising)	*I found him (to be) a good doctor*
NP-V-NP-Vpastpart (V = passive) (Obj Control)	*He wanted [the children] found*
NP-V-P-Ving-NP (V = +ing) (Subj Control)	*They failed in attempting the climb*
NP-V-Part-NP-PP	*I separated out [the three boys] [from the crowd]*
NP-V-NP-Part-PP	*I separated [the three boys] out [from the crowd]*
NP-V-S (Wh-move)	*He asked [how she did it]*
NP-V-PP-S	*They admitted [to the authorities] [that they had entered illegally]*
NP-V-Part	*She gave up*
NP-V-S (whether = Wh-move)	*He asked [whether he should come]*
NP-V-P-S (whether = Wh-move)	*He thought about [whether he wanted to go]*

with a prepositional phrase complement (*he apologized to her*), and so on. They are described in Williams (2007) using a taxonomy of verb subcategorization possibilities originally developed by Ted Briscoe, John Carroll and their colleagues (see Briscoe & Carroll 1997, Briscoe 2000, Korhonen, Krymolowski & Briscoe 2006, Preiss, Briscoe & Korhonen 2007). Table 2 lists the verb co-occurrence frames that appear first at B1. These are positive grammatical properties which are criterial for the B and C levels, i.e. for all of [B1, B2, C1, C2] distinguishing them from the A levels. This is a type 1 (positive) criterial feature for these levels as defined in section 2.1. Table 3 shows the new frames which appear at B2 and which are criterial for [B2, C1, C2] versus [A1, A2, B1].

Table 3   *New B2 Verb Co-occurrence Frames.*

NP-V-NP-AdjP (Obj Control)	*He painted* [*the car*] *red*
NP-V-NP-*as*-NP (Obj Control)	*I sent him as* [*a messenger*]
NP-V-NP-S	*He told* [*the audience*] [*that he was leaving*]
NP-V-P-NP-V (+*ing*)(Obj Control)	*They worried about him drinking*
NP-V-P-VPinfin (Wh-move)(Subj Control)	*He thought about* [*what to do*]
NP-V-S (Wh-move)	*He asked* [*what he should do*]
NP-V-Part-VPinfin (Subj Control)	*He set out to win*

Notice here and throughout this section that the example sentences given to illustrate the verb co-occurrences of English, in the tables and the main text, have been taken from the research literature and are not being claimed to be attested in the CLC in exactly this form. If a co-occurrence frame is listed in one of these tables, this means that Williams (2007) found evidence for this construction type at this level, not necessarily for this token. Research is currently in progress to determine exactly which lexical verbs are found in each construction at each level, see Hawkins & Filipovic (2011) and the English Profile Wordlists compiled by Annette Capel.

7. Relative clauses

We searched for relative clauses, i.e. structures such as *the student that I taught*, in the CLC and also in various subcorpora of the British National Corpus (BNC). The particular types we were interested in involved Relative Clause Formation on subject, direct object, and indirect object/oblique positions. Examples of the different relative clause types are given in Table 4. See Keenan & Comrie (1977) and Hawkins (1994, 1999, 2004) for discussion of the significance of these relative clause types from a grammatical, typological and processing perspective, and Eckman (1984) and Hyltenstam (1984) for their significance in a second language acquisition context.

Table 4   *Relative clause types.*

Subject Relatives	The student who/that wrote the paper
	The film which/that shocked the world
Direct Object Relatives	The student who(m)/that I taught
	The student O I taught
	The film which/that I saw
Indirect/Oblique Relatives	The student to whom I gave the book
	The student who/that I gave the book to
	The student O I gave the book to

Table 5 gives the frequencies of occurrence for relativizations on these different positions as a percentage of the total within each CEFR level. Table 6 does the same for various subcorpora of the BNC and so provides a comparison of the learner data in Table 5 with

Table 5   *Usage of different types of relative clauses as percentage of total within each CEFR level.*

	A2	B1	B2	C1	C2
Subject RCs	67.74%	61.05%	71.11%	70.35%	74.80%
Object RCs	30.65%	37.33%	26.09%	25.02%	20.90%
ind/obl RCs	1.61%	1.62%	2.80%	4.63%	4.30%

Table 6   *Usage of different types of relative clauses as percentage of total within subcorpora of the BNC.*

	The Guardian	Mils and Boon	New Scientist	The Law Report
Subject RCs	77.20%	75.63%	80.67%	58.32%
Object RCs	17.02%	20.03%	12.44%	21.98%
ind/obl RCs	5.76%	4.34%	6.89%	19.71%

data from native speakers of English. The BNC has also been tagged and parsed using the RASP toolkit (see section 5), making exact comparison between them possible.

The distributional differences between relativizations on these different positions are striking and roughly in accordance with corpus frequency studies and psycholinguistic experimental findings, which have been reported elsewhere (cf. Hawkins 1994, 1999, 2004). It must be stressed, however, that these data are preliminary and that an adequate gold standard remains to be set. Identification of relative clauses by RASP is more reliable where there is an explicit relative pronoun (*the student that I taught*) as opposed to a zero form (*the student Ø I taught*). This may explain one surprising feature of Table 5: the high proportion of object relatives (*the student that I taught*) to subject relatives (*the student who wrote the paper*) at A2–C1 is 30.65%, 37.33%, 26.09% and 25.02% respectively compared with corresponding percentages in the BNC (17.02%, 20.03%, 12.44% and 21.98% respectively). It could be that more structures are being analysed as zero (object) relatives by RASP, whereas subject relatives which do not permit zero relatives are being more faithfully recognized by the parser.

With this caveat, notice an interesting feature of Table 5 that can be regarded as criterial, by the definitions in section 2. The relative distribution of indirect object/oblique relative clauses (e.g. *the professor that I gave the book to*) to relativizations on the other clausal positions (subjects, direct objects and genitives) departs at CEFR levels A2, B1 and B2 from that of native speakers of English: 1.61%, 1.62% and 2.80% respectively in Table 5 versus at least 4.34% in the BNC (cf. Table 6). The distribution at the C levels (4.63% and 4.30%) matches the lowest figures of the BNC subcorpora (4.34%). Hence we have here a positive usage distribution feature for relatives formed on indirect objects/obliques at [C1, C2], i.e. a type 3 criterial feature, but a type 4 negative usage distribution feature for [A2, B1] and B2 respectively, each with their characteristic frequency bandwidths.

8. Errors

Criterial features of type 2 (section 2.2) refer to negative grammatical properties, i.e. errors occurring at a certain level or levels with a characteristic frequency. We can say that an error distribution is criterial for a level L if and only if the frequency of errors at L is significantly different from its frequency at the level with the next higher error count.

Error frequencies were quantified in two ways for each of the error types in the CLC, i.e. for each of the coded errors identified by our Cambridge University Press colleagues (see section 5). First of all, the number of errors was calculated (with the help of Caroline Williams) per 100,000 words of text at each level. In tables 7 and 8 this figure is given in square brackets alongside the level to which it refers. For example, B1[22.5] means 22.5 errors per 100,000 words of text at B1. Secondly, and more reliably, error counts were relativized to the total number of tokens of each linguistic category (e.g. verb or noun). The purpose here was to better quantify the ratio of errors to correct instances of each error type. The number of tokens of different parts of speech can be quite different in 100,000 words of text, and hence the actual number of errors for each category could be quite different too. By relativizing error counts to the part of speech in question we achieve a better reflection of the universe of relevant instances within which we can quantify success or failure. These relativized error percentages are shown in parentheses in Tables 7 and 8. For example, an error percentage of (5%) indicates that one in 20 instances of the category in question exhibits that error.[1]

For an error frequency to be considered significant and criterial in this context we insisted on a reduction of at least 29% compared to the level with the next higher error count (which is not always, as we shall see, the next higher proficiency level). If the percentage of errors at level L were 2.5%, for example, and the percentage for the level with the next higher count were 4%, then this would be a reduction of 1.5 relative to 4, or 37.5%. The 29% figure was chosen since it guarantees a minimum requirement for statistical significance, namely at least one standard deviation from the mean. This percentage difference in error improvements was therefore used as a minimum criterion for distinguishing error frequencies between levels. Two or more levels were grouped together for criteriality if each was not significantly differentiated from the level with the immediately higher error count (i.e. by less than 29%), using the relativized error percentage figures for each linguistic category. An error bandwidth was then calculated between each criterial level or levels by splitting the difference between criterial error distributions. For example, if the lowest error percentage for a given error type at the B levels ([B1, B2]) were 2.2% and the highest for the C levels ([C1, C2]) were 1.2% then the bandwidth between them would be set at 1.6%: lower than that would be criterial for the C levels, higher would be criterial for the B levels.

Two major patterns emerge from the CLC in these error distributions. First there are 'progressive learning patterns', i.e. error frequencies decrease at higher CEFR levels as learners improve their command of English. (A2 has insufficient quantities of words overall and is excluded from the tables in this section.) Sometimes there is a significant reduction in errors between adjacent levels. At other times there is no significant difference between two

---

[1] These error percentages were derived from approximately 5 million words of the CLC that had been parsed and tagged by RASP (cf. section 5).

Table 7   *Progressive learning errors in the CLC B1 → C2.*

1. DD **Derivation of determiner**	*Shes name was Anna*		B1 > B2 > C1 > C2
‖ B1[22.5] ‖ B2[15.9] ‖ C1[10.6] ‖ C2 [5.52] ‖			
(0.263%) (0.171%) (0.097%) (0.053%)			
0.217% 0.134% 0.075%			
2. FD **Form of determiner**	*I have an car*		B1 > B2 > C1 > C2
‖ B1[40.1] ‖ B2[29.8] ‖ C1[25.8] ‖ C2[15.5] ‖			
(0.469%) (0.321%) (0.236%) (0.148%)			
0.395% 0.279% 0.192%			
3. IV **Incorrect inflection of verb**	*I spended last weekend in London*		B1 > [B2, C1, C2]
‖ B1[108] ‖ B2[64.5] C1[46.5] C2[40.1] ‖			
(0.498%) (0.296%) (0.228%) (0.192%)			
0.397%			
4. MT **Missing preposition**	*I gave it John*		[B1, B2] > [C1, C2]
‖ B1[288] B2[299] ‖ C1[183] C2[127] ‖			
(3.297%) (2.998%) (1.684%) (1.253%)			
2.341%			
5. MQ **Missing quantifier**	*I'll call in the next days*		[B1, B2] > [C1, C2]
‖ B1[17.0] B2[23.8] ‖ C1[11.2] C2[9.12] ‖			
(1.377%) (1.250%) (0.557%) (0.514%)			
0.904%			
6. RQ **Replace quantifier**	*It all happened a lot of years ago*		[B1, B2, C1] > C2
‖ B1[27.6] B2[33.8] C1[26.3] ‖ C2[15.1] ‖			
(2.234%) (1.777%) (1.312%) (0.851%)			
1.082%			
7. RY **Replace adverb**	*He stared at her intensively*		[B1, B2, C1] > C2
‖ B1[159] B2[172] C1[160] ‖ C2[117]			
(2.598%) (3.076%) (2.957%) (1.888%)			
2.243%			
8. DY **Derivation of adverb error**	*It happened fastly*		[B1, B2, C1] > C2
‖ B1[68.7] B2[76.9] C1[60.6] ‖ C2[39.2] ‖			
(1.120%) (1.377%) (1.121%) (0.634%)			
0.877%			
9. CQ **Quantifier countability error**	*It cost him many money*		[B1, B2, C1] > C2
‖ B1[9.66] B2[16.1] C1[15.1] ‖ C2[6.2] ‖			
(0.113%) (0.174%) (0.139%) (0.059%)			
0.086%			

Table 8 *Inverted U error patterns in the CLC.*

1.AGN **Noun agreement error** *One of my friend*		B2 > [B1, C1, C2]
‖ B2[207] ‖ B1[113] C1[121] C2[79.3] ‖		
(0.978%) (0.582%) (0.527%) (0.378%)		
0.780%		
2. UV **Unnecessary verb** *I spend to be last weekend in London*		B2 > [B1, C1, C2]
‖ B2[113] ‖ B1[70.6] C1 [61.6] ‖ C2 [50.8] ‖		
(0.517%) (0.328%) (0.302%) (0.241%)		
0.423% 0.272%		
3. MN **Missing noun** *It was an interesting*		B2 > [B1, C1, C2]
‖ B2[78.4] ‖ B1[39.4] C1[62.8] C2[32.9] ‖		
(0.370%) (0.203%) (0.273%) (0.157%)		
0.322%		
4. CN **Countability error** *I don't have any monies*		[B2, C1] > [B1, C2]
‖ B2[52] C1[42] ‖ B1[22] C2[22] ‖		
(0.249%) (0.183%) (0.114%) (0.105%)		
0.149%		
5. UD **Unnecessary determiner** *There was a lot of the traffic*		[B2, C1] > [B1, C2]
‖ B2[341] C1[255] ‖ B1[181] C2[190] ‖		
(2.584%) (2.399%) (1.578%) (1.138%)		
1.989%		
6. MC **Missing conjunction** *The stripes were red green*		B2 > [B1, C1, C2]
‖ B2[100] ‖ B1[43] C1[57.8] C2[51.6] ‖		
(1.441%) (0.730%) (0.843%) (0.714%)		
1.142%		

or more adjacent levels, and these levels are grouped together for criteriality. Examples of these progressive learning patterns are shown in Table 7, with the respective error counts in square brackets and parentheses, and with error bandwidths marked off by '‖' together with their defining error frequencies measured in percentages of errors relative to total words of each category.

Second there are 'inverted U patterns' in the CLC, i.e. errors actually increase after B1 and then decline again by C2. These error types in second language acquisition are reminiscent

of similar patterns in first language acquisition, for example when children use irregular morpho-syntactic forms correctly at first (*run* versus *ran*, *mouse* versus *mice*), then get them wrong by overgeneralizing the productive forms (*run/runned*, *mouse/mouses*), and later learn the exceptions to productive paradigms and use them correctly again (cf. Andersen 1992 for a summary of relevant first language research). Examples of these inverted U patterns are shown in Table 8.

The error types in Tables 7 and 8 involve a broad range of syntactic, morpho-syntactic and lexical properties of English. Syntactic errors include the missing prepositions (Table 7 #4), unnecessary verb (Table 8 #2), and missing conjunction errors (Table 8 #6). Morpho-syntactic errors include the derivation of determiners (Table 7 #1), incorrect inflection of verbs (Table 7 #3), and noun agreement errors (Table 8 #1). Wrong lexical-semantic choices are captured by the replace adverb (Table 7 #7) code, while various subtle lexical co-occurrence restrictions are counted in the replace quantifier code (Table 7 #6).

There are a number of criterial error bandwidths in these tables, and they are criterial for and characteristic of the levels shown. Either this may involve just one level, B1, B2, C1, or C2, or different pairs of levels and triples, [B1, B2], [C1, C2], [B1, B2, C1], and so on. These error types therefore contribute a number of additional diagnostic properties, of a negative nature (i.e. involving properties of English that learners have not yet mastered), to the pool of positive grammatical and lexical properties of the L2 that are indicative of the respective levels, such as the verb co-occurrence frames of Tables 1, 2 and 3 and the relative clause types of Table 5.

9. L1-specific features

Section 4 referred to the possibility some certain criterial features for a level might hold for learners of a certain L1 or group of L1s sharing certain typological or inherited genetic features. A particularly revealing illustration of this comes from CLC data involving determiners in English, i.e. the definite and indefinite article, possessives (*my*, *your*, etc), and demonstratives (*this*, *these*, etc). Speakers of languages without definite and indefinite articles have error bandwidths for Missing Determiner ('MD') errors in L2 English that are significantly higher, for all levels, than those for speakers of languages with articles. Compare Tables 9 and 10. The MD errors here involve sentences such as *I spoke to President* instead of *I spoke to the President* and *I have car* in place of *I have a car*. The characteristic error bandwidths for article-less languages (Table 10) generally follow a B1 > B2 > C1 > C2 pattern (apart from Chinese) which makes the relevant bandwidths criterial for the relevant L1s at the respective levels.

The figures in Tables 9 and 10 indicate the percentage of errors with respect to the total number of correct uses.

10. Benefits and challenges

The grammatical and lexical features presented in this paper are designed to illustrate the theory and practice of criterial feature identification. They are a small part of a research programme that was initiated in response to the Council of Europe's (2001) more functionally oriented illustrative descriptors and 'can do' statements. The CEF and its reference levels

Table 9 *Missing determiner error rates for L1s with articles.*

	A2	B1	B2	C1	C2
			Missing '*the*'		
French	4.76	4.67	5.01	3.11	2.13
German	0.00	2.56	4.11	3.11	1.60
Spanish	3.37	3.62	4.76	3.22	2.21
			Missing '*a*'		
French	6.60	4.79	6.56	4.76	3.41
German	0.89	2.90	3.83	3.62	2.02
Spanish	4.52	4.28	7.91	5.16	3.58

Table 10 *Missing determiner error rates for L1s without articles.*

	A2	B1	B2	C1	C2
			Missing '*the*'		
Turkish	22.06	20.75	21.32	14.44	7.56
Japanese	27.66	25.91	18.72	13.80	9.32
Korean	22.58	23.83	18.13	17.48	10.38
Russian	14.63	22.73	18.45	14.62	9.57
Chinese	12.41	9.15	9.62	12.91	4.78
			Missing '*a*'		
Turkish	24.29	27.63	32.48	23.89	11.86
Japanese	35.09	34.80	24.26	27.41	15.56
Korean	35.29	42.33	30.65	32.56	22.23
Russian	21.71	30.17	26.37	20.82	12.69
Chinese	4.09	9.20	20.69	26.78	9.79

and descriptors have had a major impact on language assessment throughout Europe and on the teaching and learning of foreign languages. They have shaped teaching materials and publications and exams in many different languages in positive ways but they make no reference to detailed grammatical and lexical properties of English, or of other languages, of the kind we have identified here.

Knowing, for example, that learners at C1 'can contribute effectively to meetings and seminars within own area of work or keep up a casual conversation with a good deal of fluency' does not lead us to expect that this would be the level at which we see significant decreases in derivation of determiner errors (Table 7 #1), in missing preposition errors (Table 7 #4), in missing quantifier errors (Table 7 #5), and so on. Knowing that learners at B2 can 'follow or give a talk on a familiar topic' does not lead us to expect that this would be the level at which the new verb co-occurrences of Table 3 would appear, or that there would be characteristic error rates for derivation of determiner errors (Table 7 #1), verb inflection errors (Table 7 #3), and suprisingly high noun agreement errors (Table 8 #1).

Grammatical and lexical properties are sometimes neatly aligned with the functions that learners have learned to perform at the different levels, as outlined in chapter 5 of van Ek & Trim's *Threshold 1990*. But many of the basic grammatical and also lexical properties of a language can be used in the performance of numerous functions, and one and the same function can be performed by many grammatical and lexical forms. A functional-notional approach containing a detailed statement of the functions that learners can express at given levels needs to be supplemented by a statement of the partly orthogonal and autonomous syntactic and morpho-syntactic and lexical properties of the language that are also characteristic of the different levels.

In short, the illustrative descriptors and 'can do' statements of the Common European Framework of Reference are underspecified with regard to grammar and lexis. Milanovic (2009) explains that

> The CEFR is neutral with respect to language and, as the common framework, must by necessity be underspecified for all languages. This means that specialists in the teaching or assessment of a given language ... need to determine the linguistic features which increasing proficiency in the language entails ... A major objective of the *EP Programme* is to analyse learner language to throw more light on what learners of English *can* and *can't* do at different CEFR levels, and to assess *how well* they perform using the linguistic exponents of of the language at their disposal (i.e. using the grammar and lexis of English).

This is the background to the research programme described here. The criterial features we are identifying are based on learners' written English in the 45 million word CLC. Additional data are currently being collected as a check on the reliability of findings to date and in order to conduct new searches, and the EP Programme is being extended to include spoken data, with discourse-based searches extending beyond the sentence-level grammar and lexis. Many more details on criterial features and a new model of learning that accounts for them can be found in Hawkins & Filipovic (2011).

This project is producing potential benefits, but also raising some challenges of both a theoretical and practical nature. For computational linguistics and corpus linguistics one challenge has involved the application of sophisticated automatic tagging and parsing techniques to learner data, much of which is errorful, thereby complicating the tagging and parsing. The error coding and corrections provided by our Cambridge University Press collaborators have enabled us to overcome these challenges. Paucity of data has also necessitated the collection of new data in electronic form. Despite this the project has initiated a novel and original application of RASP to learner data, and by tagging and parsing the BNC in the same way a systematic comparison between acquisition data at different levels and native speaker data becomes possible, as exemplified in Tables 5 and 6.

The project is also contributing new patterns and principles to the field of second language acquisition (SLA). Current theories of SLA do not make such clear empirical predictions for a large set of correlating syntactic, morpho-syntactic and lexical properties across six proficiency levels in the acquisition of English. The criterial features that we are extracting from the corpus provide a new set of patterns that we can use to inform predictive and multi-factor principles of acquisition, on the basis of which we can ultimately model the time course and progression of L2 acquisition from a given L1, in the manner of a complex adaptive system using a computer simulation (Gell-Mann 1992).

With respect to practical applications of this work, it is reasonable to assume that examiners are aware of these criterial features. Since the CEFR descriptors and 'can do' statements are underspecified with respect to them, and since there is a large measure of inter-examiner agreement, examiners have clearly acquired a sensitivity to these features when assigning their levels and scores to candidates. The features identified here make explicit just a small sample of the kinds of linguistic properties that examiners have learned to look for. Making them explicit in this way can have significant benefits for test development, for validation and examiner training. It can also lead directly to improved publishing outcomes in which not only lexical features are calibrated more closely with textbook preparation at a given level, as in the past, but many grammatical features as well. Textbooks oriented to particular worldwide markets, for example Chinese or Spanish learners of English, can also benefit from the L1-specific features identified in this research.

Acknowledgements

Abbreviations

The following abbreviations are used in this paper (excluding those that are clearly defined in the text itself):

AdjP	adjective phrase
ALTE	Association of Language Testers in Europe
AT	article
AV	adverb
CEFR	Common European Framework of Reference
CJC	conjunction

CLC	the Cambridge Learner Corpus
DT	determiner
ind RC	relative clause formed on an indirect object
infin	infinitive
JJ	adjective
LOC	locative
MD	missing determiner
NP	noun phrase
Obj	object
obl RC	relative clause formed on an oblique position
P	preposition
Part	particle
PNP	pronoun
PP	prepositional phrase
PRP	preposition
PUN	punctuation
recip	reciprocal
S	sentence/clause
Subj	subject
V	verb
VM	modal verb
VP	verb phrase
Vpastpart	past participle verb
VPinfin	infinitival verb phrase
VV	main verb
Wh	wh-word
Wh-move	Wh-movement
NN	common noun
RASP	Robust Accurate Statistical Parser

References

Andersen, E. (1992). Complexity and language acquisition: Influences on the development of morphological systems in children. In Hawkins, J. A. and Gell-Mann, M. (eds.), *The evolution of human languages*. Redwood City, CA: Addison Wesley, 241–271.

Briscoe, E. J. (2000). Dictionary and system subcategorisation code mappings. Manuscript, University of Cambridge: Computer Laboratory.

Briscoe, E. J. & Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Conference on Applied Natural Language Processing, 1997*. Morristown, NJ: Association for Computational Linguistics.

Briscoe, E., Carroll, J. and Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, Sydney, Australia*. Stroudsburg, PA: Association for Computational Linguistics. Accessible at http://acl.ldc.upenn.edu/P/P06/P06-4020.pdf.

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

Eckman, F. R. (1984). Universals, typologies, and interlanguage. In Rutherford, W. E. (ed.), *Language universals and second language acquisition*. Amsterdam: John Benjamins, 79–105.

Ek, J. A. van & Trim, J. L. M. (1991). *Threshold level 1990*. Cambridge University Press.

Gell-Mann, M. (1992). Complexity and complex adaptive systems. In Hawkins, J. A. & Gell-Mann, M. (eds.), *The evolution of human languages*. Redwood City, CA: Addison Wesley, 3–18.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press.

Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. *Language 75*, 244–285.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford 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 *Proceedings of the 3rd ALTE Conference 2008*. Cambridge University Press.

Hawkins, J. A. & Filipovic, L. (2011). Criterial features in L2 English: Specifying the reference levels of the Common European Framework. Cambridge University Press.

Hyltenstam, K. (1984). The use of typological markedness conditions as predictors in second language acquisition: The case of pronominal copies in relative clauses. In Andersen, R. W. (ed.), *Second languages: A cross-linguistic perspective*. Rowley, MA: Newbury House, 39–58.

Keenan, E. L. & Comrie, B. (1977). Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry 8*, 63–99.

Korhonen, A., Krymolowski, Y. & Briscoe, E. J. (2006). A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC, 2006*. Paris: ELDA.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *University of Cambridge ESOL Examinations Research Notes 37*, 2–5.

Preiss, J., Briscoe, E. J. & Korhonen, A. (2007). A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, June 2007, Prague. Morristown, NJ: Association for Computational Linguistics, 912–919. Available online: www.aclweb.org/anthology/P/PO7/PO7-1115.

Williams, C. A. M. (2007). A preliminary study into verbal subcategorisation frame usage in the CLC. Manuscript, University of Cambridge: RCEAL.