28 Corpus Approaches to L2 Learner Profiling Research

Yukio Tono
Tokyo University of Foreign Studies, Japan
y.tono@tufs.ac.jp

There is a growing interest in profiling L2 learners' proficiency using a common framework such as the Common European Framework of Reference for Languages (CEFR). The profiling method often especially involves a large amount of learner data with CEFR levels. Features which characterize level differences are extracted by using clustering or classification techniques used in corpus linguistics. This paper describes the three domains of L2 learner profiling research—lexical, grammatical, and textual profiles—and discusses the significance, implications and issues involved in such studies. As an example of concrete studies, a case of an English grammar profile using parallel corpora of students' original and corrected writings is reported.

Keywords: corpus linguistics, L2 learner profiling, underuse/overuse

INTRODUCTION

In corpus linguistics, the compilation of learner corpora started in the early 1990s, when 'learner English' was collected as one of the varieties of English to be contrasted against regional varieties such as British or American English. The driving force was the International Corpus of Learner English (ICLE) (Granger, 1998), which was launched as one of the subdomains of the International Corpus of English (ICE) (Greenbaum, 1996). The ICLE team collected a corpus of argumentative essays written by third-year university English-majors with different L1s. Their methodological approach was called Contrastive Interlanguage Analysis (CIA) (Granger, 1998, p.12), where two types of comparison, one between native language (NL) and interlanguage (IL) and the other between different interlanguages (IL vs. IL) were proposed in order to identify those linguistic properties which were shared by all the ILs and those which were specific to a group of learners with a particular L1.

The first two decades saw a growing number of learner corpus studies and a journal dedicated to learner corpus research was launched in 2015. The learner corpus bibliography available at the website of the Learner Corpus Association now contains more than 1,000 published papers and books on learner corpora. As the field becomes more mature, there is a growing awareness that learner corpus research needs to be further refined in terms of (a) multimodal learner production data, (b) multiple annotation perspectives, and (c) more sophisticated data analysis techniques. One recent trend of learner corpus research is to examine a large amount of L2 learner data from writers of different proficiency levels to identify linguistic features characterizing the given level of proficiency from the rest. This paper will describe basic approaches of L2 learner profiling research and introduce major profiling projects including the CEFR-J.

L2 LEARNER PROFILING RESEARCH

L2 learner profiling research aims to give a comprehensive description of what L2 learners can do with language at a given proficiency level. The motivation behind this is to supplement the description of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001, 2018). The CEFR was first proposed by the Council of Europe as a common reference framework for learning, teaching and assessing a foreign language within EU member countries, but it gradually expanded its influence outside the EU. The CEFR was based on the detailed specification of language functions and general/ specific notions, which were first proposed in the form of Threshold Levels for an individual language (van Ek and Trim, 1974; 1990). The Council of Europe recommended that the framework should be shared among all the EU member countries in order to facilitate mobility of the labor force and promote plurilingual/ pluricultural competences of each as an active social agent.

Instead of describing specific language properties denoting functions and notions, the CEFR proposed a standard set of illustrative descriptors that indicate what a person can do with a language at a given level of proficiency. The illustrative descriptors were scaled using Rasch analysis based on the teachers' responses to a series of descriptors regarding whether a particular learner those teachers had in mind could perform the tasks which were described in a given descriptor. The final version of the CEFR consists of a collection of illustrative descriptors addressing various aspects of communicative competence in the target language across the levels defined in the framework.

After the release of the CEFR in 2001, more and more learner corpus studies refer to the CEFR as a standard for level specifications. The CEFR itself went on to investigate the profile of L2 learners at each of the six common reference levels, which has become one of the primary research areas in learner corpus research. I will describe in detail three projects related to the recent development in L2 learner profiling research; the English Profile, Pearson's Global Scale of English, and the CEFR-J RLD Project.

English Profile

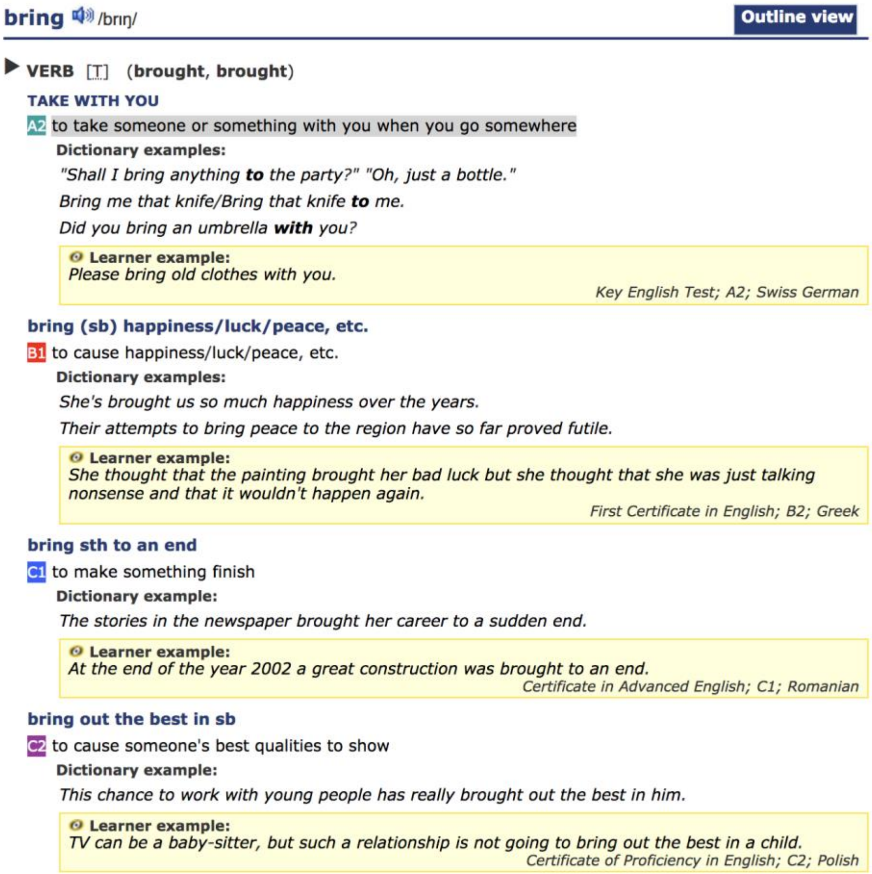
The English Profile is a project funded by Cambridge University, Cambridge English Language Assessment, the University of Bedfordshire, and the British Council. According to the official website, the English Profile Programme is the latest stage in a process dating back to the 1970s, when John Trim and Jan van Ek developed the original Threshold series, the first systematic specification of learning objectives for the English language. Initially, threshold specified objectives for a broadly intermediate level (B1 in the current CEFR) were proposed; a lower and an upper level were subsequently described, known as Waystage (A2) and Vantage (B2). Collectively known as the Threshold or T-series, they contributed to the development of the six-level scale of the Council of Europe Common European Framework of Reference (CEFR).

The English Profile aimed to identify so-called 'criterial features,' which serve as criteria for distinguishing one CEFR level from another. Criterial features can be "positive" in that the occurrence of a specific linguistic feature in the learner's performance indicates the attainment to a particular CEFR level. On the other hand, negative criterial features are primarily related to the non-target-like use of linguistic features.

Criterial features are not limited to the use/misuse of particular linguistic features only. They are also concerned with the usage distribution. For example, relative clause constructions are often underused by beginning to intermediate learners (Takahashi, 2018), but from B to C levels the entire usage distribution will become similar to that of native speakers, which indicates that the use of relative clause constructions becomes more target-like.

Researchers involved in the English Profile Programme are developing an innovative and unique methodology for describing the English language using corpus research techniques. Previous language profiles such as van Ek and Trim (1990) have been produced by language specialists largely using their insight as expert users and teachers of the language. However, English Profile's methodology is empirical, based on data provided by real learners of English, which means that it provides concrete evidence of what learners throughout the world can do at each level of the CEFR. They used two main resources for this: the Cambridge International Corpus (1.5 billion words) and the Cambridge Learner Corpus (50 million words). This corpus-based RLD project has influenced our project’s approach, which is also corpus-based.

In the English Profile programme, the Corpus Linguistics Research Team, based in RCEAL and led by John Hawkins and Henriëtte Hendriks, worked on two areas of direct relevance to English Profile: 1) a set of criterial features that characterize and distinguish the six levels of the CEFR with respect to English, and 2) the impact of different first languages on performance at each of the levels and their interaction with the criterial features (Kurtes & Saville, 2008).



bring /brɪŋ/

Outline view

▶ **VERB** [T] (**brought, brought**)

TAKE WITH YOU

A2 to take someone or something with you when you go somewhere

Dictionary examples:

*"Shall I bring anything **to** the party?" "Oh, just a bottle."*

*Bring me that knife/Bring that knife **to** me.*

*Did you bring an umbrella **with** you?*

Learner example:
Please bring old clothes with you.
Key English Test; A2; Swiss German

bring (sb) happiness/luck/peace, etc.

B1 to cause happiness/luck/peace, etc.

Dictionary examples:

She's brought us so much happiness over the years.

Their attempts to bring peace to the region have so far proved futile.

Learner example:
She thought that the painting brought her bad luck but she thought that she was just talking nonsense and that it wouldn't happen again.
First Certificate in English; B2; Greek

bring sth to an end

C1 to make something finish

Dictionary example:

The stories in the newspaper brought her career to a sudden end.

Learner example:
At the end of the year 2002 a great construction was brought to an end.
Certificate in Advanced English; C1; Romanian

bring out the best in sb

C2 to cause someone's best qualities to show

Dictionary example:

This chance to work with young people has really brought out the best in him.

Learner example:
TV can be a baby-sitter, but such a relationship is not going to bring out the best in a child.
Certificate of Proficiency in English; C2; Polish

Figure 1. The entry 'break' in the English vocabulary profile

The current English Profile website provides two useful resources online. One is the English Vocabulary Profile (EVP), which is a list of vocabulary assigned to the six CEFR levels (A1 to C2). The wordlist not only shows the CEFR level for the headwords but also for individual word senses and spoken phrases. Figure 1 shows the entry *bring* in the EVP, which tells us that the first sense "to take someone or something with you when you go somewhere" is classified as A2, whereas the phrase, *bring sth to an end*, is labelled as C1. This judgement, according to Capel (2012), is based upon the analysis of the Cambridge Learner Corpus. Thus, it is important to keep in mind that the CEFR levels indicated by EVP are based on the production of L2 learners, not their receptive knowledge.

Another important resource is the English Grammar Profile, which is a set of grammar inventories to go with each CEFR level. Hawkins and Filipović (2012) addressed the extent to which learners knew the grammar, lexicon and usage conventions of English at each level of the Common European Framework of Reference (CEFR). These levels used to be illustrated in functional terms with 'Can Do' statements in the CEFR. Greater specificity and precision can be achieved by using the tagged and parsed corpus of native and learner languages, which enables them to identify criterial features of the CEFR levels, i.e. properties that are characteristic and indicative of L2 proficiency at each CEFR level. In practical terms, once criterial features have been identified, the grammatical and lexical properties of English can be presented to learners more efficiently and in ways that are appropriate to their levels (ibid.). Table 1 shows the levels of modal auxiliary verbs specified in the English Grammar Profile.

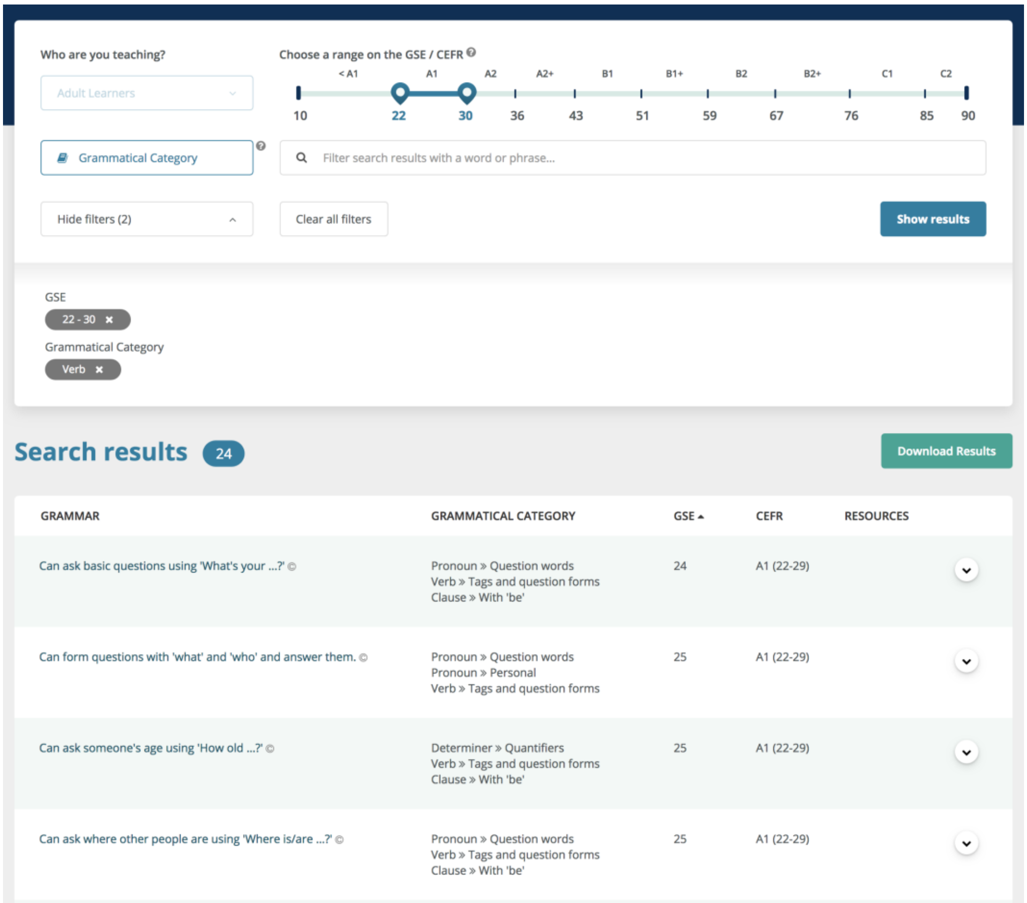
Table 1. The CEFR levels of modal auxiliaries in the English grammar profile. (Examples are partly modified)

modal verb	meaning	CEFR	Example
may	possibility	A2	Then we may go sightseeing.
may	permission	B1	May I suggest that …?
might	possibility	A2	… the paint might make our T-shirts dirty.
might	permission	C1	Might I tell you what we discuss?
can	ability/permission	A1	Can you make me a big salad?
can	possibility	A2	We can meet at our school.
must	obligation	A2	We must be there at 7 o'clock.
must	necessity	B1	She must be feeling so happy!
should	advice	A2	You should wear old clothes …
should	probability	B1	I have invited all his friends, so we should be 28 people.

Global Scale of English

Pearson has a different approach toward the scaling of illustrative descriptors and relevant vocabulary and grammar level descriptions. Their scale is called the Global Scale of English (GSE). The GSE is a standardized, granular scale which measures English language proficiency. Unlike some other frameworks which describe attainment in broad bands, the Global Scale of English identifies what a learner can do at each point on the scale across speaking, listening, reading and writing skills. The original work for the CEFR by Brian North also produced detailed theta values (θ) for each of the scaled descriptors using Rasch analysis. However, the CEFR team decided to take a broader band, A1 to C2, as a level indicator. The GSE took a different method, converting these theta values to 10 to 90 scales, thus enabling all the descriptors to be located at a certain point on the scale. They did this based on 6,000 teachers' responses on the questionnaire of learning objectives (similar to illustrative descriptors, but broader categories including grammar knowledge).

They conducted the scaling of grammar and vocabulary on the GSE in a unique way. Instead of looking at learner output, they conducted teacher surveys, in which teachers were asked whether particular grammar/ vocabulary items or expressions were necessary for given CEFR level learners. The teacher responses were then calibrated using Rasch analysis and mapped on the GSE scale. Altogether, over 450 grammar objectives, 39,000 vocabulary items, and 80,000 collocations were on the scale and can be searched by CEFR/GSE level. This online tool is called the GSE Teacher Toolkit and is freely available with registration. Figure 2 shows the results of the search for grammar items for GSE scales 22 to 30 (A1 level) with the searches limited to verbs only.



Who are you teaching?
Adult Learners

GSE
22 - 30

Grammatical Category
Verb

Search results 24

GRAMMAR	GRAMMATICAL CATEGORY	GSE ▲	CEFR	RESOURCES
Can ask basic questions using 'What's your ...?'	Pronoun » Question words Verb » Tags and question forms Clause » With 'be'	24	A1 (22-29)	
Can form questions with 'what' and 'who' and answer them.	Pronoun » Question words Pronoun » Personal Verb » Tags and question forms	25	A1 (22-29)	
Can ask someone's age using 'How old ...?'	Determiner » Quantifiers Verb » Tags and question forms Clause » With 'be'	25	A1 (22-29)	
Can ask where other people are using 'Where is/are ...?'	Pronoun » Question words Verb » Tags and question forms Clause » With 'be'	25	A1 (22-29)	

Figure 2. GSE teacher toolkit
(https://www.english.com/gse/teacher-toolkit/user/lo)

The approach taken by Pearson is a nice blend of teacher's expertise and modern psychometric analysis. Even though the results are not attested by actual language use, GSE Teacher's Toolkit provides teachers and researchers with a list of learning objectives, grammar and vocabulary items for teaching and assessment in a quite user-friendly way.

The CEFR-J Reference Level Description

As a final example of L2 learner profiling research, I will introduce my own project, the CEFR-J and its RLD work. The CEFR-J is an adapted version of the CEFR in the context of English language teaching in Japan. Based on our nation-wide survey of learning objectives for English, we decided to investigate how the CEFR should be adapted to our local context and launched the CEFR-J project in 2008. The main reason was that since the release of the CEFR in 2001, there had been a growing influence of the CEFR in many areas of language teaching, especially the common framework for comparing different language proficiency tests or proficiency levels of learners in different EFL/ESL contexts. For further details of the CEFR-J development, see Negishi, Takada & Tono (2013), Negishi & Tono (2016) and Tono (2017). Since the release of the CEFR-J in 2012, it is becoming more and more influential as a concrete example of implementing the CEFR into a local context. The Companion Volume of the CEFR (2018) adopted approximately 30 illustrative descriptors from the CEFR-J as descriptors for younger learners. Also, the GSE has a white paper on the alignment of the CEFR-J to the GSE (Mayor, et al. 2016). As a concrete example of RLDs for the CEFR-J, two projects will be described below: the CEFR-J Wordlist and the CEFR-J Grammar Profile.

The CEFR-J Wordlist

In order to develop the wordlists for the CEFR-J, a close examination was made regarding the frequency analysis of English textbooks used at primary and secondary schools in nearby Asian countries/ regions (e.g. China, Korea, and Taiwan). They were not specifically designed based on the CEFR, but we assessed the approximate CEFR levels of the textbooks by examining the learning objectives described in their national curriculums. In this way, we prepared Pre-A1 to B2 level sub-corpora, each of which comprises textbook data. In the analysis of CEFR-level textbook corpora, the texts were first tagged for parts of speech, using TreeTagger (Schmid, 1994) and then the frequency lists of lemmas with POS were created for each textbook published in each country/region as well as each CEFR level. Finally, the Pre-A1 words were determined by selecting only the words that appeared in all three regions' textbooks classified at the Pre-Al level. The A1-level words were then extracted in the same way, after subtracting all the Pre-A1 words from the texts in advance. In this way, vocabulary for each CEFR level was determined.

Interestingly, since the vocabulary growth between Pre-A1 and A1-levels was very small (only 100 words), the two levels were merged into the A1-level. Table 1 shows the breakdown of the wordlist. The 'Corpus' row indicates the initial query results of the words found across all three regions' textbooks at a given level. The third row shows our initial target number of words. Altogether we expected to have 6,000 words from the A1 to B2 levels, but after the analysis of textbook corpora, we compared our results with the English Vocabulary Profile (EVP) compiled by the English Profile team and found that while the first two levels (A1 and A2) cover a relatively homogeneous set of words, there is a larger gap in B1 and B2 level words between the two lists, so we decided to incorporate those words which are missing from our list but exist in the EVP. The row called 'Final Version' shows the number of entries in the final version of the wordlist.

Table 2. The breakdown of the CEFR-J wordlist

Level	A1	A2	B1	B2	Total
Corpus	976	1057	1884	1722	5639
Our initial target	1000	1000	2000	2000	6000
Final Version	1068	1358	2359	2785	7570

The final version of the wordlist was then annotated with the notion categories from the British Council/EAQUALS *Core Inventory for General English* and *Threshold Level* (van Ek and Trim, 1990), which enables the users to extract level-appropriate vocabulary belonging to a particular thematic category. Table 2 shows a sample list of entries from the CEFR-J Wordlist.

Table 3. The entries of the CEFR-J wordlist

Entry	CEFR level	POS	Thematic domains
activity	A1	n	Leisure activities
actor	A1	n	Work and Jobs
age	A1	n	Personal information
airplane	A1	n	Ways of travelling
airport	A1	n	Travel and services vocab
animal	A1	n	
answer	A1	n	
apple	A1	n	Food and drink
apron	A1	n	Objects and rooms

The CEFR-J Wordlist was made publicly available in 2012. One can access the wordlist at the resource page of the CEFR-J website (http://www.cefr-j.org). This wordlist will serve as one of the important resources for the CEFR-J x 27 project later on.

The English Grammar Profile

In the JSPS KAKAN project (Kiban A; No. 24242017; 2012-15), we conducted RLD research similar to previous projects, such as the English Profile or the Core Inventory. There were two reasons why we had an independent RLD project. First, the CEFR-J has many sub-levels under A1 to B2, and it was desirable to specify grammar and vocabulary to go with each sub-level. For this purpose, the resources provided by the English Profile or the Core Inventory were not sufficient. Second, past reports on RLDs did not always specify the procedure of how each item of grammar or vocabulary was assigned to a given CEFR level. Overall methods were presented but they did not make the actual data available. Thus, we had a genuine methodological interest in how to do RLDs in an objective, valid way. We aimed to be as transparent as possible throughout all the stages of RLD work and made sure that the procedure should be available as a standard for those who wish to do their own RLD research. In addition, we used corpus-based approaches similar to the English Profile, though our profiling technique was very different from theirs; thus comparison would be methodologically interesting.

In our project, identification of the CEFR levels was considered a type of classification task defined in the field of Natural Language Processing (NLP). Figure 3 illustrates this point. Basically, it involves supervised learning of features in the texts with the CEFR level information. First, a machine creates a certain model based on a set of feature vectors from training texts with some class information, such as CEFR levels. Then the model predicts a CEFR level when a new text is given.

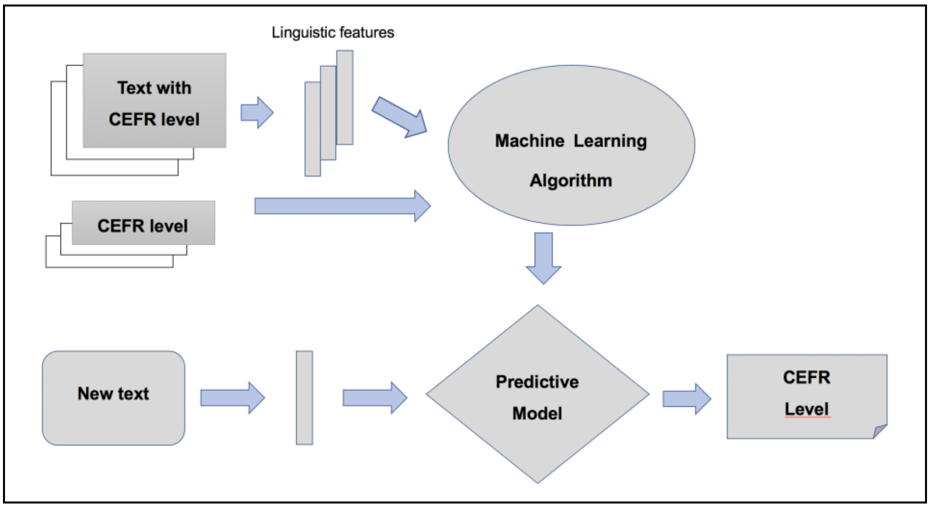


Figure 3. The supervised learning for CEFR-J RLDs

The strength of this machine learning approach is the ability to discover the relative importance of the predictive features used for the classification. In our case, this is the question for which grammatical items play an important role in classification. In the English Profile, these features are called 'Criterial Features' (Hawkins and Filipović, 2012). A feature is *criterial* when the occurrences of this feature are so prominent at the given CEFR level that it helps distinguish that CEFR level from the rest. To prove this, we need information that this feature is significantly more frequent at a given CEFR level than the others. To make matters more complicated, the CEFR level decision by humans is made not solely on a single feature but a bundle of lexical or grammatical features. Therefore, we used this machine learning algorithm not only to create a model to best predict the CEFR levels but also to select the best combination of grammatical features as predictors.

To this end, we prepared two types of corpora, ELT textbook corpora as 'input' and learner corpora as 'output'. These two types of corpora were needed in order to do RLDs for both teaching and assessment purposes. The 'input' corpus is a collection of CEFR-based course books published in the U.K. Since there is no CEFR-based English textbook published in Japan yet, course books published in the U.K. after the release of the CEFR in 2001 were collected and their content examined to see whether the textbooks were designed with appropriate CEFR levels in mind. In total, 96 textbooks were gathered, scanned with OCR, and prepared in an XML format. Each piece of textbook data in the corpus was tagged for CEFR level, section information for different skills (four skills and grammar), part-of-speech and lemma for each word. The data set (c. 1,640,000 tokens) was prepared for both normal text processing and concordancing using Sketch Engine.

The 'output' corpus consists of two sets of learner corpora: the JEFLL Corpus (Tono, 2007) and the NICT JLE Corpus (Izumi et al. 2004). The JEFLL Corpus is a collection of approximately 10,000 secondary school students' written compositions (size: 0.7 million), and the NICT JLE Corpus is a collection of oral interview test scripts by 1,280 test-takers (size: 2 million). Both sets of data were originally gathered without CEFR levels, but for this project all the sample texts were aligned to the CEFR levels.

The extraction of grammar items from the two types of corpora was mainly done by my colleague in the CEFR-J project (Ishii, 2016; Ishii and Tono, 2016). Altogether, approximately 500 grammar items were automatically extracted by using a set of pattern matching queries for each item. Frequencies and dispersion measures were obtained for each grammar category at all the CEFR levels and the matrix of [grammar category] x [each text with CEFR-levels] was used for machine learning. Several machine learning algorithms were tested, and a random forest and ranking Support Vector Machine (SVM) was used for the final analysis (Tono, 2017). The CEFR-J

Grammar Profile was released as a dataset first in March 2018, followed by a English teacher-friendly version in fall 2018.

L2 LEARNER PROFILING FOCUSING ON OVERUSE/UNDREUSE OF GRAMMAR ITEMS

So far, I have described the major projects of L2 learner profiling research in conjunction with the CEFR Reference Level Descriptions. In this final section, the research I conducted with Yasutake Ishii, the other CEFR-J project member, will be reported. The initial report of this study was published in Ishii and Tono (2018).

The corpus used in this present study is the Japanese EFL Learner (JEFLL) Corpus (Tono, 2007). It is a collection of English essays written by junior and senior high school students in Japan. The total number of essays is 10,038 and the total size of the corpus is 669,304 running words. For data elicitation, the six essay tasks were controlled in terms of text types (argumentative vs. narrative) and possible time expressions (past, present and future). The participants were asked to write an essay without the use of dictionaries in 20 minutes in class. All the compositions were hand-written.

Currently, the JEFLL Corpus has a parallel corpus version and the CEFR-based version. The former was compiled by asking native speakers to proofread all the essays so that the original and corrected versions were prepared as a parallel corpus. We also asked those who were familiar with the CEFR to re-evaluate the essays according to CEFR levels, thus producing the CEFR-based version of original vs. corrected essays. We used this version for our analysis. The total size of the original and corrected version is shown in Table 4.

Table 4. The JEFLL-CEFR corpus: subcorpus breakdown

	A1	A2	B1	B2
Original	131,525	309,561	212,158	8,658
Corrected	153,887	338,696	226,600	9,092

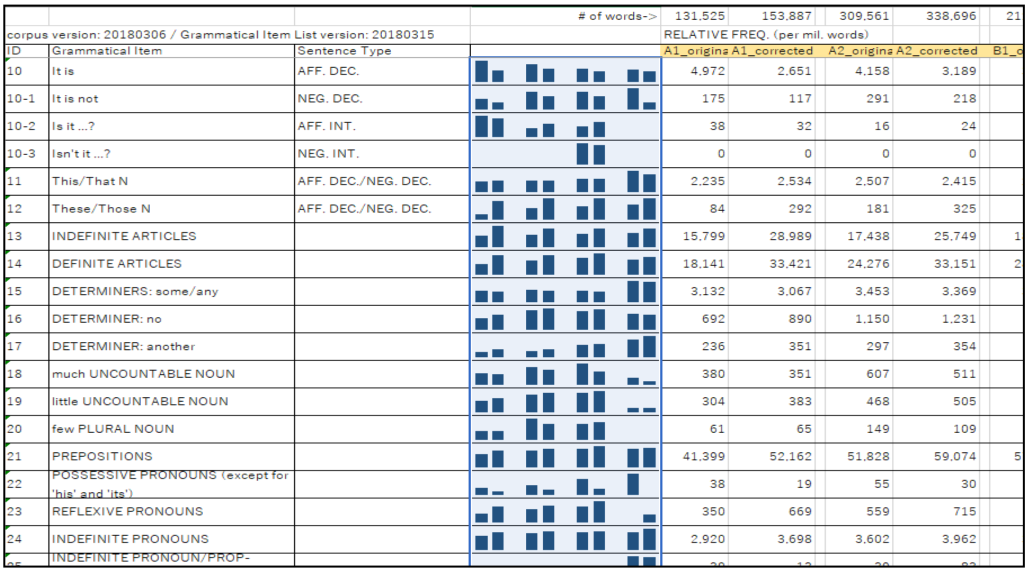
As one can see, the size of the B2-level subcorpus is much smaller than the other three, which might affect the results of less frequent grammar items. The following results provide the statistics, including the B2 level, because there is some useful information for frequent grammar items. Further research will be needed to sort out the effects of the unbalanced corpus size.

We have made query patterns for each item using regular expressions searching for combinations of word forms, lemmas and parts of speech. Table 5 shows some examples. The texts to be analyzed were processed on TreeTagger (Schmid, 1994), by which each word was morphologically analysed and displayed with its wordform, lemma and part of speech. Patterns of all 501 grammar items were automatically extracted and counted from the two sets of corpora: the original JEFLL and the corrected JEFLL, and exported into a CSV file.

Table 5. Part of the items adopted in the CEFR-J grammar profile

ID	Item	Pattern
26	INDEFINITE PRONOUN: none	\bnone_NN_none\b
49	COMPARATIVE and COMPARATIVE (the same adjective)	\b(\S+_(JJR|RBR)_\S+) and_CC_and \1
66	TENSE/ASPECT: PAST PROGRESSIVE (AFFIRMATIVE DECLARATIVE)	(was|were)_VBD_be(?! (going_VVG_go to_TO_to|gonna_VVG_gonna) \S+_V._\S+) \S+_V.G_\S+
145	AUX+PERFECT (AFFIRMATIVE DECLARATIVE)	(?!cannot\b)\S+_MD_\S+ have_VH_have \S+_V.N_\S+

To analyze how grammar categories are used by Japanese EFL learners, we compiled a frequency table, part of which is shown in Figure 4. The resulting data reveal which grammar categories are frequently or infrequently used in the learners' writings and their proofread versions.



			# of words->	131,525	153,887	309,561	338,696	21
corpus version: 20180306 / Grammatical Item List version: 20180315				RELATIVE FREQ. (per mil. words)				
ID	Grammatical Item	Sentence Type		A1_origina	A1_corrected	A2_origina	A2_corrected	B1_c
10	It is	AFF. DEC.		4,972	2,651	4,158	3,189	
10-1	It is not	NEG. DEC.		175	117	291	218	
10-2	Is it ...?	AFF. INT.		38	32	16	24	
10-3	Isn't it ...?	NEG. INT.		0	0	0	0	
11	This/That N	AFF. DEC./NEG. DEC.		2,235	2,534	2,507	2,415	
12	These/Those N	AFF. DEC./NEG. DEC.		84	292	181	325	
13	INDEFINITE ARTICLES			15,799	28,989	17,438	25,749	1
14	DEFINITE ARTICLES			18,141	33,421	24,276	33,151	2
15	DETERMINERS: some/any			3,132	3,067	3,453	3,369	
16	DETERMINER: no			692	890	1,150	1,231	
17	DETERMINER: another			236	351	297	354	
18	much UNCOUNTABLE NOUN			380	351	607	511	
19	little UNCOUNTABLE NOUN			304	383	468	505	
20	few PLURAL NOUN			61	65	149	109	
21	PREPOSITIONS			41,399	52,162	51,828	59,074	5
22	POSSESSIVE PRONOUNS (except for 'his' and 'its')			38	19	55	30	
23	REFLEXIVE PRONOUNS			350	669	559	715	
24	INDEFINITE PRONOUNS			2,920	3,698	3,602	3,962	
25	INDEFINITE PRONOUN/PROP-							

Figure 4. The comparison table of grammar items in the original and corrected JEFLL

The grammar items focused on in this study were chosen based on their frequencies in the learners' original writings, whereby out of 501 grammar items in our grammar profile, 193 items with the raw frequency of 20 or over were selected. The analysis was based upon the overall increase and decrease of the given grammar items that occurred in the essays across different CEFR levels. The frequency of each grammar item was defined as the total instances of that particular grammar item in the JEFLL-CEFR subcorpus. The overuse/underuse was determined by calculating the ratio of the number of students' original uses over that of native speakers' corrections. For example, take the case of 'I am ...' in Table 6:

Table 6. The distribution of 'I am ...' across CEFR levels between the original and corrected versions of JEFLL

I am ...	A1	A2	B1	B2
original	4,858	3,049	2,512	2,888
corrected	3,574	2,979	2,573	3,410
Ratio	1.36: 1	1.02: 1	0.98: 1	0.85: 1

At the A1-level, for example, 4,858 occurrences of '*I am ...*' were observed in the students' original writings, whereas in the corrected version, only 3,574 cases were found. It means that after native speakers' corrections, about a half of the use of '*I am ...*' was changed to some other constructions. Here the ratio of the original essays over the corrected ones was 1.36 :1, which shows that A1-users tend to overuse '*I am ...*' 136%, compared to the native speakers' corrected version. This overuse tendency gradually decreases as the CEFR level increases. At the B2 level, for instance, the original essays only contained 2,888 cases of '*I am ...*' compared to 3,410 cases in the proofread essays. Thus, the original to corrected ratio was 0.84 to 1, which means compared to the corrected version, 84% of the items occurred at the B2 level.

Table 7 shows the number of grammar items that are underused compared to the corrected version of the essays:

Table 7. The number of grammar items underused

Underuse	A1	A2	B1	B2
0 to 49%	56	16	12	59
50 to 74%	43	98	46	26
75% & above	94	79	135	108
Total	193	193	193	193

The results show that at A1-level, there were 56 grammar items that were used less than 50% compared to the corrected version, but this rate rapidly decreased to 16 at A2 and 12 at B1 respectively. The figures for B2-level increased again but this may not be very accurate due to the lack of B2-level data, which resulted in many missing grammar items. Overall, the underuse phenomena gradually disappear as the level goes up.

The ten most underused grammar items in the JEFLL Corpus are shown in Table 8.

Table 8. The ten most underused grammar items in the JEFLL corpus

Lexical forms	A1	A2	B1	B2	Average
being + PAST PARTICIPLE	0.000	0.256	0.334	0.000	0.147
PREP+RELATIVE PRONOUN	0.090	0.283	0.226	0.000	0.150
TENSE/ASPECT: PRESENT PERFECT	0.268	0.398	0.415	0.000	0.270
RELATIVE PRONOUN: NONRESTRICTIVE	0.047	0.322	0.303	0.525	0.299
MODAL/AUX: *should*	0.000	0.547	0.691	0.000	0.310
RECIPROCAL PRONOUN: *each other*	0.731	0.421	0.115	0.000	0.317
TENSE/ASPECT: PRESENT PERFECT PROGRESSIVE	0.146	0.505	0.668	0.000	0.330
whether	0.351	0.469	0.534	0.000	0.338
WH- QUESTION: *When ...?*	0.585	0.377	0.420	0.000	0.345
PASSIVE: PAST PERFECT	0.123	0.729	0.551	0.000	0.351

It is worth noting that many items involve the combination of more than one grammar item and demand a quite heavy processing load. For instance, the most underused item is the construction “being + PAST PARTICIPLE,” which is a complex combination of progressive aspect and participle construction. The original essays used this construction with the following frequencies: 0 (A1), 36 (A2), 71 (B1), 0 (B2), whereas in the corrected essays it occurred 201 (A1), 139 (A2), 212 (B1), 330 (B2) times. The correlation between the two essays was r = -0.434, which means the usage pattern shows a negative correlation. Native speakers use this pattern in corrected essays, whereas learners constantly avoided it. The same kind of tendency is observed in the case of the second most underused item, “PREP+RELATIVE PRONOUN.” This construction involves the raising of prepositions together with relative pronouns with oblique cases, such as “*of which*” or “*in which.*” The knowledge of appropriate choice of prepositions and relative pronouns poses a problem again, which leads to constant underuse by the learners, while the native speakers use this construction much more frequently, as shown in the corrected essays.

Such combinations of grammar items can be seen in other items, such as PRESENT PERFECT PROGRESSIVE (present prefect+progressive), PASSIVE: PAST PERFECT (passive+past perfect), PASSIVE: FUTURE (passive+future), and AUX+PERFECT (auxiliary verb+perfect). One of the reasons for underuse of these items may be that the structure itself is very complex, involving more than one grammar item, which gives learners a heavy processing burden and leads to underuse phenomena.

Another possible factor of underuse is related to how the item is taught. For example, there is another type of relative pronoun, ranked at No.4, which is a non-restrictive relative clause. This construction is usually introduced in a syllabus after teaching a set of restrictive relative clauses, but the treatment of non-restrictive relative clauses in class seems to be marginalized and is not very systematic. Despite the fact that this construction is quite frequently used by native speakers, Japanese EFL learners find it difficult to use in writing.

Some grammar items might involve difficulty in acquiring due to functional-semantic problems. For instance, there are many TENSE/ASPECT categories such as present perfect, present perfect progressive, and past perfect, ranked in the top ten underused items. These constructions are usually introduced around the A2-level, so it is natural that A1-level essays do not include them. There is a tendency for learners to start using the constructions from the A2 to B1 levels, but still the rate of using the constructions was consistently lower than the native speakers' corrected version of the essays. These tense/aspect markers are known to be problematic for L2 learners due to the gap regarding how to express tense/aspect between L1 and L2 (cf. Bardovi-Harlig, 2000).

There are not as many grammar items which are constantly overused across the CEFR levels. Table 9 shows some of those items.

Table 9. The most overused grammar items in the JEFLL corpus

Lexical forms	A1	A2	B1	B2	Average
These/Those are ...	5.265	2.188	2.594	1.050	2.774
he/she is ...	3.385	1.647	1.617	1.470	2.030
there + be ...	1.404	1.810	1.978	2.100	1.823
It is not ...	1.495	1.331	1.308	3.150	1.821
ELLIPTICAL ACCUSATIVE RELATIVE PRONOUN	1.209	1.245	1.434	2.071	1.489
much + UNCOUNTABLE NOUN	1.083	1.189	1.576	2.100	1.487
MODAL/AUX: *would*	0.439	0.574	0.518	4.201	1.433
It is ...	1.875	1.304	1.244	1.187	1.403
TENSE/ASPECT: PRESENT (BE)	1.791	1.320	1.248	1.214	1.393
TENSE/ASPECT: PRESENT (BE): NEGATIVE	1.390	1.346	1.258	1.470	1.366

Many overused grammar items involve the use of *be*-verb, as in "*These* [*Those*] *are* ...", "*he* [*she*] *is* ...", "*there is* [*are*] ...", "*It is not* ...", "*It is* ...", and present tense "*be*". Especially the distinction of copula *be* and lexical verbs is confusing to Japanese learners of English. The Japanese language has a topic-comment structure, which is often confused with a subject-predicate construction in English (e.g. *Boku wa* [TOPIC: 'as for me'] *Ramen da* [COMMENT: 'it's ramen']. vs. *Kare wa* [SUBJECT: 'He is'] *daigakusei da* [PREDICATE: 'a college student'].). The writings in the JEFLL Corpus at A-level contain many incorrect choices of the *be*-verb directly mapped from Japanese TOPIC-COMMENT structures, which actually should have been expressed with different lexical verbs in English.

CONCLUSION

In this paper, a general introduction to L2 learner profiling research was given with a special emphasis on the CEFR level descriptions. After describing major research projects, the RLD project by the CEFR-J team was presented in detail. The use of parallel corpora of the original students' essays and their proofread versions will especially reveal some interesting patterns of underuse and overuse of English grammar items. Analysis of the most underused items suggests important pedagogical implications. First, the underused constructions often involve a complex combination of grammar items which are usually introduced one at a time throughout the course, but as the CEFR levels go up, learners are supposed to produce complex sentences by combining two or more constructions at the same time. Unless teachers are aware of those complex, underused items, learners may not have sufficient knowledge or opportunities to use those items. Hawkins and Filipović (2012) also pointed out a similar case of combinations of different grammar knowledge, such as prepositions followed by verb-*ing* forms (gerund).

Another interesting finding is that many overused grammar items are related to the use of copula *be*, and for Japanese learners of English, the use of the copula is quite problematical in a cross-linguistic sense. The function-form mapping involving subject-predicate vs. topic-comment structures is extremely complicated, and Japanese A-level users produced errors related to mapping those two functions into proper constructions. This has much to do with English and Japanese verb semantics and their alternation patterns like the ones proposed by Beth Levin (1993). It is inspiring that a bird's-eye-view comparison between the original and the corrected essays in terms of grammar item usage clearly shows those tendencies. Pedagogically, teachers should keep in mind that some of those overused items need special attention as learners attempt to use them in output activities.

Last but not least, the approach described in this paper is made possible by the methodological innovation in corpus linguistics and natural language processing. Integrating quantitative corpus linguistics techniques into pedagogical dimensions of foreign language learning and teaching research will shed new light on possible innovations in materials and syllabus design based on empirical findings from L2 learner corpora. L2 learner profiling research in conjunction with the CEFR will be a driving force for this rigorous new field of inquiry.

REFERENCES

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use*. Malden, MA: Wiley-Blackwell.

Capel, A. (2012). Completing the English vocabulary profile: C1 an C2 vocabulary. *English Profile Journal, 3*. https://doi.org/10.1017/S2041536212000013

Council of Europe (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Council of Europe (2018). *The Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volumes with new descriptors.* Strasbourg: Council of Europe.

Granger, S. (Ed.). (1998). *Learner English on computer*. Addison-Wesley Longman.

Greenbaum, S. (Ed.). (1996). *Comparing English worldwide: The international corpus of English*. Oxford: Clarendon Press Oxford.

Hawking. J., & Filipovic, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*. Cambridge: Cambridge University Press.

Ishii, Y., & Tono, Y. (2018). Investigating Japanese EFL learners' overuse/underuse of English grammar categories and their relevance to CEFR levels. *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference 2018* (pp. 160-165).

Kurtes, S., & Saville, N. (2008). The English profile programme: An overview. *Cambridge ESOL: Research Notes, 33*, 2-4.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* Chicago: University of Chicago Press.

Mayor, M., Seo, D., de Jong, J., & Buckland, S. (2016). *Technical report: Aligning CEFR-J descriptors to GSE.* Pearson. Accessed on 29 September, 2018 at https://online.flippingbook.com/view/220811/2/.

Negishi, M., Takada, T., & Tono, Y. (2013). A progress report on the development of the CEFR-J. In E. D. Galaczi & J. W. Cyril (Eds.), *Exploring language frameworks, Proceedings of the ALTE Krakow Conference,* July 2001, 135-163.

Negishi, M., & Tono, Y. (2016). An update on the CEFR-J project and its impact on English language education in Japan. *Studies in Language Testing*, *44*, 113-133.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, 45–49.

Takahashi, Y. (2018). A corpus-based study on relative clause constructions: CEFR criterial features and error analysis. *English Corpus Studies*, *25*, 57-78.

Tono, Y. (Ed.). (2007). *Chukousei 1-man nin no eigo corpus: The JEFLL corpus* [A corpus of 10,000 Japanese secondary school students' writings: The JEFLL corpus]. Tokyo: Shogakukan.

Tono, Y. (2017). The CEFR-J and its impact on English language teaching in Japan. *JACET International Convention Selected Papers*, Volume 4, pp. 31-52. JACET.

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