Annual Review of Applied Linguistics (2025), 1–22
doi:10.1017/S0267190525100093

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

NLP-powered quantitative verification of the English Grammar Profile's structure-level assignment

Daniela Verratti-Souto, Nelly Sagirov and Xiaobin Chen

Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany
Corresponding author: Daniela Verratti-Souto; Email: daniverratti@gmail.com

Abstract
Since its inception, the Common European Framework of Reference (CEFR) has become increasingly influential in the field of second language (L2) education. In an effort to define the grammatical structures that English learners acquire at each CEFR level, the English Grammar Profile (EGP) provides a list of over 1,200 structure-level mappings derived from largely manual analysis of learner corpora. Though highly valuable for the design of didactic materials and examinations, the EGP lacks comprehensive quantitative methods to verify the acquisition levels it proposes for the grammatical structures. This paper presents an approach for revisiting the EGP structure-level mappings with empirical statistics. The approach utilizes automatic grammatical construction extraction, a large learner corpus, and statistical testing to empirically determine the level of each structure. The structure-level mappings resulting from our approach show limited agreement with that of the original EGP proposals, suggesting that frequency data alone does not provide enough evidence for the acquisition of the grammatical structures at the levels presented by the EGP.

Keywords: English Grammar Profile; CEFR; EFL; L2 grammar acquisition; automatic grammar structure extraction

The Common European Framework of Reference for Languages (Council of Europe, 2001) has become ubiquitous in the field of second language proficiency assessment, both across and outside of Europe (Harrison, 2015a; North, 2009). It defines six levels of language competence: A1 (Breakthrough), A2 (Waystage), B1 (Threshold), B2 (Vantage), C1 (Effective Operational Proficiency), and C2 (Mastery). Each of these is characterized in terms of general descriptors, outlining communicative functions that learners are able to carry out at each stage of L2 development. Due to the language- and

theory-independent nature of the CEFR, these descriptors do not make reference to specific grammar or vocabulary that is used by learners of the target language to perform said functions. It is thus up to national and regional teams of experts to specify how these communicative functions manifest in the unique characteristics of their respective languages. Such specifications are often referred to as Reference Language Descriptions (RLDs).

Consequently, with the goal of concretizing the CEFR levels for English, Cambridge University Press and Cambridge English Language Assessment, along with other institutions, initiated the English Profile Programme (EPP, https://www.englishprofile.org/) as a project that aims to develop RLDs for English. Since then, much work has been done to characterize learners' proficiency in English at each level in terms of grammar and lexis (see Capel, 2015; Harrison, 2015b; Saville & Hawkey, 2010). The English Grammar Profile (EGP) is a subproject of the EPP whose goal is to identify the grammatical skills of learners at each CEFR level. Within this subproject, over 1,200 grammatical constructions (henceforth, "EGP structures" or simply "structures") have been derived from real-life student texts and mapped to one of the six levels.

Given the implications of the EPP not only for second language acquisition (SLA) research, but also for English Language Teaching (ELT), the design of didactic materials and of high-stakes assessment, it is of great importance that the assignment of CEFR levels to itemized descriptions of grammatical knowledge provided by the EGP accurately reflects the actual abilities of learners at the corresponding stage of language development. The original assignment of CEFR levels to each EGP structure was made based on a combination of the statistical analysis of learner corpora and careful qualitative insights of SLA experts (O'Keeffe & Mark, 2017). Although the knowledge and experience of ELT experts is instrumental for the design of the EGP, a robust quantitative approach to determine the EGP structure levels might provide extra evidence for the mappings between grammatical structures and CEFR levels. To the best of our knowledge, no such quantitative study of the EGP has been carried out to date.

In the present study, we aimed to evaluate the level assignment of the EGP by utilizing an in-house Natural Language Processing (NLP) system, Pedagogically-Oriented Language Knowledge Extractor (POLKE), which is designed to annotate the use of EGP structures in plain text. As a tool that parses natural language to match spans of text to EGP structures, POLKE uses artificial intelligence in a broad sense, since it relies on rule-based algorithms for processing learners' written production.

We analyzed a subcorpus of the EFCAMDAT (Geertzen et al., 2014), a large-scale learner corpus consisting of texts written by English as a Foreign Language (EFL) and English as a Second Language (ESL) learners across various proficiency levels. We examined the level at which each EGP structure started being used significantly more frequently than at previous levels and, in doing so, attempted to empirically determine whether the original EGP structure-level assignment aligned with what we could observe from learner data.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

Literature review

Describing proficiency across languages: The CEFR

Describing learner language at different stages of acquisition is a central goal in SLA research (Corder, 1975; James, 1990). The CEFR addresses the need for a standard description of the communicative functions that learners of European languages are able to perform with varying degrees of accuracy and sophistication in the target language. Therefore, the purpose of the CEFR is to lay out a common basis for describing proficiency across genetically and typologically diverse languages, thus facilitating comparability in language assessment and teaching in Europe and beyond. The 2001 document published by the Council of Europe and subsequent revisions (North & Piccardo, 2019) define numerous functional descriptors, phrased as "can-do statements," which provide detailed explanations of the communicative abilities of language learners on different scales, across a range of domains and contexts, and characterize said competences as typical of one of the CEFR levels.

The CEFR has thus been greatly influential in SLA research focusing on proficiency development. Many projects have aimed to observe how different aspects of language use evolve as learners progress along the six levels. For example, in relation to the acquisition of syntactic patterns, Römer and Berger (2019) investigate the usage of verb-argument constructions by Spanish and German learners of English as they progress through the levels A1 to C1. Huang et al. (2023) instead observe discourse markers in spoken data and analyze how their use relates to CEFR proficiency levels, fluency, and immersion. In general, it is crucial to highlight the breadth and variety of SLA research based on the CEFR.

As a result of the diverse nature of the languages to which the CEFR might be applied, the guidelines it provides, though relatively comprehensive, are deliberately vague and underspecified in terms of the grammar and lexis used to carry out the functions outlined in the descriptors. The Council of Europe encourages teams of language and teaching experts to develop RLDs: concrete, language-specific criteria to adapt the framework to regional and national languages.

RLDs for English

Official research into RLDs for English has been carried out by Cambridge University Press and Cambridge English Language Assessment, who, along with other associated institutions, launched the EPP. EPP research takes van Ek and Trim's Threshold series (van Ek & Trim, 1991a, 1991b, 2001), a systematic specification of communicative objectives and of the grammar and lexis that may be used to fulfill them, as a starting point. This series of publications, also known as the T-series, was based on the expert knowledge of English teachers. However, resources such as the now-available large, searchable learner language corpora like the Cambridge Learner Corpus (CLC; Nicholls, 2003) were not taken into consideration.

Two EPP subprojects, (1) identifying criterial features (Hawkins & Buttery, 2010) and (2) the English Grammar Profile (O'Keeffe & Mark, 2017), focused on grammar development (Harrison, 2015a). Hence, they are of direct relevance to our present work. Both of these projects made use of authentic learner data from the CLC to

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

look into patterns in the acquisition of L2 English grammar. The corpus contains texts produced by learners in the writing section of Cambridge English exams, along with metadata about the exam candidates (first language, gender, level of education, etc.), year of exam, exam performance, and task information. Moreover, over half of the texts have been error-coded by human annotators. However, the CLC is not publicly available.

There has also been work on other projects with goals similar to the EPP. For example, The CEFR-J (Tono, 2017) is a framework for language teaching in Japan based on the CEFR and accompanied by RLDs for grammar, vocabulary, and text properties. Additionally, the Council of Europe's website lists the Core Inventory for General English (CIGE) (North et al., 2010) as a "simplified content specification for English" (Council of Europe, n.d.). A validation study of the CIGE was carried out by Jones (2015), in which it was found that only 47% of the analyzed items presented sufficient evidence of being acquired at the CEFR level that the Inventory assigned to them. The author explains that the low accuracy was partly due to the inability of their operationalization to classify an item as A1 given a lack of texts of that level in their dataset.

Criterial features

Hawkins and Buttery (2010) define the term "criterial features" as linguistic features that make it possible to discriminate between L2 competence levels. Although they specifically deal with lexical, syntactic, and morphosyntactic features, the authors mention that future research should involve the identification of phonological and semantic criterial features, as well as deal with form-function correspondences.

The authors propose a classification of criterial features as positive or negative, depending on whether they approximate the typical production of first language (L1) users or not, respectively. Furthermore, criterial features can be analyzed as properties – namely, as features that are either present or absent from writings of L1 speakers – or as usage distributions of correct features, in which case, the analysis focuses on the frequency of a structure as used by learners in comparison to general L1 corpora. Thus, when learners are found to use a linguistic property roughly as often as native speakers, they are said to present a positive usage distribution of the feature, whereas a negative usage distribution of a feature means that L2 speakers at a given level significantly overuse the feature or produce it considerably less frequently in comparison to native speakers. Finally, ranges of error frequencies within certain structures can also be considered criterial for different levels.

The English Grammar Profile

While the research into criterial features aims to describe general features of learner language, including error frequencies and frequencies of use of grammatical constructions, the EGP focuses on identifying which CEFR level corresponds to the acquisition of each structure. For their grammatical inventory, O'Keeffe and Mark (2017, p. 466) utilized the "ELT canon of grammatical structures" as a starting point. They defined this as the grammatical items that are standard in EFL/ESL syllabuses.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

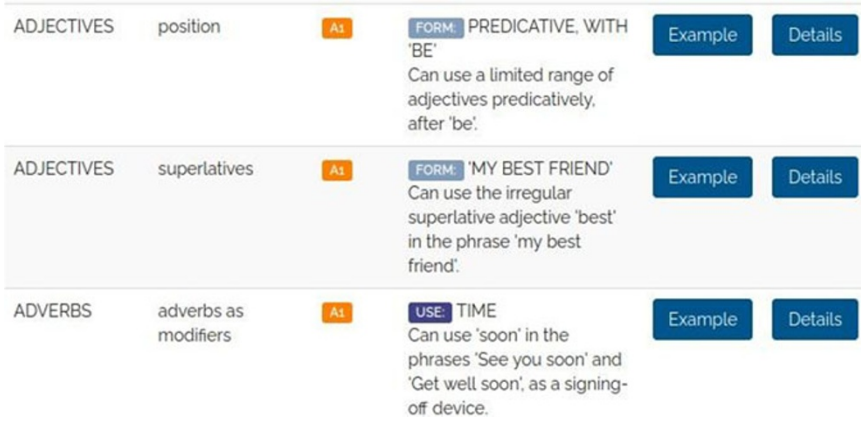

Figure 1. Example EGP form- and use-based descriptors for the A1 level.

At the time of publication of O'Keeffe and Mark's (2017) article, the CLC contained over 55.5 million words obtained from texts produced by learners taking the writing section of Cambridge English exams, consisting of open-ended tasks. They made use of the text metadata to determine whether a structure was used by a wide enough variety of learners at a certain CEFR level. Moreover, the error-coded subcorpus allowed the researchers to calculate the rate of correct uses of each construction. Based on this information, they designed over 1,200 descriptors of grammatical structures, each of them mapped to the CEFR level where it was deemed to be acquired. Figure 1 shows three examples of A1 structures taken from the EGP website (https://www.englishprofile.org/english-grammar-profile/egp-online). Two are form-based, that is, grammatical, and one is use-based, that is, focused on pragmatics.

The methodology for the development of the EGP was based on an iterative process wherein the researchers identified a range of uses of each grammatical structure from the ELT canon, queried the corpus for texts of a given CEFR level, inspected the results to determine whether they met the criteria for considering this specific use of the grammatical structure as "acquired" and, if so, they formulated a "can-do statement" for it. If the criteria were not fulfilled, the same process was repeated for the next CEFR level with the same use of the grammatical structure under analysis. In order for a structure to be considered acquired at a given level, the researchers applied the following criteria: a frequency of use above that in the British National Corpus, a 60% rate of syntactically and pragmatically correct uses, usage by a wide range of individuals from different L1 language families, and occurrences across a variety of contexts and tasks.

The process described above is painstaking and time-consuming, as well as partly reliant on manual analysis. While multi-layered work with qualitative components such as the one described above is necessary, it is very costly and difficult to replicate. As a result, potential issues in the EGP-level assignments are likely to go unnoticed, and such faults might have important implications when insights drawn from the EPP are used to inform decisions in the design of didactic materials and examinations. However, a fully quantitative approach like the one we present in this paper is not

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

without issues: the conclusions drawn from such methods might overlook nuances in the data, such as patterns that can only be identified upon careful inspection of the learner texts. Thus, biases in the corpus due to learning tasks, curricula, or formulaic use of some grammar structures may distort the findings. Therefore, our proposed fully quantitative approach, built on top of the EGP work, aimed to provide insights that are complementary to those obtained by O'Keeffe and Mark (2017).

The use of the CLC poses an issue for replicability, since the data are not freely available to all researchers. Using more automatic methods for EGP research, potentially with new, large datasets, would supplement the existing results, as well as possibly allow for future robust and reliable analyses in different directions. In Green's outline of the work required for developing RLDs for English, he concludes that "the production and validation of […] specifications is probably a necessary step towards more adequate tools in the future" (Green, 2010, p. 16). The present work is, to our knowledge, the first attempt to use a purely quantitative approach for such a validation.

CEFR level assessment: Complexity *vs.* RLDs

The Council of Europe's website states that RLDs are inventories of specific languages' grammatical rules and words which align with the CEFR descriptors in that they are instrumental for carrying out the corresponding communicative functions for a given level of the framework (Council of Europe, n.d.). The Council of Europe provides some guidelines for developing RLDs for regional and national languages (Council of Europe, 2005) and, as of now, it has approved RLDs for Croatian, Czech, English, French, Georgian, German, Italian, Latvian, Portuguese, Spanish, and Turkish.

Alternatively to RLDs, a commonly used proxy for learner knowledge is linguistic complexity (e.g., Bulté & Housen, 2012; Ortega, 2003). R. Ellis (2003) defines complexity as a measure of variedness and elaborateness of language production. Although countless complexity measures have been developed and often shown to be indicative of a learner's proficiency (e.g., Brezina & Pallotti, 2019; Casal & Lee, 2019; Kim, 2014), they are not pedagogically actionable; that is, they offer insights into language development that are too abstract for teachers to use in guiding students towards improvement. Conversely, RLDs describe L2 proficiency in a pedagogically grounded fashion, leveraging insights from SLA research and teaching tradition. Therefore, unlike complexity measures, they are easily interpreted and have direct applications in the language classroom.

In sum, previous research has resulted in the EGP, which consists of a comprehensive list of grammar structures assigned to the full spectrum of CEFR levels, making the CEFR framework operationalizable and measurable for English from the grammatical knowledge perspective. The EGP is a collection of RLDs for English. Like all other RLDs, it is required that their level assignments be validated (Green, 2010), especially in a quantitative and replicable manner. However, such validation is still missing to date, despite the EGP's wide adoption for English curriculum design and assessment. Our research attempts to fill this gap by asking the following research questions:

1. How can the EGP's grammar structure levels be determined in a fully quantitative way with large-scale learner data?

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

Table 1. Englishtown levels and their CEFR and Cambridge English counterparts

Englishtown level	1–3	4–6	7–9	10–12	13–15	16
CEFR	A1	A2	B1	B2	C1	C2

2. To what extent does a fully quantitative approach to the EGP grammar structure-level assignment agree with the original EGP structure-level mapping? In other words, can the EGP's structure-level mappings be validated with quantitative data from a large-scale learner corpus?

Methodology

Data

The data for this study were taken from the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2014). The corpus, compiled by the University of Cambridge in collaboration with EF Education First, comprises writing assignments completed by language learners enrolled in *Englishtown*, EF's online English school.

The full trajectory in *Englishtown* covers a total of 16 levels, aligned with the CEFR levels as shown in Table 1 (adapted from Geertzen et al., 2014). This alignment is instrumental for carrying out research into the abilities of learners at different proficiency levels given the widespread use of the CEFR and limited availability of other large, CEFR-tagged corpora, such as the CLC. As a result, the EFCAMDAT corpus has been used in numerous studies (e.g., Derkach & Alexopoulou, 2024; Murakami & Alexopoulou, 2016; Römer & Berger, 2019).

In the version of the corpus used for this study, every *Englishtown* level contains eight units, each culminating in a writing task which students are required to complete. Students can qualify for a level by either being assigned to it after a placement test, or by successfully completing all units of the previous level. When a student is assigned to an *Englishtown* level, they start from the first unit and work their way up. Should a student receive a failing grade in a writing task, they must retake it before moving on to further units (Alexopoulou et al., 2015).

Given that our analysis requires the comparison of normalized structure frequencies in student writings across CEFR levels (see section "Assigning CEFR levels to EGP structures"), this study is based on a subset of texts taken from the EFCAMDAT dataset. At each CEFR level, we determined the number of students who had completed a text for each of the relevant units. This resulted in a heavily unbalanced dataset, with many more beginner-level learners than advanced ones. C1 had the fewest students who met our criteria, with a total of 49 learners. To ensure a balanced sample at each proficiency level, we randomly sampled 49 students from those who had completed all units of the CEFR level. We acknowledge that this is a limited sample, much smaller than that of the CLC. However, our approach relied on comparing usage frequencies across learners at several levels, and drastically different student numbers across levels might have undermined our results. Nevertheless, a pilot study using a similar method on a much larger, yet highly unbalanced, subset of EFCAMDAT was also conducted (Verratti-Souto, 2024), yielding results that are comparable to the current study.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

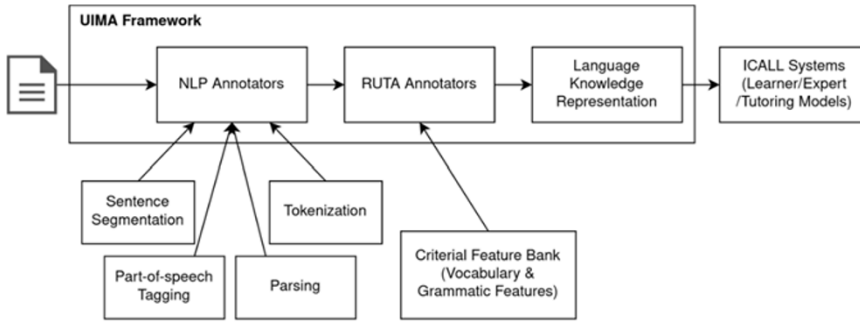

Figure 2. Obtaining L2 knowledge representations with the UIMA framework.

Automatic extraction of EGP structures

To evaluate the EGP's structure-level assignment, we wanted to see when each structure started to be used significantly more frequently in a CEFR level than in the previous level. For example, when a structure was used significantly more at the B1 level than at the A1/A2 levels, we operationalized B1 as the level corresponding to that structure. We acknowledge that this is a simplified operationalization because usage (or frequency of usage) can also be affected by many other factors, including tasks, contexts, genres, formulaic language production, etc. Nonetheless, it can also be argued that usage, especially when it is observed from large-scale learner corpora of real-life language courses, is a reflection of learning, or at least readiness to learn, because ecologically valid language courses with millions of participants should contain a variety of tasks that train learners on writing of different genres and follow reasonable curricula.

We made use of POLKE (currently available at https://polke.kibi.group/extractor.html), an in-house system that has been developed to model L2 learner knowledge by annotating the grammar structures used in language production. More specifically, the system is capable of annotating the full list of form-based grammar structures from the EGP. With this tool, we automatically extracted EGP structures from the learner texts.

Automatic criterial feature extraction is a complex issue owing to the large number of structures that need to be identified and to the challenges that working with natural language entails, including, for instance, distinct grammatical structures with identical surface forms, structural ambiguity, among others. To deal with such complexities, previous studies (Meurers et al., 2010; Quixal et al., 2021) have made use of the Unstructured Information Management framework (UIMA; Ferrucci & Lally, 2004) for NLP annotation, and Rule-based Text Annotation (RUTA) rules (Kluegl et al., 2016) for identifying the target structures (see, for example, Chen et al., 2021, for automatic extraction of subordinate clauses). A visual representation of POLKE's technological framework is provided in Figure 2.

RUTA rules make it possible to access the output from the upstream UIMA NLP annotations at the word, phrase, clause, and sentence level and to use this information to match EGP structures by designing grammar rules (Quixal et al., 2021). POLKE's

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

Table 2. Distribution of extractable structures across CEFR levels

EGP structure level	Number of extractable structures	%
A1	89	13.20
A2	193	29.29
B1	175	26.56
B2	120	18.21
C1	45	6.83
C2	39	5.92

output is provided in the JavaScript Object Notation (JSON) format as a list of structures that have been found in the input string. Each list item contains the ID of the grammatical structure, along with the beginning and ending indices that span the part of the text where the structure was identified.

The technical framework of POLKE draws heavily on Quixal et al.'s (2021) work. Currently, rules have been implemented for recognizing 659 EGP structures, which include all the form-based structures but no use-based structures, since the latter are either difficult to implement using a rule-based approach or too ambiguous. A full list of the EGP form-based structures extractable with POLKE and their IDs can be found at https://polke.kibi.group. The extractable structures are distributed across CEFR levels as reported in Table 2.

A validation study of POLKE has been carried out, with the manuscript currently in preparation. Fifteen structures per CEFR level, for a total of 90, were randomly selected from the total structures POLKE can recognize. The validation dataset was compiled from the Corpus of Contemporary American English (COCA; Davies, 2008—). For each structure, 100 sentences were manually sampled from COCA, 50 of which were specifically selected to contain the structure while the remaining 50 were randomly selected. This dataset is limited due to its costliness in terms of time and resources; nevertheless, the random structure sampling and dataset structure provides a broad overview of POLKE's performance.

As we wanted to focus on the performance of the RUTA rules on correctly formed structures, we chose to validate on native language rather than learner language. Nonetheless, in their automatic analysis of EFCAMDAT, Geertzen et al. (2014) conclude that NLP tools display robust performance on L2 English data. Since the RUTA rules rely on the output of the upstream NLP tools, we can assume that the performance on learner language would be comparable to that of native language.

Table 3 summarizes the results of the validation. POLKE was evaluated using precision, recall, and F1-score, which are evaluation metrics that are common for classification problems with unbalanced classes in machine learning. In the validation study, precision refers to the proportion of correctly identified structures among all structures which POLKE retrieved. Recall, on the other hand, is the proportion of the correctly identified structures among all actual structures. F1-score refers to the harmonic mean of precision and recall. The average precision, recall, and F1-score across the entire validation dataset are 0.91, 0.84, and 0.88, respectively.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

Table 3. Average precision, recall, and f1-score across all evaluated structures for each CEFR level

Level	Precision	Recall	F1
A1	0.91	0.82	0.86
A2	0.93	0.88	0.90
B1	0.95	0.86	0.90
B2	0.88	0.88	0.88
C1	0.92	0.87	0.89
C2	0.89	0.74	0.80

Assigning CEFR levels to EGP structures

While there is general agreement among examiners about whether a particular learner text is characteristic of a proficiency level (Hawkins & Buttery, 2010, p. 2), it can be difficult to determine at which exact point in their L2 development a learner has reached a specific milestone, such as a new level in the CEFR. This is a key difference between the CLC, on which the EGP is based, and EFCAMDAT. While the former contains single writings by learners which have been deemed to meet the requirements of a user of English at the relevant CEFR level, EFCAMDAT contains longitudinal learner data that follow the development of students' L2 skills in a gradual manner.

The continuous nature of the writings in the EFCAMDAT corpus thus posed a difficulty for determining the level of student texts in a way that was comparable with the CLC. Therefore, in order to obtain a "snapshot" of learners' proficiency at a specific CEFR level, we selected the texts from the last three units of the Englishtown trajectory for the given CEFR level and the first three units of the following one. The underlying assumption is that, as learners advance through the Englishtown lessons, their language ability approaches what the corresponding CEFR level stipulates. Similarly, it is to be expected that during the first units of the next course, students will not yet have sufficiently acquired those structures that would characterize them as English users at the next proficiency level. Thus, we consider these writings to best represent students' skills at the CEFR level. To approximate pre-A1 proficiency, we used per-learner structure frequencies from texts taken from the first six *Englishtown* A1 units. Given the limited number of C2 texts, especially at higher units, we decided to remove C2 structures from our analysis.

The correspondences between EGP structures and CEFR levels proposed in the present study were thus extracted by means of statistically analyzing the frequency with which all structures appeared in *Englishtown* students' writings. Concretely, within any given CEFR level, we obtained the frequency of each grammatical structure in the texts of every individual learner in the sample. The number of occurrences of an EGP structure in a student's writings at a given CEFR level was then normalized to obtain the structure frequency per hundred words for each student.

The ultimate goal was to compare the normalized structure frequencies at the early stages of a level with those at the end, and to determine whether they significantly

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

differed from each other. If a grammar structure is used significantly more in a level (e.g., B1) than in the previous levels (e.g., A2 and A1), we assume the structure is acquired at that level (i.e., B1). To establish usage differences across levels, we made use of hypothesis testing. The exact implementation of the tests is described in the following sections.

Determining the discriminative power of EGP structures

The first step in our corpus-based observation of EGP structures was to determine whether the whole grammatical inventory of the EGP could be used to discriminate writings from learners at different levels. Presumably, if a structure does not present a significant difference in its normalized frequencies at any level, its presence in learner output is uninformative with regard to the stage of L2 proficiency development.

In order to establish whether the frequency of a given grammatical structure varied across texts of different levels, we used the Kruskal–Wallis *H* test, a non-parametric test for detecting statistical differences in more than two independent samples. When the resulting *p*-value was greater than 0.05, we assumed that there was a lack of evidence for different structure frequencies across levels. In these cases, the structure was deemed uninformative and was therefore removed from further analysis. On the other hand, structures whose frequency in texts of at least one level was significantly different from the others were further analyzed to be mapped to the CEFR level at which they were acquired.

Assigning a CEFR level to a structure

Texts written by students of an L2 are usually produced in the context of highly specific learning scenarios and are focused on particular didactic goals in accordance with the curriculum, such as a certain grammatical structure. Therefore, it is likely that a structure will only be used extensively after being successfully acquired by the students, either incidentally or due to explicit instruction. As a result, in this study, the level of acquisition of an EGP structure was operationalized as the first CEFR level (ordered from lowest to highest) whose texts presented a significantly higher normalized frequency of the structure in comparison to the texts of the immediately preceding level.

This part of the analysis was implemented by iteratively performing one-sided Mann–Whitney U-tests between consecutive levels. The null hypothesis was rejected when there was enough evidence ($\alpha = 0.05$) to state that the normalized structure frequency was greater in texts belonging to the higher level. For instance, if one failed to prove that the frequency of a particular EGP structure at the A2 level was higher than that at A1, yet the frequency at B1 was significantly higher than at A2, the CEFR level proposed for the EGP structure at hand would be B1. The development of the structure's usage frequency after the first significant difference was deemed irrelevant. Figure 3 shows the frequency means across all levels of the EGP structure 120 (Can use adverbs of degree ["really," "so," "quite"] with an increasing range of common gradable

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

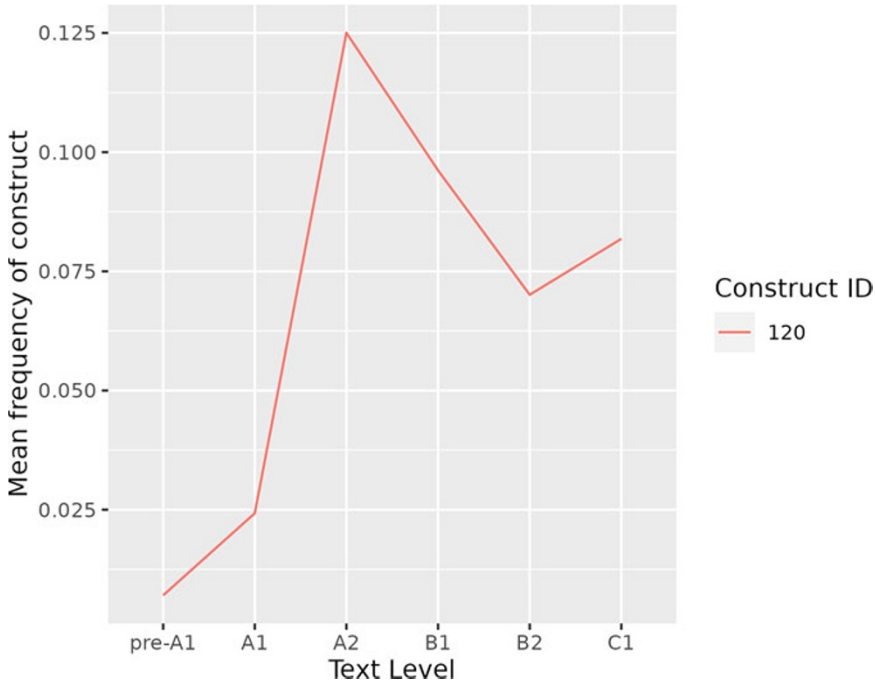

Figure 3. Example of an expected trajectory for an A2 structure.

adjectives), which would be considered an A2 structure, both according to the EGP and to the operationalization proposed in this paper.

It is worth noting that some structures did not appear at all in some levels. In such cases, we made no comparison between contiguous levels, to avoid drawing conclusions from results based on too limited data. This was because, regardless of the frequency of the structure in texts of the level with non-zero occurrences, our approach would have identified a significant difference.

It was possible for a structure not to be assigned to any level at this stage of the analysis. This might occur when significant differences were attested only between non-consecutive levels, or when no consecutive levels being compared in the Mann–Whitney U-test presented any occurrences of the grammatical feature and thus no comparisons were made.

Results

Disregarding grammatical structures to which the EGP assigned the C2 level, a total of 96 EGP structures were not found in any of the analyzed texts. Possible explanations for their absence are task effects, failure of the tool to recognize a given grammatical structure, or the rarity of the structures. The percentage of structures of each level with a frequency of zero is reported in Table 4.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

Table 4. Number and percentage of structures of each level not found in the data

EGP structure level	No. of zero-frequency structures	%
A1	3	3.45
A2	24	12.44
B1	21	12
B2	32	26.67
C1	16	35.56

Table 5. Number and percentage of structures per EGP level presenting significant differences in their frequencies across text levels

EGP structure level	Number of informative structures	%
A1	74	88.10
A2	134	79.29
B1	100	64.94
B2	55	62.50
C1	15	51.72

Table 6. Percentage of informative structures with a significant difference at their corresponding EGP level

EGP structure level	No. of significant differences at level	%
A1	27	36.49
A2	49	36.57
B1	28	28.00
B2	17	30.91
C1	3	20

Statistical analysis of structure frequencies

In examining the normalized frequencies per student of the structures at each text level, the Kruskal–Wallis test determined that 155 EGP structures showed no significant differences across text levels. Table 5 reports the numbers and percentages of non-zero frequency structures whose frequencies were found to be informative with regard to the text level after carrying out the Kruskal–Wallis test. A list of all structures ruled out at this stage are provided in the online supplementary materials.

Before making predictions about levels of acquisition based on the method outlined above, we calculated that only 124 (32.80%) of the informative structures displayed a significantly higher frequency at their EGP-assigned level than at the immediately preceding one. In general, this showed limited agreement between the EGP-assigned levels and the mappings derived here. Table 6 reports the percentage of structures of each level that presented significant differences at their assigned EGP level.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

Table 7. Precision, recall, and f1-score for structure-level assignments

Level	Precision	Recall	F1
A1	0.45	0.40	0.42
A2	0.35	0.32	0.33
B1	0.19	0.14	0.17
B2	0.16	0.18	0.17
C1	0.05	0.25	0.09
Macro average	0.24	0.26	0.24
Micro average	0.18	0.15	0.16

Structure-level mappings

For our purposes, automatically mapping the EGP structures to a level was comparable to a classification problem, where the classes correspond to each of the five proficiency levels under analysis. The levels assigned to the structures by O'Keeffe and Mark (2017) and the ones proposed here were compared by using precision, recall, and F1-score. The EGP level assignments were used as the expected classes and the level suggestions obtained with the current approach were treated as the model's predictions. Table 7 lists the values for each of the aforementioned metrics. In the present study, precision is the proportion of structures assigned to a specific level that match the EGP classification. Conversely, recall is the proportion of structures classified as a given level by the EGP that our approach also identifies as belonging to that level. Given the large differences in the number of structures across levels, these metrics were considered more appropriate than alternatives like accuracy, defined as the percentage of items (EGP structures) for which the assigned class matches the expected one.

Note that the presented values for precision, recall, and F1-score do not take into consideration the ordinal nature of the CEFR levels. For example, although the distance between A1 and A2 is smaller than between A1 and B2, these disagreements in between the expected and the observed classes are penalized equally. For more transparency, Figure 4 shows the confusion matrix for our classification in comparison with the EGP levels. Here, a greater overlap between the two approaches is represented by a darker color. The number within each cell shows the total number of grammatical structures that the approaches classified at the corresponding combination of levels.

An alternative way to measure the degree of correspondence of the statistically derived level predictions in comparison to the existing EGP mappings is through the lens of inter-rater reliability. This metric allows us to determine whether the observed agreement between two annotators (in our case, the original EGP level assignment and our frequency-based approach) is better than would be expected by chance alone. To account for the ordering of the classes, we calculated the weighted Cohen's kappa statistic ($\kappa = 0.193$, $P < 0.01$) with squared weights using the kappa2() function from the irr package (Gamer et al., 2019) in the R Statistical Software (v4.1.2; R Core Team, 2021). According to Landis and Koch (1977), this value for κ corresponds to slight agreement between the two approaches.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

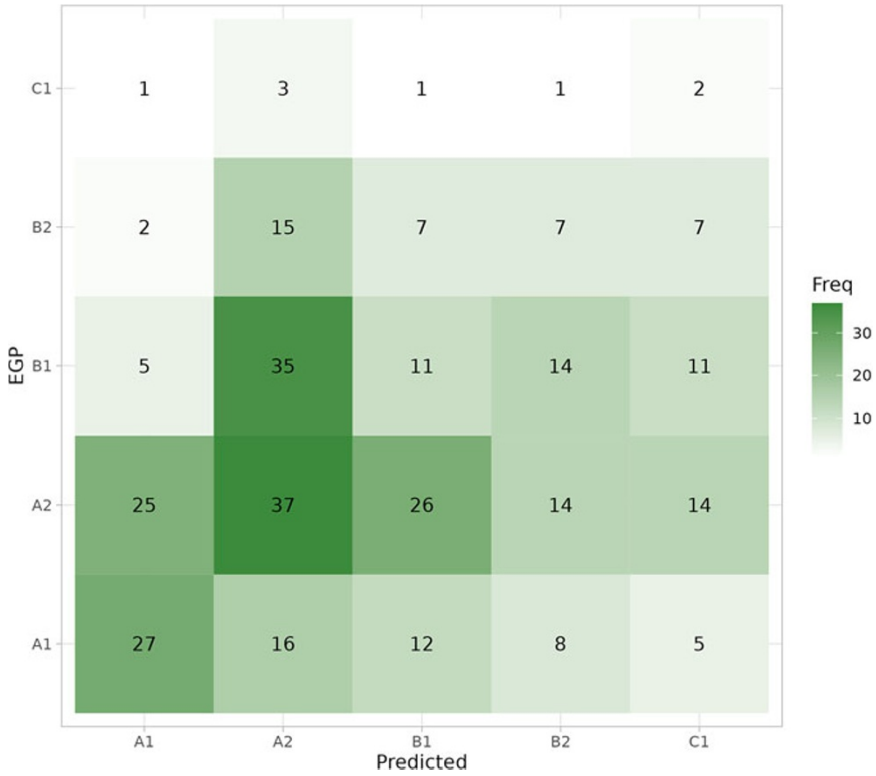

Figure 4. Classification of structures into levels: our predictions vs. the EGP-assigned levels.

Inspecting the POLKE output

The substantial disagreement between the levels assigned to the structures using our statistical approach and the EFCAMDAT data and those in the original EGP creates the need to inspect the learner texts in conjunction with POLKE's output. By conducting manual exploration of the data, we aimed to determine possible causes for these differences. First of all, a misalignment for a structure might be caused by it being differently distributed across levels in EFCAMDAT in comparison to the CLC. Secondly, it might also be due to inaccuracies in the structure extraction, formulaic sequences, or task effects (Alexopoulou et al., 2017).

We thus randomly sampled several structures for inspection. Firstly, we chose two random structures per EGP level (A1—C1) and inspected the texts where POLKE had identified them. In doing so, we aimed to determine whether the extracted structures properly matched the descriptors. Additionally, texts where the tool had not encountered the structure were randomly sampled and searched for any occurrences that the extractor might have missed. Secondly, we chose three structures for which the predicted level and the EGP level differed by three levels or more and compared their use in texts from both its EGP level and the level assigned by our operationalization. Finally, out of those structures for which POLKE found no occurrences, one structure

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

was sampled per level; we then searched the data for occurrences of these structures using regular expressions. This inspection revealed a range of sources for the misalignment between the EGP and the approach taken here. In the following paragraphs, we summarize our findings.

For the structures that were randomly sampled from all non-zero frequency structures, the general trend was towards a higher accuracy of the tool for lower-level grammatical structures, although there was considerable variation in this regard. Importantly, the structures often appeared embedded in identical sentences across writings of the same units written by different students, suggesting that they might have been copied from the task description. Even when sentences were not identical, some units seemed to elicit some structures especially frequently. Finally, while the tool found some structures in sections of the text where they were not present, the instances where our inspection revealed that the extractor had missed an annotation mostly corresponded to misspellings or ungrammatical sentences.

Examining the behavior of some EGP items that were assigned considerably different CEFR levels than expected offered various explanations for the discrepancies. Firstly, structures which are most prominent in pre-A1 can decrease in use in higher-level texts. Therefore, since assignment to pre-A1 is not possible, our approach only recognized the next level where there was a significant increase in frequency, which was often one of the more advanced ones. Furthermore, tasks played an important role in determining in which texts certain structures appeared most frequently. Finally, POLKE struggled to correctly identify certain grammar structures, which caused inaccuracies in the automatic analysis of the structure frequencies.

Searching the data for the five sampled structures which were not recognized by POLKE revealed that three of these were actually present in the data. The structures which the tool did not recognize included some plurals (structure 70, EGP A1), the past simple passive with a limited range of ditransitive verbs (structure 596, EGP B1), and use of "the best" with ellipted "can" or "could" (structure 995, EGP C1). The remaining two structures, "have (got) to" in its question form and the present perfect continuous with adverbs in the mid position, were not found in the learners' texts.

Discussion

This paper set out to describe an approach to automatically and quantitatively analyze the appearance of EGP structures in learner data and their mappings to CEFR levels based on texts obtained from the EFCAMDAT corpus. Limited agreement with the structure levels derived by O'Keeffe and Mark (2017) was found. Generally, there was greater agreement between our structure-level mappings and the EGP's for beginner levels, that is, our operationalization was much more likely to classify lower-level items at the same level as the EGP or one of the adjacent ones. This was not the case for the CEFR B and C levels. Such limited overlap might arise due to fundamental differences between the CLC and EFCAMDAT, with the former being a corpus of test compositions, while the latter student writings from an online English course. Another reason for the limited overlap was the considerable divergence between the original EGP-level assignment methods and our frequency-based approach. The original EGP levels were determined by researchers combining qualitative and quantitative methods making

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

use of the CLC corpus. Our level assignment was based purely on quantitative findings with automatic grammar structure extraction using NLP technologies.

After automatically extracting the EGP items from the texts, around 15% of the structures for which rules for POLKE have been implemented were not found. Manual exploration of a small sample of these structures revealed two main causes for this: issues with the extraction of the structure and the actual absence of the grammatical structure in the data. It is sensible to assume that the more basic structures should have a higher frequency than those presumably acquired at more advanced levels. Therefore, it is likely that those zero-frequency grammatical items assigned to lower EGP levels result from POLKE failing to recognize them. Conversely, higher-level items have a greater likelihood of simply not being present in texts. Similarly, structures that appear very infrequently in our data pose a problem to the quantitative classification of structures into levels, since their absence might result from their uncommon nature in general language use and not necessarily from the fact that learners at that level have not yet acquired them. This issue pertains not only to this study, but also to O'Keeffe and Mark's (2017) proposed structure-level mappings. Nevertheless, it is arguable that, given the central role that frequency plays in acquisition (see N. C. Ellis, 2002), less common grammatical features – as well as uncommon uses of certain grammatical forms – are less likely to be acquired at early stages.

A lack of significant differences among text levels for certain structures can be explained and interpreted in a variety of ways. Firstly, if a given grammatical structure was present in texts of all levels with relatively similar frequencies, it may suggest that this structure is acquired at A1. On the other hand, it seems that a lack of data points (structure occurrences) was often the cause for an item to be ruled out by the Kruskal–Wallis test. If scarcely any learners in a few levels used a grammatical form, this was likely an uncommon structure without enough evidence pointing towards it being acquired at any specific level.

The approach presented above was not able to make predictions for all EGP items for which Kruskal–Wallis found significant differences among the levels. In some cases, this had to do with frequencies only decreasing significantly from their use in lower levels, which was often the case for structures to which the EGP assigns the level A1. This is easily explained by students acquiring basic structures at early stages of language proficiency and using them extensively due to a lack of grammatical means to express similar concepts with a variety of structures; with the expansion of their linguistic inventory at subsequent levels, the frequency of some basic grammatical structures is likely to decrease in favor of more advanced, nuanced ones. Some other structures had no occurrences in texts of consecutive levels, which made it impossible for our approach to assign any CEFR level to them. Finally, the frequency of some EGP items does not differ across immediately consecutive levels, but when comparing two non-adjacent levels, the Mann–Whitney U-test reveals a statistical difference.

A limitation of this study is that it does not take syntactic and pragmatic correctness into account, whereas a 60% rate of correct uses was one of the criteria for the authors of the EGP to consider that a structure had been acquired. Similarly, they required a frequency of use of each structure greater than that in reference corpora with texts written by native speakers, whereas we did not use a threshold frequency in judging the acquisition of a grammatical feature.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

O'Keeffe and Mark (2017) justify the choice of using frequencies from the British National Corpus as a baseline for acquisition due to the fact that they utilized texts from high-stakes examinations, where structures deemed by learners and teachers to be advanced might be overrepresented. These are characteristics in which EFCAMDAT and CLC texts differ considerably: the writings in the CLC were produced in highly controlled environments, without access to external information and at a time likely preceded by periods of rigorous preparation on the part of the learners. On the other hand, EFCAMDAT comprises texts from low-stakes, untimed tasks that learners completed unsupervised. Nevertheless, the accuracy patterns between EFCAMDAT and CLC writings have been reported to be similar (Derkach & Alexopoulou, 2024, p. 7). As a result, it might be argued that such differences make the EFCAMDAT corpus more ecologically valid than the CLC. Even so, both corpora are of fundamentally different natures. Thus, any insights drawn from EFCAMDAT are to serve as supplementary for the purposes of EGP research and not as absolute truths overriding previous findings.

However, another important limitation is that, while a CLC text tagged as "passed" is meant to guarantee that the writing has the characteristics of a text of the corresponding level, EFCAMDAT texts result from continuous assessment tasks. This means that a text with a successful grade is likely indicative of the objectives of the unit being met, but not necessarily of the writing being appropriate for a learner at the CEFR level the student is enrolled in. We address this issue by defining the texts of any given CEFR level as those written in the last three units of the level or the first three of the next one. Nevertheless, we acknowledge that these cutoff points are arbitrary and that more advanced methods would be necessary to unequivocally assign a text to a CEFR level in a way that makes it comparable to the CLC.

Given these differences between O'Keeffe and Mark's (2017) data and methods and those employed here, it would be misguided to assert that different level assignments for a structure necessarily implies an error in the EGP. Instead, these results are to be taken as an invitation to further explore quantitative methods for large-scale, automatic grammatical structure analysis to facilitate future attempts to validate the EGP. For this purpose, the quantitative results and the frequency plots presented here are a valuable aid for future improvement of the EGP, including the level assignment of the grammar structures.

Finally, this study should be repeated once POLKE has reached a higher level of maturity, as it is currently still under development. A large-scale evaluation will be useful for finding issues in the implementation of RUTA rules and for gauging the robustness of the tool and the reliability of the results.

Conclusions

NLP technologies for extracting specific grammatical constructions have often been used for SLA research and for the development of Intelligent Computer-Assisted Language Learning (ICALL) applications. Given the widespread use of the CEFR and the influence of projects such as the EPP that aim to specify the communicative descriptors of the CEFR in terms of grammar and vocabulary, there is considerable need for validating the findings that have been widely used to inform the design of teaching materials and high-stakes examinations.

https://doi.org/10.1017/S0267190525100093 Published online by Cambridge University Press

Such is the case for the EGP, a subproject of the EPP that set out to identify which uses of different grammatical forms are acquired by English L2 learners at different stages of proficiency as defined by the CEFR. These grammatical descriptions and level mappings were designed by O'Keeffe and Mark (2017) based on manually and statistically analyzing the contents of the CLC. However, a fully quantitative validation of these results is yet to be carried out.

The goal of this study was to use an in-house NLP system, POLKE, to extract the structures defined in the EGP from authentic learner writings and propose a method to automatically derive new structure-level mappings and observe how these compare to the EGP. This project used the EFCAMDAT dataset, a CEFR-aligned corpus with texts written by thousands of L2 English students worldwide.

The approach relied on significance testing to determine whether the frequency of use of each grammatical structure differed across levels. With this procedure, we showed that around a third of the structures show significant differences between the level the EGP assigns to them and the immediately preceding one. 155 grammatical structures exhibited no significant differences across usage frequencies at different levels. This could be due to a lack of data where the structure appears or, potentially, it might suggest the existence of grammatical items whose presence is uninformative with regard to the learner's proficiency level. Finally, the automatic assignment of CEFR levels showed little overlap with the mappings that currently comprise the EGP, showing that the approach presented here does not provide substantial support for the current EGP-level assignments. However, such discrepancies between the two sets of mappings may be partially accounted for by differences between the approaches and the data used by the original EGP authors and in the present work.

In the future, an automatic, fully quantitative analysis of the EGP with revised methods is likely to still present some divergences from the EGP's assignment of levels to grammatical structures. In such cases, although a reevaluation of the EGP might be warranted, it is important that any revisions be carried out in collaboration with SLA researchers and experienced teachers, while using robust, quantitative methods for flagging potentially problematic EGP structures. In any case, since the quality of the EGP has a profound influence on SLA research and EFL/ESL teaching and assessment, it is of paramount importance that the quality of the resource be validated and improved.

Supplementary material. For more details on our qualitative and quantitative results, see the online supplementary materials at https://osf.io/u7pd6/?view_only=cbed1b08e6af4b89a6aadb37319cc51d.

Acknowledgments. This project is funded by the German Ministry of Education and Research (BMBF) under funding number 01IS22076.

References

Alexopoulou, T., Geertzen, J., Korhonen, A., & Meurers, D. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, *1*(1), 96–129. https://doi.org/10.1075/ijlcr.1.1.04ale

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. *Second Language Research*, *35*(1), 99–119. https://doi.org/10.1177/0267658316643125

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (21–46). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.02bul

Capel, A. (2015). The English vocabulary profile. In J. Harrison & F. Barker (Eds.), *English profile in practice* (9–27). Cambridge University Press.

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62. https://doi.org/10.1016/j.jslw.2019.03.005

Chen, X., Alexopoulou, T., & Tsimpli, I. (2021). Automatic extraction of subordinate clauses and its application in second language acquisition research. *Behavior Research Methods*, *53*(2), 803–817. https://doi.org/10.3758/s13428-020-01456-7

Corder, S. P. (1975). Error analysis, interlanguage and second language acquisition. *Language Teaching*, *8*(4), 201–218. https://doi.org/10.1017/S0261444800002822

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2005). *Reference level descriptions for national and regional languages (RLD): Draft guide for the production of RLD (Version 2)*. Council of Europe Language Policy Division. Retrieved March 17th, 2025, from https://rm.coe.int/090000168077c574

Council of Europe. (n.d.). *Reference level descriptions (language by language)*. Retrieved March 18th, 2025, from https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions

Davies, M. (2008—). The Corpus of Contemporary American English (COCA): 560 million words, 1990—present. Retrieved March 18th, 2025, from https://www.english-corpora.org/coca/.

Derkach, K., & Alexopoulou, T. (2024). Definite and indefinite article accuracy in learner English: A multifactorial analysis. *Studies in Second Language Acquisition*, *46*(3), 710–740. https://doi.org/10.1017/S0272263123000463

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3–4), 327–348. https://doi.org/10.1017/S1351324904003523

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement (Version 0.84.1) [R package]*. Comprehensive R Archive Network (CRAN). Retrieved November 15th, 2024, from https://CRAN.R-project.org/package=irr

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In R. T. Miller, E. Bayram, L. Heilenman, & J. Rehner (Eds.), Selected proceedings of the 2012 Second Language Research Forum: Building bridges between disciplines (240–254). Cascadilla Proceedings Project.

Green, A. (2010). Requirements for reference level descriptions for English. *English Profile Journal*, *1*, e6. https://doi.org/10.1017/S204153621000005X

Harrison, J. (2015a). What is English Profile? In J. Harrison & F. Barker (Eds.), *English profile in practice* (1–8). Cambridge University Press.

Harrison, J. (2015b). The English Grammar Profile. In J. Harrison & F. Barker (Eds.), *English profile in practice* (28–48). Cambridge University Press.

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*, e5. https://doi.org/10.1017/S2041536210000103

Huang, L. F., Lin, Y. L., & Gráf, T. (2023). Development of the use of discourse markers across different fluency levels of CEFR: A learner corpus analysis. *Pragmatics*, *33*(1), 49–77. https://doi.org/10.1075/prag.21016.hua

James, C. (1990). Learner language. *Language Teaching*, *23*(4), 205–213. https://doi.org/10.1017/S0261444800005905

Jones, G. (2015). *A validation study of the British Council–EAQUALS Core Inventory for General English*. Assessment Research Awards and Grants Research Reports Online, British Council. Retrieved March 18th, 2025, from https://www.britishcouncil.org/sites/default/files/glyn_jones_layout.pdf

Kim, J. Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, *69*(4), 27–51. https://doi.org/10.15858/engtea.69.4.201412.27

Kluegl, P., Toepfer, M., Beck, P.-D., Fette, G., & Puppe, F. (2016). UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, *22*(1), 1–40. https://doi.org/10.1017/S1351324914000114

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., & Ott, N. (2010). Enhancing authentic web pages for language learners. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications (10–18). Association for Computational Linguistics.

Murakami, A., & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*, *38*(3), 365–401. https://doi.org/10.1017/S0272263115000352

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), Proceedings of the Corpus Linguistics 2003 Conference: Technical papers (Vol. 16, 572–581). UCREL, Lancaster University.

North, B. (2009). The educational and social impact of the CEFR in Europe and beyond: A preliminary overview. In L. Taylor & C. J. Weir (Eds.), Language testing matters: Investigating the wider social and educational impact of assessment. Proceedings of the ALTE Cambridge Conference, April 2008 (Studies in Language Testing, Vol. 31, 357–378). Cambridge University Press.

North, B., Ortega, Á., & Sheehan, S. (2010). *British Council–EAQUALS core inventory for general English*. British Council/European Association for Quality Language Services. Retrieved March 18th, 2025, from https://www.eaquals.org/wp-content/uploads/EAQUALS_British_Council_Core_Curriculum_April2011.pdf

North, B., & Piccardo, E. (2019). Developing new CEFR descriptor scales and expanding the existing ones: Constructs, approaches and methodologies. *Zeitschrift Für Fremdsprachenforschung*, *30*(2), 142–160.

O'Keeffe, A., & Mark, G. (2017). The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, *22*(4), 457–489. https://doi.org/10.1075/ijcl.14086.oke

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Quixal, M., Rudzewitz, B., Bear, E., & Meurers, D. (2021). Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning (15–27). Liu Electronic Press. Retrieved March 17th, 2025, from https://aclanthology.org/2021.nlp4call-1.2.pdf

R Core Team. (2021). *R: A language and Environment for Statistical Computing*. [*Computer software*]. R Foundation for Statistical Computing. Retrieved March 17th, 2025, from https://www.R-project.org/

Römer, U., & Berger, C. M. (2019). Observing the emergence of constructional knowledge: Verb patterns in German and Spanish learners of English at different proficiency levels. *Studies in Second Language Acquisition*, *41*(5), 1089–1110. https://doi.org/10.1017/S0272263119000202

Saville, N., & Hawkey, R. (2010). The English Profile Programme–the first three years. *English Profile Journal*, *1*, e7. https://doi.org/10.1017/S2041536210000061

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

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

van Ek, J., & Trim, J. (1991b). *Waystage 1990*. Cambridge University Press.

van Ek, J., & Trim, J. (2001). *Vantage*. Cambridge University Press.

Verratti-Souto, D. (2024). *Towards a quantitative validation of the English Grammar Profile: A corpus-based approach* [Unpublished bachelor's thesis]. Universität Tübingen.

Cite this article: Verratti-Souto, D., Sagirov, N., & Chen, X. (2025). NLP-powered quantitative verification of the English Grammar Profile's structure-level assignment. *Annual Review of Applied Linguistics*, 1–22. https://doi.org/10.1017/S0267190525100093