ATLANTIS
Journal of the Spanish Association of Anglo-American Studies
42.2 (December 2020): 1-21
e-ISSN 1989-6840
DOI: http://doi.org/10.28914/Atlantis-2020-42.2.01

Determining the Level of a Language Test with English Profile: A Forensic Linguistics Case Study

Núria Gavaldà[A,B] and Sheila Queralt[B]
Universidad Internacional de La Rioja[a], Laboratorio SQ-Lingüistas Forenses[b]
nuria.gavalda@unir.net, sheila.queralt@cllicenciats.cat

This article deals with a forensic linguistics case study of the determination of the level of a B1 English multiple-choice test that was challenged in court by numerous candidates on the grounds that it was not of the appropriate level. A control corpus comprising 240 analogous multiple-choice questions from B1 exams aligned with the Common European Framework of Reference for Languages (CEFR) was compiled in order to establish a threshold for the percentage of questions of a level higher than that being tested which can be expected in such exams. The analysis was carried out following a combination of qualitative and quantitative methods, with the help of the tool English Profile, which provides Reference Level Descriptions (RLDs) for the English language within the CEFR. The results of the analysis of the control corpus established a baseline of 5 to 7% of questions that include key items classified as higher than B1, while the percentage was 68% in the case of the disputed exam. Thus, the present study proposes a further application of the tool English Profile within the field of forensic linguistics and puts forward the concept of Level Appropriateness Threshold (LAT), analogous to other thresholds established in forensic linguistics, which can serve as a baseline for determining the appropriateness of B1 English multiple-choice exams and a model for other levels and skill areas.

Keywords: language level tests; multiple-choice exams; English Profile; forensic linguistics; Level Appropriateness Threshold

. . .

Determinación del nivel de una prueba de idioma con *English Profile*: un estudio de caso de lingüística forense

Este artículo trata de un estudio de caso de lingüística forense sobre la determinación del nivel de un examen de opción múltiple de inglés con nivel B1 que fue impugnado por numerosos candidatos que consideraban que su nivel no era el adecuado. Se compiló un corpus de control compuesto por 240 preguntas de opción múltiple similares provenientes de exámenes oficiales de B1 acordes con Marco Común Europeo de Referencia para las Lenguas (MCER), con el fin de establecer un umbral para el porcentaje de preguntas con un nivel superior al del nivel examinado que sería el esperable en dichos exámenes. El análisis se llevó a cabo siguiendo una combinación de métodos cualitativos y cuantitativos, con la ayuda de la herramienta English Profile, que proporciona Descripciones de Niveles de Referencia para el inglés dentro del MCER. Los resultados del análisis del corpus de control establecieron un umbral de adecuación de nivel de entre el 5 y 7% de preguntas clasificadas como superiores a B1, mientras que el porcentaje fue de un 68% en el caso del examen impugnado. Así pues, el presente estudio propone una nueva aplicación de la herramienta English Profile en el campo de la lingüística forense y el concepto de Umbral de Adecuación del Nivel, análogo a otros umbrales establecidos en la lingüística forense, como base para determinar la idoneidad de los exámenes de opción múltiple de inglés con nivel B1 y modelo para otros niveles y competencias.

Palabras clave: pruebas de nivel de idioma; exámenes de opción múltiple; English Profile; lingüística forense; Umbral de Adecuación del Nivel

1. Introduction

Forensic linguistics is the branch of applied linguistics that deals with the analysis and description of the language used by different participants in judicial and investigative contexts. The field is often described as the interface between language and the law. John Gibbons and M. Teresa Turell distinguish three main areas in forensic linguistics: the language of the law, the language of the court and language as evidence (2008). The first area, the language of the law, focuses on descriptive studies of different kinds of textual genres typically found in legal contexts, the peculiarities of which have been of interest to jurists and linguists for a long time, such as contracts, judgements and laws (Lavery 1921; Tiersma 1999; Stygall 2010). Another major interest within this area of forensic linguistics is the promotion of a language that is comprehensible for as many users involved in legal and bureaucratic contexts as possible (Montolío 2012; Poblete et al. 2018).

The second main area, the language of the court, explores the different uses of language by the various participants in processes related to law enforcement, such as judges, magistrates, witnesses, lawyers, prosecutors, police agents, victims and suspects. These processes include, among others, trials, police interviews and summary procedures. In these contexts, the forensic linguist analyses the types of questions, the discourse strategies, the vocabulary and the appropriateness of all these in the context at hand.

The third area, language as evidence, includes a wide range of situations where the knowledge of a forensic linguist is required in judicial or investigative processes either as an advisor or as an expert witness, and it continues to expand in order to satisfy the needs of society. The sphere of language as evidence may include cases related to authorship attribution, speaker identification, speaker/author profiling, plagiarism detection and the analysis of disputed utterances in recordings or ambiguous clauses in legal texts. In many of these areas, forensic linguistics draws on the field of corpus linguistics to determine the rarity and the salience of the linguistic traits found in the comparison of two or more linguistic samples (Coulthard 1994; Grant 2007; Turell 2010). In other words, the use of existing reference corpora with data on the distribution of certain linguistic traits across a specified population—e.g., the rate of presence or omission of the subject pronoun in Spanish written texts or the average fundamental frequency of male and female speakers in English recordings—can help establish the relevance of the level of similarity or difference between two spoken or written texts.

In recent years, the Laboratorio SQ-Lingüistas Forenses in Barcelona has received an increasing number of enquiries related to what seems to be a further field of application of forensic linguistics in the area of language as evidence that, to our knowledge, has not yet been described in the literature. This new branch of research deals with the determination of the appropriateness of the level of a language test conducted outside the framework of standard, regulated language tests such as Cambridge English or the Test of English as a Foreign Language (TOEFL) exams. The enquiries that our laboratory has received are concerned with English multiple-choice tests that are being used in public service entrance exams and are generally made by candidates who claim that they had the required English level but found the exam too difficult and failed it. The moment candidates bring a lawsuit against the companies or institutions that are using these tests in their selection processes, the disputed test becomes the linguistic evidence around which the judicial process will centre. As a consequence, an analysis of the exam is required by a linguist expert witness with extensive knowledge of the English language and its assessment methods. The fundamental questions that arise from this situation are the following:

> RQ1 Does the English disputed test have the level that it intends to assess?
> RQ2 Does it have a different level—higher or lower—and is therefore an unfair test for the applicants?
> RQ3 How can the forensic linguist objectively determine whether the level of the test is appropriate?

The present article outlines the process involved in answering these questions by reference to a real case of a disputed exam. Similar to other types of cases involving written texts—such as plagiarism detection or authorship attribution cases—the present study relies on corpus linguistics to empirically establish the appropriateness of the test under discussion. Far from being uncommon, the use of corpus linguistics to evaluate language has increased (Barker 2010, 2014; Park 2014; Weigle and Goodwin 2016; Cushing 2017) since it involves an empirical analysis that can help compare a variety of features, such as the vocabulary used by an early-stage learner in comparison to an advanced learner or the vocabulary and grammatical structures used at a B1 or B2 level.

2. Case Outline

This article deals with the first case related to the appropriateness of an English test that was dealt with by our laboratory, in 2017. A group of candidates who had sat a selection exam for a state-run transport company contacted us claiming that there had been some irregularities involving the English test used in the selection process. The test consisted of 50 multiple-choice questions each with 3 options, which focused on grammatical and, particularly, lexical items. The level of English required as set out in the terms of the call for the positions was a B1. Yet a large number of candidates who had the required level—often with B1 certificates from prestigious bodies—agreed that the exam had far exceeded the intended level. After the final results confirmed that the vast majority of candidates had failed the English test and, as a result, had been eliminated from the selection process, several groups of candidates around Spain decided to take the state-run company to court on the grounds that the English exam had been unfair.

Even though the case was clearly a forensic linguistics matter, since it was essentially based on linguistic evidence, there was no existing literature within the field that could provide a suitable methodology to establish the appropriateness of the level test for judicial purposes. However, after looking at studies and methodologies related to English testing, we came across the tool English Profile (EP), which seemed to provide an adequate qualitative and quantitative framework for the empirical determination of the level of the disputed test. The tool, explained in section four, provides Reference Level Descriptions (RLDs) for English, one of the few languages for which they have been developed.[1] RLDs determine linguistic structures and words that are typically acquired at each proficiency band according to the Council of Europe's Common European Framework of Reference for Languages (CEFR) (2001a).

A preliminary analysis revealed that many of the key items in the questions were classified by EP as having a higher level than B1—not only B2 but also C1 and even C2. However, there was nothing in the literature that could answer the following two questions: is it acceptable for a level test to include questions with a higher level than the intended one? If it is, how many, or what proportion of such questions are appropriate? For forensic purposes, it was necessary to establish a threshold for the percentage of B1 questions and questions of a higher level that should be included in a B1 test in order for it to be appropriate, henceforth referred to as the Level Appropriateness Threshold (LAT).

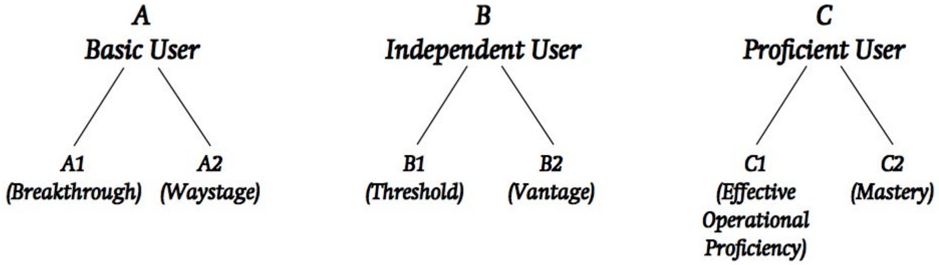
The concept of *thresholds* has been previously used in forensic linguistics in the field of plagiarism detection, especially as regards to similarity or uniqueness. The establishment of a threshold level of textual similarity between texts is crucial in plagiarism detection, since it is fundamental in determining whether the similarities between two texts raise concerns of plagiarism. As explained by Turell, the use of specific software to detect plagiarism "allows researchers to come up with the base-line or threshold level of similarity, which will establish the point at which this similarity becomes suspicious" (2008, 287). Turell established that the similarity threshold between two texts produced independently should not exceed 50% (2007). A further study by Montse Marquina and Sheila Queralt found that in the case of journalistic texts, the similarity threshold should be placed between 35 and 40% (2014). Following the same rationale, the context of disputed language tests could also benefit from an analogous threshold, in this case one that establishes the minimum percentage of questions of a certain level—in our case B1—that is necessary in order for the test to be appropriate for a B1 level. To this effect, a control corpus of comparable multiple-choice questions from B1 exams that were aligned with the CEFR was compiled to ascertain whether and to what extent they included questions requiring a higher level.

[1] RLDs need to be developed specifically for each language. The RLDs developed to date can be found on the Council of Europe webpage (Council of Europe 2020).

3. B1 Level in the Common European Framework of Reference for Languages

The Council of Europe's CEFR describes the different levels that can be achieved when learning any foreign language, as well as which specific skills should be acquired in order to achieve each level effectively. The CEFR provides clear guidance for the elaboration of anything related to language learning and testing—syllabuses, curriculum guidelines, examinations, textbooks, etc. (Council of Europe 2001b). The CEFR defines three levels, each with two sublevels, as can be seen in figure 1:[2]

Figure 1: The six levels of proficiency defined by the CEFR (Council of Europe 2001a, 23)



Apart from the case dealt with in this article, the other cases that have been the subject of consultation at Laboratorio SQ have also questioned tests of an alleged B1 level. This may not be a coincidence, since B1 (Independent User) is the level at which the learner

> can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans. (Council of Europe 2001a, 24)

Consequently, B1 is the level at which the learner has achieved the minimum effective communication skills that can be used in a professional as well as a personal sphere, and it is also the level theoretically achieved at the end of obligatory secondary education in Spain (Ley Orgánica 8/2013). Therefore, B1 is the level that is often required for many nontechnical positions in public institutions in Spain.

[2] A detailed analysis of the CEFR or its levels is beyond the scope of this article. For further information, see Council of Europe (2001a).

Despite the fact that the CEFR sets these six standard proficiency bands that are independent of the language and the context of the learner, several authors have claimed that the definition of each band is rather vague (e.g., Carlsen 2012). In response to this perceived vagueness, the Council of Europe has encouraged the creation of RLDs for national and regional languages, which are designed to provide language-specific details about the content acquired at each CEFR band (Council of Europe 2020). These RLDs need to be created for each language separately by transposing the general descriptors from the CEFR into specific vocabulary and grammatical items that "should be set as objectives for teaching, or used to establish whether a user has attained a particular proficiency level" (Leńko-Szymańska 2015, 121).

4. English Profile

EP was created in 2008 within the framework of the English Profile Programme by Cambridge English and it is endorsed by the Council of Europe. The tool is the result of an interdisciplinary collaboration between numerous prestigious bodies in the areas of research and teaching English as a Second Language: Cambridge English, Cambridge University, Cambridge University Press, Cambridge English Language Assessment, the British Council, the University of Bedfordshire and English UK. The main objective of EP is "to 'transpose' the Framework descriptors that characterise the competences of users or learners at a given level into the linguistic material which is specific to a given language (i.e. grammar, lexical items etc.) and considered necessary for the implementation of those competences" (University of Cambridge Local Examinations Syndicate (UCLES) and Cambridge University Press 2011, 5). In other words, EP describes the specific linguistic structures and vocabulary items that learners need in order to perform the "can do" statements in each level of the CEFR, which in turn informs teachers, curriculum developers, course-book authors and test developers about which grammatical and lexical contents are suitable at each level (Cambridge University Press 2015).

What makes EP innovative and unique is that its conclusions are based on objective data extracted systematically from the largest existing corpus of English learners' exams, the Cambridge Learner Corpus (Nicholls 2003). This corpus compiles data on more than two hundred thousand exams from learners from more than 148 language backgrounds (Cambridge University Press 2015). According to Julia Harrison and Fiona Barker, the fact that EP uses such a large reference corpus with data drawn from real English learners means that the conclusions reached are data driven, that is, based on "concrete evidence of what language learners throughout the world actually know at each level of the CEFR" (2015, 4). The corpus also contains data on learners with very different first languages (L1s), which makes this tool is objective and language independent. Consequently, "English Profile's findings can be referenced with far greater certainty than anything preceding it" (2015, 4).

EP makes it possible to carry out various tasks related to the alignment of English skills with the CEFR and one of its main applications is indeed developing exams. Ben Knight claims that exam alignment with the CEFR is extremely important and that EP is a suitable tool not only for writing and editing materials, but also for developing appropriate questions and tasks for exercises and exams by aligning the vocabulary and grammar involved with the level being taught or assessed (2015).[3] In this sense, EP has recently been evaluated in the context of classifying learner output according to CEFR levels, and the results reveal that there is a strong positive correlation between the alignment of student essays with CEFR proficiency levels as made by human expert raters and the analysis carried out by means of EP (Leńko-Szymańska 2015). Similarly, Abdullah Arslan and Ali Erarslan successfully employed EP and its related software, Text Inspector, in determining the appropriateness of the level of textbooks used in secondary schools in Turkey (2019).

EP consists of two main search engines. On the one hand, English Vocabulary Profile (EVP) defines the CEFR level at which certain lexical items, semantic nuances or idiomatic expressions could be expected to occur, while English Grammar Profile (EGP) indicates which grammatical skills learners acquire at each learning stage. All in all, it is clear that EP is a reliable tool that can help assign English written texts such as learner's output texts or textbooks to their corresponding CEFR proficiency band. In the context of a disputed exam, it would appear to be a suitable tool for forensic linguists to empirically assess the appropriateness of the grammatical and lexical content of the questions with respect to the intended level.

5. Objectives and Methodology

5.1. Objectives

The objective of the present study is twofold. On the one hand, a set of B1 multiple-choice questions from the control corpus are analysed by means of EP in order to establish a LAT, i.e., the percentage of questions that are classified as B1 in these tests. On the other, the same EP analysis is applied to the questions in the disputed test so as to ascertain whether it meets the previously prescribed LAT. The ultimate goal is to determine to what extent the disputed exam is an appropriate B1 test.

5.2. Control Corpus

Since the disputed exam consists of multiple-choice questions, the control corpus compiled includes the multiple-choice parts of two B1 exams: the Preliminary English

[3] The reader is referred to the document *Using the CEFR: Principles of Good Practice* for guidelines on using the CEFR to develop exams (University of Cambridge-ESOL Examinations 2011).

Test (PET) from Cambridge Assessment English and the B1 test from the Escola Oficial d'Idiomes (EOI) in Catalonia. Cambridge Assessment English is one of the most prestigious international bodies involved in the certification of English levels, while the EOIs are public language centres under the control of the Education Department of each regional government in Spain that "issue language proficiency certificates through unified and validated tests that match up with the levels established in the Common European Framework of Reference" (Escola Oficial d'Idiomes Barcelona Drassanes).[4] The rationale for choosing these two sources was to compile data not only from an international body, but also from a Spanish body, since the disputed exam had been designed from a Spanish perspective.

The multiple-choice questions from the Cambridge PET exams were extracted from four books from different publishers that include authentic examination papers (Adams 2006; *Cambridge English PET 7* 2011; Hashemi and Thomas 2013; *Cambridge English PET 8* 2014), where the multiple-choice questions occur in Part 5 of Paper 1 (Reading and Writing), which consists of a short text with ten gaps each of which has four answer options testing both vocabulary and grammatical items—this format is the same in all Cambridge exams, regardless of the level. In total, 180 multiple-choice questions from Part 5 of Paper 1 of eighteen PET exams were analysed.

The B1 exam from EOI has a two-part section devoted to Use of English. Part 1 has texts with gaps and three choices per gap targeting grammatical and lexical constructions, while Part 2 contains gapped sentences with four options each. The present study analyses two B1 EOI exams, one paper sample and one online sample, each of which contains 30 multiple-choice questions so that the total number of EOI questions analysed was 60 (Generalitat de Catalunya 2019). The reason for the imbalance between the number of PET (240) and EOI (60) questions is that these were the only EOI B1 samples available on the Internet and requests to EOIs for more samples were not successful. Table 1 shows a summary of the sources from which the multiple-choice questions were extracted and the total number of questions that made up the control corpus.

TABLE 1. Summary of questions that make up the control corpus and their sources

Source	Year	Publishers	Exams	N Questions
Adams	2006	Cengage Learning	Part 5 of Paper 1 from Tests 1 to 6	60
Hashemi and Thomas	2013	Longman	Part 5 of Paper 1 from Tests 1 to 6	60
Cambridge English PET 7	2011	Cambridge UP	Part 5 of Paper 1 from Tests 1 to 4	40

[4] More information about the respective validation processes can be found on Cambridge Assessment English (2020) and Escola Oficial d'Idiomes Barcelona Drassanes.

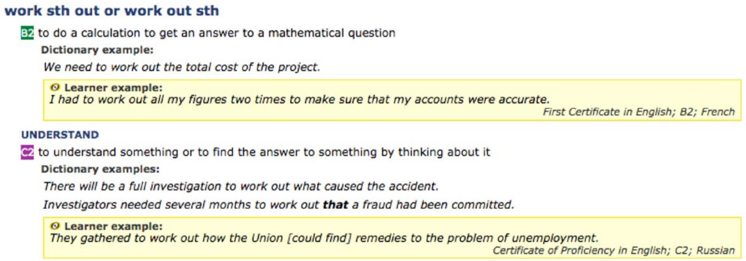
Source	Year	Publishers	Exams	N Questions
Cambridge English PET 8	2014	Cambridge UP	Part 5 of Paper 1 from Tests 1 to 2	20
EOI English B1 test	Paper sample	EOI	Use of English, Part 1 and Part 2	30
EOI English B1 test	Online sample	EOI	Use of English, Part 1 and Part 2	30
Total Multiple-choice Questions Analysed				240

5.3. Methodology

The analysis of the questions of both the control corpus and the disputed exam combines a qualitative and a quantitative approach. The thorough qualitative analysis carried out is characterised by the exploration of the grammatical structures and vocabulary items that are key to successfully answering each of the questions analysed. It is supported by the use of EP, which determines the level of the grammatical constructions and vocabulary items involved in terms of the CEFR bands.

Figure 2 shows a sample of an EVP output. In this case, the expression that was examined is *to work something out*, which has two main meanings. The first one, "to calculate," students are expected to know at B2 level, as shown in figure 2. The second main meaning is "to understand," which students are not expected to know until C2 level. This demonstrates that EVP not only determines which lexical items are acquired at each stage, but also which semantic nuance of a given lexical item is acquired at different stages. According to Agnieszka Leńko-Szymańska, this is one of the main innovative aspects of EVP, since "CEFR levels are assigned not to words, but to their individual meanings and to the recurrent expressions these words are part of. In this way, the EVP does justice to the current models of L2 vocabulary acquisition, which clearly point to the fact that vocabulary learning is incremental and a word is not learned with all its meanings and other pertinent information on a single encounter" (2015, 122).

FIGURE 2. A sample of the output provided by EVP



work sth out or work out sth

B2 to do a calculation to get an answer to a mathematical question

Dictionary example:

We need to work out the total cost of the project.

Learner example:

I had to work out all my figures two times to make sure that my accounts were accurate.

First Certificate in English; B2; French

UNDERSTAND

C2 to understand something or to find the answer to something by thinking about it

Dictionary examples:

There will be a full investigation to work out what caused the accident.

Investigators needed several months to work out ***that*** *a fraud had been committed.*

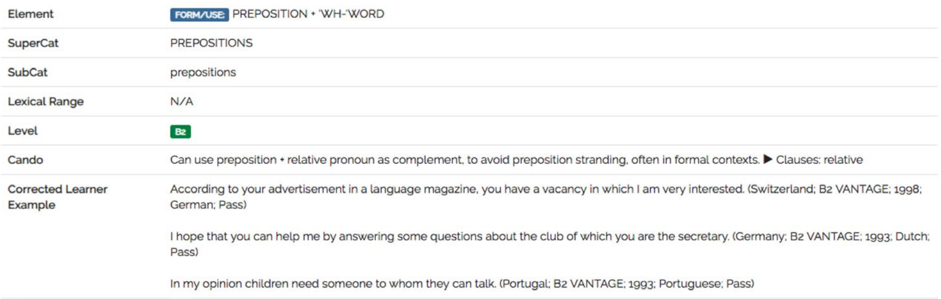
Learner example:

They gathered to work out how the Union [could find] remedies to the problem of unemployment.

Certificate of Proficiency in English; C2; Russian

Figure 3 shows an example of an EGP output. It corresponds to the grammatical construction "preposition followed by *wh*-word" —e.g., *a vacancy in which you are interested* —which is expected to be correctly used at B2 Level.

FIGURE 3. A sample of the output provided by EGP



Element	FORM/USE: PREPOSITION + 'WH-'WORD
SuperCat	PREPOSITIONS
SubCat	prepositions
Lexical Range	N/A
Level	B2
Cando	Can use preposition + relative pronoun as complement, to avoid preposition stranding, often in formal contexts. ▶ Clauses: relative
Corrected Learner Example	According to your advertisement in a language magazine, you have a vacancy in which I am very interested. (Switzerland; B2 VANTAGE; 1998; German; Pass) I hope that you can help me by answering some questions about the club of which you are the secretary. (Germany; B2 VANTAGE; 1993; Dutch; Pass) In my opinion children need someone to whom they can talk. (Portugal; B2 VANTAGE; 1993; Portuguese; Pass)

The determination of the level of the questions in the data analysed is based on two main criteria. Firstly, the CEFR level of the correct option, that is, whether the correct option—grammatical construction or lexical item—is classified by EP as B1, even if some of the distractors have a higher level than B1. We considered that a candidate with a B1 level should be able to answer correctly when the correct answer is of B1 level or lower, even if they do not know the meaning of some of the distractors because they are of a higher level. Secondly, the vocabulary used in the text of the question, that is, whether the items in the question text— rather than the answers—were also categorised as B1 or lower by EVP. In this sense, if the question contained a key item of a higher level that was essential for its comprehension, then it was considered to be of a higher level, since the learner would not be able to choose appropriately if they were not able to understand the question.

Following the qualitative analysis, a quantitative approach involving descriptive and inferential statistics was applied, the aim of which was to identify and quantify the questions that require a level higher than B1. As regards the inferential statistical analysis, forensic linguists operate on the basis of two hypotheses: the null hypothesis, which states that there are no differences between the disputed sample and the known sample—in this case the control corpus—and the alternative hypothesis, which determines that there are statistically significant differences between both samples.

One of the most frequently used statistical methods in forensic linguistics whenever both the dependent and the independent variables are categorical is the chi-square test (Garayzábal Heinze et al. 2019, 53). In the present study, the corpus is the independent variable and the CEFR level of each question is the dependent variable. The objective of the chi-square statistic in this particular analysis is to test whether any observed differences between the number of questions of a higher than B1 level in the control corpus and the disputed exam can be attributed to variation produced by chance or to the effect of the independent variable, i.e. the corpus. The formula for the chi-square statistic is as follows:

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

f_o is the observed frequency and f_e is the expected frequency if no relationship existed between the variables. Thus, the chi-square statistic is based on the difference between what is actually observed in the data and what would be expected if there was no relationship between the variables. The greater the difference between the observed and expected frequencies, the less likely it is that the observed differences are due to chance.

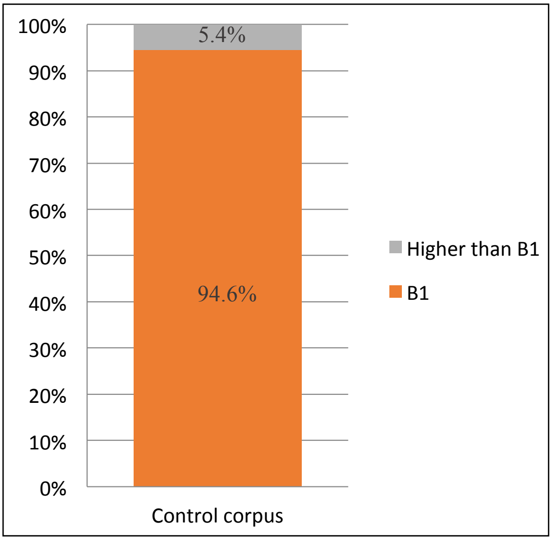
A quantitative approach makes it possible to reliably compare the proportion of beyond B1 level questions in the corpus and the disputed exam, and in turn provides a visual representation by means of figures and graphs, which is essential in a forensic context where judges require results from expert witness reports to be clear and easily interpretable. The combination of both a qualitative and a quantitative approach is fundamental in forensic linguistics, as it is in many other disciplines, since quantitative results cannot provide enough information by themselves and rely on the experts' interpretation on the basis of their knowledge, an interpretation that is necessarily qualitative.

6. Results

This section presents the results of the analysis of the questions included in the control corpus and in the disputed exam. First, the percentages of questions from the control corpus that are classified into the different CEFR levels by EP are provided. These results allow the LAT of B1 questions that should be found in a B1 multiple-choice test to be established. Secondly, the results of the analysis of the disputed exam are presented, which brings to light the considerable difference in the percentage of questions of a level higher than B1 in the disputed exam in comparison with the control corpus. Thirdly, a comparison of the results of the analysis of the two data sets is carried out, and the LAT established through the analysis of the control corpus is used to determine whether the disputed exam was appropriate for its intended level or not. Finally, a chi-square test is used to test the validity of the qualitative analysis.

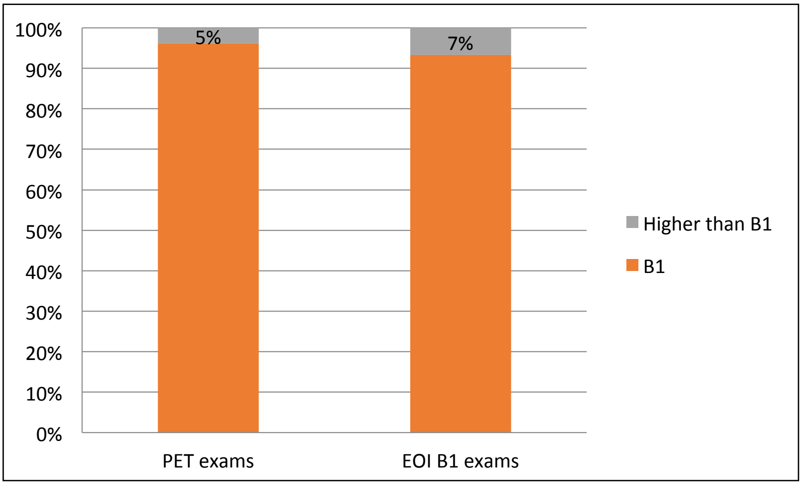
Figure 4 shows the percentage of questions from the control corpus that are classified by EP as being either B1 or lower, or higher than B1. The analysis reveals that out of the total 240 questions, the vast majority are classified as B1 or lower—227, 94.6%—whereas 13 questions—5.4%—are above the intended level of the test, of which 12 were classified as B2 and 1 as C1.

FIGURE 4. Control corpus: percentage of questions of B1 level or lower and of a level higher than B1



If we analyse the two types of exams that make up the control corpus separately—PET exams on the one hand and EOI exams on the other—these percentages are very similar across them: 5% in questions from the Cambridge PET exams and 7% in the EOI B1 exams (figure 5).

FIGURE 5. Control corpus: percentage of questions of B1 level or lower and of a level higher than B1 in the two types of exams



In light of these results, it can be stated that for a B1 multiple-choice English test to be appropriate, between 93 and 95% of questions should be classified by EP as B1 or lower when analysing their core grammatical and vocabulary items, with only 5 to 7% of questions of a higher level being acceptable. This provides evidence for a LAT that can have forensic implications in disputed exam cases.

The subsequent step is to apply the same methodology to the questions from the disputed exam. The analysis of the 50 questions in the disputed test shows that two of them contain key grammatical constructions that are classified by EGP as B2, either the target construction or one of the constructions in the text providing the context: in (1) *I wish* in the context of an imaginary past with a past perfect verb tense, and in (2) future continuous used in *wh*-questions:[5]

(1) We will have to wait for seven hours for our train.
I wish we … an earlier train!
a) booked b) had booked c) would book

(2) Who will you be … to in your new position?
To the Customer Agent directly.
a) involving b) reporting c) liaised

Before reporting the results from the analysis of the vocabulary by means of EVP, it is important to mention that the specific context for this exam was taken into consideration when conducting the analysis. As mentioned earlier, the test was used by a transport-related state-run company in Spain, so it is reasonable to expect applicants to be familiar with some specific terminology related to the transport sector, even if some of the technical lexicon is of B2 level. Examples (3) and (4) show the two questions in the exam that contain transport-related vocabulary regarded as B2 by EVP: the term *container* in (3) and the term *upgrade* in (4). Despite their higher level—one proficiency band—these terms are considered acceptable for a B1 exam.

(3) Before walking through the metal detector, you must take all the metal items out of your pocket and put them on a …
a) board b) flat container c) plate

(4) Is it really the first time you travel first class?
Yes, that's right. But a couple of years ago I was … to business class.
a) exchanged b) upgraded c) promoted

[5] For reasons of confidentiality, the questions have been partially modified without any significant changes.

However, when the level of a transport-related term is more than two proficiency bands above the test's supposed level, it is considered inappropriate. (5) and (6) are examples of such questions, where the terms *to retrieve* and *to divert* are classified by EVP as C2. We consider that B1 candidates cannot, and should not, be expected to know such expressions even when working in a transport-related context. Successful communication in a nontechnical job involving a B1 English level should be able to rely on simpler and more level-appropriate equivalent expressions such as *to get* or *to change destination*.

(5) Passengers are required to … their luggage as soon as it is available on the conveyor belt.
a) revise b) relieve c) retrieve

(6) Because of the fog, they had to close the airport and … the relief flight elsewhere
a) distract b) deflect c) divert

The results of the EVP analysis show that the disputed test has a considerable number of questions including key vocabulary of a level above B1—in some instances even corresponding to vocabulary expected to be known at C1 or C2 levels—that most often appears in the correct option. Examples (7) to (12) show some of these questions.[6] (7) and (8) contain the B2 expressions *to make up for* and *to go off*. (9) and (10) include verbal constructions and expressions that are classified as C1: *give rise to concern* and *to unwind*. Finally, (11) and (12) display C2 vocabulary items: *not to have the faintest idea* and *to put something down to*. None of these expressions—nor many others in the disputed questions—are justified by the transport-related context of the test.

(7) Please, kindly accept this present. We hope it will make … for some of the inconvenience the company has involuntarily caused you. (B2)
a) up b) over c) away with

(8) You'd better not drink that soda; it's … (B2)
a) been off b) become off c) gone off

(9) The latest results gave rise to … that there would be more redundancies. (C1)
a) concern b) nuisance c) trouble

(10) I get so stressed at work that sometimes it is hard to… (C1)
a) undergo b) unwind c) untie

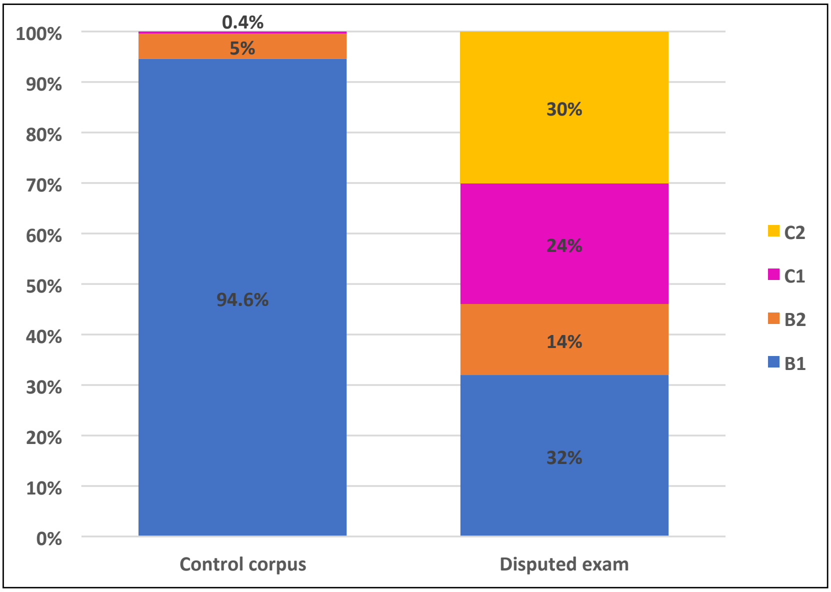
[6] For reasons of confidentiality, the questions have been partially modified without any significant changes.

(11) I'm afraid I haven' [*sic*] the… idea why all trains are being delayed today. (C2)
a) smallest b) weakest c) faintest

(12) After the investigation, the accident was put down to human … (C2)
a) mistake b) slip c) error

The quantitative analysis of the questions reveals that a total of 34 out of 50 are classified by EP as having a level higher than B1, which is 68 % of the whole exam. More specifically, 7 questions are classified as B2, 12 as C1 and 15 as C2. As a matter of fact, there is about the same number of B1 questions (16) as C2 ones (15). The corresponding percentages can be seen in figure 6, which shows the distribution of the questions in the disputed exam according to the CEFR levels in comparison with the questions in the control corpus:

FIGURE 6. Control corpus and disputed exam: percentage of questions of level B1, B2, C1 and C2



Bearing in mind the LAT established in the analysis of the control corpus, which was a percentage of 93 to 95 of questions classified as B1 or lower, it is clear that the percentage of B1 questions in the disputed exam (32%) is considerably lower.

The chi-square test of independence performed to examine the relationship between type of corpus—control corpus and disputed exam—and the CEFR level of questions was found to be statistically significant, (X^2 (1, N = 290) = 119.342, *p* = 0.0000). This therefore confirms that such a high percentage of questions in the disputed exam that were above the CEFR level that it was meant to be testing cannot be attributed to chance alone.

7. Discussion and Conclusion

The context of the present study is a new area of application of forensic linguistics that has arisen over the past few years in Spain—determining the appropriateness of the level of English multiple-choice tests used in selection processes for jobs, sometimes in the public sector, in order to test whether the candidates have the English level stipulated as a requirement. In the case study presented in this article, the English test was disputed in court by numerous applicants on the grounds that it was not appropriate for its intended level.

The results of the EP analysis of the control corpus showed that only between 5 and 7% of questions were classified as higher than B1 in the two types of exams analysed. From these results, we can establish that a LAT of between 93 and 95% of questions with a B1 level should be found in a B1 multiple-choice test that is correctly aligned with the CEFR. In contrast, the results from the analysis of the disputed exam show only 32% of the questions to be classified as B1. In fact, as well as several questions being classified as B2 and C1, a rather surprising 30% were classified as C2—almost the same percentage as B1. A comparison between the LAT established for the control corpus and the percentage of B1 questions in the disputed test objectively demonstrates that the disputed exam is far from appropriate to test its intended level, as the majority of questions could not be answered correctly by candidates with a B1 level.

The implications of the results of the present study are twofold. On the one hand, they provide further empirical evidence of the reliability of EP to align the level of a written text with a CEFR proficiency band, confirming previous claims about its usefulness in determining the level of language learning and testing tools (Harrison and Barker 2015; Leńko-Szymańska 2015; Arslan and Erarsland 2019). Moreover, whereas these studies applied EP to a language teaching and assessment context, ours, to our knowledge, is the first and only one to apply EP to an exam level being disputed in court. On the other hand, the conclusions of the present study also have a very practical application, since it has been demonstrated that EP can be a valuable tool for forensic linguists acting as expert witnesses in cases involving a dispute over the level of an English test. In this sense, the particular nature of forensic contexts, where linguist expert witnesses need to present their results to nonexpert judges in a clear and easily interpretable way, promotes the use of data-driven baselines or thresholds. In the area of plagiarism detection, a similarity threshold (Turell 2008; Marquina and Queralt 2014) has already been established, which helps the linguist expert witness to determine the point at which the similarity found between two texts begins to suggest plagiarism. In the area of exams being disputed in court, the LAT can help the linguist expert witness to ascertain the acceptable percentage of questions in a test for a particular level of English that can be of a higher level. Ultimately, this threshold can determine whether a specific English test has been correctly aligned with the CEFR and is appropriate to test its intended level.

The present study is not without its limitations. The control corpus contains questions from two types of exam and the number of questions from each type is quite unbalanced given the lack of availability of samples for the second type. The reason for the inclusion of the limited EOI data was the fact that they emanated from a similar context to that of the disputed exam. The results, however, show that the LAT is very similar regardless of the type of the exam, which suggests that the LAT is practically the same in any test, whether developed for an international or a national target, provided they are aligned with the CEFR. Therefore, the inclusion of the EOI exams may not have been as critical as initially envisaged.

Moreover, the present study provides a LAT for a very specific type of English exam, a B1 test based solely on multiple-choice questions—the type of questions involved in the disputed exam—whereas other aspects of assessment, such as listening or reading activities, were not explored. In this sense, further research should be carried out to establish the LATs for these types of activities too. As regards the CEFR bands, the results obtained here are only generalisable to B1 multiple-choice tests, since it is possible that the LAT might be different for other CEFR bands. More research is needed to establish whether the LAT is, in fact, the same for all the CEFR bands or whether it needs to be adapted to each of them. Finally, further research could also be conducted in other languages for which RLDs—such as EP for English—have been developed in order to establish language-specific LATs that would likely prove useful to linguist expert witnesses working in those languages.

Works Cited

Archer, Dawn et al., eds. 2003. *Proceedings of the Corpus Linguistics 2003 Conference*. UCREL Technical Paper, 16. Lancaster: Lancaster University.

Arsland, Abdullah and Ali Erarslan. 2019. "Lexical Analysis of a Textbook Based on the EVP." *International Journal of Languages' Education and Teaching* 7 (1): 1-12.

Banerjee, Jayanti and Dina Tsagari, eds. 2016. *Contemporary Second Language Assessment*. London: Bloomsbury.

Barker, Fiona. 2010. "How Can Corpora Be Used in Language Testing?" In O'Keeffe and McCarthy 2010, 633-45.

—. 2014. "Using Corpora to Design Assessment." In Kunnan 2014, 1013-28.

Callies, Marcus and Sandra Götz, eds. 2015. *Learner Corpora in Language Testing and Assessment*. Amsterdam and Philadelphia: John Benjamins.

Cambridge Assessment English. 2020. "Validity and Validation." [Accessed online on October 30, 2019].

Cambridge University Press. 2015. *English Profile – What the CEFR Means for English*. [Accessed online April 2, 2019].

Carlsen, Cecilie. 2012. "Proficiency Level: A Fuzzy Variable in Computer Learner Corpora." *Applied Linguistics* 33 (2): 161-83.

Coulthard, Malcolm. 2004. "Author Identification, Idiolect and Linguistic Uniqueness." *Applied Linguistics* 25 (4): 431-47.

Coulthard, Malcolm and Alison Johnson, eds. 2010. *The Routledge Handbook of Forensic Linguistics.* London and New York: Routledge.

Council of Europe. 2001a. *The Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge UP. [Accessed online on October 30, 2019].

—. 2001b. *Structured Overview of all CEFR Scales.* [Accessed online on October 30, 2019].

—. *Reference Level Descriptions (RLDs) (Language by Language).* [Accessed online on October 30, 2019].

—. 2020. *Reference Level Descriptions (RLDs) Developed so Far.* [Accessed online on October 30, 2020].

Cushing, Sara T. 2017. "Corpus Linguistics in Language Testing Research." *Language Testing* 34 (4): 441-49.

Escola Oficial d'Idiomes Barcelona Drassanes. "Mission, Vision and Values." [Accessed online on October 30, 2019].

Garayzábal Heinze, Elena, Sheila Queralt Estévez and Mercedes Reigosa Riveiros. 2019. *Fundamentos de la lingüística forense.* Madrid: Síntesis.

Gibbons, John and M. Teresa Turell, eds. 2008. *Dimensions of Forensic Linguistics.* Amsterdam and Philadelphia: John Benjamins.

Grant, Tim. 2007. "Quantifying Evidence in Forensic Authorship Analysis." *The International Journal of Speech, Language and the Law* 14 (1): 1-25.

Harrison, Julia and Fiona Barker, eds. 2015. *English Profile Studies 5. English Profile in Practice.* Cambridge: CUP.

Knight, Ben. 2015. "Applications of English Profile." In Harrison and Barker 2015, 93-105.

Kunnan, Antony John, ed. 2014. *The Companion to Language Assessment.* Vol 2, *Approaches and Development.* Hoboken, NJ: Wiley-Blackwell.

Lavery, Urban A. 1921. "The Language of the Law". *American Bar Association Journal*, 7 (6), 277-83.

Leńko-Szymańska, Agnieszka. 2015. "The English Vocabulary Profile as a Benchmark for Assigning Levels to Learner Corpus Data." In Callies and Götz 2015, 115-40.

Ley Orgánica 8/2013. *Boletín Oficial del Estado*, December 10.

Marquina, Montse and Sheila Queralt. 2014. "Similarity Threshold to Detect Plagiarism in Spanish." *RAEL: Revista Electrónica de Lingüística Aplicada* 13 (1): 79-95.

Montolío, Estrella. 2012. "La modernización del discurso jurídico español impulsada por el Ministerio de Justicia. Presentación y principales aportaciones del Informe sobre el lenguaje escrito." *Revista de llengua i dret* 57: 95-121.

NICHOLLS, Diane. 2003. "The Cambridge Learner Corpus: Error Coding and Analysis for Lexicography and ELT." In Archer et al. 2003, 572–82.

O'KEEFFE, Anne and Michael McCarthy, eds. 2010. *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge.

PARK, Kwanghyun. 2014. "Corpora and Language Assessment: The State of the Art." *Language Assessment Quarterly* 11 (1): 27-44.

POBLETE, Claudia, Lisbeth Arenas, Alejandro Córdova, Emmy González and Daniela Tapia. 2018. *Estrategias en comprensión del discurso escrito en contextos jurídicos*. Valparaíso: Ediciones Universitarias de Valparaíso.

STYGALL, Gail. 2010. "Legal Writing: Complexity". In Coulthard and Johnson 2010, 51-64.

TIERSMA, Peter. 1999. *Legal Language*. Chicago, IL: U of Chicago P.

TURELL, M. Teresa. 2007. "Plagio y traducción literaria." *Vasos Comunicantes* 37 (1): 43-54.

—. 2008. "Plagiarism". In Gibbons and Turell 2008, 265-99.

—. 2010. "The Use of Textual, Grammatical and Sociolinguistic Evidence in Forensic Text Comparison." *The International Journal of Speech, Language and the Law* 17 (2): 211-50.

UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE (UCLES) and Cambridge University Press (CUP). 2011. *English Profile: Introducing the CEFR for English Version 1.1*. [Accessed online on April 2, 2019].

UNIVERSITY OF CAMBRIDGE ESOL EXAMINATIONS. 2011. *Using the CEFR: Principles of Good Practice*. [Accessed online on April 2, 2019].

WEIGLE, Sara Cushing and Sarah Goodwin. 2016. "Applications of Corpus Linguistics in Language Assessment." In Banerjee and Tsagari 2016, 209-23.

SOURCES USED IN THE STUDY

ADAMS, Dorothy. 2006. *Cambridge PET: Practice Tests for the Preliminary English Test*. Boston, MA: Cengage Learning.

Cambridge English Preliminary English Test 7: Authentic Examination Papers from Cambridge ESOL. 2011. Cambridge: Cambridge UP.

Cambridge English Preliminary English Test 8: Authentic Examination Papers from Cambridge ESOL. 2014. Cambridge: Cambridge UP.

GENERALITAT DE CATALUNYA. 2019. "Idiomes a les EOI: Anglès." *Estudiar a Catalunya*. [Accessed online on October 30, 2019; no longer available].

HASHEMI, Louise and Barbara Thomas. 2013. *PET Practice Tests Plus*. Harlow: Longman.

Received 25 October 2019 Revised version accepted 25 February 2020

Núria Gavaldà holds a PhD in Linguistic Communication and Multilingual Mediation (specialisation in Forensic Linguistics) from Universitat Pompeu Fabra. She is currently a lecturer in English phonetics and linguistics at Universidad Internacional de la Rioja, and she carries out research on L2 speech, forensic phonetics and forensic linguistics. She has published on phonetics and forensic linguistics and has collaborated in several forensic linguistic cases.

Sheila Queralt holds a PhD in Translation and Language Sciences from Universitat Pompeu Fabra. She is currently the director of the Laboratorio SQ-Lingüistas Forenses. She is author of *Atrapados por la lengua* and the *Decalogue for Requesting a Linguistics Expert Report* (2020), and coauthor of the books *Soy lingüista, lingüista forense* (2019) and *Fundamentos de la lingüística forense* (2019). She collaborates with different Spanish police forces as an expert witness and she is mentor of the National Cyber League organised by the Guardia Civil.