



UNIVERSIDAD DE JAÉN

Facultad de Humanidades y Ciencias de la Educación

Creation of descriptors for the assessment of advanced levels of oral production

Alumna: María Arance Rusillo

Tutores: Prof. D. Antonio Bueno González

Prof. D. Joaquín Manuel Cruz Trapero

Dpto.: Filología Inglesa

Centro de Estudios Avanzados en
Lenguas Modernas (CEALM)

Table of Contents

1. Introduction	1
2. Statement of Purpose	2
2.1 Objectives.....	2
2.2 Specific Research Questions.....	3
3. Theoretical Background	4
3.1 Definition of Concepts	4
3.2 Literature Review.....	7
3.2.1 Annotated Bibliography.....	7
3.2.2 Testing and Correction	7
3.2.3 Introducing Rating Scales.....	10
3.2.4 Descriptors	11
4. Methodology.....	11
4.1 Description of the Research Paradigm.....	12
4.2 Research Tools.....	13
5. Study	14
5.1 Rating Scales.....	14
5.2 The Concept of Native Speaker	17
5.3 Different Proposals of Rating Scales	18
5.4 Descriptors: An Introduction.....	22
5.5 Characteristics of Descriptors.....	24
5.6 Possible Solutions to the Under-definition of Higher-level Descriptors.....	26
6. Conclusion.....	33
6.1 Concluding Remarks	33
6.2 Our Own Contribution	34
6.3 Limitations of the Study	34
6.4 Lines for Further Research	35
7. References	35

Abstract

In standardized language proficiency tests, advanced levels of proficiency have been normally ignored or underdefined, perhaps due to the fact that not many speakers achieve this level of proficiency. However, in any comprehensive assessment continuum it is important to define such levels to allow language learners to obtain a more objective perspective of their actual language abilities. The aim of this Degree Dissertation is to find sources for the definition of the higher levels of proficiency, chiefly those levels that go beyond C2 level in the *Common European Framework of Reference* (Council of Europe, 2001), henceforth CEFR, and to provide this definition as objectively as possible. Accurate definitions of high proficiency levels will be beneficial both for raters, who will have a clearer idea of the target levels, and for the assessed speakers, who will obtain accurate feedback on their performance that is potentially beneficial for the improvement of their language skills.

Key words: Analytic scales, assessment, bands, CEFR, descriptors, dimensions.

Resumen

En los exámenes de competencia los niveles avanzados se han ignorado y han estado poco definidos, lo que puede deberse a que no muchas personas alcanzan estos niveles. Sin embargo, en cualquier tipo de evaluación es importante definirlos ya que la creación de estas rúbricas permite al alumno obtener una perspectiva más objetiva de sus habilidades lingüísticas. El objetivo de este Trabajo de Fin de Grado es encontrar posibles fuentes para definir los niveles superiores de competencia, aquellos niveles que superan el nivel C2 en el *Marco Común Europeo* (Consejo de Europa, 2001), y proporcionar esta definición de la manera más objetiva posible. Este proyecto se considera útil para los examinadores que tienen que calificar a estudiantes avanzados para que estos estudiantes obtengan una puntuación clara en un examen oral, lo que permitirá a los evaluadores proporcionar una retroalimentación más beneficiosa a los estudiantes y esto puede ayudar al progreso de las habilidades de habla inglesa.

Palabras clave: Escalas analíticas, evaluación, bandas, CEFR, descriptores, dimensiones.

1. Introduction

Speaking is a fundamental skill which we use in order to talk to others or even to ourselves. It helps us to express our opinions and to define our identity in the eyes of others. Speaking is arguably at the core of the human ability of language. Nowadays, certain advanced levels of oral proficiency in foreign languages are essential for many activities such as getting access to higher education programs, employment and also for international mobility. Currently, the teaching and assessment of speaking has become a central aspect in language tests, due to the rise of the importance of communicative skills in the educational area. However, speaking is considered to be challenging to assess in a reliable manner. This can be due to the difficulty around the creation of consistent methods that could help to the assessment of the oral skills. The difficulty of creating consistent procedures increases when we are assessing advanced proficiency levels.

The assessment of oral proficiency can be difficult because multiple aspects have to be considered, such as the context, the characteristics of the examinee, the register, and the rating scale, which is also known as rubric or scoring rubric, that the examiner is provided with. For the assessment of oral proficiency, raters normally rely on analytic scales. A quick review of the literature or of existing analytic scales proves that there is a lack of description when it comes to evaluating higher levels of proficiency (Council of Europe, 2001, 2018; Thompson, 2006; O'Sullivan and Dunlea, 2020). When creating these scales, it is important to consider various aspects in order to allow the scoring to be as objective as possible. We will focus on the CEFR scales that the Council of Europe developed (2001, 2018) and we will focus on the problem with the scales in the C2 level, concentrating on the under-defined upper bands of this level. This dissertation will provide different possible solutions to defining higher levels of proficiency. We believe that a complete understanding of the higher levels of proficiency can guarantee a more objective rating process for both the raters and the examinees.

One of the first attempts to develop a more sophisticated system to evaluate learners of languages was produced by the Foreign Service Institute in the United States in 1958 and this has functioned as a prototype for many rating scales created since then (Fulcher and Davidson, 2007: 94). It postulates that learners have a quality which they labelled as 'proficiency' that starts at an elementary level, advances through three intermediate levels and concludes in bilingual or native proficiency, which is defined as the capacity to use the language as it would

be expected from an educated native speaker. This notion at the top end of the scale was employed as the ultimate benchmark in the measurement of language proficiency until very recently (Fulcher and Davidson, 2007: 94).

Our aim in this dissertation is to offer possible solutions with which we can define levels of proficiency beyond C2. Within this project, we do not provide a final solution to this problem, given that a study must be accomplished. Our intention is to raise awareness about it and to state that this issue can be eradicated by offering some possible descriptors that can define or that can be adapted to define the CEFR scales.

For this, we will first describe our objectives in section 2. In section 3 we will provide the reader with some definitions of the most important concepts that we will use within the project. In section 4 we will state what made us think about this problem and how we arrived at this specific research question, we will also provide a description of the research paradigm, and the research tools that allowed us to create this dissertation, in which we include the different scales that will be analyzed. Finally, section 5 contains the core of the project, which is divided into different subsections, in which we develop various aspects such as the concept of native speaker, what characteristics should a descriptor have to be reliable, and the possible solutions that have been analyzed to define higher-level descriptors, among other different things. To conclude, in section 6, we will provide various possible solutions to the shortcomings analyzed, our conclusion to the dissertation, the limitations that the study presents and some lines of research that will go beyond this dissertation. Section 7 contains the bibliographical references that will put an end to this dissertation.

2. Statement of Purpose

2.1 Objectives

As previously mentioned, a common problem of analytic scales for the assessment of language proficiency tests is the definition of levels higher than C2. The CEFR (Council of Europe, 2001: 181-182) recommends building 5-band scales by

Selecting or defining a descriptor for each relevant category which describes the desired pass [...]. That descriptor is the named 'Pass' or '3' and the scale is norm-referenced around that standard [...]. The formulation of '1' & '5' might be other descriptors drawn or adapted from the adjacent levels on the scale.

This means, for example, that in a 5-band scale aimed at evaluating a C2 level, the band numbered as 3 is the cut-off score and therefore this is the band which should contain C2 descriptors. Higher bands, namely bands 4 and 5, must still be defined, obviously with descriptors depicting levels of competence higher than the C2 level.

In the following lines we will list the characteristics that descriptors higher than C2 should contain to be reliable and useful. Furthermore, we will address the need for higher-level descriptors in the CEFR (Council of Europe, 2001) scales. In the absence of such descriptors in the CEFR (*ibid.*), many examining boards find different solutions to this problem. We will analyze these possible solutions to provide valid descriptors for bands 4 and 5 in C2 5-band scales.

The general aim of this dissertation is to raise awareness about the problem of defining higher-level descriptors and try to provide the reader with some feasible options that could be used to solve this problem.

Some of the specific objectives of this dissertation are to provide the reader with a general view of how the higher-level descriptors above C2 could be described by means of providing various options that can help to define these descriptors. We will also give an account of the different characteristics that descriptors should have to be reliable and understandable. We will deal with some controversial topics such as the term native speaker and why it should not be part of the definition of descriptors, and the presence of self-repair in descriptors.

2.2 Specific Research Questions

In language proficiency tests, objective, direct scoring coexists with other forms of scoring. Listening and reading comprehension can be marked through dichotomic decisions (correct or incorrect) which can be even made mechanically through optical recognition devices. However, the scoring of oral and written production normally requires other approaches. Oral and written production cannot simply be deemed as correct or incorrect and it is at this stage when analytic scales become useful.

In this dissertation we will analyze the relevance of proficiency exams and their importance in the learning process. We will consider the differences between objective and subjective scoring, and we will focus on the subjective type of correction, with special attention

to the subjective marking of oral skills (vs written skills), because of their importance in the learning process.

We will address the controversial topic of the native speaker and if this term should be present in the definition of descriptors for rating scales, and we will also deal with the different characteristics that descriptors should fulfil in order to be useful and reliable.

The main research question in this paper is “how can we define higher-level descriptors?” To answer this, we will provide the analysis of the different options available to define these levels, particularly in 5-band scales aiming at evaluating C2, according to the CEFR (Council of Europe, 2001, 2018).

3. Theoretical Background

3.1 Definition of Concepts

There are several concepts which are paramount for our research, namely the definition of rating scales, either analytic or holistic and the differences between them. Other important concepts are those of scoring rubric, band, dimension, native speaker, language proficiency, and descriptor.

The term ‘scale’ and ‘rubric’ are frequently interchangeable. Richards et al. (2002: 471) for example, describe a scoring rubric as “a set of scoring guidelines or criteria used in scoring a test taker’s performance or response to an assessment task (e.g., writing an essay in L2) to make raters’ subjective judgements more reliable.”

Scales (or rubrics) can be either analytic or holistic. An analytic scale, according to the CEFR, focuses on the different aspects of the performance of the learner, for example, Accuracy, Pronunciation, Range, (Council of Europe, 2001: 38). According to Galti et al. (2018: 6), the “analytic scales are rubrics that include explicit performance expectations for each possible rating, for each criterion”. He further states that “analytic scales are especially appropriate for complex learning tasks with multiple criteria”.

In contrast to an analytic scale, the CEFR (Council of Europe, 2001) defines a holistic scale as the scale that offers one descriptor for each level (Council of Europe, 2001: 38). Thus, in this sense, a holistic scale is different from the analytic one.

The scales that are used in the CEFR (*ibid.*) are analytic, because with this type of assessment, we can look at different aspects of the level of a learner separately (Council of Europe, 2001: 190). In this sense, it is argued that the examiner will have to assign separate results to different and independent aspects of the performance of the learner, and this type of assessment is said to encourage the examiner to observe closely (Council of Europe, 2001: 190).

A rating scale also known as scoring rubric or proficiency scale, has been defined by Davies et al. (1999: 153) as:

A scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged. [It] provides an operational definition of a linguistic construct such as proficiency. [They] range from zero mastery through to an end point representing the well-educated native speaker. [They] are commonly characterized in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency, and cohesion) [...]

Scales are chiefly built of bands interacting with dimensions. A band is each of the independent levels of proficiency we can identify withing a scale. A band defines the different levels that the user can achieve depending on the knowledge that the person has of the language; in the CEFR (Council of Europe, 2001) they are known as A1, A2, B1, B2, C1 and C2.

Dimensions are the different columns that we can find in a given scale. The dimensions in the CEFR (Council of Europe, 2001) are used to determine the parameters of communicative activity and language competence in each level. The CEFR (*ibid.*: 7) states that

[The] CEF should differentiate the various dimensions in which language proficiency is described, and provide a series of reference points (levels or steps) by which progress in learning can be calibrated. [...] (T)he development of communicative proficiency involves other dimensions than the strictly linguistic (e.g. sociocultural awareness, imaginative experience, affective relations, learning to learn, etc.).

The term ‘descriptor’ is also a very important concept to consider. According to the CEFR (*ibid.*), a descriptor is the tool that specifies the progressive mastery of the different skills present in the various levels of proficiency (A1, A2, B1, B2...). The descriptors define these levels; in the CEFR (*ibid.*) they were created without references to a specific language, this is what guarantees their relevance and their applicability (Council of Europe, 2001: 17-25). The CEFR (Council of Europe, 2001: 21) further states that descriptors “need to be context-relevant and appropriate for the function they are used for in that context”.

Fulcher (2015: 199) further states that descriptors are interconnected, and they are what define the whole scale. The creation and definition of descriptors has been addressed as a difficult task by many experts (North and Schneider, 1998; Fulcher, 2015).

Richards et al. (2002: 152) defines the term descriptor as:

A description of the level of performance required of a test taker for a specific level or band on a rating scale. [It] can be general, consisting of a short sentence, or fairly detailed, consisting of a paragraph with several sentences.

Moving now to the concept of language proficiency Richards et al. (2002: 292) define it as the “degree of skill with which a person can use a language, such as how well a person can read, write, speak, or understand language”. The language proficiency of a speaker can be measured by proficiency tests. The structure of the descriptive scheme of the CEFR (Council of Europe, 2001) divides overall language proficiency into four different categories (Council of Europe, 2018: 30): general competences, communicative language competences, communicative language activities and communicative language strategies and, ideally, the descriptors of our scales should draw definitions from all of them to make them as comprehensive as possible.

Dealing with language proficiency, and focusing on the higher-level descriptors, it is important to give a definition to the term native speaker, given the controversy around this term when dealing with the assessment of oral production. Richards et al. (2002: 351) define the term as a:

(p)erson who learns a language as a child and continues to use it fluently as a dominant language. Native speakers are said to use a language grammatically, fluently and appropriately, to identify with a community where it is spoken, and to have clear intuitions about what is considered grammatical or ungrammatical in the language (Richards et al., 2002: 351).

Regarding the term native speaker, it is important to consider what the Companion Volume (Council of Europe, 2018: 35) states in the following lines:

The top level in the CEFR scheme, C2, has no relation whatsoever with what is sometimes referred to as the performance of an idealized ‘native-speaker’, or a ‘well-educated native speaker’ or a ‘near-native speaker’. Such concepts were not taken as a point of reference during the development of the levels or the descriptors.

3.2 Literature Review

3.2.1 Annotated Bibliography

The data collected for this study have been retrieved chiefly from the CEFR (Council of Europe, 2001), which is a basic instrument for the elaboration of language syllabi, textbooks, guidelines, and examinations across Europe, created in 2001 as a way to defeat the difficulties between experts that work with modern languages.

The CEFR (Council of Europe, 2001) contains the description of different skills of a lower level, such as A1, A2; that can be suitable for the learner who has a more restricted knowledge of the target language. These lower levels will promote plurilingualism because they encourage the students to learn more European languages by offering formal recognition to their abilities even though they do not possess a high level of proficiency.

We will also use its *Companion Volume* (Council of Europe, 2018), which was published in 2018 and contains extensive descriptors, representing another phase in this process of internalization of the teaching of modern languages. This volume is a complement to the CEFR (*ibid.*) and it is important to mention that the publication of this new volume does not change the significance of the CEFR (*ibid.*).

3.2.2 Testing and Correction

An important book that was used to collect information for this dissertation is Bachman (1990). This book discusses the most important issues concerning the beginning of any language testing struggle. One of the objectives of the book is to offer a conceptual underpinning that helps to answer questions concerning the use and creation of language tests, the foundation includes the context that regulates the uses of language tests, what the disposition of the language abilities that we will measure is, and the nature of the measurement itself. Another important objective for Bachman is to investigate the problem of language being both the instrument and the object of measurement. In this respect, Bachman developed a framework that will help to a better understanding of the aspects that affect language performance of a learner in a language exam.

We have also collected information from Luoma (2004). This book contains a research into language testing and the most appropriate ways of assessing speaking. It is a highly practical book, and it will help us to develop assessment tools for oral skills. It offers advice

and leads to practical solutions to this type of assessments. The author of this book has wide experience in test development particularly in the assessment of speaking, as she has taught testing and assessment to students and practitioners. She acknowledges that assessing speaking is challenging because of the many factors that influence the impression of how well someone can speak a language and recognizes that we expect test scores to be precise, fair, and suitable for the purpose of the examiner (Luoma, 2004: 1).

Useful ideas and keys concepts were also collected from Bueno (1996). This article contains information about English testing, together with the different approaches that have been used in the testing and assessment of languages, focusing on the communicative language testing and how we can create a reliable test. It provides us with the different types of tests that we can use and different ways of assessing these tests. As Bueno (1996: 17) states, testing is very controversial, but it is necessary in the teaching process. It proves the usefulness of the process of teaching. However, when dealing with foreign language testing the problem that we find is that language functions as the instrument and the object of measurement, in other words, we are measuring the language ability that a learner has and for this we are also using language. The tests that we use to assess a learner should be as authentic as possible. Furthermore, when creating and evaluating a test, we should refer to and combine the linguistic components (vocabulary, grammar, pronunciation, phonetics) along with the language skills (speaking, listening, writing, reading, and interaction).

In Bueno (1996) we find that proficiency tests are the ones that measure the overall competence that a learner has in a language. They try to evaluate the ability that a learner has to employ what they have learnt in real-like situations. Thus, they are said to be focused on the future of the learner. Within this type of tests, we find objective tests which are very useful to test at the reception level, and they are easy to mark, we can even use computers in order to save us time; however, objective tests are difficult to write. And subjective tests which, on the contrary, are better for the production level, they are easy to prepare but grading them can be difficult (Bueno, 1996: 33).

Subjective correction is further divided into written production and oral production. Written production, as it names states, consists in the production of a written text by the examinee. According to Bueno (1996: 35), it is one of the easiest assignments to set as a testing exercise. It is within subjective correction since we cannot fully evaluate a written text with a

computer as there are many sociolinguistic and cultural factors that have an important influence the assessment of a written texts.

Another important source for this dissertation Knoch (2009), also mentions the design of some programs created to rate essays, some of them are 'Exrater', 'Project Essay Grader'. 'e-rater' and so on. They were created given the difficulty of finding consistency and reliability in the scores of human raters. Furthermore, computer scoring is quicker than human scoring, something which would provide immediate feedback to the students. However, computers will rate texts objectively and written production texts should be evaluated considering other things apart from grammatical structures, vocabulary, and the errors that they may contain.

Regarding the division between 'subjective' and 'objective' tests, in Pilliner (1968) it is emphasized that 'subjective' tests are differentiated from 'objective' tests entirely in terms of the scoring procedure. In an objective test, the accuracy of the response of the examinee is fully decided by predetermined criteria so that no judgment is necessary on the part of examiners (Pilliner, 1968: 21). The multiple-choice system is the clearest example of an objective test, although other tests can be scored objectively as well. Dictations and cloze tests (a text in which there are gaps every six or seven words and the student has to complete these gaps), for instance, can be scored objectively by providing examiners with scoring answers that accurately enumerate which words are adequate and which are not.

Pilliner (1968) further states that, in subjective tests, the examiner must make a judgment about the correctness of the response based on their subjective interpretation of the scoring criteria. Tests such as the oral interview or the written composition that entail the use of rating scales are essentially subjectively scored, given that there is no possible way to make objective the scoring procedure. Subjective tests are normally evaluated by the use of rating scales or scoring rubrics. These scales help the examiner to identify the strengths and weaknesses of the examinee and they are further defined with descriptors, which will help the examiner to determine the level of a given examinee.

Oral production is also part of subjective correction; however, in this case, the assessment that the examiner must provide has a specified time frame, this means that examiners make decisions in real time, whereas the examiners for written texts do not have this difficulty. Due to this time-bound inconveniency, the methods for assessing this type of tests should be defined and clear enough for the examiner to make decisions. The most common

method is the development of rating scales or scoring rubrics, in which the examiner is provided with a scale, depending on the level that they are assessing.

Fulcher and Davidson (2006) also helped in this study. The book offers us a pioneering and detailed review of an extensive variety of matters such as the practical niceties of test development. In the book, Fulcher and Davidson also explore test development. The organization of the book helps the reader to concentrate on particular topics, such as the assessment of tests.

3.2.3 Introducing Rating Scales

We have compiled information from Fulcher (2015). This article states that the assessment of oral tests is quite recent. Fulcher provides us with a chronological framework which shows the origins and history of speaking assessments.

An important linguist for the creation of this dissertation was Berger (2016, 2020). Just as she states in this article “It is unclear to what extent rating scales describe an implicational continuum of increasing language proficiency that corresponds to what speakers actually do in real speech” (Berger, 2020: 26). Berger (2020) is an investigation of a speaking construct in two different analytic rating scales and to investigate whether these scales show an intensification in the level of speaking proficiency of the learners as they advance in their learning path.

Another reference relevant for the creation of this dissertation is the above-mentioned Knoch (2009). This book tried to develop and evaluate a rating scale for the assessment of diagnostic writing tests and to compare and see if this scale was more reliable and valid than an already-existing scale which contained less detailed descriptors. Even though this book is focused on the assessment of writing, it was of great help, given that writing and speaking are both part of subjective correction. The aim of the study was to “establish if an empirically created rating scale for writing assessment with detailed band descriptors would ensue more valid and reliable ratings than the pre-existing more traditional rating scale” (Knoch, 2009: 14).

One important book that we will include in this section is Isaacs and Trofimovich (2017). This is an edited volume which contains different research discussing relevant issues about the assessment of second language pronunciation. There are several contributors in the book including such important linguists as Davies, Fulcher, Galaczi and Harding, the chapter

“What Do Raters Need in a Pronunciation Scale? The User’s View” by Harding (2017), published in this book, has been of great help for the creation of this dissertation. The study presented the investigation and development of a pronunciation scale.

Harding (2017: 14) states that “field testing of scales [...] may be considered a key step in the scale development or revision cycle, particularly in large-scale testing programs”, and one of the aims of the study was to explore the experiences that examiners had when using a pronunciation scale in order to establish the requests of examiners in the creation of new pronunciation scales.

3.2.4 Descriptors

The works by Trim (2007), and van Ek (1975), which listed the different situations where a learner would be in the need of using the foreign language, were paramount for the development of this project. And, particularly, Wilkins (1978) was quite important for this dissertation, given his contributions to the development of the rating scales in the CEFR (Council of Europe, 2001), because they were part of the group in charge of the Major Project of the Council of Europe in the 70s and 80s.

The first time a possible set of levels or stages was presented in the Council of Europe was in a presentation by Wilkins in 1977 at the Ludwigshafen Symposium, and this represented the first try to advance towards what later was known as the *Common European Framework* in the form of a unit/credit scheme related to joint levels, even though it was not successful.

North and Schneider (1998) report results from a Swiss National Science Research Council project which tried to develop scales of language proficiency. Brian North is a highly recognized linguist, he participated in the creation of the CEFR (Council of Europe, 2001) and on its *Companion Volume* (Council of Europe, 2018) and is known for his work on the development of descriptors.

4. Methodology

The idea for crafting this dissertation and the necessity for the creation of these bands emerges from a part of a subject of the Degree in English Studies, in the Humanities and Educational Sciences Faculty of the University of Jaén. In this subject, the students are assigned to an area within the University in order to put their knowledge into practice.

The students are assigned to different departments and the investigation section of the Center for Higher Studies in Modern Languages (CEALM) is one of them. Here, we were given the opportunity to help in different areas, one of which was the creation of the bands for a 5-band scale of level C1 of oral production.

While defining the bands for the C1 scale (bands from 1 to 5) we found that the more we progressed in the creation of these bands, and the higher the levels we were defining, the harder it was to find suitable descriptors that fitted the level that was required in the band. (See section 2.1)

This problem made us think about the difficulties that one could face when approaching the creation of the bands that are supposed to define the competences of the higher levels of the C2 scale, given that it is the highest level of proficiency present in the scales.

This is why we decided to write this dissertation, in which the different options will be presented and provided to the reader with possible solutions to this problem according to the CEFR (Council of Europe, 2001).

4.1 Description of the Research Paradigm

Concerning the general framework of this dissertation, this is an applied project dealing with the problem of the definition of higher-level descriptors. According to the sources from which we have compiled the information it is a secondary research project given that we have used secondary sources such as articles, books, and electronic publications. The approach that this project presents is analytic, since it focuses on a concrete factor when developing rating scales, which is the definition of higher-level descriptors in oral proficiency scales.

Moving to the data presented in this project it is important to mention that the method of collection of the different sources was non-experimental, since we have used secondary sources. Dealing with the creation of this project, as has been stated above, the decision of focusing on this research question appeared from our own experience while dealing with high-level rating scales (See section 4).

4.2 Research Tools

For this project we have used resources such as the bibliography already indicated in section 3.2. In addition to this section, this project contains a reference section in which we include every article or book that has been used for the purpose of the project (see section 7).

Furthermore, we have used different scales to give several possible solutions to the main research question in this dissertation. The scales presented are the following ones:

The scales by North (2007) define levels ‘D1’ also labelled ‘ambilingual proficiency’, ‘D2’ or ‘genuine bilingual’ and ‘E’ or ‘WENS: Well-Educated Native Speaker’. These levels present features that are above the C2 level in the CEFR scales.

The following scale was part of the PETRA (2016) project, the reason behind the development of this scale was the creation of descriptors for the different levels of competences that a literary translator can fulfill, in which levels 4 and 5 are above the C2 level. These descriptors can help us to define the higher-level descriptors in the CEFR (Council of Europe, 2001) scale, given that they deal with the description of a quite advanced level of proficiency.

Another scale which could be used for the definition of higher-level descriptors is the ILR (Interagency Language Roundtable, 1985) scale, whose third level corresponds with the C2 level in the CEFR (Council of Europe, 2001). This scale presents 5 levels and a plus level in between each of them. Consequently, we will find a 4+ level between levels 4 and 5.

This scale deals with higher levels of proficiency, given that level 3 is interpreted as a level difficult to obtain even for a native speaker, who will have to learn their own language and refine their skills, develop an educated vocabulary (see section 5.6).

The ACTFL Proficiency Guidelines (Swender and Vicars, 2012) could also help to the definition of these descriptors, these guidelines are not presented in the form of a scale; however, they are divided into 5 sections, the name of the highest one being ‘distinguished’, which can help to define bands 4 and 5 of the CEFR scales. When examining this section further we will state whether it could or could not be used as a definition for higher-level descriptors.

The two analytic scales created by the Austrian English Language Teaching and Testing (hence ELTT) are also offered as a possible solution in this project (Berger, 2020). These scales present five levels, two of which are above the C2 level in the CEFR (Council of Europe, 2001).

These levels are labeled as ‘advanced academic’ and ‘full academic proficiency’ and they can help to define the higher-level descriptors of CEFR (*ibid.*) C2 scale.

5. Study

This section will contain a brief introduction in which we will highlight the importance of carrying exams which state the level of proficiency in which we can find students of a language. Following this, we will focus on proficiency tests and their importance in the learning process. An exam can be corrected objectively and subjectively, thus we will define these two terms and we will focus on the subjective correction of tests, which is further divided into written and oral production. Finally, we will focus on the assessment of oral production.

The data collected for this study is part of the target of the Council of Europe to guarantee the right to inclusive education to all nations. The data present in the CEFR (Council of Europe, 2001, 2018) eases the recognition of skills that a student can acquire in different learning backgrounds and this helps European mobility. The two volumes try to separate language proficiency into different constituents which allow communication among human beings. They describe what a language learner has to acquire for them to use a language for communication and define the different levels of proficiency which determine the progress of the learners at the different stages of the learning progression, they also provide the way for teachers, course designers, teacher trainers etc. to show how they are working. In this way, they promote the precision and clarity of syllabuses, aptitudes, and courses, which facilitates to the encouragement of international collaboration through modern languages.

5.1 Rating Scales

In this section we will focus on rating scales and how to develop them by using a certain number of steps, we will also discuss the number of levels that a scale should present in order to be feasible and realistic. Our main concern will be the CEFR (Council of Europe, 2001) scale which presents six different levels, and we will focus on the top level, the C2 level. We will discuss the controversy that has been present around the idea of native speaker and its use on descriptors. Following this discussion, we will focus on the development of higher levels of proficiency going beyond the C2 level, concentrating in the scale created in 1978 by Wilkins in which there are seven levels, and the controversy around the top level on this scale. This project, just as other different articles such as North (2007), will confirm the existence of life

beyond the C2 level and we will see an example of this. We will comment on one of the functions that the *Companion Volume* (Council of Europe, 2018) has.

Furthermore, a distinction of the different scales will be given, and we will comment on important features that descriptors should have, such as being objective, including the use of self-repair in the learner, the issue of pronunciation. This dissertation also contains the different characteristics that the Council of Europe considers important to contemplate when defining descriptors and as a final part of this section we will provide the reader with possible solutions for the definition of higher levels of proficiency for the C2 level in the CEFR (Council of Europe, 2001) scale.

There are different rating scales that are divided into the different levels that a learner can achieve, the scales created by the CEFR (Council of Europe, 2001) have 6 levels which go from A1 to C2. We will focus our attention in the C2 level, which is further separated in bands. These bands mark the different levels that the test taker can achieve, going from 1 to 5, and they are filled with descriptors. In the case of the CEFR (*ibid.*) the descriptors for oral proficiency are range, accuracy, fluency, coherence, and interaction. Descriptors define the competences and the features of a learner's performance; they will define a higher or lower level of proficiency.

The main question in this project will be the analysis of the different options available to define higher level bands of oral proficiency, the bands 4 and 5 of C2 scales according to the CEFR (Council of Europe, 2001). The scale contains 5 bands for each level of proficiency, 1 being the lowest and 5 being the highest, 3 being the mark that a student needs to pass the exam as mentioned above; we will focus on the highest levels of the C2 scale given that they have not been defined yet, and we will try to provide a solution to this problem.

When creating a rating scale, as Weigle (2002: 122-125) describes in a number of steps that can be helpful to define it, we have to bear in mind the type of rating scale that we will use, who is going to put the rating scale into practice, how the scores should be reported and the question which we will focus on in this study, how we should formulate the descriptors of the scale.

Regarding the creation of rating scales, the protocol created and tested by Cruz (2016: 154) is the following one. First of all, we find the previous considerations that we must bear in mind when creating a rubric, such as deciding if it will be holistic, analytic, or holistic trait,

identify and describe the linguistic features it will assess, the number of bands it will contain, and how the scores will be reported. After this stage is completed, we should compose the descriptors, in this concrete case, we would have to define the higher-level descriptors of the CEFR C2 level, once the descriptors are defined, we must validate them, first in a qualitative way and secondly in a quantitative way. This will allow the descriptors to be formulated in a clearer and more reliable manner. Once this validation process is complete, we should implement the descriptors by making use of the rubric in a real exam and collect the information and draw some conclusions. After this step is completed, a revision should take place in order to improve the rubric if it is necessary. As a final part, Cruz (2016) suggests the repetition of the implementation stage as many times as needed.

The type of scales used by the CEFR are analytic scales, they are very useful for diagnostic testing because they provide more information about the abilities of the examinee. This type of scales, just as Galaczi (2011: 221) affirms, are scales that take advantage of the strengths of a range of intuitive, qualitative, and quantitative approaches. Galaczi (2011: 221) further comments that the best methods for rating scale development are said to take advantage of the strengths of these three approaches. And the CEFR advocates “a complementary and cumulative process” (Council of Europe, 2001: 207) that brings together all the three.

We will now focus on the number of levels that a rating scale is supposed to have, the levels that are normally developed by applied linguists or teachers are intended to be learning objectives in a class, or a specific course. Thus, the decision about the total amount of levels that a given examiner has to refer to depends generally on the context in which they would be employed. If we look at this process with a certification perspective, these levels should be suitable to comprise the extent of competencies and mandatory aptitudes for the target population, however, just as the CEFR states, they “should not exceed the number of levels between which people are capable of making reasonably consistent distinctions” (Council of Europe, 2001: 21).

In the CEFR (Council of Europe, 2001) there is a six-level system labelled from the lowest level, A1 to C2, but C2 is not the highest imaginable level for proficiency in a foreign language. As the CEFR (*ibid.*) mentions, the C2 level can “be extended to include the more developed intercultural competence above that level which is achieved by many language professionals” (*ibid.*: 23). ‘Mastery’, as the C2 level is frequently referred to in the CEFR

(*ibid.*), is not intended to imply native-speaker or near native-speaker competence. What is intended is “to characterize the degree of precision, appropriateness and ease with the language which typifies the speech of those who have been highly successful learners” (Council of Europe, 2018: 35). The descriptors that are calibrated in the CEFR (*ibid.*) for the C2 level include:

[The speaker is able to] (c)onvey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of modification devices; has a good command of idiomatic expressions and colloquialisms with awareness of connotative level of meaning; backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it. (Council of Europe, 2001: 36)

Berger (2015: 289) claims that another feature of the C2 level is the capacity that a learner has for actively managing the interaction between the speakers which are present in a conversation.

5.2 The Concept of Native Speaker

This leads us to the controversy that has been around the definition of the term native speaker. In the CEFR (Council of Europe, 2001), we can find the use of the expression native speaker in the descriptors that define the superior levels of proficiency in the scales; this occurred in 13 descriptors in the C2 level, and it was erased when its *Companion Volume* (Council of Europe, 2018) was published.

Some scales, such as Foreign Service Institute scale of the United States (Interagency Language Roundtable, 1985), “consider their higher level, level 5, as ‘functionally equivalent to that of a highly articulate, well-educated native speaker’” (Clark and Clifford, 1988: 131). This reference to the native speaker is not only used in this level, but it is also present in the rest of the levels of the scale. However, it could be argued that this scale makes a distinction between higher levels of speaking proficiency by referring to the term native speaker. The use of this term is controversial because as Bachman (1990: 39) states, native speakers cannot be considered as homogenous, even though there is often an assumption that they are. Native speakers do not have all the same level of proficiency although they are considered as native.

The Interagency Language Roundtable defines the term as:

[The term] ‘Native speaker’ refers to native speakers of a standard dialect. ‘Well-educated,’ in the context of these proficiency descriptions, does not necessarily imply formal higher education; however, in cultures where formal higher education is common, the language-use abilities of persons who have had such

education is considered the standard. That is, such a person meets contemporary expectations for the formal, careful style of the language, as well as a range of less formal varieties of the language (Interagency Language Roundtable, 1985: 1).

There are important problems in defining what kind of language use we should consider as the native speaker norm, while the uncertainty of what establishes a native speaker, or whether we can speak of human beings who are native speakers, is the topic of many discussions. Furthermore, to determine who is a native speaker of the English language we should first determine which variation of the English language we choose as the native speaker standard. This is particularly tricky in English, given its international use. The standards of a language are used in different countries or even regions and this complicates the choice of a particular standard to be the native speaker norm.

It is also important to mention the fact that a native speaker is not a level of proficiency that someone can achieve but an identity. A non-native speaker of a certain language, even though they are really fluent, advanced, and they use the language in a natural manner cannot be a native speaker simply because they were not born into that language and they will hardly be accustomed to the body language, or cultural references that a native speaker of the language has. Thus, even if a learner of a language has a high level of the language and a very educated vocabulary, they can never be native speakers because this is not a level that they can achieve, but an identity that someone is born with.

5.3 Different Proposals of Rating Scales

Moving now to the development of higher levels of proficiency we have to mention Wilkins who proposed a seventh level in 1977. Wilkins was a pioneer in the sense that he was one of the first to use the term ‘communicative approach’ to later define his seven-level scale. The seventh level was part of the scale created as an option to develop a European unit/credit system or scheme. The model that was constructed tried to fulfil all the requirements of a unit/credit scheme, thus, it tried to define all the objectives that can be present in the process of language learning. In the creation of this scheme, Van Ek (1975: 16) states the following:

There are learners who will be satisfied with the ability to maintain themselves physically in a foreign-language environment and to establish superficial social contacts, whereas for others nothing short of native-like command of the foreign language will do. Between these two extremes we may distinguish various levels of ability, all corresponding to the needs of particular classes of learners. The number of

levels to be ultimately distinguished in our system is "largely a matter of practical convenience", as Trim puts it.

It is important to mention that this project was oriented towards adult language learning. This seventh level was thought to be beyond the extent of ordinary education and that is why the CEFR only adopted the first six levels. It was proposed at a 1977 symposium; however, the scale was not adopted by the Council of Europe. Although the unit/credit scheme was never implemented the principles on which it was based were considered to be generally applicable.

The levels that Wilkins created, from which the Council of Europe took only the six first levels, are the following ones (Wilkins, 1978: 72):

- Level 1 Level of Formulaic Proficiency
- Level 2 Level of Survival Proficiency
- Level 3 Level of Basic Operational Proficiency (Threshold Level)
- Level 4 Level of Limited Operational Proficiency
- Level 5 Level of Adequate Operational Proficiency
- Level 6 Level of Comprehensive Operational Proficiency
- Level 7 Level of Ambilingual Proficiency

If we deepen into the contents with which Wilkins defined the seventh level, we find the following descriptors (Wilkins, 1978: 74-78):

[The speaker is] able to understand all forms of speech normally understood by native speaker of similar background, including rapid informal conversation. [The speaker] speaks with facility, fluency and accuracy that is indistinguishable from that of native speaker. This level is normally reached only by people who have lived for a substantial period in the country concerned. Can cope with different styles and handle language needs within same limits as those encountered by native speakers. [The speaker] possesses the writing ability of the educated native speaker in relation to specific objectives. Complete confidence in meeting all demands places on spoken language ability. Facility in switching levels, topics [...] the speaker is able to reformulate some ideas in differing linguistic forms for emphasis, polemics, disambiguation [...] as effective in communication as a native speaker. Not at any disadvantage in use of second language.

Even though this is quite important for the development of higher-level descriptors, these descriptors could not be used nowadays for the definition of the C2 CEFR scale, since Wilkins makes clear references to concept of the native speaker, and this is not a reliable manner to define descriptors.

The CEFR Working Party and the Swiss Project adopted the scale developed by Wilkins except for the top level, also known as 'ambilingual proficiency' and they added a lower level

at the bottom of the scale which later became A1 level. We can find a table containing the differences between the levels created in the Symposium by Wilkins and the CEFR levels, together with other scales such as the Cambridge scale. This table was taken from an article written by North (2008: 32).

Wilkins 1977/8, Ludwigshafen	Cambridge 1992	Council of Europe/ Swiss Project 1992–97	CEFR Levels
Ambilingual proficiency			
Comprehensive operational proficiency	CPE	Mastery	C2
Adequate operational Proficiency	CAE	Effective operational proficiency	C1
Limited operational proficiency	FCE	Vantage	B2
Basic operational proficiency (Threshold Level)	PET	Threshold	B1
Survival proficiency	KET	Waystage	A2
Formulaic proficiency		Breakthrough	A1

Table 1: Differences between levels by Wilkins, CEFR and Cambridge scale. (North, 2008: 32)

We can see that the framework presents in one of its figures a proficient user branch in which the C2 level is divided in two, C2.1 and C2.2, as we can observe in the figure below. The question here is what descriptors we should include in the C2.2 sublevel to define this level which is superior to C2 level.

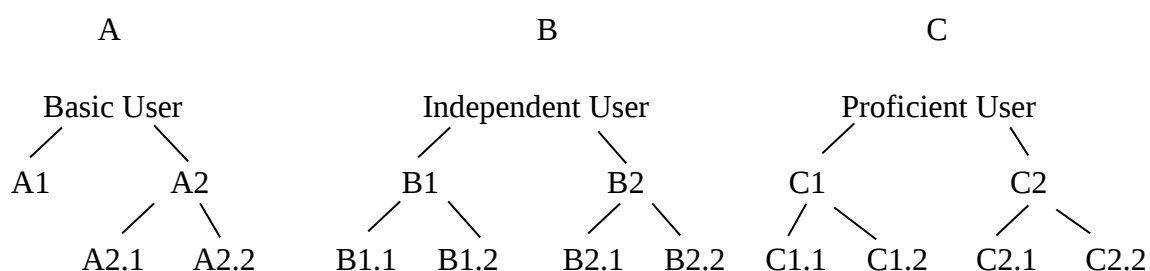


Figure 1 (Council of Europe, 2001: 33)

The Swiss National Research Project confirmed the existence of the seventh level of proficiency and affirmed that there are learners who have clearly a level of proficiency beyond the C2 level, such as simultaneous interpreters and professional translators. This seventh level, named by Wilkins in a symposium in 1977 as stated above, was not thought to be a realistic

objective for formal education and so it was not adopted in the creation of the CEFR (Council of Europe, 2001).

The C2 level of the CEFR (*ibid.*) occupies the third of five levels in the PETRA (2016) project for literary translation. Many writers support this seventh level of ‘ambilingual proficiency’. The term refers to the learner of a language who has a very high level of proficiency without being bilingual from birth. However, in the PETRA (2016) project we find that most of the descriptors that deal with a level of proficiency superior to the equivalent to the C2 level do not present a description. Nonetheless, this project showed that in fact there are superior levels to that of the C2 level. In some cases, the C2 level of the CEFR (*ibid.*) does not correspond with the third level presented in this project, but appears in the second level, reinforcing the fact that the C2 level is not the highest level of proficiency that a learner or user can achieve. One of the possible solutions for the creation of the superior descriptors in the C2 level in the CEFR (*ibid.*) is the use of the PETRA (2016) project. We will discuss the possibility of using these scales in the following sections of this dissertation.

One of the functions that the *Companion Volume* (Council of Europe, 2018) has is that of enriching the description of the level A1 and the C levels, particularly the C2 level. In it, we can appreciate how in some descriptors of C2 there is the entry ‘No descriptor available: see C1’. Here we are recommended to consider whether the learner is making a more demanding use of the descriptor given for the level C1. Even though the *Companion Volume* (Council of Europe, 2018) provides us with a wider picture of descriptors, especially for the C2 descriptors which were adapted to better reflect the competence of these users, we still find some problems in the creation of descriptors for levels which present a superior stage of proficiency than C2.

If we take into consideration the words that North (1996: 20) used to define the challenge of developing rating scales, we can see how he rightly labels it as “trying to describe complex phenomena in a small number of words on the basis of incomplete theory”. Moreover, as Brindley (1998: 116) remarks, it is not always “easy to tell what scale descriptors are meant to describe, what learners ‘ought’ to be able to do at each of the scale levels or what they ‘do’ do”.

The problem that Fulcher (2015) posed in his study is how many statements the student must be able to do before achieving a certain level. In the case of the C2 level, the problem is

that the level is under-defined because it does not contain enough information for us to match descriptors to the level.

This leads us to the challenge that language test developers face when creating a rating scale. The difficulty is divided into two parts as Harding (2017: 12) states in his article, the first problem they encounter is knowing what information they should include in scale descriptors at different levels, and the second problem would be ensuring that the scale created is interpreted appropriately and consistently by examiners. He further affirms that “much of the literature on rating scales and rater behavior has been framed around these two challenges” (Harding, 2017: 12).

5.4 Descriptors: An Introduction

Rating scales are divided into three subtypes according to Alderson (1991: 78) and depending on which one we use the descriptors may be formulated differently.

The first type is *user-oriented scales*, the ones that we can find in the CEFR (Council of Europe, 2001), which are used to report information about typical behaviors of an examinee at a certain level. Here, descriptors are usually formulated as ‘can do’ statements, they also are positively worded, even when we are dealing with low levels. The second type is *constructor-oriented scales*, which help the examiner to select tasks for the test by describing what the examinee can do at a given level, these scales also focus on what the learner can do. The third type is *assessor-oriented scales*, which guide the examiner to rate the examinee, focusing on the quality of the examinee’s performance.

It is important to ascertain which variables are appropriate and practical to assess in a given test situation, just as McNamara (1996) states in his work. This leads us to the question posed above, how we should formulate the descriptors of a scale.

As Fulcher (2015: 199-200) states:

Some scales are a patchwork quilt created by bundling descriptors from other scales together based on scaled teacher judgments. How we define the speaking construct and how we design the rating scale descriptors are therefore interconnected. Design decisions therefore need to be informed by testing purpose and relevant theoretical frameworks.

North and Schneider (1998: 223) affirm that descriptors should not be affected by the people who are creating them and their opinions. The rating scale has to be accepted as more than a description of the people who constructed the scale.

The concept of ‘objectivity’ is very important, given that descriptors should not be determined by the opinion of the author, but they should be established through data. The rating scale has to be objective in the sense that it can be used by different people and with different instruments. As stated by Luoma (2004: 25), assessment developers can use the SPEAKING framework created by Hymes (1974) to make the initial plans for their test, because using this framework can make the analysis of differences more systematic and help to make judgements more informed and avoid making value judgements.

To formulate the descriptors needed for this scale the CEFR (Council of Europe, 2001) offers us the use of descriptors containing ‘can do’ statements. According to the CEFR, these are provided for reception, interaction, mediation, and production (Council of Europe, 2001: 25). These descriptors state what the learner can do at a given level.

If we want to achieve a truly learner-orientated assessment system, as Luoma (2004: 154) states, it is also important in the defining of scale descriptors to consider the learner interpretations of the different descriptors and the differences that appear between different descriptors, even though we must remember that they are very meaningful and important for teachers and applied linguists.

As an important remark in the field of fluency and pronunciation, it is important to say that they have been regarded as the most difficult components to evaluate in a learner, and many of the raters that have taken part in studies of creation of rubrics for the evaluation of oral production have referred to this difficulty (Brown, 2006; Isaacs and Trofimovich, 2017; Luoma, 2014).

Examiners comment on the fact that normally scales miss one important element, they notice the absence of ‘self-repair’ in the descriptors (Harding, 2017: 25). Examiners find this aspect salient in their perceptions of the tests that they carry out. There are reasons why we should include self-repair in the higher-level descriptors of an oral production scale, starting with the fact that self-repair is thought to be a higher-level skill, because the learner has to be aware of their mistakes and has to be able to correct them in a short amount of time.

Nevertheless, they should be reasonably considered, which means that the amount of self-repair present in an oral test should refer only to isolated incidents.

Finally, the score that raters give to a learner should not be dependent on their pronunciation, the learner should not be penalized if they attempt to pronounce words which contain difficult consonant clusters, the ability of self-repair should be lexically dependent, meaning that it should be focused on grammatical errors that the learner is able to mend quickly while having a conversation. We should provide a concrete formulation of the descriptors, because this will allow people who read the scale to interpret the scale as it is supposed to, whether it is read by a professional or an outsider, unlike abstract formulation, which makes the descriptors depend on one another by using words like ‘extensive’ or ‘large’, ‘small’ and ‘limited’. Furthermore, if we define the descriptors with a concrete approach, they will say something, the descriptors would be converted into a checklist of different statements in which the rater would answer ‘yes’ or ‘no’, depending on the performance of the learner.

5.5 Characteristics of Descriptors

When creating new descriptors, it is important to consider different things that the developers of rating scales are concerned with. The Council of Europe provides us with the characteristics that they should have to be clear and useful for the examiner: positiveness, definiteness, clarity, brevity, and independence (Council of Europe, 2001: 205-207).

Positiveness is a characteristic that should be present in the descriptors of scales because it creates the objectives that the learner has to achieve to succeed in a specific level with positive formulations of the descriptors. It can be difficult to provide the lower levels with positive descriptors because it is more difficult to describe what the learner can do in these levels rather than what they cannot do.

Definiteness refers to the concrete tasks or the specific degrees that the descriptors should contain in each level. The descriptors should not be vague and use impressionistic terminology, they should be as objective as possible, given that we are addressing a subjective skill, in other words, they should not contain such words as, ‘almost perfectly’, ‘adequate’, ‘extensive’, ‘few’ or by replacing ‘moderate’ with ‘good’ and ‘excellent’ at the levels above. The problem with vagueness is that these descriptors are accepted and read nicely but this may be covering the fact that examiners are giving them their own interpretation. Descriptors should

be clearly distinguished and to achieve this we may find gaps in the scale because we cannot provide that level with concrete distinctions.

As Knoch (2009: 238) affirms, when defining descriptors for different levels we have to bear in mind that descriptors could mean several things to different examiners. Moreover, we have to be cautious not to mix various aspects of speaking into one scale category, since this can complicate the differentiation of the scale categories from each other. On the other hand, they should not be too precise or specific, the examiner has to be able to use and understand what the descriptor is referring to in order to allow them to apply that descriptor and evaluate correctly. This leads us to the third feature, *clarity*, which refers to the fact that descriptors should be understandable, and they should be written in simple sentences and with a logical organization.

Brevity is concerned with the length that descriptors should have, a descriptor which is longer than a two-clause sentence cannot be fully assessed during a test, given that the examiner does not have much time to read the descriptors and assess the learner at the same time if this descriptor is too long. Raters prefer short definition, not longer than twenty-five words. This is connected with the over-specification of a descriptor, which could provoke the loss of information because they are too specific. Descriptors should not be too long because this may cause some confusion to the examiner and they will not have time to read them and mark the learner simultaneously, while trying to listen to the learner. Descriptors need to be concrete and practical at the same time.

The last characteristic that the CEFR (Council of Europe, 2001) distinguishes is *independence* and it is concerned with short descriptors, which have two advantages, the first one is that they can describe a behavior which can be answered with a yes/no question, therefore definite descriptors can be used as independent statements in a scale. This also means that the descriptor can act as an objective rather than having meaning only by relating a concrete descriptor to other descriptors on the scale.

These characteristics are highly relevant. However, Berger (2016: 41) suggests another three characteristics, which are *uniqueness*, which refers to the fact that descriptors have to be unique in their own level, instead of occurring at different levels in the scale. The second criterion is that of *directness*, which is concerned with the ‘can do’ statements, and she further states that descriptors would function better if they were ‘do’ statements, they would be

understood better if they referred to the appearance of the ability, rather than the principal ability itself. Lastly, the feature of *specificity*, which refers to the need of the scale to be as specific as possible, for this to occur we have to create descriptors which indicate the language activity and the circumstances under which the learner can use it.

This last feature is very important and is related to one of the characteristic present also in the CEFR (Council of Europe, 2001): *definiteness*. Defining descriptors that are as specific as the descriptors that we will try to provide is really difficult because of this level of specificity that we have to achieve, given that the majority of scales are very broad, scales normally focus on the overall performance and on the level itself without defining the different bands that are within those levels and do not concentrate in specific levels to develop them with more details.

5.6 Possible Solutions to the Under-definition of Higher-level Descriptors

Moving to the main research question in this dissertation, we will proceed to provide some possible solutions to the under-definition of higher-level descriptors. We will provide six proposals that could be considered when defining higher-level descriptors. It is important to mention that the solutions that we will offer here are not a definite solution to this problem and that a complete qualitative and quantitative study should be carried out in order to prove the effectiveness of the descriptors.

The first proposal is that of North (2007), who suggests the creation of three new higher levels, which would be placed above the C2 level. These levels are addressed as D1, D2 and E, and they will be commented on in the following pages.

The second proposal is the one developed by the PETRA project (2016), which is concerned with the definition of high-level descriptors for simultaneous and literary translators. It deals with the competences that a translator has.

The ILR (Interagency Language Roundtable, 1985) also presented a scale which could be helpful when defining the higher-level descriptors. This scale was created to measure language proficiency in the United States and shows a high level of proficiency, since, as stated by Brooks and Mackey (2011: 297) even native speakers would need to learn their own language in order to perfectionate and refine the use of vocabulary and the different skills of the English language.

Swender et al. (2012) also provided some guidelines that could be helpful for the definition of higher-level descriptors in the CEFR (Council of Europe, 2001). They offered a description of what learners can do in real-life situations, with guidelines which contain 5 levels, and the more advanced one could help to develop higher-level descriptors in the CEFR C2 scale.

Berger (2020) published a summary of a study containing two different analytic rating scales which added specific details to the reference levels in the CEFR (*ibid.*). The study gave definitions to prominent characteristics at every level and created reference points from the descriptors. We will analyze these descriptors and see if we could use them to define the higher-level descriptors of the CEFR C2 level.

As mentioned above, we will start with the ideas provided by North in a presentation (2007: 14) where he states that there is “life beyond the C2 level”. He created a scale that provided us with levels that were superior of the C2 level. These levels were addressed as ‘D1’, which he also named as ‘ambilingual proficiency’, also known as ‘Language professionals: interpreters, translators, some university professors.’ The next level for him would be that of ‘D2’, which he also addressed as ‘genuine bilinguals.’ And the final level of proficiency that he offers us is ‘E’ or ‘WENS: Well-Educated Native Speaker.’

The apparent ambilingualism or ‘D level’ that North (2007: 18) proposes is defined by him with the following characteristics:

[The speaker can] (c)onvey, elaborate, or translate to explicit expression the nuances and subtleties of their own and of others’ meaning by exploiting a comprehensive knowledge of the language to do so. The speaker can function in all situations to all intents and purposes exactly as the mother tongue; they can use the language in a sophisticated, natural, accurate manner apparently indistinguishable from the performance of a native speaker.

He further states that the C2 is under-defined, given that there are very few descriptors calibrated in the CEFR and also in the Swiss project. He suggests the inclusion of suitable descriptors from DIALANG, ALTE, Catalonia and the Portfolio bank. He proposes trying to avoid the native speaker attributes frequently present in the C2 level, as the Council of Europe proceeded in the *Companion Volume* (Council of Europe, 2018). And he suggests incorporating “insights from qualitative analysis of samples, for instance, Cambridge” (North, 2007: 32). Finally, he states that it is important to “define Level D, at least as an outline, to give upper

boundary” (North, 2007: 32) and the lower bands of Level D, bands 1 and 2 respectively, would be what we could include in the superior bands of Level C2, bands 4 and 5.

There is a problem with the definition of levels such as ‘D’ and ‘E’: this could be challenging in the sense that there would not be many people who could achieve this level of proficiency, given that it is a very advanced level and only simultaneous translators, and some highly educated native speakers may be able to achieve this level. Thus, the creation of these two levels may be useless because they will not be put into practice with frequency. Still, at least one of them should be defined because this would give high-advanced learners of the language the opportunity to obtain a higher level of proficiency than that of C2, and consequently this will help with the development of the higher bands of the C2 level.

Notwithstanding this, the definition of this level should still refer to the non-native accent that the learner can maintain and allowing this feature not to affect the score of the test, given that even though the level of proficiency is very high, the learner can still maintain their accent, and this should not be considered as a negative aspect, but as a sign of identity of the test taker. These more advanced levels should also refer to the use of cultural references by the speaker, and reward them, as well as the use of colloquialisms when pertinent, and idioms. For the descriptors to be concrete enough, we must first investigate some samples of the performance of advanced learners to make sure that they share this concreteness.

The second option that we will present is the above-mentioned PETRA Project (2016), which was concerned with the definition of high-level descriptors for advanced learners (See section 5.3). However, this project is concerned with the definition of descriptors for the different levels of competences that a literary translator can fulfill. Thus, if we were to use these descriptors to define bands 4 and 5 of the C2 level it could pose a problem, given that this project deals with the different competences that a translator is capable of doing. The project is concerned with the competences or ‘can do’s’ of a literary translator and therefore the descriptors are concerned with the different competences that a translator has. Thus, these descriptors would not be adequate to define the descriptors of bands 4 and 5 of the C2 level dealing with oral production.

Another possible solution to this problem would be the adaptation of the descriptors defining the Interagency Language Roundtable scale, henceforth ILR (Interagency Language Roundtable, 1985). In this scale level 3 is equivalent to the C2 level in the scales of the CEFR.

As stated by Brooks and Mackey (2011: 297), even for a native speaker to receive a score superior to that of level 3 in the ILR scale would be a difficult task and they would have to learn their own native language, and develop an educated and defined vocabulary, refining the multiple skills and making use of pragmatic and sociolinguistic strategies that must be culturally suitable and extremely efficient, and exercising the skills in order to expand the smoothness of speech across a wide variety of situations.

The superior levels of the scale are the 4th and 5th levels, also labelled as Speaking 4 (Advanced Professional Proficiency), Speaking 4+ (Advanced Professional Proficiency Plus) and Speaking 5 (Functionally Native Proficiency), in which we find the following descriptors:

The 4th level is defined as being “able to use the language fluently and in an adequate manner on all levels relevant to professional requirements. The use and ability of the language of the learner are effective to perform” (Interagency Language Roundtable, 1985: 4) The learner controls discourse, using suitable rhetorical speech devices, native cultural references, and comprehension. The learner would hardly ever be perceived as a native. They speak without much effort and fluently and are able to use the language with an advanced degree of usefulness and exactitude for all the personal and professional experience and extent of obligations. The learner can be an informal interpreter in some circumstances, they can accomplish language tasks even though they are extensive or complex.

Some examples are that the speaker can talk about “concepts in detail which are essentially different from the target culture” (Interagency Language Roundtable, 1985: 4), they can also make those concepts understandable and manageable to the native speaker. They can understand the details and difficulties of concepts which are culturally different from their own. They can “perform in conferences, lectures, and debates on topics of discrepancy with native speakers” (Interagency Language Roundtable, 1985: 4) and they understand and make changes of the area under discussion and tone. They are “able to understand native speakers talking in major dialects in face-to-face interaction” (Interagency Language Roundtable, 1985: 4).

The 4th level plus has been defined normally as an equivalent to that of a well-educated, highly articulate native speaker, but the speaker would not be recognized as native. Some examples of the actions that the learner can perform in this level are: Learners apply the social and circumstantial knowledge of native speakers; however, they cannot maintain it under all situations. Learners have an “extensive range and control of structures, but an occasional non-

native mistake may happen” (Interagency Language Roundtable, 1985: 4). Learners have an advanced control of vocabulary and terminology that is seldom inaccurate, yet there are sporadic “weaknesses in idioms, popular expressions, pronunciation, cultural references, and some mistakes when performing in a totally native manner” (Interagency Language Roundtable, 1985: 4).

In the level described above, we can observe the use of negative sentences such as ‘weaknesses in idioms’, ‘cannot sustain the performance’. The problem with the use of this negative formulation can lead the rater to mark the learner with a mark below the learner’s proficiency and, given that we are in a very high level of proficiency we should try to avoid the negative formulations at all costs.

The higher level of proficiency of this scale, level 5, is concerned with the speaking proficiency of the learner when this is equivalent to the performance of a highly articulate “well-educated native speaker and reproduces the cultural canons of the people who speak the language natively” (Interagency Language Roundtable, 1985: 4). As stated in the ILR, speakers with this level of proficiency can do the following in the target language:

Speakers handle the language with flexibility and intuition, their speech is entirely accepted by well-educated native speakers in all of its features, including range of vocabulary and idioms, popular expressions, and relevant cultural references. Their pronunciation is normally consistent with that of well-educated native speakers of a non-stigmatized dialect. (Interagency Language Roundtable, 1985: 4)

The problem with this scale is the use of the term native speaker in the definition of the levels. This can be problematic because, as we commented above, we cannot give a concrete definition of this term. Therefore, we should avoid the use of this term in descriptors.

Another option for the definition of the higher bands in the CEFR (Council of Europe, 2001) scales is the *ACTFL Proficiency Guidelines* (Swender et al., 2012). These are not actual scales, but they are descriptions of what learners can do with language in real-life situations in terms of listening, writing, speaking, and reading, we will focus on the speaking guidelines. These guidelines have 5 levels of proficiency, and we will focus our attention on the last and more advanced level, which was labelled as ‘distinguished’. This level will correspond to the higher level of proficiency and it can help to develop the descriptors in the higher bands of the CEFR scale. All the sentences quoted below are found in Swender et al. (2012: 4).

The first sentence that can be suitable for the CEFR (Council of Europe, 2001) scale is: “Speakers are educated and articulate users of the language”, which can be adjusted to define the aspect of fluency in the higher bands of C2. Continuing with “they can reflect on a wide range of global issues and highly abstract concepts in a culturally appropriate manner”, which can perfectly define the aspect of accuracy in the CEFR speaking scale. The following sentence says: “speakers can use persuasive and hypothetical discourse for representational purposes, allowing them to advocate a point of view that is not necessarily their own”, which could be used to define the aspect of range.

The sentence “they can tailor language to a variety of audiences by adapting their speech and register in ways that are culturally authentic” can be used to define the aspect of accuracy. Moving now to the sentence “speakers can produce highly sophisticated and tightly organized extended discourse”, which can define the higher band of the aspect of coherence in the CEFR scale. The next sentence “they can speak succinctly, often using cultural and historical references to allow them to say less and mean more” can be used to define the aspect of range if we adapt the descriptor to the CEFR scale. The last sentence “a non-native accent, a lack of a native-like economy of expression, a limited control of deeply embedded cultural references, [...] occasional isolated language error may still be present at this level” could be used in the aspect of fluency if we adapted it to the CEFR scale.

The first problem that we find when trying to adapt these guidelines into the CEFR scale is that it is not an actual scale so we will try to analyze each sentence and adapt it to the aspect more suitable for the higher levels of the CEFR scale. Another problem that we find in this guideline is the use of negative words such as ‘lack of’, ‘limited’, and ‘error’. This can give the examiner a negative impression in a very high level of proficiency, so we will try to avoid the use of those expressions in the definition of the higher-level descriptors.

The study created by the Austrian English Language Teaching and Testing (hence ELTT) group is another possible solution to the definition of descriptors above the C2 level. It defined two analytic rating scales for assessing academic presentations and interaction. The results of the study proposed a specification of academic speaking; thus, they add specific details to the reference levels in the CEFR. The study was summarized in Berger’s (2020) paper.

The study ended with the validation of a pre-defined range of academic speaking proficiency, and it covered the highest levels only. It confirms that the possibilities that a

descriptor has of being useful were highest when descriptors were linked to linguistic aspects rather than to task aspects. They created reference points from the descriptors and defined prominent characteristics at every level, which were further labelled as ‘effective operational’, ‘full operational’, ‘general academic’, ‘advanced academic’, and ‘full academic proficiency’. The last two groups will help us to expand the vague and under-defined descriptors for the higher bands of C2 and, in this way, these descriptors will provide greater, meaningful, and specific band distinctions.

The level labelled as ‘advanced academic’ is defined with the following set of descriptors:

Learners can speak fluently and accurately on all levels. They can use a very broad range of language, which allows them to speak without having to restrict the effectiveness of their contribution. They can speak naturally, fluently and with a consistently high degree of accuracy, even when their cognitive resources are being directed towards non-linguistic matters. There are two new features: Firstly, there is an unprecedented degree of flexibility and ease, and secondly, the descriptors reflect a new degree of meta-cognitive awareness that allows speakers to perform their tasks more consciously and effectively. In functional terms, this band is characterized by an increased capacity for persuasion rather than just argumentation and for interaction management rather than just participation (Berger, 2020: 96).

Full Academic Proficiency reproduces “the highest degree of control and naturalness in all respects” (Berger, 2020: 96). In this level “the speaking process has become completely automatic so that the full attentional capacity is available for the communicative effect” (Berger, 2020: 96). Whereas the CEFR (Council of Europe, 2001) defines the established common reference points at the C2 level as “degree of precision, appropriateness and ease with the language which typifies the speech of those who have been highly successful learners” (Council of Europe 2001: 35-36), the study summarized here provided a greater distinction for the upper proficiency levels of the scale.

If we compare the descriptors created to define the ‘full academic proficiency’ level with the descriptors that we can find in the scales for oral production in the CEFR (Council of Europe, 2001) in the C2 level, we can observe that the CEFR (*ibid.*) scales are divided into five different aspects which makes the assessment process much more specific. This division is lacking in the ELTT descriptors; however, in Berger’s (2015: 289) study we find a table in which the different levels for academic speaking at C1 and C2 are grouped. Bands were

organized from the lowest, band 5, to the highest level of proficiency, band 1. In this table we can see a distinction of the different tasks that the learner can perform in each level.

Bands		Common reference points
1	Full Academic	• high degree of automaticity and flexibility in language use • consistent phonetic/phonological control • deliberate use of vocal features • full use of presentation skills • flexibility and ease even with abstract, complex unfamiliar topics • skillful and effective use of sophisticated collaboration strategies
2	Advanced Academic	• consistent accuracy even when the focus is not on language • fluency, flexibility and ease • meta-cognitive awareness to enhance communicative effect • initiative and interaction management • effective persuasion • effective use of collaboration strategies
3	General Academic	• disfluencies merely for communicative enhancement • effective argumentation, logic and valid reasoning • attributes of professional public speakers begin to emerge • adaptive responsiveness • use of advanced collaboration strategies
4	Full Operational	• few disfluencies for linguistic planning and repair • errors are insignificant • appropriate use of basic task-specific activities • discernible awareness of contextual factors • use of basic collaboration strategies
5	Effective Operational	• clarity and appropriateness of language • disfluencies for linguistic planning and repair • appropriate use of essential academic functions

Table 3: Different levels for academic speaking at C1 and C2. (Berger, 2015: 289)

6. Conclusion

6.1 Concluding Remarks

As we observed throughout this investigation, we can state that the construction of rating scales is not frequently reported. Turner (2000) states that although rating scales play such an important part in the rating process and ultimately represent the construct on which the performance evaluation is based, there is surprisingly little information on how commonly used rating scales are constructed and what the criteria are based on. It is difficult to write a scale for two reasons, the first one is the insufficiency of reliable evidence about language learning and the second one is the necessity to encapsulate descriptors into short sentences to make them

easy to use. Consequently, hardly any rating scale employed by examination boards is published.

It is normal for scales on oral production to vary depending on the main objective and the aimed audience and it is practical to make different versions for the different audiences, in this way we will avoid misconceptions even though it may take more time. Experts have also highlighted several issues which enlightened the creation of higher-level descriptors for this scale development stages, such as the separation of the ‘grammar and vocabulary’ descriptors, which were united in the lower and mid-levels of the scale; however, they are divided into two independent criteria at C1 and C2 levels (Galaczi et al. 2011).

6.2 Our Own Contribution

As a possible and more definite solution to this problem, the explanation that we propose in a final instance in this project would be the analysis and election of the most suitable descriptors for the CEFR (Council of Europe, 2001) of the different scales present in this dissertation. Once this process of analysis and selection has been completed, we should carry out a qualitative and quantitative study which will prove the usefulness and reliability of the new descriptors.

Finally, it is important to mention that it is necessary for the reliability of rating scales to be in permanent contact with reality and with the progression that our modern society experiences constantly. That is to say, we should contemplate the different technological, scientific, linguistic developments and include the ones that are suitable to define the descriptors at the most appropriate level.

The truth is that we live in a society where people want to get closer, to discover new cultures, countries, and people because of the globalization process, which increases continuously in our modern civilization. This fact should also be reflected and improved in the descriptors with the passing of time, given that the learning of languages, in this case English, due to the desire present in society of feeling closer and being able to communicate with different speaker from a varied number of cultures.

6.3 Limitations of the Study

It is worth mentioning that rating scales are created before the actual test of proficiency, and normally by an expert or a group of experts by means of their intuitive judgment regarding

the type of language competence. Therefore, after we have defined the scale with descriptors, it is essential to make validity studies of the descriptors to confirm that the descriptors are meaningful, and they indicate the proficiency of a given learner in a precise language modality. Thus, we have to test the different solutions provided in this dissertation to verify that the scales are adequate and valid for their purpose.

It is important to consider that this project has many limitations, given that it offers only some possible solutions to the problem posed and for the scales to be proved effective we will have to carry out a complete qualitative and quantitative study where we could prove the effectiveness of the descriptors by using them in real testing situations and with experienced testers for the scale to be reliable. As a final remark of this project, we want to state that, just as the CEFR states, “we are raising questions, not answering them” (Council of Europe, 2001: Notes to the User, xi).

6.4 Lines for Further Research

Future investigations are essential to confirm the reliability of the solutions that can be drawn from this study, given that, as has been already mentioned in this project, there is little information that could help the development of rating scales and the definition of their descriptors.

Future research on this field should be concerned about the problem of vague descriptors and confirm that there is in fact “life beyond the C2 level”, just as Brian North suggests in a presentation (2007). Therefore, future investigations should be devoted to the creation of higher-level descriptors to allow for the development of learner proficiency to go beyond the current levels.

7. References

- Alderson, C. (1991). Bands and scores. In C. Alderson & B. North (Eds.), *Language Testing in the 1990s: The Communicative Legacy*. London: Modern English Publications/British Council/Macmillan, pp 71-86.
- Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

- Berger, A. (2015). *Validating Analytic Rating Scales: A Multi-method Approach to Scaling Descriptors for Assessing Academic Speaking*. Frankfurt am Main, Germany: Peter Lang.
- Berger, A. (2016) Rating scales for assessing academic speaking: A data-based definition of progression. *Views*, 25: 25-44. Vienna: University of Vienna.
- Berger, A. (2020) Specifying Progression in Academic Speaking: A Keyword Analysis of CEFR-Based Proficiency Descriptors. *Language Assessment Quarterly*, 17/1: 85-99. Vienna. Routledge. Doi: 10.1080/15434303.2019.1689981
- Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman and A. D. Cohen (Eds.), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press, pp. 112–140.
- Brooks, L.R and Mackey, B. (2013) Who, what, where, WENS: the native speaker in the Interagency Language Roundtable Scale. In Galaczi, D. E. and Weir, J. C. *Exploring Language Frameworks. Proceedings of the ALTE Kraków Conference*. Cambridge. UCLES, pp 288-301.
- Brown, A. (2006) An examination of the rating process in the revised IELTS Speaking Test. *IELTS Research Reports*, 6: 41-70. Canberra and Manchester: IELTS Australia and British Council.
- Bueno González, A. (1996) Testing English as a Foreign Language: An Overview and some Methodological Considerations. *Revista Española de Lingüística Aplicada*, 11: 17-49.
- Clark, J. L. D. and Clifford, R. T. (1988). The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status and needed research. *Studies in Second Language Acquisition*. 10: 129–147.
- Council of Europe (1947). John Trim Collection: box 9, document 22. In Kirchberger, V. (2016) *John Trim and the Communicative Approach, From the 1960s to Threshold Level*. PhD.
- Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Cambridge University Press.

- Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe.
- Cruz, J. (2016) *A Protocol to Design a CEFR-linked Proficiency Rating Scale for Oral Production and its App Implementation*. Doctoral Thesis. Universidad de Jaén.
- Davies, A et al. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.
- Fulcher, G (2015) Assessing second language speaking. *Language Teaching*, 48: 198-216. doi:10.1017/S0261444814000391
- Fulcher, G and Davidson, F. (2006) *Language Testing and Assessment*. London and New York: Routledge. Taylor and Francis e-Library.
- Galaczi, D.E. et al. (2011) Developing assessment scales for large-scale speaking tests: a multiple-method approach. *Assessment in Education: Principles, Policy and Practice*, 18/3: 217-237. Cambridge: Routledge.
- Galti, A. et al. (2018) Rating scale in writing assessment: Holistic vs. Analytical scales: A review. *International Journal of English Research*, 4/6: 4-6.
- Harding L. (2017) What Do Raters Need in a Pronunciation Scale? The User's View. In Isaacs, T and Trofimovich, P. (Eds.) *Second Language Pronunciation Assessment. Interdisciplinary Perspectives. Second Language Acquisition* 107. Bristol: Multilingual Matters, pp 12-34.
- Hymes, D. H. (1974) *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.
- Interagency Language Roundtable (1985) *Interagency Language Roundtable Language Skill Level Descriptions*. Washington: ICLS
- Isaacs, T and Trofimovich, P. (2017) Key Themes, Constructs and Interdisciplinary Perspectives. In Isaacs, T and Trofimovich, P. (Eds.) *Second Language Pronunciation Assessment*. Bristol: Multilingual Matters, pp 3-11.

- Knoch, U. (2009). *Diagnostic assessment of writing: The development and validation of a rating scale*. Frankfurt: Peter Lang
- Knoch, U. (2011) Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*. 16/2: 81-96. Victoria, Australia: Elsevier Ltd.
- Luoma, S. (2004) *Assessing Speaking*. Cambridge, UK: Cambridge University Press.
- Luoma, S. (2004) Self-assessment in DIALANG. An account of test development. In Milanovic, E and Weir, J. C. (Eds.) *European Language Testing in a Global Context. Studies in Language Testing 18*. Cambridge. Cambridge University Press, pp 143-156.
- Madsen, H. S. (1983) *Techniques in Testing*. London. Longman.
- McNamara, T. (1996). *Measuring Second Language Performance*. London & New York: Longman.
- North, B and Piccardo, E. (2019) Developing new CEFR descriptor scales and expanding the existing ones: constructs, approaches and methodologies. *Zeitschrift für Fremdsprachenforschung*, 30: 2, pp. 142-160
- North, B and Schneider, G. (1998) Scaling descriptors for language proficiency. In Alderson, C. J and Bachman, F. L. (Eds.) *Language Testing*. London: Hodder Headline Group.
- North, B. (1996/2000). *The Development of a Common Framework Scale of Language Proficiency*. PhD thesis, Thames Valley University, London, UK. Published in 2000 as *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.
- North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. TOEFL Monograph, 24: 1-106
- North, B. (2007) *The CEFR Levels: Key Points and Key Problems*. Sevres, France: Presentation. Eurocentres: Language Learning Worldwide.
- North, B. (2008) *The CEFR Levels and Descriptor Scales*. Switzerland. Eurocentres and EAQUALS. In Weir, J.C. and Taylor, L. (Eds.) *Multilingualism and Assessment. Studies in Language Testing*, 27: 21-66. Cambridge, UK: Cambridge University Press.

- O'Sullivan, B and Dunlea, J. (2020) *Technical Report APTIS Advanced Technical Supplement*. Manchester: British Council.
- Petra Project. (2016) *Petra-E Framework of Reference for the Education and Training of Literary Translators*. Colophon. <https://petra-educationframework.eu/> Last accessed 26 March 2021.
- Pilliner, A. E. G. (1968) Subjective and objective testing in Davies, A. (ed.) 1968. *Language Testing Symposium. A Psycholinguistic Perspective*. 19-35. London: Oxford University Press.
- Richards. C. J. et al. (2002) *Longman Dictionary of Language Teaching and Applied Linguistics*. Malaysia: Pearson Education.
- Swender, E, Conrad, J. D and Vicars, R. (2012) *ACTFL Proficiency Guidelines*. New York: ACTFL INC.
- Thompson, G. (2006) Reevaluating the Test Specifications of an Oral Proficiency Test. *The journal of Kanda University of International Studies*, 21: 233-260.
- Trim, J. L. M. (2007) *Modern Languages in the Council of Europe 1954-1997*. Strasbourg. Council of Europe.
- Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review*, 56/4: 555–584.
- University of Cambridge ESOL Examinations. (2008) *Certificate of Proficiency in English Handbook for Teachers*. Cambridge: UCLES.
- University of Cambridge ESOL Examinations. (2011) *Using the CEFR: Principles of Good Practice*. Cambridge: UCLES.
- Van Ek, A. J. (1975) The Threshold Level in a European unit/credit system for modern language learning by adults. *Systems Development in Adult Language Learning*. Strasbourg: Council of Europe.
- Weigle, S. C. (2002) *Assessing Writing*. Cambridge: Cambridge University Press.

Wilkins, A. D. (1978) Proposal for levels definition. In Trim J. L. M. *Some Possible Lines of Development of an Overall Structure for a European Unit/credit Scheme for Foreign Language Learning by Adults*, pp 71-78. Strasbourg: Council for Cultural Cooperation of the Council of Europe.