811.111:371.275
81'243:371.26
Претходно саопштење

Милан Д. Милановић[1]
Ана Д. Милановић
Филолошко-уметнички факултет
Универзитет у Крагујевцу

Наслеђе 28 • 2014 • 57-77

USING THE CEFR TO PROVIDE TEST SPECIFICATIONS FOR ASSESSING VOCABULARY FOR ESL/EFL ACADEMIC WRITING: POTENTIALS AND LIMITATIONS[2]

The Common European Framework of Reference has caused much debate ever since it was published in 2001. It was developed with the aim of being used as a descriptive tool in language teaching, learning and assessment. In this paper we will explore its potentials in helping language practitioners utilize its descriptors and guidelines for development of test specifications for ESL/EFL academic vocabulary. We will examine the existing vocabulary-related descriptors and reflect on their potentials to be used in two ways**:** as a basis upon which test specifications for assessing academic writing will be developed, and as descriptors for rating scales in test rubric. At the same time, we will reflect on two aspects which have been a subject of criticism in terms of language assessment. First, it has been claimed that the CEFR lacks a strong link to any theoretical models (apart from, maybe, a model of communicative competence), which hinders its potentials to be used as a basis for test specifications. Second, descriptors within the Framework do not provide enough contextual clues, as is necessary for developing a language test.

Keywords: assessment, test specifications, the CEFR, scales, reference level descriptors, rating rubrics

Introduction

According to Palmer and Bachman, test development process consists of three phases: test design, test operationalization, and test administration (1996: 86). In this paper, we will focus on the second stage, the one resulting in test specifications based on which concrete tests are developed and administered. More specifically, we will examine whether the Common European Framework and publications related to it can be used as a basis for writing test specifications, in our case in writing assessments where vocabulary knowledge is assessed as a part of a broader construct. Before we provide a short analysis of three current models of test specifications (and by this we do not

1 milan.milanovic@kg.ac.rs

2 Краћа верзија овог рада изложена је у виду усменог саопштења на „11th Conference of the European Society for the Study of English", одржаној у Истанбулу у септембру 2012. године.

imply that these are exclusive models for writing test specifications), we will discuss components of vocabulary knowledge that are of interests to language examiners and researchers, and provide a brief overview of the process of developing a vocabulary assessment. Considering the fact that test specifications normally provide information on scoring method and criteria for correctness, rubrics for rating vocabulary in written assessments are given additional consideration in this paper. Namely, they are considered from the perspective of assessor-oriented scales provided by the Framework, which are argued to be of use to test assessors, i.e. test raters. For this reason, we will analyze the illustrative scales found in the main document (i.e. the CEF document) in order to ascertain whether reference level descriptors provide sufficient information for holistic or analytic scoring rubrics. At the same time, both potentials and limitations to the use of the Framework will be discussed, the former in terms of resources that the Framework and documents related to it provide to test developers, users, and validators, and the latter in terms of the lack of contextual clues necessary for writing test specifications.

1 Academic writing and (academic) vocabulary assessment

Academic writing refers to a range of writing activities that take place at various educational settings, including various tasks administered and completed for most different purposes. In this paper we will focus on lexical contents of academic writing, or academic vocabulary which can be assessed as a part of a broader construct of academic writing. Since the writing tasks students face in their academic disciplines are so diverse it seems impossible to provide a full account of all the possible academic writing tasks in this paper (see Weigle for more on designing writing assessment tasks, 2000: 77-107). Instead, we would like to point out that lexical contents of academic writing is often rated, either by the means of holistic or analytic rating rubrics, and for this reason it is interesting to know if test takers' performance can be described and assessed using descriptions provided in the CEFR scales.

Lexical units whose presence and usage is observed in assessments are usually classified as **high-frequency** and **low-frequency** words (including specialized and sub-technical vocabulary). It is therefore worth saying that academic writing tends to contain a fairly large number of low-frequency words as compared to general writing tasks, and among these words technical, and sub-technical vocabulary will account for a significant number of total running words. The so-called sub-technical vocabulary refers to "words which occur quite frequently across a range of registers or topic areas in academic and technical language" (Read, 2000: 159), and this is where **academic vocabulary** steps in. The term refers to words and phrases which are found at educational settings, and also to the words which involve some specific, content knowledge. Some of these words may have one meaning in general vocabulary, whereas their meaning changes in accordance with a specific content-filled context in which they are used. These words are often collected and published in wordlists, so these may be valuable sources for test developers (see Cox-

Наслеђе 28 • 2014 • 57-77

head's Academic Word List, 2000). Before we proceed to our discussion of the process of test development, we will consider what can be in the focus of vocabulary assessments.

Throughout the better part of the twentieth century vocabulary was assessed in content-independent tests where lexical items were tested as discrete objectively-marked units. The communicative approach to language ability has made a point of using vocabulary as a resource for various communicative purposes through integrative tasks in context-dependent assessments (Read, 2000). However, both approaches are still widely used to complement each other, because there are arguments supporting both discrete and contextualized assessments. Whatever approach they adopt, test developers create vocabulary testsin order to determine the following aspects of vocabulary knowledge:

- vocabulary size (**breadth**), or "the number of words a learner knows regardless of how well he or she knows them" (Daller et al, 2007: 7, in Milton, 2010: 218);
- the quality of vocabulary knowledge (**depth**);
- productive vocabulary knowledge, or "the ease and speed with which words can be called to mind and used in communication", and this is what Milton refers to as lexical fluency (Milton, 2010: 219).

Considering our intention to examine the potentials of the CEFR to facilitate (academic) vocabulary assessment, it can be observed that the Framework distinguishes between three aspects of vocabulary knowledge, termed as *vocabulary range, vocabulary size* and *vocabulary control*. These are considered to be important aspects of language acquisition, and for this reason they are important for assessing language proficiency. To develop tests in which vocabulary is assessed, test developers may consider the following:

- select key words and phrases in thematic areas required for the achievement tasks relevant to learner needs,
- refer to high-frequency words in general word counts or opt for assessing low-frequency, specialized, and sub-technical vocabulary lists to select the vocabulary fitting to their testing purposes,
- select authentic input materials and identify specific vocabulary, and encourage test takers to use the words in their response to input materials.

1.1 On (academic) vocabulary test development

The purpose of assessment will play a crucial role in the way vocabulary tests are developed and their results used. Discussing the process of test development, Read (2000)emphasizes the difference existing between proficiency tests in which vocabulary is seldom assessed and classroom tests where vocabulary testing still has an important role. He says that vocabulary as such tends not to be tested in the proficiency tests (2000: 186), and we would like to add that some standardized language proficiency tests such as TOEFL i-BT, do test vocabulary, but in context rather than in isolation, in the sub-sections of this test, for example in the Reading and Listening sections (for more see

ETS, 2010). Classroom tests, on the other hand, do make use of vocabulary assessment, especially in educating EFL and ESL learners. However, vocabulary assessment may refer to assessing discrete lexical units and to assessing lexical content of spoken and written texts. We will discuss the process of test development with respect to discrete assessment of vocabulary in the following paragraph. Developing a vocabulary assessment within a spoken or written test will follow a similar procedure, however the tasks and items may take a number of forms, and it goes beyond the scope of this paper to cover them all.

Since the purpose of assessment will guide test development, it is important to know what it is that teachers and language testers want to find out about language learners, and what their decisions based on test results will be used for. Most commonly, vocabulary is assessed for the purposes of placement in placement tests, diagnosis in diagnostic tests, and for measuring achievement or progress in achievement tests. After the purpose of testing has been determined, test developers need to provide a construct definition which may be based on a syllabus when vocabulary assessment takes place within a course of study, or on theory, when it is intended for research purposes (for more on syllabus- and theory-based construct definitions see Bachman and Palmer, 1996: 117-120, and Read, 2000: 153). Following the Bachman and Palmer's test development model, Read focuses on two aspects in development of test tasks which follow the process of defining the ability to be measured – the input and expected response.The input refers to the materials and information that test takers need to process in order to provide responses to test items and tasks – the prompt, instructions, materials to read and process before completing the task, etc. Hughes suggests that in proficiency tests lexical items be specified by referring to one of the published wordlists which indicates the frequency with which the words are found to be used in the real world situations (Hughes, 1989, in Read, 2000). Once the target words have been selected test items are created in line with the purpose of assessment and test specifications, test writers need to make decisions whether they want to test words in isolation or in context (or both) before they proceed to designing test items. When it comes to test items in discrete vocabulary assessment, they will require test takers to respond to them in the envisaged manner, e.g. to match, select, paraphrase, define, explain, provide synonyms and antonyms, use the target word in a sentence of their own, etc. In contextualized vocabulary assessments, vocabulary breadth and/or depth are measured. It goes beyond the scope of this paper to discuss the characteristics of expected response, but readers are advised to refer to the Bachman and Palmer's test task characteristics framework which details characteristics of input and expected response, as well as the relationship between the two (Bachman and Palmer, 1996). Finally the responses need to be evaluated, and for this reason the scoring method has to be specified at the stage of writing test specifications. When vocabulary is assessed in terms of discrete units, objective marking can be applied. However, when it is embedded into a piece of written or spoken text, a more elaborate means of scoring and interpreting of results is needed.Bachman and Palmer (1996) sug-

Наслеђе 28 • 2014 • 57-77

gest using holistic and analytic rating scales to rate test takers' performance whenever vocabulary is incorporated into the construct of a written or spoken production. Both types of scales use descriptors to refer to the performance in assessments which are at risk of being subjectively rated, and this is avoided/mitigated by using descriptors and trained raters who link the performance to one of descriptors (and corresponding score) in the scale.

1.2 Rating (academic) vocabulary knowledge in writing assessments

In writing assessments rating scales are most commonly used as a means of enhancing objectiveness of scoring procedures. This often requires that raters be qualified and trained for applying rating scales, which, depending on the purpose of assessment, can be classified as those used for **primary trait scoring**, **holistic scoring**, and **analytic scoring**. In **primary trait scoring**, the rating scale focuses on a particular writing assignment, with descriptors developed for each and every task in the assessment. As such, this kind of scoring is time-consuming and constructing a rating scale is difficult, while at the same time the rating process can take a lot of time, because every task that a student completes has to be measured against a specific rubric created to rate the performance on that task. **Holistic scoring** refers to the assigning of a single score to a piece of writing produced by a test taker, with test takers' performance being judged against criteria explicitly stated in the scoring rubric. The scoring guide for the Test of English as a Foreign Language (TOEFL) is an example of a holistic rubric, and for the sake of illustration let us see what descriptor for the rating 5 looks like at the 0 to 6 rating scale (with zero point assigned when an essay/paper contains no response) in Example 1:

5 An essay at this level

- may address some parts of the task more effectively than others
- is generally well organized and developed
- uses details to support a thesis or illustrate an idea
- displays facility in the use of language
- demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors

Example1: TOEFL writing scoring guide (in Weigle, 2002: 113)

Weigle argues that this kind of holistic scoring is more reliable than its predecessor known as *general impression marking* (Weigle, 2002: 112). The rubrics in holistic rating scales are accompanied by benchmark samples of writing linked to certain scales within the rubric, and their purpose is to facilitate the rating process (however, it should be noted that other types of rating scales are accompanied by writing samples for the very same reason (for more see Weigle 2002: 112). The advantages to holistic scoring lie in its practicality, but Weigle maintains that in second-language contexts such rubrics fail to help users distinguish between "various aspects of writing such as control of syntax, depth of vocabulary, organization and so on" (p. 114). **Analytic** rubrics,

on the other hand, may feature a desired number of aspects of writing, such as content, organization, mechanics, grammar, vocabulary, etc. Let us illustrate this by the criteria for rating vocabulary in a test of writing, developed for the Test in English for Educational Purposes (TEEP) in Example 2 (it should be noted that apart from this one, the whole rubric contains the following aspects of writing performance: *Relevance of adequacy of content, Compositional organization, Cohesion, Adequacy of vocabulary for purpose, Grammar, Mechanical accuracy I (punctuation), and Mechanical accuracy II (spelling)*):

D. Adequacy of vocabulary for purpose

0. Vocabulary inadequate even for the most basic parts of the intended communication.
1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.
2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/ or circumlocution.
3. Almost no inadequacies in vocabulary for the task. Only rare inapropriacies and/or circumlocution.

Example 2. TEEP attribute writing scales (Weir, 1990 in Weigle, 2002: 117).

Each of them is assigned certain weighing and raters need to go through writing scripts several times to provide rating for every segment in the rubric. Regardless of practicality issues associated with this kind of rating, analytic rubrics provide a better picture of a test taker's writing profile.

2 Test specifications

Test specifications are often considered to be essential to the process of test development (Coombe, 2007), and some authors define them as "generative blueprints for test design" (Davidson and Lynch, 2002 in Coombe, 2007). The role of test specifications is also outlined in the Manual for Language Test Development and Examining, where test specifications are recognized to be of importance for both high-stakes and low-stakes assessments (Council of Europe, 2011). In the case of the former, test specifications are seen as an instrument for ensuring quality of a test and validity of inferences made on the basis of test results. Similarly, low-stakes assessments benefit from test specifications as well, especially in terms of ensuring that "all test forms have the same basis and that a test correctly relates to teaching syllabus " (Council of Europe, 2011: 23).As suggested in the Manual, sample test specifications can be found in the works of Alderson, Clapham and Wall (1995), Bachman and Palmer (1996), and Davidson and Lynch (2002), so in the following chapter we will discuss these three models.

2.1 Test Specification Models

The sample test specifications mentioned above will be discussed here as three widely used models which share some common characteristics, but it should be noted that they also differ in various features. However, these models are not to be taken for the only possible and exclusive test specification models, although it can be argued that they provide test developers, test takers, and test users with useful pieces of information.

Alderson, Clapham and Wall (1995) Model

Although they are aware that some other authors use terms *test specifications* and *syllabus* interchangeably, Alderson *et al.* find differences between them. They argue that test specifications provide "the official statement about what the test tests and how it tests it" (1995: 9) and these can serve internal purposes of the examining body, which means that they are sometimes confidential, whereas the test syllabus, as a public document, contains information useful to teachers and test takers. Consequently, the former often contain valuable information for test and item writers, but they also provide test users, test takers and test validators with essential information for establishing test validity and usefulness (1995: 9).The stakeholders interested in test reliability and validity may have varying needs, so that Alderson *et al.* advocate using different forms of test specifications according to the type of audience that will be using them. Accordingly, they discuss test specifications developed for *test writers*, *test validators*, and *test users* respectively. Given the essential role of test and item writers in the process of test development, test specifications created to suit their needs is in the focus of our discussion here. As cited in Coombe (2007:11-12), Alderson *et al.* include the following features into their model of test specifications intended for test and item writers:

- General statement of purpose
- Test battery (list of components and the time allowed for each)
- Test focus (description of the sub skills/knowledge areas to be tested)
- Source of texts (where appropriate text materials can be found)
- Test tasks (range of tasks to be used on the test)
- Item types (range of item types and number of items)
- Rubrics (form and content of instructions given to test takers).

Apart from test specifications developed for test writers, there is a recognized need for test specifications developed specifically for test validators and test users. Test validators' role is to provide arguments supporting validity of test results and inferences based on them, which means that they should be aware of the constructs the test intends to measure, as well as of the model of language ability these constructs are based on (Coombe, 2007). Test users, however, vary in their types of needs, although it is fairly easy to recognize several common types of users of test results: test takers, teachers (or educators), school/university officials, and employers. Alderson *et al.* suggest that test users should be made aware of what "the test measures, and what the test should be used for" (Alderson *et al.*, 1995: 20). Test specifications intended

for test users are termed as "user specifications" and authors state that they should contain descriptions of a typical performance at each level, and also" a description of what a candidate can be expected to be able to do in the real world". This is where the CEFR's "can do" statements step in, because they are developed in such manner that they reflect a learner's ability to use a target language (including grammar, vocabulary, and language functions) appropriately, while at the same time their performance can be linked to the corresponding levels on proficiency scales.

Bachman and Palmer (1996) Model

The second model we discuss in this paper is that developed by Bachman and Palmer (1996). In this model they introduce a test *blueprint* which consists of a two-part test specification. The blueprint is a detailed test plan which can serve a number of purposes: (1) to permitthe development of parallel forms of a test with the same characteristics, (2) to evaluate the work of test writers, (3) to evaluate the correspondence between the final product and the original intentions, and (4) to evaluate test (tasks) authenticity (Bachman and Palmer, 1996: 176-7). The two-part specifications include the *structure* of a particular test, while the second part is what authors term as the *test task specifications*. It can be observed that the former includes information on the number and order of parts in a test (in the case when a test consists of subtests), the weighing of tasks and items and their respective numbers per test/sub-test. This part of the blueprint corresponds to some extent to the model discussed above, whereas the second part, that of the test task specifications is developed in more detail.

Palmer and Bachman argue that *a task* is the elemental unit of a language test, and for this reason test operationalization stage should focus on development of test tasks (1996: 171). Test tasks are developed with respect to target language use (TLU) task types in order to provide information on a test taker's ability to perform desired language functions in the real world. The starting point in test tasks development refers to identifying TLU task types which will provide a basis for development of test tasks. The characteristics of test tasks should correspond to TLU task characteristics, and for this reason the latter should be identified and taken into consideration in the process of test development. The TLU characteristics identified here are accompanied by the specific purpose and construct definition for each type of task which finds its way in a particular test, within a document known as *test task specifications* (Bachman and Palmer, 1996: 172). These authors claim that test specifications need to include all of the following characteristics (not necessarily in the same order): (1) the purpose of the test task, (2) the definition of the construct to be measured (by a particular task), (3) the characteristics of the setting of the test task, (4) time allotment, (5) instructions for responding to the task, (6) characteristics of input, response, and relationship between input and response), and (7) scoring method.

Davidson and Lynch Model (2002)

The third model we discuss here is that of Davidson and Lynch (2002). As the authors point out, their model is somewhat similar to that of Bachman and Palmer, although some components of the two models are organized and labeled differently, with the significant differences referring to Bachman and Palmer's explicitly stated time allotment, instructions and scoring method (Davidson and Lynch, 2002: 30). The model presented by Davidson and Lynch builds on the earlier one, developed by Popham (1978), consisting of the following five components:

- General description (a brief summary statement about what is being tested and measured)
- Prompt attributes
- Response attributes
- Sample item
- Specification supplement

Davidson and Lynch state that test specifications are aimed at creating tests which measure the same skill(s) as specified in this document, through a set of similar test tasks and items. The information contained in test specifications helps teachers, test administrators, test takers, test writers, and test users understand what is tested by the test and how results may be appropriately used (Davidson and Lynch, 2002).

The three models discussed above are not the only possible models of test specifications. Douglas, for example, says that test specifications should contain, at minimum, the following components:

- a description of the test content, including the organization of the test, a description of the number and type of test tasks, time allotment for each task, and specifications for each test task/item type,
- the criteria for correctness
- sample tasks/items (Douglas, 2000: 110-113).

As can be seen above, there are many possible ways of writing specifications that cover the essential elements identified by Douglas (Douglas, 2000 in Weigle, 2002: 83) depending on the purpose of assessment and intended audience for who specifications are developed.

3 The CEFR and language assessment

Before we explore the potential use of the CEFR in the process of test development, we need to consider its intended uses, which include the following:

- the planning of language learning programs,
- the planning of language certification, and
- the planning of self-directed learning.

The planning of language certification refers to specifying the content of syllabus of examinations, and to determining assessment criteria in terms of positive achievement (COE, 2001: 6). The scales of descriptors provided by

the Framework can be of use to the process of language assessment on condition that there is an accurate identification of the purpose the scale is to serve (COE, 2011). For this reason, there is a functional distinction made between three types of scales of proficiency (Alderson, 1991): *user-oriented* (they report typical behaviors of learners at any given level focusing on **what a learner can do**), *assessor-oriented* (they guide the rating process, and although they are often negatively worded, descriptions of reference levels can follow the example provided in Table 3 of the Framework and employ positive wording with necessary limitations in establishing **how well a learner performs)** (COE, 2001: 28-29), and *constructor-oriented* (they inform the process of test development at appropriate levels of proficiency by providing statements expressed in terms of specific communication tasks the learner is to perform in a test, demonstrating what they **can do**). A problem may occur if proficiency scales designed for one function is used for another (2001: 37), for example if user-oriented scales are used by raters to evaluate performance.

The Framework is concerned with language assessment in terms of providing solid basis for ensuring validity, reliability, and feasibility of assessments, so its authors suggest it be used in the following three ways:

- for the specification of test contents and examinations;
- for stating the criteria to determine the attainment of learning objectives; and
- for describing the levels of proficiency in existing tests and examinations for the purpose of their mutual comparisons across different systems of qualifications.

In other words, the Framework may help test developers, administrators, secondary and higher education officials to determine what is assessed, how performance is interpreted, and how comparisons can be made. In this paper, we will focus on the first two intended uses, because they can be of use to test developers and test raters.

3.1 Using the Framework to develop the specification of the content of tests and examinations

As outlined above, developing test specifications is not only recommendable but often a necessary and valuable step in developing language assessments. In this chapter we will explore the possibilities of using the CEFR in developing test/task specifications. It can be noticed that the three models of test specifications discussed above are very much in consensus as to what test specifications should include, although they use different terminology and ordering to list and describe test specification components. What interests us here is whether the CEFR and publications related to it can help test developers (or "constructors") in the process of developing test specifications for assessing academic vocabulary within a piece of academic writing.

First of all, it should be noted that the CEFR was developed so it could meet a number of purposes, and language assessment is but one of them. The

Chapter 4 of the Framework provides descriptions of language use and users, and more specifically it focuses on communicative language activities in terms of spoken and written interaction and production. For this reason, test developers need to adapt the CEFR to their own needs and the first step in this process is to specify the domain of language use and the purpose of their test (ESOL, 2011: 19). The CEFR offers some help as to the specification of different domains (personal, public, occupational, and educational) within which language use is set in the contexts of various situations (COE, 2001:45). The users of the Framework are advised to select domains with respect to the needs of the learners who will have to operate in them, but it is to be noted that, depending on a situation in which language is used, more than one domain may be involved (COE, 2001: 45). When it comes to situations, they can be termed as *target language use* (TLU) situations where various language tasks can be identified, which is of much use in defining constructs which will be measured in language tests. Table 5 of the Framework provides examples of domains, including a number of variables that can be found within them: locations, institutions, persons, objects, events, operations, and texts. Communicative themes, tasks and purposes, communicative language activities and strategies are illustrated as well. However, the authors of the table state that this table is just an illustration of situations that may arise in each of the domains they identify, and therefore it has no claims to be exhaustive or final (see COE, 2001: 46, 48-49, and ESOL, 2011: 18).Consequently, test developers will have to work out the TLUs of their choice, and identify important characteristics they want to incorporate in their test specifications or test task specifications (Bachman and Palmer's test task characteristics framework could also be of help in this process, 1996). Decisions regarding time allotment, instructions for responding, test rubrics and sample items and tasks have to be made by test developers, considering the purpose of assessment and the audience for which test specifications are developed. However, the Framework provides test developers with some hints in the section 4.6 which deals with "texts" (page 93) and in the section 7.3 related to tasks and their characteristics (page 157). These can be made use of together with "the growing "toolkit" designed to help designers exploit the CEFR" (ESOL, 2011: 19). This refers to an increasing number of publications related to utilizing the CEFR, including the *Manual for Language Test Development and Examining. For the Use with the CEFR* (COE/ALTE, 2011*), Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual* (2009);the validated *Can Do statements* provided by the Association of Language Testers in Europe (ALTE);the publications and resources of the *English Profile Programme* (including the validated English Vocabulary Profile wordlists, and the Can-Do statements for C levels of language proficiency- which are still the work in progress). To sum up, it can be noted that the CEFR can provide valuable resources for test developers but it does not contain all the answers test developers may ask themselves in the process of developing a communicative language test.

3.2 Using the Framework to specify the criteria for the attainment of learning objectives

It is argued that scales provided in the Framework and descriptors can be of use in developing scales to rate performance. However, care must be taken to distinguish between descriptors of communicative activities and descriptors of aspects of proficiency related to particular competencies. The former can be useful for reporting results to test users (employers, university officials and administrators, etc.), whereas the descriptors of aspects of proficiency related to particular skills and competences may be used for specifying criteria for performance assessment. The latter can be done in three ways:

- descriptors can be presented as a scale in the form of a holistic paragraph per any given level,
- descriptors can be presented as a checklist where descriptors are grouped under categories, and
- descriptors can be presented as a grid of selected categories, which makes it possible to give a diagnostic profile. The grid of sub-scales can take the form of proficiency scale, where relevant levels are defined for certain categories, and it can take the form of an examination rating scale, where descriptors are defined for each relevant category (COE, 2000).

4 The CEFR scales and (academic) vocabulary assessment

In this chapter we will explore resources, in terms of illustrative scales and descriptors that the CEFR document(s)[3] offers to test developers and test raters in their attempt to develop test specifications for testing academic vocabulary within assessment of academic writing, and rate performance in such assessments respectively. Of course, test developers and test raters are not necessarily the same people, i.e. they may constitute different audiences, and consequently the former will make use of constructor-oriented scales, whereas the latter will find assessor-oriented scales more useful.

4.1 Using the CEFR to provide test specifications for assessing academic vocabulary in tests of academic writing

Test purpose / General description

As we can see in the models of test and test task specifications provided above, test developers need to determine and specify the purpose of assessment, and identify what it is they want to test and measure in their assessment. The construct definition can be based on a syllabus or on a model of language ability, and since the CEFR is based on communicative language ability it may be argued that it offers some information on communication themes, communicative tasks and purposes, communicative language activities and strategies,

3 By this we have in mind the text of the Framework and associated documents, often found as appendices to the main document, e.g. the ALTE's *Can Do statements*.

communicative language processes, communicative language competences. However, the information provided in the main document is fairly general, though illustrative examples are provided throughout the document. To identify test purpose for assessing academic writing, test developers may consult the main document, but their decision will more likely be based on a specific language learning syllabus, particularly at educational settings, though a model of language proficiency may be consulted as well if test results are to serve the purpose of linguistic research.

Prompt attribute / Characteristics of the input / Source of texts

The text of the CEFR contains a section on *Texts*, where a "text" refers to "any piece of language […] which users/learners receive, produce or exchange (COE, 2001: 93). Texts here are described in terms of text types (e.g. news broadcasts; memoranda, essays and papers, etc.) and activities where texts are used as input or output of communication processes. The media used to transfer texts are also covered (e.g. voice, manuscript, videotape, etc.) with the purpose of explaining how physical properties of media affect the processes of reception and production of texts in this sense (p.93). The Manual of 2009, on the other hand, suggests that Grids provided on the website of the Council of Europe be used for "profiling the features of tasks, expected performances (answer length, discourse types, register, etc.), rating instruments and feedback given to candidates", and such profiles are intended for linking particular assessments to the CEFR (Manual, 2009: 30). The information on text types and activities contained within the main document, and the Grids found on the COE's website could be of some use to test developers who could use this data to create a sort of a checklist to help them specify the characteristics of the input/output. However, since the Framework and the Manuals are not intended to be used as a blueprint for any assessments (including the assessment of academic writing), it cannot be expected of them to provide more than resources which are to be consulted in particular assessment projects. Namely, the authors of the Manual are explicit in the claim that information provided in the Manual is by no means "a recipe for a test blueprint, it is a rather a resource to the examples of good practice" (Manual, 2009: 13). The ALTE's Can Do statements, on the other hand, are of no use here, as they describe what learners can do at certain levels of proficiency, but they do not contain any clues as to the specific input materials and other prompt attributes related to academic writing or academic vocabulary.

Response attributes/ Characteristics of the response

When developing test specifications item writers need to specify what the expected response will be like (e.g. whether it includes selection, limited, or extended production). These characteristics can be described as suggested in Bachman and Palmers' Framework of test task characteristics, or in any other way as deemed most suitable by test developers. This component of test

specifications is closely related to test purpose and the input, and needs to be very specific. The CEFR and the related documents are, on the other hand, very general, and they do not focus on any specific situation within any given of language use. Consequently, the CEFR fails to be of any use in determining the characteristics of a specific expected response, in any given, specific context. The Grids aimed at linking assessments may be of use as checklists outlining the possible characteristics of a response to the prompt (the CEFR Content Analysis Grid for Writing and Speaking Tasks is provided in Council of Europe, 2009: 159), but there are other, more comprehensive checklists to be used for this purpose (see, for example, Bachman and Palmer, 2006).

Scoring method/ Criteria for correctness

Test specifications, according to Bachman and Palmer's model provide information on how performance will be rated and scored. In the case of writing assessments, the kind of rating scale (holistic or analytic) is provided. Also, the criteria of correctness is often included to familiarize test takers with what will be considered as correct/ sufficient response to the prompt. The CEFR scales might be of some use to test developers, because they sometimes provide holistic descriptions of learners' language ability (this is discussed in more detail in the following chapter).

Item Types/ Sample item

This component of test specifications is context-specific and will depend on all the components discussed above. The illustrative examples of test items in the CEFR and the CEFR-related documents, though they exist, are not provided for all communicative activities and functions. However, even if they were, sample items are so diverse, and test developers need to devise their own.

The reasons why test developers may try to use the CEFR to write test specifications is because the CEFR claims to be comprehensive and because it is supposed to facilitate language assessment by providing information what learners can be expected to know at six levels of language proficiency scale. However, as Weir notes, though it claims to be comprehensive and introduces many concepts the CEFR fails to put them into scales (Weir, 2007: 7), and the same applies to the Manual for Relating Examinations to the CEFR which leaves it to test developers to decide for themselves what is (in terms of language functions, grammar, vocabulary, and structures) appropriate at each of the six levels. The analysis of the CEFR and the documents related to it, including ALTE's Can Do statements, has shown that the Framework provides little information that test developers may find useful to the process of writing test specifications, especially in terms of distinguishing between performances at different (let alone adjacent) levels. Weir notices that although functional competences are well described, there are still some contextual parameters which are insufficiently described and specified in the CEFR: *purpose, response format, time constraints, channel, discourse modes, text length,*

topic, lexical competence, and structural competence. The presence and explicitness of these parameters may help test developers to write test specifications for assessments linked to different levels of the Framework. For example, the purpose of fulfilling tasks is not made explicit for any tasks and activities described in the Framework, and it is left for exam providers to specify the content and task purposes for different levels of the proficiency scales. Alderson et al. notice that there is nothing in the CEFR that would give test developers any hints about response formats which could be used at different levels of the CEF scales (Alderson et al, 2004 in Weir, 2005: 10). Time allotment is missing for all sorts of tasks, and Weir argues that different processing time is needed for dealing with texts and carrying out activities and completing tasks at different levels. The channel and discourse modes are not made specific for any tasks and activities, and Alderson points out that apart from the descriptions provided in the scales, there is nothing much as to the content of any given level, especially in terms of what texts (written or spoken) are appropriate for each level (Alderson, n.d. in Weir, 2005:11). The topic is one more thing that a test developer needs to consider when designing test tasks and choosing input materials, because the topic will react with a test taker's general and specific knowledge and in that way affect their performance and test results. Weir argues that the CEFR is of little use as to determining what topics are relevant or appropriate at different levels, although it provides some illustrative examples of communication themes (COE, 2001: 51). Lexical competence is of particular interest in this paper, and apparently the CEFR provides little help in identifying level-appropriate vocabulary, regardless of whether it is receptive or productive. Moreover, the CEFR fails to provide examples of typical vocabulary and structures found at each level of the Framework, which means that typical structures, grammar, and vocabulary need to be identified and defined for any language which is to be a subject of assessment that is to be linked to the CEFR. The English Profile Programme is an international collaborative project aimed at providing a set of Reference Level Descriptions for the CEFR levels, showing the specific vocabulary, grammar, and functional language that students can be expected to master at each level in English (for more visit www.englishprofile.org).Weir notices, for example, that it is almost impossible to assess vocabulary depth or breadth by using the CEFR and for this reason one strand of research within the English Profile has resulted in the English Vocabulary Profile – an online vocabulary resource which uses empirical data to provide information of the CEFR levels of "words, phrases, phrasal verbs and idioms" for just under 7,000 headwords (Capel, 2010, and Capel, 2012). This is more than a wordlist, because it is easily searchable according to several criteria, and at C1 and C2 levels it comprises both General and Academic vocabulary linked to the CEFR by following a 'can-do' rationale, and as such it can be of use to test developers, test users, and test takers (especially those who assess vocabulary breadth) because the entries show what learners can do at a particular level. Generalized characterizations, for example, those referring to vocabulary in Tables 1 and 2 (see Tables 1 and 2 in the following chapter) do

not contain sufficient information (in terms of TLU, activity, task, etc.) for developing test specifications or rating test takers' performance. For this reason, a TLU should be identified, specific activities and tasks appropriate to the TLU need to be selected, so that test items can be developed. If this specific target language use situation refers to academic writing, test purpose will determine whether vocabulary range, vocabulary control, or vocabulary size will be assessed, or, on the other hand, any combination of these three components of vocabulary knowledge. The CEFR-related publications are of some help to test developers because they can be used as a resource, or they can provide useful links to other resources. Manual of 2011, for example, provides useful links to test developers, while at the same time it maintains the connection to the Framework document advising test developers to consult the following sections of the Framework in their attempt to create a test: Overviews of the Common Reference Levels, Overviews of Communicative activities, Overviews of aspects of communicative language competence, Communicative activities particularly relevant to the occupational and educational domains (Council of Europe, 2011: 13).

Critics of the Framework claim that there are several problems associated with the use of the CEFR in language assessment. First of all they claim that the CEFR is not a framework but a model of language proficiency, because it is too abstract to be a framework on the basis of which test specifications can be made (Weir, 2005, Fulcher, 2007). Fulcher argues that "true frameworks need to mediate between the abstract and the context of a particular test" with the purpose of operationalizing the components of a model which are in line with a specific purpose of a test, and as such the framework enables test developers to produce test specifications (Fulcher, 2004: 259). Another problem is related to the problem with formulating descriptors which are sometimes found to be vague and inconsistent. Alderson *et al.* found similar descriptors occurring at different levels, different verbs describing apparently one and the same cognitive process, etc. (Alderson *et al*, 2004 in Weir, 2005: 16-17).

4.2 Using the CEFR to develop rating scales for rating use of academic vocabulary in academic writing assessments

In this part of our paper we will analyze illustrative scales provided in the CEFR in order to see if they have potentials for being used as a basis for rating scales in writing assessments, with a special focus on the academic vocabulary components. Test developers and test users may be interested in mapping test takers' performance on the CEFR scales, because it seems grounded to assume that vocabulary depth and sophistication of its use will increase with increase of communicative competence. However, the question here is whether descriptors in the Framework are detailed enough to help test raters tell the difference between performance at different, and especially at adjacent levels of the CEFR scales (e.g. what depth and breadth of vocabulary, or what range and vocabulary control places a test taker at B2 or C1 levels). To do this, we

will analyze the illustrative scales which have specifically been developed to describe vocabulary knowledge, in terms of its two components – vocabulary range and vocabulary control.

First of all it should be noted that the main document, i.e. the CEFR itself, provides two illustrative scales for the range of vocabulary knowledge and the ability to control that knowledge (Tables 1. and 2, adapted from Council of Europe, 2001: 112). These scales comprise descriptors provided for almost all levels (the only exception is a missing descriptor for the level A1, in the scale for vocabulary control). Vocabulary range and control, as described by the means of reference level descriptors here, refer to a general knowledge of vocabulary, and as such, they fail to provide any information on academic vocabulary range and control. However, even for the general vocabulary knowledge, descriptors in the Framework lack specificity of any kind. For example, what does *abroad lexical repertoire* xactly refer to in the descriptor found at C1 level of the Vocabulary range scale (Table 1)? If we assume that raters use this scale as a rubric for rating a piece of writing, how will they distinguish between a *broad lexical repertoire* at the level C1 and *a **very** broad lexical repertoire* at the level C2? What *is a basic vocabulary repertoire*?

VOCABULARY RANGE
C2 *Has a very good command of **a very broad lexical repertoire** including idiomatic expressions and colloquialisms, shows awareness of connotative levels of meaning.*
C1 *Has a good command of a **broad lexical repertoire** allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.*
B2 *Has a good range of vocabulary for matters connected to his or her field and most general topics. Can vary formulation to avoid repetition, but lexical gaps can still cause hesitation and circumlocution.*
B1 *Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel and current events.*
A2 *Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.* *Has a sufficient vocabulary for the expression of basic communicative needs.* *Has a sufficient vocabulary for coping with simple survival needs.*
A1 *Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations*

Table 1. Vocabulary range criteria from Council of Europe (2001, p. 112)

It may be concluded that descriptors lack specificity, and as such they are of little use to language assessors who are supposed to rate performance and assign scores which will place a performance to a corresponding level in the CEFR.

If we take a look at the table containing descriptors related to vocabulary control, we can notice that descriptors are fairly general as well. For example, how can a rater know how many words constitute a narrow repertoire? What can be considered to be a *minor* as opposed to a major slip at C1 level? To answer these questions one has to be aware of test purpose, as well as of the intended use of test results and inferences based on test takers' performance. Holistic descriptions like these do have a potential of being used by test raters, but to suit this purpose well they need to be complemented by benchmark responses which would help train raters to rate academic writing. Apart from that, every descriptor needs to be elaborated in more detail, so that it could offer more clarity and precision, particularly in terms of differentiating between the levels of proficiency. Academic vocabulary assessed in a writing assessment will also depend on

VOCABULARY CONTROL
C2 *Consistently correct and appropriate use of vocabulary.*
C1 *Occasional minor slips, but no significant vocabulary errors.*
B2 *Lexical accuracy is generally high, though some confusion and incorrect word choice does occur without hindering communication.*
B1 *Shows good control of elementary vocabulary but major errors still occur when expressing more complex thoughts or handling unfamiliar topics and situations.*
A2 *Can control a narrow repertoire dealing with concrete everyday needs.*
A1 No descriptor available

Table 2. Vocabulary control criteria from Council of Europe (2001, p. 112)

test purpose. For example, test developers may want to check the breadth of academic, highly technical vocabulary related to mechanical engineering in the field of automotive industry, by assessing a piece of academic writing, for example a report discussing a plan for introducing innovations to a passenger car engine. The vocabulary of interest will probably include a range of words such as: *analyze, argue, pertain*, and so on, but it will probably include words such as *cylinder, injection, compressor*, etc. If such report is rated by using a *holistic rating scale*, it may include one of vocabulary-related descriptors such as one of these found in the CEFR (with holistic, more general descriptions), and this scale will probably be accompanied by benchmark responses demonstrating difference in performance among different levels of the scale. If, however, test raters are interested in vocabulary breadth and depth, and/or vocabulary control, more elaborate descriptions need to be provided in an analytic rating rubric, and benchmark writing samples demonstrating differences among the levels will complement the rating rubric to ensure rating and construct validity of the scores.

Vocabulary- related descriptors as they stand in the current version of the CEFR are fairly general and they need a degree of specificity to be of any use to test constructors, but the wording in the illustrative vocabulary-related

scales bear a potential of their being used as assessor-oriented scales. The fact that these descriptors are general is recognized in the very CEFR document (2001), but it is explained by the fact that the Framework is language neutral, and for this reason it should be used as a reference tool for describing different languages. As such the scales provided in the CEFR can apply across languages, but context-specific information need to be added so that descriptors could refer to specific communicative goals and activities. The same applies to their intended use for rating written academic performance. Context need to be considered, and wordlists and benchmark responses need to be provided so that test raters can use these scales in holistic or analytic rating.

Conclusion

The reason why test developers may want to use the CEFR as a basis for test specifications lies in the notion that reference level descriptors demonstrate the difference in performance at different levels. However, the analysis of the CEFR document and other publications related to it shows that there are not enough contextual clues provided in the illustrative scales or validated Can Do statements as they stand in the existing publications. The Framework is language neutral, which means that it has to be applied to specific languages, in terms of illustrating what specific grammar, vocabulary and functions learners can be expected to demonstrate at each of the CEF levels. Furthermore, although the documents we analyzed provide illustrations and examples of communicative events and situations, they do not offer a whole range (if that is possible at all) of target language use situations, accompanied with data needed for writing test specifications. When it comes to vocabulary knowledge, there are no examples of (academic) vocabulary knowledge that test takers can be expected to demonstrate at any levels of the Framework. Without these examples, accompanied by benchmark writing samples demonstrating vocabulary knowledge linked to different levels, test developers are not able to provide test specifications for any given writing assessment.

The CEFR scales on the other hand, might prove to be of some use in developing rating rubrics for assessing productive skills. Let us remind the reader of Alderson's distinction between constructor-oriented scales, assessor-oriented scales, and user-oriented scales (Alderson, 1991). Whereas many authors feel that the CEFR has limited potentials for test assessors and test constructors, scaled can-do statements "are ideal for reporting a generalizable meaning of test scores to test users, in terms of what a test taker with a particular score on a given test may typically be able to do" (Fulcher, 2004: 264). The ALTE's validated *Can Do* statements in their original conception are user-oriented (Jones, 2000: 11), and as such they may provide test users with valuable information on test takers' ability to use grammar, structure, vocabulary and language functions, and what is more important, they provide them with information on how well they can be expected to use the language in a particular (communicative) situation. However, the can-do statements provided by the Association of Language Testers in Europe do not cover the full

range of target language use, and new can-do statements have to be developed to suit the specific purposes of individual assessments.

References

Alderson et. al. 1995: C. Alderson, Clapham, C., & and D. Wall. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

ALTE 2012: The Alte Can Do Project. Retrieved July 14, 2012, from http://www.alte.org/attachments/files/alte_cando.pdf

Bachman and Palmer 1996: L. F. Bachman and A. S. Palmer. *Language Testing in Practice*. Oxford: Oxford University Press.

Capel 2010: A. Capel. A1-B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal, 1*(e3). Retrieved July 3, 2012

Capel 2012: A. Capel. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal, 3*(e1). Retrieved July 10, 2012

COE 2011: COE. Manual for Language Test Development and Examining. Retrieved April 5, 2012

Coombe 2007: C. Combe. Developing Test Specifications. *Perspectives, 15* (1), 10-13.

Council of Europe 2012. *Council of Europe*. Retrieved July 11, 2012, from http://www.coe.int

Council of Europe 2011. Council of Europe. *Manual for Language Test Development and Examining. For use with the CEFR*.

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual*. Strasbourg.

Coxhead, 2000: A. Coxhead, The Academic Wordlist. Retrieved from http://www.victoria.ac.nz./lals/resources/academicwordlist/

Davidson and Fulcher 2007: F. Davidson and G. Fulcher, The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language teaching* , 231-241.

Davidson, and Lynch 2002: F. Davidson and B. K. Lynch,*Testcraft*. New Haven: Yale University Press.

Douglas 2000: D. Douglas,*Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

ESOL 2011: ESOL, Using the CEFR: Principles of Good Practice. Cambridge. Retrieved March 23, 2012

Fulcher 2004: G. Fulcher, Deluded by Artifices? The Common European Framework and Harmonization. *Language Assessment Quarterly, 1* (4), 253-266.

Hughes 1989: A. Hughes, *Testing for language teachers*. Cambridge: Cambridge University Press.

Jones 2000: N. Jones, The ALTE Framework and the Can-do project. *Research Notes*, 11-14.

Milton 2010: J. Milton, In: Bartning, M. Martin, & I. Vedder (eds.),*Communicative proficiency and linguistic development* , pp. 211-232.

Powers 2010: D. E.Powers, The case for a comprehensive, four-skills assessment of English-language proficiency. *R&D Connections 14* .

Read 2000: J. Read, *Assessing Vocabulary* . Cambridge: Cambridge University Press.
Read 2007: J. Read, Second Language Vocabulary Assessment: Current Practices and New Directions. *International Journal of English Studies, 7* (2), 105-125.
The English Profile 2012: *The English Profile*. (2012). Retrieved July 10, 2012, from http://englishprofile.org
Weigle 2002: S. C. Weigle, *Assessing Writing.* Cambridge: Cambridge University Press.
Weir 2005: C. Weir, Limitations of the Common European Framework for develping comparable examinations and tests. *Language testing, 22* (3), 1-20.

Наслеђе 28 • 2014 • 57-77

Милан Д. Милановић
Ана Д. Милановић

КОРИШЋЕЊЕ ЗЕРОЈ-А ЗА ИЗРАДУ ТЕСТОВНЕ СПЕЦИФИКАЦИЈЕ ПРИЛИКОМ ПРОВЕРЕ ЗНАЊА ВОКАБУЛАРА У АКАДЕМСКОМ ПИСАЊУ НА ЕНГЛЕСКОМ КАО СТРАНОМ / ДРУГОМ ЈЕЗИКУ: МОГУЋНОСТИ И ОГРАНИЧЕЊА

Резиме

Заједнички европски референтни оквир за језике (ЗЕРОЈ) предмет је бројних полемика још од 2001. године када је објављен. Оквир је развијен са циљем да се употребљава као описно средство које ће помоћи у реализацији наставе језика, учењу језика, и провери језичког знања. У овом раду испитујемо могућности да се на основу Оквира произведе тестовна спецификација за тестирање академског вокабулара у тестовима енглеског језика у којима се тестира писана продукција. Постојеће дескрипторе употребе вокабулара анализирамо како бисмо утврдили да ли могу да се користе као: (а) основа за развој тестовне спецификације за тестирање академског вокабулара у оквиру писане продукције, и (б) дескриптори у скалама у оквиру рубрика за оцењивање. Истовремено, коментаришемо два аспекта Оквира која су предмет критике у области тестирања језичког знања. Први се односи на тврдњу да Оквиру недостаје јасна веза са неким од теоријских модела језичких компетенција (сем можда са моделом комуникативне језичке компетенције), чиме се умањује његова потенцијална улога у формирању тестовне спецификације. Други се односи на тврдњу да дескриптори у Оквиру не пружају довољно параметара за контекстуализацију задатака у тестовима језичког знања.

Кључне речи: провера језичког знања, тестовна спецификација, ЗЕРОЈ, скале, дескриптори референтних нивоа, рубрике за оцењивање

Примљен у мају 2014.
Прихваћен у јуну 2014.