STUDIA UBB PHILOLOGIA, LXIV, 4, 2019, p. 157 - 168
(RECOMMENDED CITATION)
DOI:10.24193/subbphilo.2019.4.09

DEVELOPING A CEFR-BASED ANALYSIS GRID FOR LISTENING TASKS AND ITEMS

ADRIANA TODEA[1]

ABSTRACT. ***Developing a CEFR-Based Analysis Grid for Listening Tasks and Items.*** Assessing listening abilities in a foreign language on CEFR (*Common European Framework of Reference for Languages*) standards requires that clear correlations be drawn between CEFR listening comprehension descriptors and the features of the input text and items of specific listening tasks. A CEFR-based analysis grid designed to assess listening tasks and items provides a useful tool of consistent measurement of the conformity of a specific task and its items to the level of proficiency they claim to assess. The present paper analyzes components of the listening comprehension construct in language test situations and builds discriminating scales based on graduating features isolated from CEFR relevant descriptors.

Key words: *comprehension threshold, input text, task, item, problem-solving operations*

REZUMAT. ***Evaluarea competenţelor de înţelegere-ascultare într-o limbă străină pe baza standardelor CECRL*** (*Cadrul European Comun de Referinţă pentru Limbi*) necesită corelări clare între descriptorii specifici competenţei de ascultare şi caracteristicile textului sursă şi a itemilor exerciţiului de ascultare. O grilă de analiză a cerinţelor unui test de ascultare poate fi un mijloc de măsurare consecventă a conformităţii cerinţei şi a itemilor aferenţi cu nivelul de competenţă lingvistică pe care prezumăm că aceştia îl testează. Acest articol analizează componentele proceselor de comprehensiune a ascultării în situaţii de testare lingvistică şi sunt construite grile graduale pe baza caracteristicilor extrase din descriptori relevanţi ai CECRL.

Cuvinte cheie: *prag de comprehensiune, textul sursă, cerinţă, item, operaţii de rezolvare a problemei*

[1] **Adriana TODEA**, PhD, Babeş-Bolyai University; lecturer at the English Department; language test developer at the ALPHA Centre of Modern Languages of the Faculty of Letters. E-mail: atodea@yahoo.com

Differences between listening and reading constructs

Classifications and subcategorizations are useful means of organizing our understanding of investigated phenomena but they can prove to be particularly misleading if we overgeneralize by ignoring the specific criteria based on which they were generated in the first place. As we are well aware that reading and listening processes are categorized as receptive skills in contrast to productive skills such as speaking and writing, the greatest mistake that could be made in test design is to assume that listening and reading are equivalent constructs with some input variation (acoustic vs. printed). It is the nature of the input, though, that significantly shapes the construct of cognitive processes involved in processing linguistic information, and although the acoustic feature is prominent when we listen, but negligible when we read, it is the ephemeral nature of the input that shapes most significantly the listening construct itself. Spoken language occurs in real time and the listener cannot get a second chance by revisiting it for an extensive analysis as one can do with written texts. Even when speakers are asked to repeat themselves, they will partially or fully rephrase the message to communicate the idea rather than reproduce the exact wording (Buck 2001:6) and, consequently, the listener's understanding is confined to the immediate for short segments or relying on reconstructive memory (often imprecise) for longer segments.

Features of spoken discourse and test design limitations

Language testing has significant limitations when it comes to recreating naturalistic communicative situations and listening tasks seem to be the most impaired ones. Oral discourse relies heavily on the interaction between speaker and addressee, and the latter is by no means a passive listener. Speakers often shape their discourse in response to their addressee's back-channeling to achieve their communicative goals. Language testing severs that connection between speaker and addressee, relegating the listener to a completely passive role, most of the time with no access to vital visual information, which may support comprehension, of speaker's attitudes, emotions or framing. As such the communicative situations reconstructed in listening tasks are severely reduced to those in which the listener has a non-collaborative role: attend a lecture, listening to a radio programme, or overhearing a conversation between two or multiple participants, which significantly hampers the listener's use of a full range of listening strategies. Such restrictions though can be justified by positing that the listening process elicited by listening tasks is specifically controlled to measure the procedural language knowledge of L2 with as little as possible opportunity for using compensatory strategies for language gaps. Such an approach to language testing, though, raises the question of whether

the purpose is to measure abstract linguistic competence or make predictions about success of real life linguistic performance.

Listening comprehension in CEFR scales and descriptors

The *Common European Framework of Reference for Languages: Learning, teaching, assessment.* (CEFR) of 2000 and its revised version *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors* of 2018 describe the L2 listening construct in a series of scales of grades of general and specific linguistic performance that conform to a non-collaborative, passive listener understanding of the skill as discussed above. The *Overall Listening Comprehension* scale mainly provides a graded description of input text with some compensatory strategies for language gaps provided for lower levels of proficiency.

Speech rate

A comfortable speech rate is essential to the ability of the listener to recognize the phonological form of words and process meaning in real-time. The faster the speech of a syntactic segment, the more likely it is that phonological forms of words are distorted by assimilation, elision, intrusion or weak pronunciations in unstressed positions (Buck 2001:33). Moreover, speech rate interacts with other text variables, such as vocabulary, syntax, topic and accent. It is therefore predicable that gradable thresholds of comprehension depend on speech rates from "very slowly and clearly" (PreA1); "with long pauses … to assimilate meaning" (A1); "slowly and clearly" (A2) to "fast natural speed" (C2)[2]. Absent on the scale are speech rates for B1, B2 and C1, which contrastively could be inferred as *normal* speech rate.

Accent

The familiarity of the accent is again essential to phonological perception. Native speakers in general have contact to a larger variety of accents than L2 speakers who most of the time are exposed mainly to standard pronunciations in the classrooms and have little opportunity to extensive interactions with native speakers. Moreover, when exposed to an unfamiliar accent, native speakers are more likely to adjust faster by virtue of their higher linguistic proficiency that helps them compensate by drawing phonological analogies. It is therefore expected that *familiar accent* is a comprehension threshold component that fades off at levels of proficiency

[2] CEFR 2018, p. 55.

closer to native-speaker levels. In the *CEFR Overall Listening Comprehension* scale an *unfamiliar accent* factors as a mild impediment to comprehension: "can understand … may need to confirm occasional details especially if the accent is unfamiliar" (C1), and no impediment at all at C2: "can understand with ease virtually any kind of spoken language"[3].

Topic and context

Another factor in establishing comprehension thresholds is the predictability of content. Familiarity with the topic and/ or context can enable the listener to anticipate and predict information and therefore compensate for language knowledge gaps. Gradually such compensatory strategies are dropped at C1 ("Can understand ... follow extended speech beyond his/ her own field...") and graded from pre-A1 to B2 ("clearly defined, familiar, everyday context" (pre-A1); "familiar topics encountered in everyday life" (A1); "areas of most immediate priority" (A2); "on familiar matters regularly encountered" (B1); "the topic is reasonably familiar", "in his/her field of specialisation" (B2)[4].

Linguistic knowledge

Scales of procedural language knowledge should make a clear distinction between the passive knowledge as a component of receptive skills and the active knowledge of productive skills. When it comes to spoken language, speaking presupposes spontaneous production of language, whereas listening requires the activation of passive knowledge. In other words, when we listen we understand more complex language than we are able to spontaneously produce. The listening construct is particularly dependant on our ability to understand individual words in order to process idea units. On highly familiar topics or in highly predictable interactions, syntactic relations between perceived words can often be inferred based on background knowledge, conversational expectations or common sense (Buck 2001: 16-17). It is the more challenging topics and contexts that make such a compensatory strategy fail and require syntactic knowledge to disambiguate meaning.

Moreover, the real-time nature of processing spoken discourse separates passive language knowledge in listening comprehension thresholds from reading comprehension thresholds. The language segments that generate idea units tend to be shorter and have simpler syntax and looser text organisation, and may contain non-standard morpho-syntactic and lexical variants, although it is true that literate texts (planned discourses) may display some of the features that are specific to written language (Buck 2001: 10-11).

[3] Ibid., p.55
[4] Ibid., p.55

Graduated scales of input text assessment

The five CEFR scales of descriptors of the listening skill put together complex combinations of various factors to describe listening comprehension thresholds. Aside from speech rate and accent discussed above, Table 1 compiles key terms from the CEFR descriptors in an attempt to generate a graduated picture of listening comprehension thresholds on the following criteria:

– *input text type and size*
– *linguistic complexity* (vocabulary/ syntactic range);
– *language domains* (general vs. idiomatic, general vs. specialized);
– *background knowledge* (predictability of content);
– *content complexity* (abstract vs. factual, information complexity).

Isolating a specific degree of a variable of comprehension from a specific descriptor comes, of course, with the risk of invalidation by interpreting it in isolation, that is why every entry has been checked vertically in the table for interrelations consistent with the descriptor from where it was extracted. Also, some extrapolations have been made, such as *concrete topics*, from B2 level to C1-C2 levels by virtue of the hierarchical inclusiveness of lower level abilities by higher level abilities.

Table 1.[5]

	A1-A2	B1	B2	C1-C2
input text size and type	*simple conversation; simple illustrated presentation; instructions for familiar activities; simple messages and announcements; simple directions; basic instructions;*	*short talks; conference presentation with visual support; straightforward clearly structured standard lecture; detailed directions; (highly predictable) public announce-ments;*	*extended speech; animated conversation; lecture; talk; report; detailed instructions; announcements and messages;*	*extended speech; specialized lectures, discussions and debates;*
idioms	--	--	--	*idiomatic expressions; regional usage (C2);*
linguistic complexity	*simple language*	--	*propositionally and linguistically complex speech;*	*unfamiliar terminology (C2);*
concrete information	*concrete information;*	*straightforward factual information;*	*concrete topics;*	*concrete topics;*

[5] The entries in the table have been compiled from descriptors in *Overall Listening Comprehension, Understanding Conversation between Other Speakers, Listening as a Member of a Live Audience, Listening to Announcements and Instructions,* and *Listening to the Radio and Audio Recordings* scales from CEFR 2018, pp. 55-59.

	A1-A2	B1	B2	C1-C2
abstract information	--	--	*abstract topics;*	*abstract and complex topics*
specialized information	--	*familiar matters; regularly encountered at work; simple technical information;*	*technical discussions in his/ her field of specialisation; forms of academic/ professional presentation*	*complex technical information;*
content complexity	*expression of (dis)agreement; personal information; general outline;*	*main ideas; general messages; specific details;*	*main ideas; general messages and specific details; complex lines of argument; main reasons for and against an argument; identify point of view;*	*identify the attitude of each speaker; follow complex interactions in group discussions and debates; jokes, allusions, inferences (C2)*
content predictability	*familiar; predictable; areas of most immediate priority;*	*common everyday or job related; everyday conversation and discussion;*	*reasonably familiar; both familiar and unfamiliar; topics of current interest;*	*beyond his/ her own field.*

We believe that the CEFR descriptors compiled above can be successfully used to create a listening task and item analysis grid for test development. Input text eligibility should be measured on gradable criteria such as *text type*, *text size*, *predictability of content*, *information complexity* and *linguistic complexity*.

The type of text should conform to a one-way non-collaborative passive role the test-taker assumes as listener according to CEFR listening skill descriptor scales and should be ascribed to specific CEFR levels with the understanding that higher level text type subsume the lower level text type. Therefore, the text type scale provides a minimal standard for selecting the type of input text eligible for the CEFR level targeted by the test developer and should not be used as a means of measuring the CEFR level of the text itself.

Text type:

□ discussion; □ debate; C1
□ report; □ detailed instructions; □ messages; B2
□ lecture/ talk; □ conference presentation; B1
□ illustrated presentation; □ instructions; □ directions;
□ announcement; □ message; A1-A2

Text size is particularly relevant in listening comprehension tasks not so much in terms of length of discourse, but in terms of length of syntactic and prosodic segments and the length of convenient pauses following a speech segment for assimilating meaning. Speech comprehension is a real-time process and its success is measured in the listener's ability to process language segments of a size compatible with average working memory. Once the meaning of a segment is processed, it is added in a cognitive build-up that generates the mental model of the text, its textual form completely forgotten (Buck 2001:26-29). The CEFR descriptors are not particularly explicit on this component of the listening construct, but inferences can be drawn from speech rate descriptors (A1-A2: *slowly and clearly*; B1: *clearly articulated*; B2: *clear standard speech*; C1: *natural speed*; C2: *fast natural speed*) and from text organisation descriptors (A-A2: *simple*; B1: s*traightforward clearly structured*; B2: *extended speech*). Very importantly, if written texts (articles, essays, etc) are selected as source for the input text, they need to be adapted specifically in relation to clause and sentence size.

Average prosodic/ syntactic segment size:

- extended segment at natural speech rate; C1
- long segment at standard speech rate; B2
- medium size segment at didactic speech rate; B1
- short segment followed by long pauses A1-A2

Accessing background information and relying on it to reconstitute meaning is an essential compensating strategy for lower levels of language proficiency in both reading and listening constructs. But the listening process relies more heavily on familiarity of topic as the construct compensates for the ephemeral nature of the input by drawing contextual inferences from sentences that allows assumptions to be made that will guide the interpretation of subsequent utterances (Buck 2001: 25). The degree of predictability facilitates comprehension at all levels of proficiency, of course, but listeners with higher levels of language knowledge can perceive incongruities and inconsistencies in the process of generating mental representations of the text and redress more readily interpretations that are the result of wrong assumptions. Topic familiarity and content predictability are components of the listening process that are very well represented in the CEFR listening descriptors based on which a graduated scale for the topic component can be reliably drawn.

Topic

- unfamiliar topic; low predictability; C1
- rather familiar; some unpredictability of content; B2
- mostly familiar topic; predictable content; B1
- very familiar topic; highly predictable content; A1-A2

Procedural linguistic knowledge, that is the ability to process phonetic, lexical and morpho-syntactic information, is at the core of any language skill, but pinpointing with accuracy the degree of complexity relevant to each listening comprehension threshold is rather more elusive. In real-life spoken interactions visual cues or requests for clarification can be used to compensate small failures of procedural linguistic knowledge, but in language test settings it is virtually impossible. The passive knowledge of language which is so relevant in reading constructs where lexical items and syntactic structures have to only be recognized, and not spontaneously produced, might be expected to play the same role in listening processes generated by an equally receptive skill. But L2 listeners may find the perception comprehension stage problematic because their listening vocabulary may be underdeveloped (weak phonetic form–meaning associations) or because they are too slow in accessing lexical knowledge which, in turn, may lead to them becoming acoustically overwhelmed and completely miss subsequent parts of the text (Goh 2000:61, 63). Therefore in judging the level of the lexical and mopho-syntactic knowledge component necessary to achieve comprehension, listening and reading constructs are not equivalent as presented in *The Dutch CEFR Grid Reading/ Listening.* Unfortunately the CEFR listening scales are deficient in their description of linguistic knowledge and no functional listening linguistic scale can be built on rudimentary contrasts such as A2: *simple language* to B2: *complex language* to C1: *idiomatic expressions* to C2: *highly specialized unfamiliar words; regional usage*. One solution to providing a relevant scale might result from conflating linguistic complexity with information complexity components such as *language domains* (concrete vs. abstract; general vs. specialized) and *background knowledge* (familiarity/ predictability).

Information complexity of input text

- concrete or abstract unfamiliar/ unpredictable, possibly idiomatic, possibly highly specialized information; C1
- complex, concrete or abstract, rather familiar and possibly specialized information; B2
- simple, factual, mostly familiar, mostly predictable and possibly somewhat specialized information; B1
- simple, concrete, very familiar and highly predictable information; A1-A2

Nonetheless, in the absence of clear linguistic knowledge descriptors of the listening construct it might be acceptable to use, as a minimum standard, the graduated language knowledge scales associated with productive skills, such as *General Linguistic Range, Vocabulary Range, Grammatical Accuracy, Vocabulary Control* and *Overall Phonological Control*[6]*.*

[6] CEFR 2018, pp. 131-136.

Graduated scale for test item assessment

Creating test items in listening comprehension tasks is inherently problematic. Firstly, they are problem solving tasks inherent to the listening construct and to processing linguistic information receptively in general, such as following instructions, directions, a line of argumentation, identifying gist, main idea, relevant details, comparing data, facts, evidence, ideas, etc, which are made linguistically explicit. Its linguistic explicitness can take either a spoken or a written form. If its form is spoken, the test item can be processed either before and/ or after the exposure to the input text, and its solving would rely exclusively on the reductive mental model generated by the interpretation of the input text stored in the long-term memory and not on its textual form. Ultimately, solving a spoken test item makes it harder to isolate listening abilities from long-term memory performance. Contrastively, a real-time processing of test items would require that they be made explicit in written form, thus compounding the listening construct itself by integrating a reading comprehension component. The advantage, though, is that real-time processing of the written test item is more successful in isolating the listening construct from the effects of long-term memory storage and it captures the ephemeral effects of processing the textual form of the input.

Testing listening abilities through specific operations of information processing, like the ones listed below in Table 2, are subject to two criteria that can narrow down the selection of an operation in item design: the first is its eligibility as a factor ensuring that comprehension thresholds at specific proficiency levels are achieved; the second has to do with its relevance to the specific goals of linguistic learning or certification that are designed to meet the requirements and standards of stakeholders who may require general language proficiency or academic/ professional language proficiency, with various degrees of specialization.

Table 2.

	Problem solving operations	CEFR descriptors for content complexity [7]
C2-C1	• identify underlying theme or concept; • infer attitudes; feelings, moods, purpose; motivation;	*identify the attitude of each speaker;* *follow complex interactions in group discussions and debates;* *understand jokes, allusions, draw inferences (C2)*
B2	• identify gist, supporting details, viewpoints, opinions; • compare and relate ideas;	*main ideas;* *general messages and specific details;* *complex lines of argument;* *main reasons for and against an argument;* *identify point of view;*

[7] as compiled in Table 1.

	Problem solving operations	CEFR descriptors for content complexity [7]
B1	• identify main ideas, points in a line of argumentation, relevant evidence; • compare and relate evidence; • reach a conclusion; • draw logical inferences;	*main ideas;* *general messages;* *specific details;*
A2-A1	• follow instructions, directions; • confirm information; • identify specific details; • compare and relate data, facts;	*expression of (dis)agreement;* *personal information;* *general outline.*

The CEFR listening comprehension scales are particularly explicit when it comes to the underlying operations of information processing of the listening construct by virtue of their performative approach in phrasing descriptors as *can do* statements. A scale for establishing the eligibility of an information processing operation underlying a task item therefore can be drawn. Note again that the scale is hierarchically inclusive and therefore it can be used to judge the eligibility of an operation to assess a specific proficiency level and not to make an assessment of the proficiency level of the item itself.

Item's underlying problem-solving operations

follow
- instructions; A1-A2
- directions; A1-A2
- points in a line of argumentation; B1

identify
- gist of (part of) text; B1
- main idea; B2
- supporting detail(s), B2
- specific detail(s); A1-A2
- relevant data A1-A2
- viewpoints; B2
- opinions; B2
- purpose; C1-C2
- motivation; C1-C2
- underlying theme or concept; C1-C2

infer/ evaluate
- attitudes; C1-C2
- feelings; C1-C2
- moods; C1-C2
- purpose; C1-C2
- motivation; C1-C2

compare
- data; A1-A2
- facts; A1-A2
- evidence; B1
- ideas; B2

identify information relationships:
- cause; B1
- effect; B1
- solution; B1
- purpose; B1
- draw logical inferences; B1
- reach a conclusion. B1

If the scale above can be used only to judge the eligibility of the test item's underlying operation, then the CEFR- based measurement of the test item itself must rely on the informational complexity and size of the input segment to which the operation applies. Both scales can be derived from judgments made based on the same criteria and previously applied to the entire input text.

Information complexity of input segment targeted by test item

- concrete or abstract unfamiliar/ unpredictable, possibly idiomatic, possibly highly specialized information; C1
- complex, concrete or abstract, rather familiar and possibly specialized information; B2
- simple, factual, mostly familiar, mostly predictable and possibly somewhat specialized information; B1
- simple, concrete, very familiar and highly predictable information; A1-A2

Size of the input text segment targeted by the test item

- extended segment at natural speech rate; C1
- long segment at standard speech rate; B2
- medium size segment at didactic speech rate; B1
- short segment followed by long pauses A1-A2

Conclusion

The goal in this paper was to break down the dynamic interdependence between underlying operations, strategies and linguistic knowledge of the L2 listening processes in relation to input text specificity and to correlate listening construct components to the proficiency level graded descriptors of CEFR in an attempt to create CEFR- based scales for assessing listening tasks and items. Such scales could constitute the basis for developing reading test specifications, test design, and expert test verification and

validation. This goal has only been partially achieved. The CEFR listening comprehension scales, whereas fully proficient in grading underlying operations and features of the input text, such as accent, speech rate, type, size, predictability of content, proved particularly deficient in providing a graduated description of procedural linguistic knowledge, which contrasts significantly with both productive linguistic knowledge and reading procedural linguistic knowledge. Further more, the CEFR also lacks a specific scale for the scanning for information strategy in the listening construct which is specifically relevant in designing test items. Nonetheless, with the above caveats in mind, the CEFR proves to be an adept tool that can be used successfully in language test development.

BIBLIOGRAPHY

Amič Irena, Darja Šorjanc Braico, Zdravka Godunc, Nataša Elvira Jelenc, Ana Marija Muster, Branka Petek, and Tatjana Žlindra. *The European Language Portfolio for Adults*, Slovenia, 2010, accessed February 10, 2014, elp.ecml.at/Portfolios/tabid/2370/PublicationID/59/language/en-GB/Default.aspx

Buck, Gary. *Assessing Listening*, Cambridge University Press, 2001

Common European Framework of Reference for Languages: Learning, teaching, assessment. Council of Europe, 2000, accessed February 28, 2017, www.coe.int

Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors. Council of Europe, 2018, accessed November 15, 2019, www.coe.int

Goh, Christine C.M. "A cognitive perspective on language learners' listening comprehension problems." *System* vol. 28, issue 1, 2000, pp.55-75

The Dutch CEFR Grid Reading/ Listening, accessed November 28, 2016, lancaster.ac.uk/fss/projects/grid