Assessing EFL Speech: A Teacher-Focused Perspective

Gaëtanelle Gilquin, Yves Bestgen, Sylviane Granger
Université catholique de Louvain

Abstract

With the aim of better understanding the difficulties that non-native teachers of English as a foreign language (EFL) face when assessing oral skills, we conducted an educational training activity for in-service teachers, involving action research and reflective practice. In the first part, 27 non-native teachers were asked to use the scales of the Common European Framework of Reference for Languages (CEFR) to assess a number of authentic EFL speech samples taken from a learner corpus. Their assessment was examined quantitatively as well as qualitatively and compared to that of two native professional raters. While the analyses highlighted a good degree of agreement among the teachers as well as between the teachers and the experts, they also confirmed the often-observed tendency for non-native raters to be more severe in their evaluation of L2 performance than native raters. The results also indicated that teachers and native experts do not base their overall assessment on the same aspects of the spoken performance. For the second part of the study, we designed group activities and discussions to help the teachers reflect on their own practices and learn from those adopted by others. The analyses showed that the teachers did not feel well-equipped to assess speech and that they would benefit from appropriate training in this area.

Keywords: language assessment; EFL speech; CEFR rating; learner corpus; action research; reflective practice

Introduction

Although oral communication skills are very much at the forefront in current foreign and second language (L2) teaching and despite the fact that 'many language learners regard speaking as the most essential skill to be mastered (…), its assessment has often been neglected in many L2 teaching and testing contexts' (Amengual-Pizarro & García-Laborda, 2017, p. 24). This lack of attention is problematic as assessing speech is extremely challenging. Actually, according to Friginal (2005), speech is the most difficult skill to assess. On the research front, there is an impressive body of work aimed at identifying the impact of the different features of speech (speech rate, number of filled and unfilled pauses, repair phenomena, etc.) on oral proficiency and improving the rating scales (Riggenbach, 1991; Iwashita, Brown, McNamara, & O'Hagan, 2008; Schoonjans, 2012; Kang & Yan, 2018; Tavakoli, Nakatsuhara, & Hunter, 2020). However, these studies mainly involve professional rating bodies and testing experts, most of whom are native speakers of the assessed language, and the voices of language teachers tend not to be heard. This is unfortunate because it is language teachers who carry out the bulk of assessment activities, often without the benefit of proper assessment training, and predominantly in a language that is not their own.

The need for more practitioner-focused research prompted us to organize an educational training activity around the assessment of oral skills for in-service

teachers, involving two key dimensions: action research (Wyatt, 2011) and reflective practice (Walsh & Mann, 2015). Within the teaching context, *action research* refers to any research activity in which teachers are involved and which aims to address some relevant educational issues with a view to developing more effective practices. Wyatt (2011, p. 417) deplores that "many teachers only rarely engage in action research" and highlights the benefits of introducing an action research element into in-service language teacher education. *Reflective practice* refers to activities that foster reflection, which Boud, Keogh, and Walker (1985, p. 19) use as a generic term for "those intellectual and affective activities in which individuals engage to explore their experiences in order to lead to new understandings and appreciations". Walsh and Mann (2015, p. 351) argue that, while reflective practice is widely accepted in the field of second language teaching, it is less clear how it should be conducted. They underline the need to provide better descriptions of reflective practice so as to prompt more "teachers and teacher educators to fully engage with its possibilities" (Walsh & Mann, 2015, p. 351). The educational training activity we conducted for in-service teachers involved these two aspects. For the action research part, the teachers were asked to assess a number of authentic speech samples of English as a foreign language (EFL), taken from a learner corpus, on the basis of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001).[1] Their assessment was examined quantitatively as well as qualitatively, and compared to that of two native professional raters, who rated the samples independently. These results served as a starting point for the reflective practice part, which consisted in group activities and discussions designed to help the teachers reflect on their own practices and learn from those adopted by others.

In this article, we will report on both parts of the training activity. Our first objective is to investigate how the teachers coped with the CEFR-based assessment task and to compare the results to those obtained from the two professional raters. For this comparison, we will compute the inter-teacher agreement and determine the aspects of speech that have the greatest impact on the global score given by the teachers and the native experts. A more qualitative analysis will allow us to relate the scores attributed by the teachers to the linguistic features of the speech samples. By investigating the teachers' ratings, we will learn more about their practices, and how they compare with those of native professional raters. A second objective is to share the teachers' views on the assessment of speech in general and CEFR-based assessment in particular. These findings will provide insights into the difficulties that the teachers may encounter when assessing learner speech and will help us identify their needs in terms of speech assessment. This will lead us to make pedagogical recommendations concerning teachers' training in the assessment of learner speech and the necessity of empowering them as well as allowing their voices to be heard.

Action research

Oral proficiency descriptors

The action research part of the activity relied on the oral proficiency descriptors of the CEFR. The CEFR is a language-neutral reference tool for teaching, learning and assessing languages which provides a description of what learners can do at six

proficiency levels: beginner (A1 & A2), intermediate (B1 & B2) and advanced (C1 & C2). Although it has been criticized for its lack of empirical foundation and the vagueness of its descriptors (Alderson, 2007; Wisniewski, 2018), it has become the de facto standard resource for grading language proficiency in Europe and beyond. Its use can therefore lead to more standardization in assessment practices, including classroom-based assessment. Since "teachers are becoming increasingly responsible for the assessment of their pupils and students at all levels, both formative and summative" (Council of Europe, 2001, p. 20), it is worthwhile investigating how teachers manage with the standardized grids of the CEFR. In pursuing this objective, our study is in line with projects that work to promote wider and better use of the CEFR in language assessment, especially by teachers, such as the English Profile project for writing (cf. Harrison & Barker, 2015).

For our activity, we used the CEFR chart for 'qualitative aspects of spoken language use' (Council of Europe, 2001, pp. 28-29) but adapted it to our purposes in three ways: (i) we left out the 'interaction' descriptors, which could not be assessed in view of the reduced size of the samples; (ii) we added the descriptors for 'phonological control' (Council of Europe, 2001, p. 117), which is an important competence in L2 speaking; and (iii) we included the global oral assessment scale taken from Council of Europe (2009, p. 42), which is "a simplified, holistic assessment scale" derived from the CEFR chart for 'qualitative aspects of spoken language use'. The resulting grid (see Appendix 1) is made up of the CEFR descriptors for the evaluation of five linguistic competences, namely range (mainly lexical), accuracy (mainly grammatical), fluency (capacity to maintain a natural flow of speech), phonological control (skill in the production of sound units and prosody) and coherence (well-structured speech), as well as the CEFR descriptors of the global oral assessment scale.

Participants

The participants were 27 Belgian non-native (mostly French-speaking) secondary school teachers (20 females, 7 males), teaching English in the French-speaking Community of Belgium. They had an average of 18 years of experience as teachers of English (range = 34) and taught an average of over 12 hours per week (range = 16). Although they had no specific experience in the use of the CEFR scales, they all had previous experience in the assessment of oral skills, often on the basis of some (homemade) assessment grid (see the section on Reflective Practice below). Table 1 provides an overview of the teachers' profiles.

Table 1: *Distribution of teachers' profiles*

Feature	**Distribution**
Gender	Female: 20 (74.1%) Male: 7 (25.9%)
L1	French only: 25 (92.6%) French and Dutch: 2 (7.4%)
Degree	Master in English language and literature: 6 (22.2%) Master in English language and literature + teacher training certificate for upper secondary education: 14 (51.9%) Master and PhD in English language and literature: 1 (3.7%) Master in translation: 1 (3.7%) Master in translation + teaching certificate: 3 (11.1%) Master in European studies + teacher training certificate for upper secondary education: 1 (3.7%) Teacher training certificate for lower secondary education: 1 (3.7%)
Years of teaching experience	Average: 17.8 Range: 34
Teaching hours per week	Average: 12.7 Range: 16
Type of education in which they teach	Lower secondary education: 2 (7.4%) Upper secondary education: 19 (70.4%) Lower and upper secondary education: 3 (11.1%) Lower/upper secondary education and higher education: 2 (7.4%) Higher education: 1 (3.7%)

The teachers volunteered to participate in the experiment after receiving an email invitation that we sent via our teacher networks. They were asked to rate authentic speech samples by giving a CEFR score for each of the five linguistic competences as well as a global assessment. Following the procedure advocated by Thewissen (2013), raters were allowed to distinguish sublevels within each main CEFR level by using + or - increments. For instance, B2+ represents stronger performance within the B2 level, but nevertheless insufficient to reach the C1 level, while C1- represents weaker performance within the C1 level. The evaluation of the samples relied on audio files only; no transcripts were provided. The participants also had the opportunity to add comments related to any of the samples or to the task as a whole. The precise instructions they received can be found in Appendix 2. It should be emphasized that we aimed to capture the participants' personal experience of the rating process and therefore did not provide any training in the use of the CEFR scales as part of the activity.[2] This was meant to reflect the teaching reality in francophone Belgium, where teachers generally have to manage with the resources that are made available to them, without any opportunities for external training. Besides, the CEFR descriptors "[h]ave been found transparent, useful and relevant by groups of non-native and native-speaker teachers from a variety of educational sectors with very different profiles in terms of linguistic training and teaching experience" (Council of Europe, 2001, p. 30),

which suggests that teachers should be able to use them without any outside help.

Samples

One of the innovative aspects of our study is that it relies on learner corpus data. The samples consisted in spoken extracts from the French component of the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin, De Cock, & Granger, 2010), a one-million-word corpus of informal interviews with upper intermediate to advanced (i.e. B2 to C2) learners of English, corresponding to over 130 hours of recording. The learners who contributed data for the French component of LINDSEI were 50 Belgian French-speaking university students who learned English as a foreign language. A five-minute sample was taken from each of the 50 interviews, and more precisely from the beginning of the most natural and spontaneous part of it, namely a free discussion between the learner and the interviewer during which the interviewer asked questions about various topics such as life at university, hobbies or travels. All 50 samples were evaluated by two highly experienced professional raters, working as raters for standardized English proficiency tests and training other people to use standardized grids. Both of them were males and native speakers of English. These expert raters were required to work with the CEFR descriptor scales for linguistic competence described above and, like the teachers, they only had access to the sound recording of the speech samples. Among the evaluated samples, we selected ten which had been assigned the same CEFR global score by the two raters and which were spread across the CEFR levels, from B2 to C2 (lower scores were not attributed by the expert raters). The material used in this study is made up of these ten samples, which correspond to 48 minutes and 17 seconds of recording (including both the learners' and the interviewer's utterances, as well as silences) and a total of 8,552 words (including filled pauses). Of these, 5,352 words were uttered by the learners, for a duration of 33 minutes and 34 seconds, that is, an average of over 3 minutes per learner. Table 2 provides information about each of the ten samples.

Table 2: *Information on the ten samples*

Sample	**Duration all (min:sec)**	**Duration learner (min:sec)**	**Word count all**	**Word count learner**	**Global expert rating**
Sample 1	04:59	03:02	914	511	C1
Sample 2	04:43	03:04	850	495	B2
Sample 3	04:33	03:16	909	638	C2
Sample 4	04:46	03:38	772	523	B2
Sample 5	04:23	02:12	750	331	B2
Sample 6	05:01	04:13	796	624	B2
Sample 7	05:09	03:16	927	571	C1
Sample 8	04:44	04:00	878	673	C2
Sample 9	05:02	03:01	957	481	C1
Sample 10	04:57	03:52	799	505	C1
Total	48:17	33:34	8,552	5,352	C1

By way of illustration, Figure 1 shows the transcription of the beginning of a randomly chosen sample evaluated by the experts and the teachers. The whole sample was assigned an average B2 global score by both the experts and the teachers.

I: so whereabouts do you come from in Belgium
L: (em) Dinant
I: Dinant
L: you you see where it is
I: oh yes I know where it is and (er) have you always studied here
L: yeah . yeah
I: and what do you think of it
L: oh it's nice I think there are well <laughs> always bad times and good times but (er) well yeah <X> overall it's nice but I'm .. I'm really: . depressed of my: Dutch studies because I think (er) .. the system is not quite good . yeah in English well the the maîtrise course is quite better in English for example I think it's a bit weak in Dutch and I . it's my only regret because I think the: the level <overlap/> is quite
I: <overlap/> what you mean your regret is that is that you don't feel capable of of learning enough or
L: n= (er) .. no I think if you want to get a a good level in Dutch well you can get it from the candidature but (er) in licence well nothing you can forget . you have <X> there are no exercises anymore and so and I feel . that if you don't work by yourself you forget everything

Figure 1: *Extract of a sample evaluated by the participants (I = Interviewer; L = Learner)*

In this sample, *I* stands for Interviewer and *L* for Learner. The transcript reproduces some of the typically spoken features that the participants could hear when listening to the audio file: unfilled pauses, indicated by dots (one dot for short pauses, two dots for medium pauses and three dots for long pauses); filled pauses, displayed in parentheses; truncated words, marked by an equals sign; syllable lengthening, represented by a colon; nonverbal vocal sounds (<laughs>); and overlapping speech (<overlap/>). <X> corresponds to an unclear syllable or sound.

Quantitative findings

In this section, we approach the data quantitatively, looking at the inter-teacher agreement, the agreement between the teachers and the experts, as well as the link between the linguistic competences and the global score. As a reminder, the quantitative data available are made up of ten interview extracts, each produced by a different learner, which were assessed on five linguistic competences and a global scale by 27 secondary school teachers and by two experts. Since these data are clearly insufficient to perform multivariate statistical analyses such as multiple regressions or to use a Rash model (Chen et al., 2014), we rely on bivariate correlations to measure the degree of agreement among the raters. It should also be pointed out that we have chosen to present all the results in a descriptive way, rather than by means of inferential statistical tests, because the sample size on which each correlation is computed (N = 10) is very small and because the level of agreement between raters is much more important than its statistical significance (Howell, 2007,

p. 159).

Inter-teacher agreement

In order to assess inter-teacher agreement, we calculated the Spearman's correlation coefficient for ranked data between the evaluations of the ten samples by each possible pair of teachers for each of the six scales (five linguistic competences + global assessment), for a total of 2,106 correlations. The boxplots shown in Figure 2 graphically summarize all these correlations.[3]

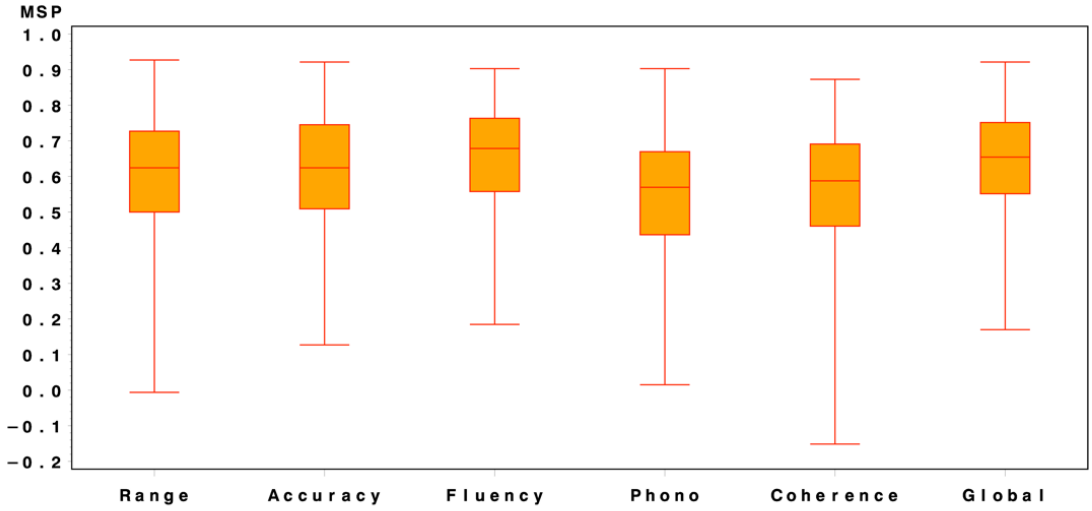

Figure 2: *Boxplots for correlations between all pairs of teachers*

There is a reasonably good inter-teacher agreement for the global score, with a mean Spearman correlation (MSP) of 0.65. However, it must also be recognized that some pairs of teachers show a very low degree of correlation, of 0.2 or lower and even negative, which suggests that assessing speech by means of a CEFR grid, and in particular "condens[ing] … possibly complex impressions of an L2 performance into a single … score" (Isaacs, 2016, p. 138), is far from being a straightforward task. If we turn to the correlations for the different scales, we see that most of them display the same reasonably good correlation, although with some variability. The lowest median correlation is found for the scales of phonological control (0.57) and coherence (0.58). As regards coherence, it is interesting to note that some teachers commented on the difficulty of assessing coherence on the basis of the CEFR descriptors, with one of them writing: 'I found it difficult to evaluate the coherence. The notion of connectors and cohesive devices is rather vague'. More generally, variability can arguably be accounted for by the degree to which raters understand the rating scale categories, the degree to which they comply with the grid, the degree of severity/leniency they exhibit, and/or the degree to which they are consistent across students and tasks. As convincingly demonstrated by Eckes (2008), even experienced professional raters differ significantly in the importance they attach to scoring criteria, which has led Eckes to categorize them into distinctive types according to their dominant scoring focus (the syntax type, the correctness type, the fluency type, etc.).

Agreement between teachers and experts

We also looked at how the ranks of the teachers' scores compared with those of the experts. In order to do so, we calculated the correlation between each teacher's evaluations and the average of the experts' evaluations for each of the six scales. As appears from the boxplots in Figure 3, most scales show a similar – and reasonably good – correlation between teachers and experts, of about 0.6-0.7. The most problematic scale is that of phonological control, which was also slightly problematic in terms of inter-teacher agreement (see above). It looks as if the teachers and the experts may have used very different criteria to evaluate phonological control, despite the common descriptors. Our hypothesis is that the teachers have evaluated aspects like word stress and phonemes, which correspond to what is usually meant by phonological control, whereas the native experts may have examined intonation and sentence stress, as advocated in the CEFR grid (see Kang & Yan, 2018, p. 27, on the importance of suprasegmentals in the assessment and perception of non-native speech by native speakers).[4]

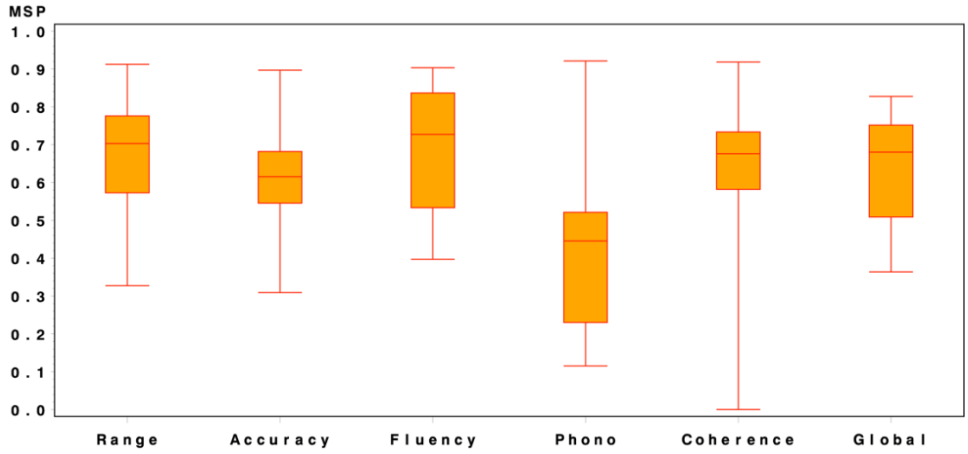

Figure 3: *Boxplots for the correlations between teachers' and experts' evaluations*

While the correlations between the teachers and the experts are relatively high (except for phonological control), it turns out that the teachers tend to attribute much lower scores than the experts, as appears from Figure 4. On average, there is a one-band difference for each scale, with the (non-native) teachers assigning a mean score of B2 and the (native) experts a mean score of C1. These results seem to confirm the tendency, highlighted in the literature (e.g. Y.-H. Kim, 2009; A.-Y. Kim & Gennaro, 2012), for non-native raters to be more severe than native raters in their evaluation of L2 performance. However, the difference may also be related to the raters' degree of expertise, as the native raters in this study are professional raters specially trained to use standardized grids, unlike the non-native raters.

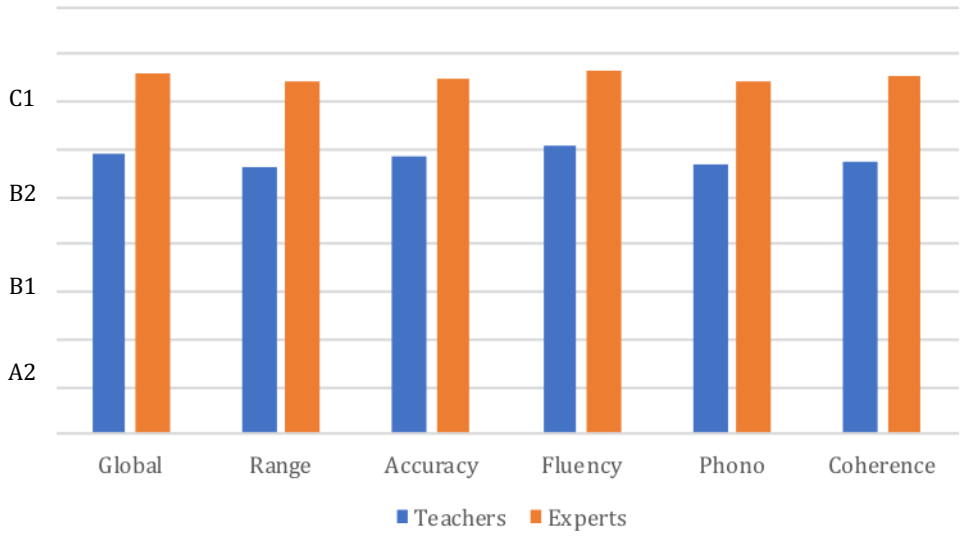

Figure 4: *Average scores per scale for teachers and experts*

Linguistic competences and global score

The results of the evaluation also allowed us to determine which linguistic competences seemed to be predominant when the raters assigned the global score. For each of the five competences, we calculated how much the average score diverged from the average global score, using absolute values. The less divergent from the global score the score of the competence was, the more it was supposed to have contributed to the global score. We did this separately for the teachers and for the experts. Table 3 lists the different competences, from the least divergent to the most divergent, that is, from the most important to the least important in assigning the global score, both for the teachers and for the experts. Among the teachers, range seems to be of utmost importance, followed by accuracy. It is thus item-based aspects of language that seem to predominate when teachers assign a global score. The experts, by contrast, appear to predominantly rely on more global aspects of language such as fluency and coherence.

Table 3: *Linguistic competences ordered according to their closeness to the global score*

TEACHERS	**EXPERTS**
1. Range	1. Fluency
2. Accuracy	2. Coherence
3. Fluency	3. Accuracy
4. Coherence	4. Phonological control
5. Phonological control	5. Range

Qualitative findings

The speech samples are too short to look for correlations between the CEFR scores and linguistic measures extracted from the samples such as speech rate or lexical sophistication. However, a more qualitative approach to the data can help look for possible links between the linguistic features of the samples and the scores attributed to them. This is what we did by comparing some linguistic features in a given sample to the average score attributed to this sample for the relevant scale (e.g. the 'fluency' scale for pauses or the 'accuracy' scale for errors). Our focus was exclusively on the scores given by the teachers. In what follows, we illustrate this approach by giving examples of rated sample transcripts and examining their linguistic characteristics against the CEFR descriptors.

Examples (1) and (2) are taken from samples that, on average, were given a score of B1 and C1, respectively, for fluency. The first extract is very short, but it includes as many as four unfilled pauses (two short ones, represented by one dot, and two longer ones, represented by two dots) as well as three filled pauses (two *em*s and one *er*), which might be said to correspond to the B1 fluency descriptor according to which 'pausing for grammatical and lexical planning and repair is very evident'. Repeats like *I'm I'm I'm* and *I I feel* also contribute to an impression of disfluency. In (2), on the other hand, only one unfilled pause and one filled pause occur in the extract. While repeats are quite common (*yeah yeah*; *an an another*; *I I have*) and syllable lengthening is found once, as indicated by the colon (*to:*), these do not appear to lead the teachers away from the evaluation that the student '[c]an express him/herself fluently and spontaneously, almost effortlessly', as stated in the C1 descriptor for fluency.

(1) *I'm . I'm I'm fond of em . old em houses in rock and .. I I feel it's a bit er .. a pity* (average fluency score: B1)
(2) *yeah yeah . well I'm thinking about doing erm an an another year but I'm not sure where in Brussels or I want to go to: England but I I have to try and convince my boyfriend to do so* (average fluency score: C1)

We also computed the number of errors in the samples and looked at how these figures relate to the average scores for accuracy.[5] Interestingly, some of the samples rated differently by the teachers display similar numbers and types of errors. Thus, one sample rated B1 and another one rated C1 have 4.1 and 3.7 errors per 100 words, respectively. They both include incorrect prepositions (*in* instead of *to* for a direction in both samples), misused connectors (*on the contrary* instead of *on the other hand* in the B1 sample and *even if* instead of *even though* in the C1 sample), as well as problems of tenses (use of a present tense instead of a past tense in both samples) or of verb agreement (*the camp are* in the B1 sample and *he become* in the C1 sample). While such errors might possibly correspond to the B1 accuracy descriptor of a repertoire of routines and patterns used 'reasonably accurately', they do not seem to match the C1 descriptor of '[c]onsistently maintain[ing] a high degree of grammatical accuracy'. This could partly be explained by the fact that students do not always make actual errors when they do not know how to express themselves. Sometimes, they use different strategies, such as the use of vague language or unfinished statements, which

do not count as errors but might still have a negative impact on the score attributed. This is very clear in one of the samples, which was given an average accuracy score of B1 despite a very small number of errors (1.3/100 words). This sample includes three occurrences of *something like that* and five occurrences of *I don't know*, as illustrated in (3) and (4).

(3) *they don't mind if I come late or something like that but they want to know where I I am* (average accuracy score: B1)
(4) *I I don't know maybe here it's a catholic university and in Brussels not or .. I don't know* (average accuracy score: B1)

Among the ten samples evaluated by the teachers, only one reached an average score of C2. This score was reached on each of the six scales. Surprisingly, however, this sample does not necessarily obtain the best results if we compute some linguistic measures traditionally used to assess the quality of texts, such as speech rate, frequency of filled and unfilled pauses, or lexical sophistication. A major exception is the number of errors, which is the lowest of all samples (only 0.3 errors per 100 words). This seems to confirm the important role of accuracy among teachers who evaluate learner speech (see above), perhaps even to the extent that they attribute higher scores on all scales, even if objectively the linguistic quality of the sample for these other competences is not so high. It might also be that some of the traditional measures of text quality do not really capture the full picture, as already suggested about the proportion of actual errors. Lexical sophistication is another case in point. It is typically measured by calculating the proportion of types and tokens belonging to different frequency levels. Such an approach, however, only takes single words into account, not combinations of words (phrasal verbs, collocations, idioms, etc.). The C2 sample may not include the highest number of sophisticated individual words, but examples (5) to (7), all taken from this sample, reveal the presence of relatively simple words which, combined together, form idiomatic phrases testifying to the lexical sophistication of this learner's oral production (cf. boldface).

(5) *he said well that's too bad you had succeeded you had passed all the exams but I cannot* ***take*** *you* ***on*** (average range score: C2)
(6) ***the long and short of it*** *erm I . first graduated as . as a: régent so from a: . a teacher's training college* (average range score: C2)
(7) *it was . kind of difficult sometimes I had to* ***cut corners*** (average range score: C2)

Reflective practice

The teachers who took part in the action research were invited to participate in the reflective practice part of the study, which took the form of a one-day training session whose main objectives were to gather feedback from the teachers on the CEFR-based assessment experiment and to prompt more reflection among them on their own assessment practices. Of the 27 teacher-raters, 13 attended the session, together with 15 other teachers or researchers involved in teacher training.

The session started with a short introduction which emphasized the importance of speech assessment, but also the difficulty it represents. Research findings from the literature were reported which demonstrated the high degree of rater variability and provided some explanations for this variability. The action research part of the training activity was then briefly summarized, especially for the teachers and researchers who had not participated in the experiment, and the main results were described. Next, the participants were divided into small groups and asked to fill in a worksheet together, which consisted in listing the positive aspects / good practices, the negative aspects / bad experiences, and the comments / questions related to several aspects, viz.

- the assessment project: what did the teacher-raters find easy or difficult? what (could have) facilitated the process? what did the participants learn from the results? etc.
- the CEFR grid: do the teachers use it? what do they think about it? is it useful/easy to use? if they use other grids, how do these grids compare with the CEFR grid? etc.
- the linguistic competences: what competences should be distinguished/assessed? are some competences more important than others? are some of them more difficult to assess? etc.
- speech assessment: how do the teachers assess students' speech (task, scores, grids, etc.)? do they receive training in the assessment of spoken skills? etc.

The last part of the session was a general discussion where the participants shared their views and ideas. In what follows, we summarize the main points of the discussion and, where relevant, we establish links with the quantitative observations made above.

As far as the *assessment project* is concerned, although some positive elements were highlighted, such as the length of the audio samples which was sufficient to give a good overview of the learners' competence or the fact that not knowing the learners helped evaluate the samples more neutrally, most of the teacher-raters' comments pointed to the difficulty of the task they had been asked to complete, e.g. 'I found the global rating difficult' or 'I found it quite difficult to make a distinction between C1 and C2 for some competences'.[6] Evaluating a student's spoken performance without seeing him/her also turned out to be difficult, among other reasons because body language could not be taken into account. In addition, teachers seemed rather insecure, as shown by comments such as 'Assessment of speech being really difficult, I hope I was in the average of the other examiners' or 'I tried not to let myself be influenced and to remain objective'. This could partly account for the lower average scores attributed by the teachers in comparison to those given by the experts (see Figure 4), which, in effect, corresponded to teachers' scores being close to B2: their insecurity may have led them to avoid extreme scores (C1-C2) and instead opt for scores towards the middle of the scale. The decontextualized nature of the experiment also seemed to pose a problem for certain participants, who observed that the evaluation had to be carried out outside any learning environment.

Several participants underlined the problems linked with the *CEFR grid*, including the lack of precision of the linguistic competences and descriptors. For example, one

teacher pointed out that the descriptor referring to 'differentiating finer shades of meaning precisely' was extremely vague. These problems, combined with the difficulty of assessing speech reported above, probably explain the discrepancies between raters outlined in the section on Action Research. One thing that appeared clearly from the discussion was that the CEFR grids are not commonly used in secondary schools of the French-speaking Community of Belgium. Teachers tend to use their own homemade assessment grids, which they consider to be better suited for their specific needs. A comparison of the grids provided to us by some of the participants highlighted great diversity in the number and types of criteria used. For example, phonology and fluency were combined into one criterion in some grids and treated separately in others. Vocabulary and grammar displayed similar discrepancy. These differences cause problems for standardization. As pointed out by Sundqvist, Wikström, Sandlund, and Nyroos (2018), a non-standardized test may be adequate for formative assessment (assessment *for* learning) but it is problematic for summative assessment (assessment *of* learning). Particularly relevant in this connection is the fact that secondary schools are increasingly expected to bring students to a minimum CEFR-based proficiency level as part of school leaving or university entrance requirements (Plo, Hornero, & Mur-Dueñas, 2014).

As for the *linguistic competences*, the participants were asked to assign a weight to each of the five competences included in the experiment, according to how important they considered them to be when assessing EFL speech. The heaviest weight was assigned to range (average of 3.5 out of 4), followed by accuracy (3.25), then fluency and phonological control (2.9 each) and finally coherence (2.8). These results largely confirmed those based on the scoring in the experiment (see Table 3), with item-based aspects of language predominating. Interestingly, some of the teachers' comments suggested that even more global aspects of language such as coherence may in fact rely on the assessment of individual items, with one teacher, for example, pointing out that he assessed coherence on the basis of lists of connectors to be used by the learners. In addition to the five competences included in the experiment, some participants mentioned other competences that they thought could or should be taken into account, such as interaction (left out of the grid on purpose; see above), contents (ideas expressed by the learner), appropriacy/relevance (did the learner fulfil the task as required?) and flexibility (did the learner adapt to the circumstances?). The difficulty of evaluating many competences at the same time was mentioned, as well as the question of how the global score was to be determined (independently of the specific competences? by averaging the different competences? on the basis of differently weighted competences?) and when (before or after the evaluation of the specific competences?). The participants also related the linguistic competences to their teaching, noting for example the paradox of evaluating intonation when intonation tends not to be taught in secondary schools. This could possibly explain the weaker correlations among the teachers and between the teachers and the expert raters for phonological control (see Figures 2 and 3): some teachers may have used criteria that correspond to what is taught in class rather than the criteria included in the CEFR grids, thus leading to disagreement with the other raters who followed the CEFR more closely.

When asked to comment on *speech assessment* in general, the participants pointed out that ranking students was easier than rating them, and that the strongest and poorest performances were the easiest to evaluate, leaving the majority of performances in a difficult-to-evaluate middle ground. They also observed that, because of teachers' various practices in terms of speech assessment, and in particular their preference for their own homemade assessment grids, students in different classes, different years, different schools, etc. tended to be evaluated differently. It was suggested that an external exam, based on a common grid (perhaps a CEFR grid), would be more objective and would help estimate each student's standardized level. Finally, the teachers admitted that they did not feel well-equipped to assess speech. They regretted the lack of appropriate training and underlined the harsh realities of day-to-day teaching which prevented them from implementing recommended practices such as collaborative assessment or the use of audio recordings for subsequent re-listening.

Conclusion

The main objective of this study was to explore teachers' experience in applying the CEFR scales to assess EFL speech and to collect feedback from these teachers on the assessment of oral skills in general.

The quantitative results of the CEFR-based action research reported above highlighted a reasonably good degree of agreement among the teachers as well as between the teachers and the experts. However, the study showed that the use of the CEFR grids is insufficient to cancel the often-observed tendency for non-native teachers to be more severe in their evaluation than native experts. It also suggested that the teachers seemed to attach more importance to grammar and vocabulary, i.e. item-based aspects of language, while the native raters were arguably more sensitive to variables such as fluency and coherence, i.e. more global aspects of language. A more qualitative analysis pointed to some correspondence between the linguistic features of a text (such as the number of pauses) and the average score attributed to it by the teachers, while emphasizing that there was no one-to-one relationship between the two. A key aspect of our study is that it relied on authentic learner corpus samples, thus combining learner corpus research and assessment research. It also involved a large number of teacher-raters, whose rating behaviour is assumed to be representative of a larger population of teachers. However, it is important to bear in mind that our study was based on the evaluation of ten samples and that this represented a relatively small number of words. The study should therefore be replicated with more and longer extracts, and possibly two extracts per learner, given the potential variability within one and the same interview (García-Amaya, 2009). The study should also be replicated with different tasks or genres, since several of the linguistic aspects we have examined are task-sensitive.

The reflective practice part of the activity highlighted the teachers' general lack of confidence with the assessment of speech and their wish for training in this area. They had very little knowledge of the CEFR grids and reported using homemade assessment grids in their daily activities. A number of teachers pointed out that the

difficulty of the assessment task was compounded by the vagueness of certain CEFR descriptors. Some of their comments could be related to quantitative findings from the action research part, such as the possible link between the failure to teach intonation in class and the weak correlations among teachers and between teachers and experts for phonological control. Despite these difficulties, it is interesting to note that teachers did quite well, as on the whole there are relatively good correlations among the teachers and also between the teachers and the expert raters (albeit with a one-band difference). It might be that thanks to their experience in oral evaluation, teachers have developed an intuitive notion of oral proficiency, and that this compensates for their unfamiliarity with the CEFR and the vagueness of some descriptors. This is still a very speculative idea at this stage and we need further analyses on more data to elaborate on this and other aspects of CEFR-based assessment.

An important pedagogical implication of our study is that teachers would greatly benefit from training in the assessment of spoken skills, which would help them achieve a higher degree of professionalism in their assessment practices and generally contribute to boosting their confidence in this area. We agree with Huang, Kubelec, Keng, and Hsu (2018, p. 13) that "rating with the CEFR descriptors incurs a great deal of subjective judgment from assessors unless they are trained". Their study shows that the provision of a rating training activity based on spoken learner corpus data and involving a thorough introduction to the CEFR scales improves the proportion of correctly assigned CEFR levels. Training activities of this type can contribute to the standardization of assessment practices, a particularly desirable outcome as standardization is gaining importance in education. Fortunately, tools have recently been developed to help teachers gain a higher level of language assessment literacy, such as the online Moodle-based training course designed by experts from six different European countries (Tsagari et al., 2018).

The main strength of our study lies in its focus on teachers, and especially non-native teachers. They are the ones who are most strongly confronted with the difficulty of evaluating learner speech and it is therefore important to empower them by providing them with clear guidelines. However, success will only be guaranteed if teachers' daily practices are duly analysed. It is essential to make teachers' voices truly heard in pedagogical development and educational policy, and the combination of action research and reflective practice implemented in the current study seems like a particularly effective instrument to achieve that objective.

Notes

1 The updated version of the CEFR (Council of Europe, 2018) was not available at the time we conducted the study. One of the main changes to the 2001 descriptors concerns phonological control, which focuses on intelligibility and no longer contains any reference to the native speaker norm.

2 For a study that seeks to investigate the impact of training on CEFR rating, see Huang, Kubelec, Keng, and Hsu (2018).

3 In these boxplots, the horizontal line indicates the median correlation, i.e. the value such that half of the correlations is smaller than or equal to it and the other half greater or equal. The box indicates where the 50% of the correlations closest to this median lie. The ends of the vertical lines (whiskers) indicate the most extreme correlations observed.

4 Regarding the coherence dimension, the lower whisker indicates that a correlation between one of the teachers and the experts was equal to zero. It seems as if this teacher had a specific problem with this dimension since his evaluations were not only uncorrelated with those of the experts, but also very weakly correlated with those of the other teachers, whereas for the other dimensions this teacher was within the average range.

5 The first and last authors of this paper identified all the grammatical and lexical errors in the sample transcripts independently, and then discussed the few cases of disagreement together. This made it possible to reach an agreement in all cases.

6 The difficulty of differentiating between C1 and C2 is a very common one, enhanced by the fact that the C2 descriptors for phonological control are the same as those at the C1 level. In fact, a test such as Aptis General does not distinguish at all between C1 and C2 (see O'Sullivan, Dunlea, Spiby, Westbrook, & Dunn, 2020, p. 25).

Acknowledgements

This research was carried out within the frame of the ARC-project "Fluency and disfluency markers: A multimodal contrastive perspective", funded by the Fédération Wallonie-Bruxelles (grant nr.12/17-044). We are grateful to the teachers who participated in the project, for generously giving of their time and for sharing their experience and expertise with us.

Biodata

Gaëtanelle Gilquin is Professor of English Language and Linguistics at the University of Louvain, Belgium, and a member of the Centre for English Corpus Linguistics. She is the coordinator of the *Louvain International Database of Spoken English Interlanguage* and one of the editors of *The Cambridge Handbook of Learner Corpus Research*. Her research interests include speech and writing fluency as well as the pedagogical applications of corpus linguistics.

Yves Bestgen is a Research Associate of the National Fund for Scientific Research (F.R.S.-FNRS) and part-time Professor at the University of Louvain, Belgium. He is a member of the Institute for Psychological Sciences. His main research interests focus on the development of techniques for automatic text analysis, especially in the field of multilingualism and opinion mining. He also develops statistical methods for corpus linguistics.

Sylviane Granger is Professor Emerita of English Language and Linguistics at the University of Louvain, Belgium. She is the founder of the Centre for English Corpus Linguistics, of which she was Director for over 25 years. In 1990 she launched one of the first large-scale learner corpus projects, the *International Corpus of Learner English*, and since then has played a key role in defining the different facets of the field of learner corpus research. Her current research interests focus on the integration of corpus data into a range of user-oriented tools, in particular electronic dictionaries and writing aids, with a special focus on phraseology. She has written widely on these topics and gives frequent invited talks, seminars and workshops to stimulate learner corpus research and promote its application to materials design and development. Her latest book publications include *The Cambridge Handbook of Learner Corpus Research* (Granger et al., 2015), *The International Corpus of Learner English* (Granger et al., 2020) and *Perspectives on the L2 Phrasicon: The View from Learner Corpora* (Granger, 2021).

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659-663.

Amengual-Pizarro, M., & García-Laborda, J. (2017). Analysing test-takers' views on a computer-based speaking test. *Profile: Issues in Teachers' Professional Development, 19*, Suppl. 1, 23-38.

Boud, D., Keogh, R., & Walker, D. (1985). What is reflection in learning? In D. Boud, R. Keogh, & D. Walker (Eds.), *Reflection: Turning experience into learning* pp. 7-17. London & New York: Routledge.

Chen, W., Lenderking, W., Jin, Y., Wyrwich, K.W., Gelhorn, H., & Revicki, D.A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality Life Research*, *23*, 485-493.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual*. Strasbourg: Language Policy Division.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors.*

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Friginal, E. (2005). Review of S. Luoma, Assessing speaking. *TESL-EJ, 9*(3). Retrieved from http://www.tesl-ej.org/pdf/ej35/r6.pdf

García-Amaya, L. (2009). New findings on fluency measures across three different learning contexts. In J. Collentine, M. García, B. Lafford, & F. Marcos Marín (Eds.), *Selected proceedings of the 11th Hispanic Linguistics Symposium* pp.68-80. Somerville, MA: Cascadilla Proceedings Project.

Gilquin, G., De Cock, S., & Granger, S. (2010). *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-la-Neuve: Presses universitaires de Louvain.

Harrison, J., & Barker, F. (Eds.) (2015). *English Profile in Practice*. Cambridge: Cambridge University Press.

Howell, D. C. (2007). *Statistical methods for psychology* (6th edition). Wadsworth: Cengage Learning.

Hu, X. (2021). Predicting CEFR levels in L2 oral speech, based on lexical and syntactic complexity. *Asia Pacific Journal of Corpus Research, 2*(1), 35-45.

Huang, L. F, Kubelec, S., Keng, N., & Hsu, L-H. (2018). Evaluating CEFR rating performance through the analysis of spoken learner corpora. *Language Testing in Asia, 8*, 1-17.

Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* pp. 131-146. Berlin: De Gruyter.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24-49.

Kang, O., & Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing & Assessment, 1*, 24-39.

Kim, A.-Y., & di Gennaro, K. (2012). Scoring behavior of native vs. non-native speaker raters of writing exams. *Language Research, 48*(2), 319-342.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K. (2020). *Aptis General Technical Manual. Version 2.2*. Technical Report. British Council. Retrieved from https://www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf

Plo, R., Hornero, A., & Mur-Dueñas, P. (2014). Implementing the teaching/learning of oral skills in secondary education in Aragón: Gauging teachers' attitudes, beliefs and expectations. *International Journal of English Studies, 14*(1), 55-77.

Ramsey, P. H. (1989). Critical values for Spearman's rank order correlation. *Journal of Educational Statistics, 14*(3), 245-253.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14,* 423-441.

Schoonjans, E. (2012). A contextual perspective on oral L2 fluency. In L. Roberts, C. Lindqvist, C. Bardel, & N. Abrahamsson (Eds.), *EUROSLA yearbook 12* pp. 135-163. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Sundqvist, P., Wikström, P., Sandlund, E., & Nyroos, L. (2018). The teacher as examiner of L2 oral tests: A challenge to standardization. *Language Testing, 35*(2), 217-238.

Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169-191.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *Modern Language Journal, 97*, 77-101.

Tsagari, D., Vogt, K., Froehlich, V., Csépes, I., Fekete, A., Green, A., Hamp-Lyons, L., Sifakis, N., & Kordia, S. (2018). *Handbook of Assessment for Language Teachers.* Retrieved from http://taleproject.eu/. ISBN 978-9925-7399-1-2

Walsh, S., & Mann, S. (2015). Doing reflective practice: A data-led way forward. *ELT Journal, 69*(4), 351-362.

Wisniewski, K. (2018). The empirical validity of the Common European Framework of Reference scales. An exemplary study for the vocabulary and fluency scales in a language testing context. *Applied Linguistics, 39*(6), 933-959.

Wyatt, M. (2011). Teachers researching their own practice. *ELT Journal, 65*(4), 417-425.

Appendix 1: CEFR descriptor scales (Council of Europe, 2001, pp. 28-29, 117; Council of Europe, 2009, p. 184)

Linguistic Competence	A2	B1	B2	C1	C2
Range	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocuti-ons on topics such as family, hobbies and interests, work, travel, and current events.	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Shows great flexibility reformulat-ing ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquiali-sms.

Accuracy	Uses some simple structures correctly, but still systematical-ly makes basic mistakes.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunders-tanding, and can correct most of his/her mistakes.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).
Fluency	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulat-ion are very evident.	Can keep going comprehens-ibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can express him/herself fluently and spontane-ously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can express him/herself spontane-ously at length with a natural colloquial flow, avoiding or backtrack-ing around any difficulty so smoothly that the interlocutor is hardly aware of it.

Phonological control	Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.	Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.	Has a clear, natural, pronunciation and intonation.	Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.	As C1
Coherence	Can link groups of words with simple connectors like “and”, “but” and “because”.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some “jumpiness” in a long contribution.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.

Global assessment	***Relates basic information on, e.g. work, family, free time etc.*** Can communicate in a simple and direct exchange of information on familiar matters. Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can describe in simple terms family, living conditions, educational background, present or most recent job. Uses some simple structures correctly, but may systematically make basic mistakes.	***Relates comprehensibly the main points he/she wants to make.*** Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Can link discrete, simple elements into a connected, sequence to give straightforward descriptions on a variety of familiar subjects within his/her field of interest. Reasonably accurate use of main repertoire associated with more predictable situations.	***Expresses points of view without noticeable strain.*** Can interact on a wide range of topics and produce stretches of language with a fairly even tempo. Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Does not make errors which cause misunderstanding.	***Shows fluent, spontaneous expression in clear, well-structured speech.*** Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Can give clear, detailed descriptions of complex subjects. High degree of accuracy; errors are rare.	***Conveys finer shades of meaning precisely and naturally.*** Can express him/herself spontaneously and very fluently, interacting with ease and skill, and differentiating finer shades of meaning precisely. Can produce clear, smoothly-flowing, well-structured descriptions.

Appendix 2: Instructions received by the teachers for the CEFR rating

Here are a few instructions to guide your rating of the extracts. If you have any queries do not hesitate to contact us. Please bear the following information in mind throughout the rating procedure.

1 The extracts

- A batch of 10 learner interview extracts (not classified according to any feature);
- Each extract corresponds to about 5 minutes of recording (total: +/- 50 minutes);
- Topics: a film the learner has seen, a country s/he has visited, life at university, etc.;
- The interviews are supposed to be relatively informal;
- The learners are French-speaking (Belgian) learners of English (English as a foreign language).

2 The rating procedure and the descriptors

- Overall rating procedure
    - The rating of the learner interviews is based on the *Common European Framework of Reference* (CEFR) descriptors.
    - The rating procedure consists in 3 successive steps: (1) CEFR grades for 5 specific competences, (2) CEFR grade for global assessment and (3) additional comments (optional).
- The *Common European Framework of Reference* descriptors
    - The CEFR gives the descriptors for five competences at each level (the descriptor table you have received is the one you should use):
        1. range (= lexical complexity/richness/sophistication)
        2. accuracy (= grammatical accuracy)
        3. fluency
        4. phonological control
        5. coherence
    - The descriptors target five distinct CEFR grades: A2, B1, B2, C1 and C2, corresponding to basic (A), intermediate (B) and advanced (C) levels of proficiency.

- 1st step: detailed rating
    - For each extract please give a **CEFR grade to each of the five competences**. You can give the same grade for each competence or a different one, as illustrated in the following example:

      e.g. for extract x:

      1. range = B1
      2. accuracy = C1
      3. fluency = B2
      4. phonological control = C1
      5. coherence = B2

      In some cases, you may feel the need to further distinguish sublevels. We suggest you do this by using + or – signs. For example, C1- would represent weaker performance within the C1 band while C1+ would represent stronger performance within that band. The same goes for A2, B1, B2 and C2. The marks for each linguistic competence should be inserted in the Excel table you have been sent (columns B to F).

- 2nd step: global rating
    - Once you have done this, please also award **one global CEFR score** (a holistic score) to each of the 10 extracts, either A2, B1, B2, C1 or C2. This score should be based on your overall impression of the proficiency displayed in each extract and on all the descriptors taken overall. Concerning the global score, you may again wish to further distinguish between sublevels: we suggest you do this by using + or – signs. For example, B2- would represent weaker overall B2 performance while B2+ would represent stronger overall B2 performance. The global score should be inserted in column G in the Excel table.

- 3rd step: comments
    - Any **additional comments** you may wish to make on a particular extract are welcome and should be included in the Excel table in column H.

- Please make sure you listen to the audio files in numerical order so that you respect the sorting of the extracts.

Thank you very much for your collaboration!