TAEC CORPUS REPORT

The report is part of a project is co-funded by the Erasmus+ programme of the European Union.

Erasmus+

TAEC
Transnational Alignment of English Competences for University Lecturers

1. Introduction

The number of English medium instruction (EMI) courses and programs at universities in countries where English is used as a foreign language (FL) has been growing in recent years. This development has raised many questions about the language abilities of EMI lecturers who are FL speakers of English and the quality of instruction in the EMI courses. Researchers have examined EMI content lecturers' classroom communication in English from linguistic, pedagogical, and intercultural communication perspectives, but their analyses tend to focus on a small number of lecturers from one EMI program or university.

As part of the project *Transnational Alignment of English Competences for University Lecturers (TAEC)*, funded by ERASMUS+ program, the ***TAEC Corpus*** was developed to provide an opportunity for researchers and teacher trainers to compare EMI content lecturers' classroom communication across different universities and countries. The corpus delivers transcripts of content lecturer English use in real EMI classrooms in different contexts. These lecturers use English on a daily basis in their academic lives, so it is worth investigating different aspects of their language uses for teaching purposes.

2. The TAEC Corpus

The corpus comprises transcripts of naturally occurring, non-scripted face-to-face interactions in EMI classrooms. The video recordings collected for the ***TAEC Corpus*** are keyboarded and annotated by trained transcribers and stored as a computerized corpus. The transcription and annotation system is provided in the appendix (see Appendices 1 and 2). Currently the ***TAEC Corpus*** comprises 30 transcripts from 30 different lecturers at five universities. The recorded lectures are not held in the first or the last week of the semester in order to avoid lecturers' unfamiliarity with students or large focus on exams.

The lecturers recorded in the ***TAEC Corpus*** are experienced lecturers from seven different language backgrounds (Afrikaans, Catalan, Croatian, Danish, Dutch, German, and Italian). The corpus includes lectures at undergraduate (n=17) and graduate (n=13) level across three broad disciplinary fields, Social Sciences and Humanities (SH), Life and Medical Sciences (LS), and Physical Sciences and Engineering (PE).

In order to contextualize the transcripts, the classroom settings, the type of delivery, and the level of interactivity are described. The classroom settings are described in terms of classroom size (auditorium, computer lab, and small or large classroom) and in terms of seating (fixed seating or movable chairs). Type of delivery is operationalized as dynamic, static, or mixed, depending on to what degree the lecturer moves in the classroom. Interactivity is operationalized as the degree to which the lecturer interacts with the students through classroom discussions, questions, or other classroom activities that involve students. The characteristics of the recorded lecturers and lectures are presented in Tables 1 and 2.

Gender	Female (n= 12) Male (n=18)
Position	Assistant professor (n=6) Associate professor (n=9) Full professor (n=8) Full-time lecturer (n=5) Part-time lecturer (n=2)
Language	Afrikaans (n=1) Catalan (n=6) Croatian (n=6) Danish (n=6) Dutch (n=3) German (n=1) Italian (n=7)

Table 1. Lecturer characteristics

Level	Undergraduate (n=17) Graduate (n=13)
Disciplinary area	SH (n= 18) LS (n=10) PE (n=2)
Room size	Auditorium (n=2) Computer lab (n=1) Large classroom (max. 120 seats) (n=6) Small classroom (max. 60 seats) (n=21)
Room type	Movable chairs (n=17) Fixed seating (n=13)
Interactivity	Highly interactive (n=4) Mostly interactive (n=5) Mostly monologic (n=9) Highly monologic (n=5) Mixed (n=7)
Delivery	Mostly dynamic (n=15) Mostly static (n=6) Mixed (n=9)

Table 2. Lecture characteristics

3. Challenges

In the initial stages of corpus development, the intention was to develop a balanced corpus by minimizing contextual and disciplinary variation. In other words, attempts were made that 1) all video recorded lectures are from related disciplines (e.g., economics, accounting), 2) an equal number of undergraduate and graduate level lectures from each university is represented, 3) the class sizes is similar, and 4) all classes have international students.

However, maintaining all criteria for the selection of lectures across the universities was challenging because of the local contextual differences. The EMI programs at the universities were not offered in the same disciplines, or the faculties and the study programs structure varied. At some universities a wider variety of EMI courses was offered at undergraduate, at others they were offered at graduate level. Moreover, different conceptualizations of class sizes existed across the contexts. For example, in some contexts a small class included 40 to 60 students, while in other contexts it included 10-15 students.

4. Transcription, annotation, and validation

The 30 video recorded lectures were transcribed and annotated following specific guidelines (see Appendices 1 and 2). The goal was to produce a written version of the interaction taking place in the classroom which would be easy to read and sufficiently detailed to permit an adequate comprehension of the speech event without having to watch the video. Only the words uttered by the lecturer and the students interacting with him/her were transcribed.

The guidelines comprise instruction for transcription proper and annotation (or mark-up). Transcription norms have to do with aspects such as spelling, the use of capital letters, acronyms, numbers and formulae. Annotation concerns features such as speaker turns, pauses, hesitations, language mistakes, contextual events, and the use of languages other than English.

The guidelines were initially developed based on the conventions used for existing spoken corpora, namely the *Michigan Corpus of Academic Spoken English* (MACASE), the *British Academic Spoken English* (BASE) corpus and the *English as a Lingua Franca in Academic Settings* (ELFA) corpus.

Three activities were organized to train the transcribers:

(1) Completion of a multiple-choice quiz;
(2) Analysis of a transcribed excerpt;
(3) Transcription of a 5-minute passage from a video recorded lecture.

The quiz included 15 questions on different aspects of the guidelines. Figure 1 provides an example of a multiple-choice question. The quiz average score was 94% indicating that the guidelines were well understood by the transcribers.

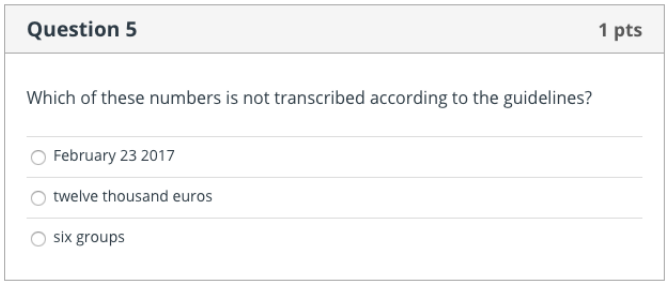

Figure 1. Multiple-choice question from the quiz

The analysis of a transcribed excerpt consisted in editing the passages that transcribers would have handled differently. Transcribers were not required to propose alternative conventions, but to verify whether they would have transcribed in the same way using the guidelines given.

The third step was the transcription of an excerpt from a video recorded class applying the set conventions.

The results of the activities were discussed among the transcribers to reach an agreement on how to deal with the aspects of the guidelines that proved most challenging to operationalize, such as the marking of pauses and language mistakes. The guidelines were hence revised taking into account the transcribers' feedback as well as specific issues emerged in the EMI lectures.

In order to ensure the consistent application of the transcription and annotation conventions, a validation process was put into place. Of each transcript, 10% was checked by two raters assigned randomly to the video recorded lectures. The chunks for validation were taken from the beginning, middle and end of each lecture. Slips or mistakes, including tags for contextual categories, were revised using the track changes in MS Word. Specific comments were added for the following categories suggesting insertions, deletions or replacements:

a) The use of the <SIC> and <PRON> labels (i.e. grammar and pronunciation mistakes)
b) Pauses
c) False starts
d) Truncated words
e) Question marks
f) Content (e.g. missing word or wrong word)

The validity of the transcriptions was assessed according to two measures: agreement with the current transcription and coverage. The agreement value was determined calculating the number of deletions and replacements suggested by the raters. The coverage value was determined based on the raters' new insertions. Table 3 presents the average values for agreement and coverage by raters and categories in the whole corpus.

	Agreement with the transcription			Coverage of the transcription		
	Average raters 1	**Average raters 2**	**Total average**	**Average raters 1**	**Average raters 2**	**Total average**
<SIC>	95.90%	97.71%	**96.80%**	*79.08%	84.29%	**81.69%**
<PRON>	93.06%	96.32%	**94.69%**	*76.98%	93.61%	**85.29%**
Pauses	93.81%	94.33%	**94.07%**	90.91%	89.41%	**90.16%**
False starts	98.14%	95.18%	**96.66%**	91.55%	97.31%	**94.43%**
Truncated words	96.30%	95.93%	**96.11%**	85.30%	89.16%	**87.23%**
Question marks	98.87%	99.63%	**99.25%**	95.16%	94.75%	**94.95%**
Content	99.49%	99.62%	**99.56%**	99.57%	99.58%	**99.57%**

Table 3. Agreement and coverage rates

The maximum divergence rate for acceptability was 20%. As can be seen from the totals, for no category was the average value below 80%. Considering the agreement with the transcription, the percentages are very high, the lowest being 94.07% for pauses and the highest 99.56% for content. These figures indicate that the corpus annotation and transcription guidelines were applied consistently, and that users can rely on the transcripts available in terms of the accuracy and regularity of the implemented conventions. As regards the values for transcription coverage, the highest figure is for content, i.e. 99.57%. This shows that the corpus reports the words uttered by the lecturers in a systematic way. The following categories are, in decreasing order, question marks (i.e. utterances with rising intonation) (94.95%), false starts (94.43%), pauses (90.16%), truncated words (87.23%), and the labels <PRON> (pronunciation inaccuracies or mistakes) (85.29%) and <SIC> (language mistakes) (81.69%).

The coverage of the labels <PRON> and <SIC> is on average acceptable. However, looking at the values for the raters 1 group (marked with an asterisk), the figures are slightly below 80%. This means that there may be occurrences not included in the current version of the corpus. Reasons for this result are the higher subjectivity in the recognition of language mistakes/inaccuracies compared to other categories and, most of all, the need for a more refined definition of what counts as a pronunciation or lexico-grammatical mistake. The <SIC> tag was initially devised to mark morphological mistakes, thus avoiding the risk for users of interpreting such mistakes as transcription inaccuracies rather than spoken performance ones. However, the tag was in fact adopted for a wider range of lexico-grammatical issues. Without a robust definition of what counts as a lexico-grammatical mistake, this resulted in a lower coverage rate compared to other categories. The aim of the tag <PRON> was mark the most evident forms of non-standard pronunciation. In this case, too, a more fine-grained definition of 'pronunciation mistake' would have led to higher coverage rates. It should be pointed out, however, that the purpose of the corpus transcription was to offer a written version of 30 spoken events that could be read and understood by users without having to watch the videos. Our goal was not to provide and error-tagged corpus. Hence, the agreement and coverage rates obtained can be considered satisfactory for the goals of the TAEC project.

5. Lecturer proficiency levels based on CEFR

Given that language proficiency is frequently used as an important variable in the analysis of EMI lecturers' communicative competence and classroom behavior, the 30 lecturers' performances were rated based on different aspects represented in the Common European Framework of Reference (CEFR) scales that are relevant for the EMI context. The rated aspects were: overall level, addressing audiences, range, accuracy, fluency, interaction, coherence, phonology, and mediation (see the scales in the Appendix).

5.1. Raters

All 30 video recordings of lecturers were rated by eight raters. Five raters were non-native speakers of English with different L1 backgrounds, and the other three were native speakers of English. All raters were teacher-trainers and/or university lecturers with background in linguistics, English for academic purposes, or EMI.

5.2. Procedures

All raters participated in a two-day norming session in which they watched benchmark performances at different CEFR levels, and then they discussed the performances referring to the different scale descriptors. Then, the raters watched and rated 10 more videos with EMI lecturer performances and then discussed their ratings in groups in order to reach agreement.

In the period of two weeks after the norming session, the raters rated the lectures in the 30 video recordings from the corpus. They were instructed to follow this procedure:

TAEC LECTURERS RATING PROCEDURE

Please follow these steps when rating the videos from the TAEC corpus.

You should rate all 30 videos individually, i.e. you should decide on the level without any discussions with others.

If you like to discuss your ratings with others, then you should do the following:

1. Rate first the six videos from your university **individually**. It is important that the levels you assign are not influenced by other rater opinions.
2. Once you are done rating the six videos, you can discuss your ratings with others. However, **DO NOT** change your ratings retroactively based on the discussion. The purpose of these discussions is to calibrate your rating and adjust for the next set of videos. Therefore, when you discuss the videos, please refer to the CEFR descriptors to justify you rating.
3. Please inform me about when the discussions happened (e.g., which performances you discussed) if you decided to have any.

<u>Before rating</u>

1. Go over the CEFR descriptors (Speaking, Mediation, Audience awareness, Overall speaking)
2. Watch the benchmark videos (C2, C1, B2) and read the descriptions of the performances.
3. You can revisit the benchmark videos any time you think you need to.

Rating

1. While watching the video, decide on the level by referring to the benchmark performances (C2, C1, B2) you saw. You do not have to listen to the entire video, but you should watch about 10-15 minutes. You should watch parts that include both monologue and interaction.
2. After you decide on the overall level, start thinking about the levels related to each of the other criteria. You can listen to the video again, or re-run parts of it, if you need to. If there is no evidence of certain aspects (e.g., no repair), please enter N/A. Do not change the overall level after you have entered the levels for each of the other criteria.

After rating

1. Every time you rate six videos, you could discuss your rating with other colleagues, if you think you need a discussion.
2. Do not change your ratings during or after the discussion even if you realize that your ratings are lower or higher.
3. When you are done with all 30 videos, please send the 1) excel sheet to me and 2) a rating report. The rating report should include information about a) the order in which you listened to the videos, b) whether you had discussions with others, c) which videos were part of the discussion, d) when the discussion happened (after which video), as well as a e) short reflection on the process.

5.3. Results

When more than two raters are used, Chronbach's alpha coefficient and interclass consistency (ICC) are appropriate consistency estimates of interrater reliability, i.e. how consistently the raters use the rating scale. Cronbach's alpha helps examine the degree to which the ratings from a group of judges are similar when measuring a common dimension (Stemler & Tsai, 2008). Interclass correlation coefficient (ICC), on the other hand, is a more conservative estimate of interrater reliability because it confounds two ways in which raters differ: consensus (mean differences) and consistency (association). When the Chronbach's alpha coefficient and ICC are closer to 1, it means that the raters' agreement is very high (Stemler & Tsai, 2008). A high level of interrater consistency was found, Chronbach's alpha=.936 and ICC=0.896-0.965 at 95% Confidence Interval. As can be seen from Table 4, the lecturers obtained higher scores for mediation, interaction, and audience awareness, while phonology was rated lowest.

	Mean	Std. Deviation
overall speaking	6,7083	1,35204
audience awareness	7,4583	1,16616
range	6,9167	1,36574
accuracy	6,2917	1,74987
fluency	6,6875	1,66511
interaction	7,2766	1,55622
coherence	6,6596	1,49312
phonology	5,9583	1,91254
mediation	7,1957	1,32698

Table 4. Means and standard deviations for all CEFR scales

Among the 30 lecturers, the overall scores ranged between B2 and C2 level on the CEFR scale. The overall scores were calculated by averaging the scores assigned by each of the eight raters. Given the restricted proficiency range, the raters used +/- to indicate the higher and the lower ends of the proficiency bands. The 30 lecturers were normally distributed: four were at C2 level, two were at C2- level, three were at C2+ level, 12 were at C1 level, three were at C1- level, two were at B2+ level, and four were at B2 level (see Table 5).

B2	B2+	C1-	C1	C1+	C2-	C2
4	2	3	12	3	2	4

Table 5. Distribution of CEFR levels

6. Fluency

Given that fluency correlates strongly with perceptions of oral language proficiency, five fluency variables [mean syllables per run (MSR), speech time (ST), speech rate (SR), silent pause time (SPT), and filled-pause time (FPT)] were analyzed for 10 of the 30 recorded EMI lecturers (two per university). These fluency variables were selected because they were found to be best predictors of oral proficiency in previous literature (Ginther, Dimova, & Yang, 2010). MSR represents the average numbers of syllable produced between pauses (both filled and silent). MSR is calculated by dividing the total number of syllables by the number of runs. ST is the amount of time spent producing utterances without the time spent on fillers and silent pauses, while SR measures the overall speed of production by dividing the number of syllables produced within the 180 seconds. SPT is the total time in seconds of all silent pauses, while FPT is total time in seconds of all filled pauses. The speech samples used for analysis were taken from two parts of each lecture, each 180 seconds in length. The first part (A moments) occurs within the first 10 minutes of the lecture, whereas the second part (B moments) occurs between minute 25 and minute 45 of the lecture. We selected moments when lecturers were in lecturing mode, i.e. providing explanations, examples or definitions, rather than interacting with students. The analysis of these fluency variables in both speech samples from each lecturer (A and B moments) suggests strong alignment with the EMI lecturers' proficiency levels

Figure 2 shows the distribution of MSR in Moments A and B based on the CEFR level of the 10 lecturers. As can be seen from the figure, L20, who is at C2 level, has the highest MSR score. He is followed by five lecturers at C1 level, and L15 and L30, who are at B2 level. Only L08 and L22 seem to diverge from the general alignment; although L08 is at C2- level, the lecturer's MSRs are lower than most lecturers at C1 level. L22's MSRs, on the other hand, are lower than the ones of the B2 level performances although the lecturer is at a C1 level.

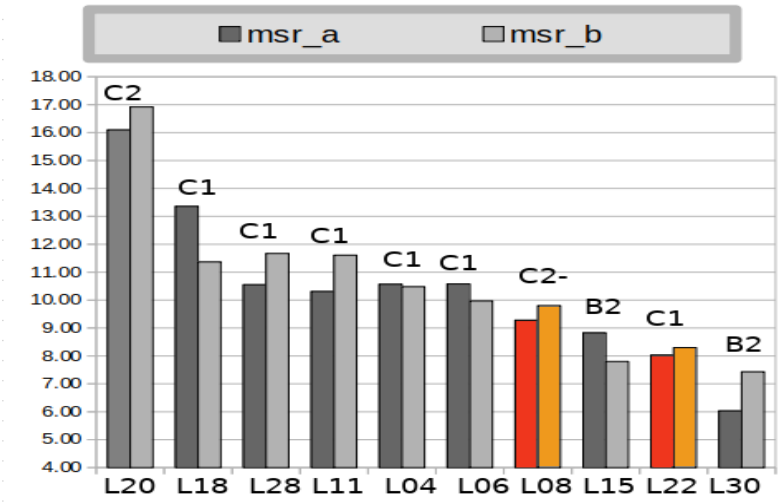

Figure 2. MSR in A and B moments of the lectures

Figures 3 and 4 represent MSR's evolution throughout the 180 seconds in each of the two speech samples (Moments A and B). In these figures, the number of syllables appears in the y axis, whereas the runs appear in the x axis. For instance, the EMI lecturers L15 and L30 (in their A moments) start with a similar pattern. However, between runs 27 and 32 (see y axis), L15's production of syllable increases more rapidly. The production flattens out for a few runs, but it goes up again in runs 44-45 and in runs 56-57. In the end, L15 produces 530 syllables in 60 runs (MSR of 8.83 syllables/run). L30 maintains the same pace, and ends up producing 392 syllables in 65 runs (MSR of 6.03 syllables per run). In the B moments of these two lecturers, though, both follow a very similar pace and in fact end up producing almost the same number of syllables in the same number of runs; L15 produces 476 syllables in 61 runs (7.80 syllables per run) and L30 produces 446 syllables in 60 runs (7.43 syllables per run).

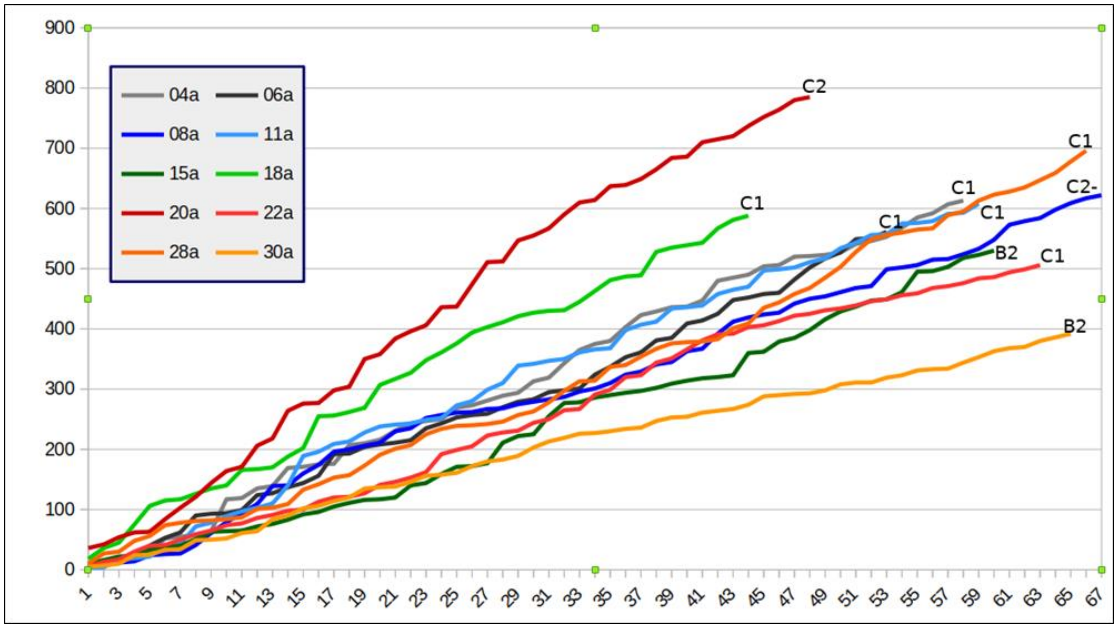

Figure 3. Syllables and runs for A moments in the lecture

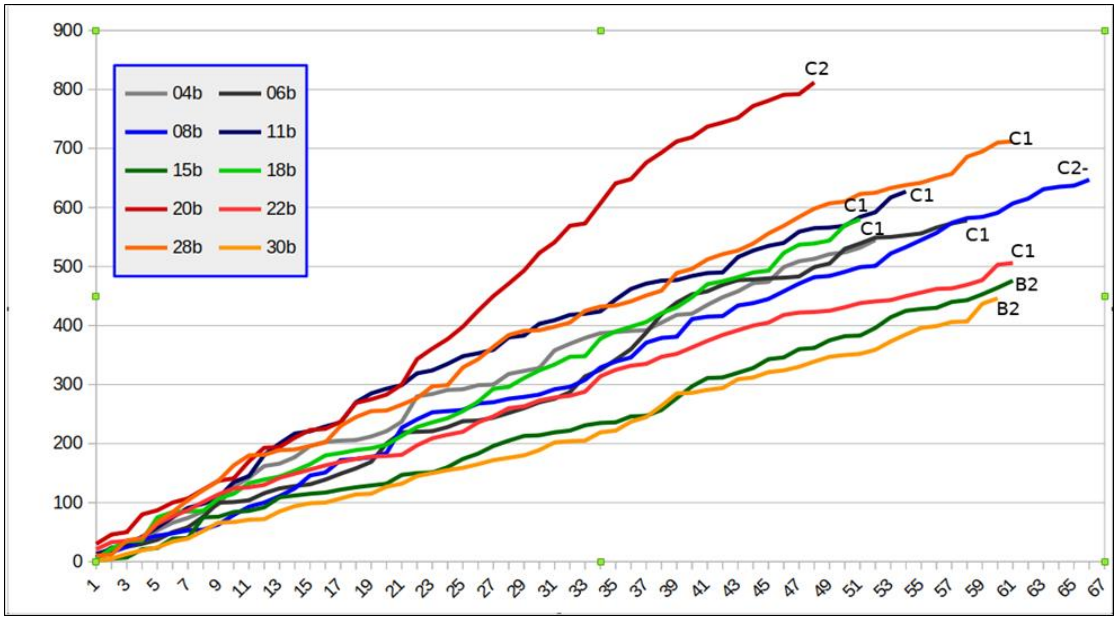

Figure 4. Syllables and runs for B moments in the lecture

Figure 5 shows the distribution in percentage of the time spent producing meaningful syllables (Speech Time), in silent pauses (SPT) or in filled pauses (FPT). According to the figure, EMI lecturers spent 78.5% of the time producing utterances on average. Only L22's production was under the 70% line. The average for FPT is 5.8%, but the filled pauses varied across lecturers.

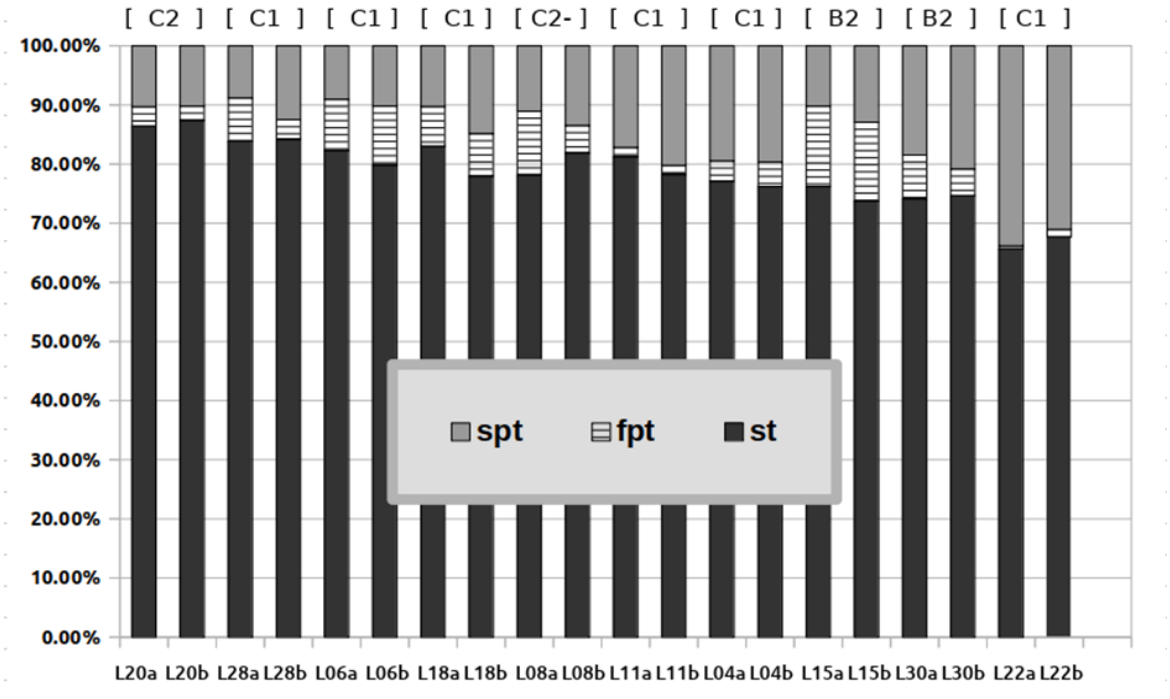

Figure 5. Speech Time, Filled-Pause Time and Silent-Pause Time

In terms of SR, Figure 6 suggests that SR is aligned with the CEFR proficiency levels of the EMI lecturers. The EMI lecturer L08, who is at C2- level, produces speech faster than all C1 lecturers, while the speech rate of the lecturer L22 is in line with that of the lecturers rated C1.

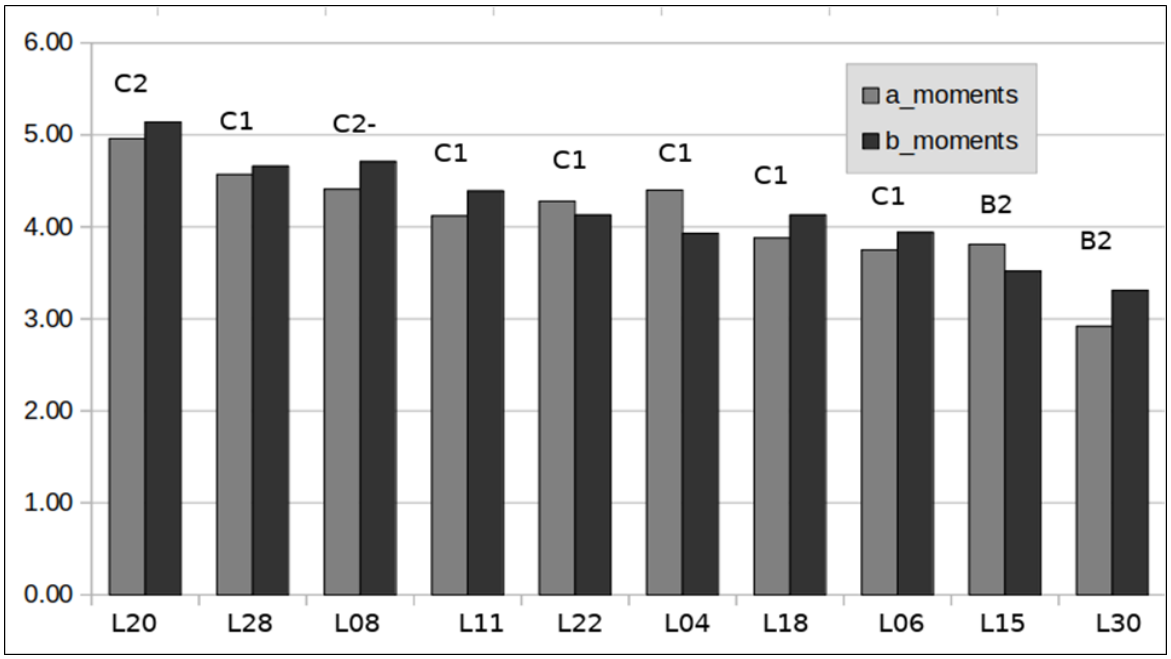

Figure 6. Speech rate in syllables per second

Overall, all five fluency measures align well with the CEFR ratings. The results suggest that one fluency measure may not be sufficient to understand the role of fluency in lecturers' proficiency. For example, MSR is considered an important indicator of proficiency because of the assumption that lower proficiency level speakers cannot produce many syllables between two pauses because their speech production is not automatized. In other words, they need to pause often in order to retrieve and articulate the necessary linguistic structures. If only MSR is taken into consideration, then the lecturer L08's assessment at C2- level seems inconsistent with his MSR, which seems lower than most lecturers rated at C1 level. However, when SR is considered, then it becomes apparent that L08 is able to produce more utterances within the same time slot than the lecturers at C1 level. The speed of production suggests that the lecturer's speech is automatized, with no need for time to retrieve the linguistic structures needed. One may hypothesize that L08 uses frequent pauses to allow the listeners to process information before proceeding. Therefore, it is recommended that fluency analyses include more than one fluency variable.

References

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379-399.

Stemler. S.E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J.W. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 29 – 49). California: Sage Publications, Inc.

Appendices

Appendix 1: Transcription guidelines

	Guidelines	Examples
Spelling	• Use standard orthography, even when words are pronounced with a foreign accent. When in doubt about spelling, look the word up in the *Oxford English Dictionary* (OED online): http://www.oed.com/	*released heat* (even if you hear /ri'lized i:t/ instead of /rɪ'li:st hi:t/)
	• Use British spelling conventions.	*centre* (not *center*), *travelling* (not *traveling*);
	• Use *-ize, -yze, -ized, -ization* (international orthography)	*realize, analyze, standardized, globalization*
Capitalization	• Capitalize proper nouns and words that in English are normally spelt with capital letter, e.g. languages, days of the week, months (also see 'acronyms and abbreviation' and 'letters used as variables').	*the University of Copenhagen* *Croatian* *Monday* *April*
	• Do not capitalize the beginning of turns	*<S1> okay, okay good evening everybody*
	• Do not capitalize the 1st person pronoun *I*	*i've counted most of you*
Acronyms and abbreviations	• Write initialisms and acronyms in capital letters	*USA, EU, NATO, IELTS*
	• If the speaker intentionally spells out a word or acronym, use a hyphen between the letters	*P-K-A* *W-W-W-dot-NATO-dot-org*
	• Write abbreviated titles according to standard orthography (no periods)	*PhD, Dr, Mr, Mrs*
	• OK must be fully spelt out	*okay*

Letters used as variables	• Letters used as variables are written in capital letters. If pre- or post-modified, use a hyphen.	*F-of-X* (reading of the function f(x)) *Z-zero*
	• Greek letters used as variables are fully transcribed. If pre- or post-modified, use a hyphen.	*alpha-squared, beta-squared*
Contractions	• Transcribe all standard contractions of *is, am, are, have, had, can, could, should, would, might, not* in normal orthography.	*i'm, you're, she's, they've, we'll, it'll, i'd, i've, couln'd, wouldn't, should've,* etc.
Lexicalized reduced forms	• Transcribe lexicalized reduced forms as they appear in the OED.	*'cause, dunno, gonna, gotta, kinda, sorta, wanna,* etc.
Hyphenated words	• Follow standard hyphenation rules. When in doubt, check the spelling of lemmas in the OED.	*socio-economic* *email* (this is the lemma in the OED, not *e-mail*)
Numbers and dates	• Numbers are written as numbers, except those smaller than 10. • Dates are written as numbers, except centuries.	*65,000; two* *2018; twentieth century*
Punctuation	• The only punctuation mark used in a syntactic sense is the question mark (see 'pauses' below). It signals utterances that function pragmatically as questions, including comprehension checks and rhetorical questions. No capital letter after the question mark.	*so my question is what is this circuit doing?* *so what is the behaviour of a T-flip-flop?* *and this is my main concern okay?*
Filled pauses, hesitations, backchannels, exclamations	• Filled pauses, hesitations, backchannels and exclamations are spelled out.	Filled pauses/hesitations: *eh* Backchannels: *mhm* Exclamations: *aha, oho, oops,* etc.
Repetitions	• All repetitions of a phrase or word are transcribed.	*so you don't need you don't need to issue another clock cycle* *let's pick eh the the result we got last week*
Self-repairs	• A self-repair is an error-correction sequence involving amendments of unintended form or meaning. All words are transcribed.	*it's better to thinks to think*
False starts	• A false start occurs when the speaker does not conclude what he/she is saying but starts again with a new sequence. An underscore is used at the end of the last word of the abandoned string.	*this_ every group groups will be formed and assignments will be let's say fixed* *so the whole_ those who bring you the energy*

Truncated words	• Cut-off words are transcribed with a hyphen at the end of the last audible sound.	*paid by the European com- commission*
Mispronunciation	• MISPRONOUNCED WORDS are transcribed in the standard form (when it is possible to understand the intended word): the <PRON> tag is added before the mispronounced words, while the </PRON> tag is added after the mispronounced word.	and you were more <PRON> involved </PRON> or less <PRON> involved </PRON> than, than watching TV?
	• THIS/THESE: It may be hard to tell whether a speaker says *this* or *these*. 1) When it is impossible to distinguish precisely what they say, transcribe the correct form that should come with the subsequent word (in singular or plural form respectively).	1) *so i have these sensors i have this technology i have this plant* (here it is not clear whether the speaker uses the singular or plural form; hence the correct form is transcribed)
	2) When it is clear that grammatical cohesion is missing, then add both <PRON> and <SIC> tags	2) *so i have <PRON> <SIC> this </SIC> </PRON> sensors i have this technology i have this plant* (here it is clear that the speaker uses the singular form, which creates a grammatical mistake; hence the wrong word is tagged)
	• SLIPS OF THE TONGUE: When a slip of the tongue occurs and is not self-repaired, transcribe what you hear and enclose it between the labels <SIC> and </SIC>, thus avoiding making the word appear a transcribing error.	*what is written is exactly the same <SIC> think </SIC> okay?*
	• MORPHOLOGY: Transcribe lexical words containing morphological errors as you hear them and enclose it with the labels <SIC>, </SIC>.	*<SIC> grammatic </SIC>*

Appendix 2: Annotation (mark-up) guidelines

	Guidelines	Examples
Speaker ID	• Speaker IDs are assigned in the order participants speak and places in angular brackets. • Use the tag <SS> for two or more speakers in unison (mostly for laughter)	*<S1>, <S2>, <S3>, etc.* *<SS>*
Speaker turns	• When a new speaker turns begins, start a new line, thus displaying turns one below the other. When the speaker turn starts, write the speaker ID in angular brackets (<S1>); when the turn ends, mark it with a slash followed by the same speaker ID (</S1>).	*<S4> (xx) is there an option for making shift to the right an- [<S1> louder please </S1>] there's an option for making shift to the right and then for for making shift to the left, another one for parallel loading [<S1> yeah </S1>] for making (xx) the ehh </S4>* *<S1> you simply_ it is easy i mean eh you can simply do not issue another clock cycle do not send another clock cycle </S1>*
Pauses	• Short pauses (under 2 seconds) are marked with a comma (,) when they occur in the middle of the utterance and have a level or rising intonation (do not have a phrase-final intonation contour).	*we are able to translate this into a eh directly into, a C programme*
	• Longer pauses (2-3 seconds, e.g. the lecturer is thinking) is indicated with ellipses (...) (No space before the ellipses; space after the ellipses). As in short pauses, these may come after level or rising intonation.	*so... what are counters?*
	• Pauses (under 3 seconds) are marked with a full stop (.) when accompanied by an utterance final (falling) intonation contour; the full stop is not used in a syntactic sense to indicate complete sentences, although it tends to occur at the end of utterances	*then F-of-X belongs to this, neighbourhood. do you agree with this?*

	• For pauses of 4 seconds and more, use the mark-up tag <P> and time them: <P04> (No spaces between the tag and the seconds).	*what is the solution? three bits step two <P06> i mean you should be very happy if i ask such a question in the written exam*
Overlaps and backchannels	• Overlapping speech and backchannels are embedded within the current speaker's turn. They are approximate, written between square brackets and placed next to the nearest word. Do not split words with overlap tags. Speaker ID is marked at the beginning and end.	*<S6> i don't understand if eh if one clock arrive when (xx) [<S1> mhm </S1>] and i want to get ehmm...*
Laughter	• All laughter is marked. Speaker ID is not marked if the current speaker laughs. Use an underscore between speaker ID and tag.	text <LAUGH> text text <S8_LAUGH> text text <SS_LAUGH> text
Contextual events	• Contextual (non-speech) events are noted **when they affect comprehension of the discourse dynamics** in a way that would not be understandable without them. They are approximate and inserted next to the nearest word. Use underscores between words in tags.	text <CHANGE_OF_SLIDE> text
	• If the action is continuous and overlaps with speech, one tag is placed at the beginning and one at the end of the activity. The closing tag features a slash.	text <WRITING_ON_BOARD> text </WRITING_ON_BOARD> text
	• If the action is continuous but does not overlap with speech, include the length of silent pauses before the tag.	text <P08> <SCROLLING_SLIDES> text
	• Debates among students are not transcribed. Use the tag <GROUP_DISCUSSION> and include information about when the discussion started and ended, as in the example.	text text text <START_01:25:02> <GROUP_DISCUSSION> <END_01:47:12> (format for time: hour:minutes:seconds)

	Common tags in spoken academic corpora: <CHANGE_OF_SLIDE> <SCROLLING_SLIDES> <POINTING_ON_SLIDE> <POINTING_ON_SCREEN> <WRITING_ON_BOARD> <DRAWING_ON_BOARD> <REWINDING_VIDEO> <FAST-FORWARDING_VIDEO> <STOPPING_VIDEO> <TURINIG_PAGES> <TAPPING_MICROPHONE> <BACKGROUND_NOISE> <AUDIO_DISTURBANCE> <BREAK_IN_RECORDING> <PHONE_RINGING> <APPLAUSE> <HAND_GESTURE> <WALKING>	
Reading	• Mark utterances read verbatim. Say if the lecturer is reading slides.	text <READING> text </READING> text <READING_SLIDE> text </READING_SLIDE>
Foreign words	• Spell non-English words and expressions as in the original language. Mark them with the tags <FOREIGN_language code> </FOREIGN_language code> Codes: CA (Catalan) IT (Italian) LA (Latin) NE (Dutch) DA (Danish) DE (German) ES (Spanish) FR (French) GR (Ancient Greek) HR (Croatian)	*and in Italian is <FOREIGN_IT> l'ultimo uomo scimmia del pleistocene </FOREIGN_IT> it's it's ehm it's a book*

Uncertain or unintelligible speech	• If a word (or more) is unintelligible, replace it with (xx). • If the transcription is uncertain, put the words in parentheses.	*it provides ammonium ions and nitrates (xx) to the ground* *such a view (might) be created programmatically* (here the lecturer says /mai/)
Non-verbal vocal sounds	• Non-verbal vocal sounds are marked when they affect speech. Common tags: <COUGHS> <BLOWS_NOSE> <CLEARS_THROAT> <SNEEZES> <WRONG_ANSWER_NOISE> <CALLING_FOR_SILENCE> (in this case the interjection *sh* is transcribed and tagged: *sh* <CALLING_FOR_SILENCE>)	*Kelvin statement says that <COUGHS> the the eh work is zero or negative okay?*
Anonymization	• Transcripts should be anonymous (names of famous people can be kept). Omit the name and replace it with descriptors like (Name), (Surname), (Name Surname), (Surname Name), etc. Each omitted word should be replaced by a label, so that the total word count is not affected.	*as you'll see with professor (Surname)*

Appendix 3: CEFR mediation scale (Companion Volume)

C2	Can mediate effectively and naturally, taking on different roles according to the needs of the people and situation involved, identifying nuances and undercurrents and guiding a sensitive or delicate discussion. Can explain in clear, fluent, well-structured language the way facts and arguments are presented, conveying evaluative aspects and most nuances precisely, and pointing out sociocultural implications (e.g. use of register, understatement, irony and sarcasm).
C1	Can act effectively as a mediator, helping to maintain positive interaction by interpreting different perspectives, managing ambiguity, anticipating misunderstandings and intervening diplomatically in order to redirect talk. Can build on different contributions to a discussion, stimulating reasoning with a series of questions. Can convey clearly and fluently in well-structured language the significant ideas in long, complex texts, whether or not they relate to his/her own fields of interest, including evaluative aspects and most nuances.
B2	Can work collaboratively with people from different backgrounds, creating a positive atmosphere by giving support, asking questions to identify common goals, comparing options for how to achieve them and explaining suggestions for what to do next. Can further develop other people's ideas, pose questions that invite reactions from different perspectives and propose a solution or next steps. Can convey detailed information and arguments reliably, e.g. the significant point(s) contained in complex but well-structured texts within his/her fields of professional, academic and personal interest. Can collaborate with people from other backgrounds, showing interest and empathy by asking and answering simple questions, formulating and responding to suggestions, asking whether people agree, and proposing alternative approaches. Can convey the main points made in long texts expressed in uncomplicated language on topics of personal interest, provided that he/she can check the meaning of certain expressions.
B1	Can introduce people from different backgrounds, showing awareness that some questions may be perceived differently, and invite other people to contribute their expertise and experience, their views. Can convey information given in clear, well-structured informational texts on subjects that are familiar or of personal or current interest, although his/her lexical limitations cause difficulty with formulation at times. Can play a supportive role in interaction, provided that other participants speak slowly and that one or more of them helps him/her to contribute and to express his/her suggestions. Can convey relevant information contained in clearly structured, short, simple, informational texts, provided that the texts concern concrete, familiar subjects and are formulated in simple everyday language.

Appendix 4: CEFR addressing audiences scale (Companion Volume)

C2	Can present a complex topic confidently and articulately to an audience unfamiliar with it, structuring and adapting the talk flexibly to meet the audience's needs. Can handle difficult and even hostile questioning.
C1	Can give a clear, well-structured presentation of a complex subject, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples. Can structure a longer presentation appropriately in order to help the audience follow the sequence of ideas and understand the overall argumentation. Can speculate or hypothesize in presenting a complex subject, comparing and evaluating alternative proposals and arguments. Can handle interjections well, responding spontaneously and almost effortlessly.
B2	Can give a clear, systematically developed presentation, with highlighting of significant points, and relevant supporting detail. Can depart spontaneously from a prepared text and follow up interesting points raised by members of the audience, often showing remarkable fluency and ease of expression. Can give a clear, prepared presentation, giving reasons in support of or against a particular point of view and giving the advantages and disadvantages of various options. Can take a series of follow up questions with a degree of fluency and spontaneity which poses no strain for either him/herself or the audience.
B1	Can give a prepared presentation on a familiar topic within his/her field, outlining similarities and differences (e.g. between products, countries/regions, plans). Can give a prepared straightforward presentation on a familiar topic within his/her field which is clear enough to be followed without difficulty most of the time, and in which the main points are explained with reasonable precision. Can take follow up questions, but may have to ask for repetition if the speech was rapid.

Appendix 5: CEFR Table 3: Qualitative features of spoken language

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating 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 colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	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.
C1	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.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2+	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.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	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 initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	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.
B2					

The TAEC project is a collaboration between the following partners:

UNIVERSITY OF COPENHAGEN

Universitat de Lleida

Maastricht University

UNIVERSITÀ DEGLI STUDI DI TORINO