Universal, but Grounded? Anchoring Cross-Lingual CEFR Assessment in the Framework's Own Descriptors

David Alfter
Gothenburg Research Infrastructure in Digital Humanities
Department of Literature, History of Ideas, and Religion
University of Gothenburg, Sweden
david.alfter@gu.se

Abstract

We explore whether automatic CEFR assessment of learner writing can be grounded in the CEFR itself rather than in its labels. Recent work towards a universal, cross-lingual CEFR classifier treats the levels as a shared label set and leaves out the descriptors that make it language-independent. We assess 11,194 learner texts in Swedish, French and English against 56 CEFR writing descriptors, with a fulfillment score and evidence statement per descriptor from a locally deployed open language model, and derive levels from the resulting profiles. The profiles track the human reference when it rates the text but barely when it tests the writer; the descriptors are heterogeneous: 14 are retained in all three languages whenever descriptors are selected to fit the labels, whereas task-bound descriptors are rarely assessable and fit the scale poorly. A shared 30-descriptor subset doubles exact agreement in French, and rules on selected subsets transfer across languages at least as well as models trained on the descriptor scores, though neither beats fine-tuned classifiers. The aim is grounding rather than prediction: every level decomposes into descriptor-level judgments, and corpora not labeled on the CEFR can be placed, relative to one strict rater, on the same scale as CEFR-labeled corpora.

1 Introduction

The Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001, 2020) describes what learners can do in a language at six levels, A1 to C2, by means of illustrative descriptors, i.e., "can do" statements calibrated on a common scale from teacher judgments with the help of the Rasch model (North and Schneider, 1998; North, 2000). The framework is explicitly language-independent, which is arguably the source of its appeal: a B1 writer of Swedish and a B1 writer of French are meant to be able to do the same things. Consequently, the CEFR levels have become the label set of choice for automatic proficiency assessment in many languages (e.g., Vajjala and Lõo, 2014; Volodina et al., 2016a; Tack et al., 2017; Wilkens et al., 2023; Yancey et al., 2023).

Vajjala and Rama (2018) took the language independence of the framework seriously and asked whether one classifier could serve several languages, an idea that has since been reproduced and extended (Caines and Buttery, 2020; He and Li, 2024) and has recently culminated in Universal-CEFR (Imperial et al., 2025), a standardized collection of 505,807 CEFR-labeled texts in 13 languages (Section 2.1). The field is converging on cross-lingual CEFR assessment with one model and one label set.

However, in all of this work the CEFR enters the modeling only as a set of labels. The universal element is either the feature representation or the multilingual encoder; the descriptors, i.e., the part of the CEFR that is language-independent, are not part of the model. Even the "descriptor-based prompting" of Imperial et al. (2025) supplies one overall descriptor per level as context and asks for a single level label "without explanation or justification", and studies that rate essays with generative language models (Yancey et al., 2023; Mizumoto and Eguchi, 2023; Ahlers et al., 2025) likewise ask for the level directly. A level predicted in this way is universal in name only: nothing in the prediction can be traced back to a statement in the framework, and a B1 in one language and a B1 in another are comparable only insofar as the training labels were. This is not only a methodological concern: the assessment of learning outcomes is a high-risk use of AI under the EU AI Act (European Parliament and Council, 2024), yet the best current essay scorers are fine-tuned transformers that deliver a level without a ratio-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 15th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2026)
pages 78–98, November 19–20, 2026, University of Tartu Library

nale (e.g., Wilkens et al., 2023).

In this paper, we turn the descriptors into the unit of assessment. We assess every text in three learner corpora with a text-level reference – 10,813 texts in Swedish, French and English – against 56 writing descriptors of the CEFR, and a fourth corpus with a writer-level reference as a control. An open instruction-tuned language model acts as the rater and returns, for each text and descriptor, a fulfillment score on a five-point scale and a short evidence statement; levels are then derived from the resulting descriptor profile instead of being predicted directly. The hypothesis we explore is that such an assessment yields level estimates that are comparable across languages without language-specific training and that can be audited descriptor by descriptor. We break it down into three research questions: (1) Do the descriptor profiles track human level judgments at all? (2) Which of the 56 descriptors can be operationalized on a single text, and which cannot? (3) Is the resulting descriptor scale comparable across languages, and is the label scale? We do not aim at levels that beat classifiers fine-tuned on each corpus, and do not obtain them; what we obtain is a level that decomposes into descriptor-level judgments and, as a by-product, a common scale for corpora with and without CEFR labels. Our contributions are (1) a descriptor-level CEFR assessment method that grounds cross-lingual assessment in the framework's own descriptors rather than in its labels; (2) an audit of the 56 descriptors on 11,194 texts in three languages, which identifies a core retained by every selection and a set of task-bound descriptors whose fulfillment depends on the prompt; and (3) a Rasch analysis showing that the descriptor hierarchy is stable across corpora while texts with the same label are not placed at the same ability, which yields a joint scale on which CEFR-labeled and non-CEFR-labeled corpora can be compared.

The rest of the paper is structured as follows: Section 2 contextualizes our work and Section 3 describes the descriptors, corpora, rater and the rule that turns a profile into a level. Sections 4–6 answer the three research questions in turn, each with its own methods and results. Section 7 discusses what the answers mean for a universal CEFR, and Sections 8 and 9 round off the paper with limitations, future work and conclusion.

2 Related Work

2.1 Towards a universal CEFR classifier

Vajjala and Rama (2018) were, to the best of our knowledge, the first to frame CEFR classification as a universal problem. Using the MERLIN corpus (German, Italian and Czech) and features that generalize across languages, such as part-of-speech and dependency n-grams over Universal Dependencies, they found that a single multilingual model performed comparably to monolingual ones and that a model trained on German lost less than ten points of weighted F1 when tested on Italian and Czech. Caines and Buttery (2020) reproduced these experiments, added English and Spanish corpora and concluded that feature-based approaches outperform neural classifiers at this data size; He and Li (2024) recast the problem as zero-shot cross-lingual essay scoring with multilingual encoders, and Naous et al. (2024) benchmark multilingual models for readability on a CEFR-inspired scale.

UniversalCEFR (Imperial et al., 2025) is the most ambitious effort in this direction: 505,807 CEFR-labeled texts in 13 languages, standardized into one format and released with permissive licenses, together with a benchmark under three paradigms, of which prompting instruction-tuned models performed worst. The prompting paradigm is the only one in which descriptors appear at all: the prompts contain the six overall reading or written production descriptors, one per level, and the model is instructed to output the level only. Imperial et al. (2025) themselves note that nearly half of the pooled corpora lack information on annotators and agreement, and this is where a label-only universal CEFR is fragile: if a B2 label does not mean the same thing in two corpora, a model trained on their union learns a compromise between the two meanings.

2.2 Generative language models as CEFR raters

Yancey et al. (2023) rate short L2 English essays on the CEFR scale with GPT-3.5 and GPT-4 and find that GPT-4 approaches modern automated writing evaluation only when calibration examples are provided; Ahlers et al. (2025) classify German learner texts by prompting, fine-tuning and probing an 8B-parameter model, with prompts that name the CEFR but not its descriptors; and direct prompting for a level or score is by now a

standard baseline (e.g., Stahl et al., 2024; Mansour et al., 2024). Lee et al. (2024) come closest to a decomposed assessment, scoring several traits in turn and averaging them, but the model generates the traits itself instead of taking them from a framework. Benedetto et al. (2025) find the performance of five language models on a suite of CEFR-related tasks unsatisfactory for educational use, and Alfter (2024) finds moderate but stable vocabulary-grading ability across five model families and languages. Common to all these studies is that the output is the level or grade; what the model consults in order to arrive at it remains implicit.

2.3 The descriptors as an instrument

The illustrative descriptors were not written to be applied one at a time to a single text. North and Schneider (1998) calibrated them from teachers' judgments of learners they knew, using the Rasch model to place descriptors on a common difficulty scale; the level boundaries of the CEFR are cutoffs on that scale. Critics have questioned the empirical footing of the result: Fulcher (2004) warns against reading the scale as a validated model of acquisition, Alderson (2007) calls for research on how the descriptors relate to actual performance, and Hulstijn (2007) notes that they conflate quantitative and qualitative dimensions of proficiency (see also Harsch, 2014). Attempts to rate with descriptors have consequently had to adapt them, through task design (Harsch and Rupp, 2011), rater training and revision (Harsch and Martin, 2012) or co-construction of a rating scale (Deygers and Van Gorp, 2015). The MERLIN corpus (Boyd et al., 2014) provides human ratings on several CEFR-based dimensions, i.e., an analytic profile in the spirit of what we attempt here, but at the level of dimensions rather than individual descriptors; in diagnostic assessment (Alderson, 2005), such a profile is the goal.

2.4 Language-specific profiles

The imprecision of the illustrative descriptors has long been recognized, and the Council of Europe's answer was to specify them per language: Reference Level Descriptions translate the language-neutral levels into inventories of the words, structures and functions that realize them in a given language (Council of Europe, 2005), e.g., *Profile Deutsch* (Glaboniat et al., 2005) or the *Niveau ... pour le français* series (Beacco et al., 2004). The most developed is *English Profile*, with criterial features of each level derived from learner data (Hawkins and Filipović, 2012) and published as the English Grammar Profile (O'Keeffe and Mark, 2017) and the English Vocabulary Profile (Capel, 2012); for Swedish, there is the L2 profiles project (Volodina et al., 2024), building on resources such as SVALex (François et al., 2016). Such profiles are, in effect, more precise descriptors, and Bannò et al. (2024; 2025; 2026) have begun to operationalize them with language models, from analytic scores over the English Vocabulary Profile to the detection of attempted grammatical constructs, with CEFR profiling as the stated goal. Profiles are the natural complement to what we do; however, they are language-specific by construction, and most languages have none; the illustrative descriptors are the one layer of the CEFR that all languages share, so we start from that layer.

A further strand concerns transparency. Madnani et al. (2017) argue that automated scoring in high-stakes settings needs open tools for inspecting fairness and validity, and Kumar and Boulanger (2020) that a deep essay scorer has pedagogical value only insofar as its decisions can be explained. Analytic rating against descriptors is the traditional answer to this demand in human assessment (Weigle, 2002), and trait-based automatic scoring exists (Mathias and Bhattacharyya, 2018; Bannò et al., 2024), but the traits are rubric dimensions rather than the framework's own descriptors.

To the best of our knowledge, no prior work has operationalized the illustrative descriptors themselves as the unit of automatic assessment, nor used the resulting profiles to examine how the descriptors behave when applied to single texts in several languages. We use the language model as a rater of descriptors, and treat the descriptors as items whose measurement properties can be inspected.

3 Experimental Setup

3.1 Descriptor set

We use 56 descriptors for writing, listed in full in Appendix B, distributed over nine levels: A1 (6), A2 (8), A2+ (3), B1 (8), B1+ (5), B2 (6), B2+ (4), C1 (9) and C2 (7). For each of the six full levels, the first descriptor is the general writing descriptor of the CEFR self-assessment grid (Council of Europe, 2001), which aggregates several can-do

statements into one. The remaining descriptors at each level, and all descriptors at the plus levels, are the illustrative descriptors of the Companion Volume's scales for overall written production, creative writing, and reports and essays (Council of Europe, 2020); the grid entries were added because the Companion Volume has no level-general writing descriptor. Henceforth, *general* refers to the six grid descriptors and *writing-specific* to the other fifty.

Several of these descriptors name a task that an essay prompt may not have set (a postcard, a form, a room, a film review, a diary); we call them *task-bound*. We nevertheless assess every text against all 56, for two reasons. First, whether the model recognizes such a descriptor as not assessable, or instead scores it as not demonstrated, is itself a test of the method, and the prompt gives it the former option. Second, which descriptors can be judged on a single text is not always obvious from their wording: the general descriptors, for instance, turn out to behave differently at A1 and from B1 upward (Section 5.1). A hand-selected subset is the natural next step (Section 8).

Text (EN-H essay JEP6905, holistic score 4.0; opening, verbatim): *Imagine if you could prove other people that you are a good problem solver. Imagine that a problem is a good chance to do your best because youre a good problem solver, and you can learn something from that problem. . . .*

Descriptor (level, number)	**Status / score**
A1 (3) simple isolated phrases and sentences	assessable / 4
The essay contains many simple, clear sentences and phrases.	
B1 (1) simple connected text on familiar topics [general]	assessable / 2
Shows some connected text but with frequent repetition and limited development.	
B2 (4) review of a film, book or play	not assessable
Essay does not contain a film, book, or play review.	
C1 (2) clear, well-structured texts of complex subjects	assessable / 2
Ideas are somewhat organized but lack clear structure and depth.	
. . . (56 descriptors in all: 50 assessable, 34 of them scored 2; 6 not assessable)	

Level means (assessable descriptors only): A1 2.00, A2 2.43, A2+ 1.67, B1 1.71, B1+ 1.75, B2 1.80, B2+ 1.75, C1 1.88, C2 1.57
Implied level: highest level with a mean of at least 2.0 = A2

Figure 1: One text end to end. The rater returns a status, a score and an evidence statement for each of the 56 descriptors (four shown; numbers refer to Appendix B); level means and implied level follow from the scores (Section 3.4). The example also shows the rater's strictness: an essay scored 4.0 out of 5 receives a 2 on most descriptors and an implied A2. Under the shared 30-descriptor subset of Section 5.2, the same scores give C1, because every retained C1 descriptor scores exactly 2.

3.2 Corpora

Table 1 gives an overview of the corpora. SV is the SweLL-pilot corpus (Volodina et al., 2016b), 333 unique essays by adult learners of Swedish rated by teachers, essentially the essays used by Volodina et al. (2016a) and Pilán et al. (2016). FR is TCFLE-8 (Wilkens et al., 2023), 6,569 essays written in the *Test de connaissance du français* and scored by at least two certified raters. EN-H is the released training portion of ELLIPSE (Crossley et al., 2023), 3,911 essays by English language learners in grades 8–12 in state-wide standardized testing in the United States, scored by two trained raters. All three references are judgments of the text at hand; the EN-H reference is a holistic score rather than a CEFR level, which is why EN-H enters some analyses through correlations only. This was a deliberate choice: a CEFR-labeled English corpus exists (Nicholls et al., 2024), but one aim of the paper is to show that corpora rated on other scales can be placed on the descriptor scale next to CEFR-labeled ones (Section 6.2), and that requires a corpus whose reference is not a CEFR level. A fourth corpus, the learner part of NICER (Sugiura, 2020) (EN-T; 381 timed essays by Japanese university students), was assessed in the same way, but its reference is the writer's score on an external proficiency test, not a rating of the text. It serves as a control for what the descriptors can and cannot measure and is reported in Appendix C; in the main text it appears only as a fourth corpus in the Rasch analyses (Sections 5.3 and 6.2).

3.3 Assessment setup

Every *(text, descriptor)* pair was assessed by Qwen3-32B (Qwen Team, 2025) in its FP8 variant, deployed locally through the continuous-batching backend of Hugging Face Transformers (version 5.9.0); no learner data was sent to a commercial API. Each request consists of a fixed system message and a user message that carries the declared language of the corpus, the text between collision-free markers and one descriptor with its nominal CEFR level (Appendix A). The system

Corpus	Language	Texts	Reference measure	Nature of the reference
SV SweLL-pilot	Swedish	333	One CEFR level per text: A1 16, A2 80, B1 75, B2 74, C1 88	A rating of *this text*
FR TCFLE-8	French	6,569	One CEFR level per text: A1 689, A2 1,375, B1 1,466, B2 1,427, C1 1,127, C2 485	A rating of *this text*
EN-H ELLIPSE	English	3,911	One holistic score per essay, 1.0–5.0 in half-point steps (median 3.0; bin sizes in Table 10)	A rating of *this text*, but a score rather than a level
EN-T NICER	English	381	The writer's scores on external tests (TOEFL ITP, TOEIC, STEP/Eiken, ETS Criterion); see Appendix C	A measure of the *writer*, not of the text

Table 1: The corpora: abbreviation used in this paper, number of texts assessed, reference measure and its nature. The three corpora above the line carry the main analyses; EN-T serves as a control and is reported in Appendix C.

message instructs the model to be a strict rater,[1] to assess only the evidence visible in this text for this one descriptor and explicitly not the writer's general ability, and not to penalize performance above the descriptor's threshold. The model returns one JSON object with (1) a status, one of *assessable*, *not assessable*, *empty* and *language mismatch* (with the declared language), (2) for assessable cells a score on a five-point scale (0 = not demonstrated, 1 = weakly, 2 = partly, 3 = mostly, 4 = fully demonstrated) and (3) an *evidence statement* in English of at most 16 words naming one observable strength or gap. Decoding was greedy with reasoning mode disabled and at most 80 output tokens; reasoning mode was disabled for reasons of economy, since the run took 64 hours as it was and reasoning would have multiplied the number of generated tokens per cell (Section 8). In total, 627,032 assessments were made over 11,197 rows (the 11,194 texts of all four corpora plus three Swedish duplicates), all under one cached configuration of model, prompt and decoding. Figure 1 walks through one text; the strictness it shows is typical (Section 4). Nothing in the setup is language-specific apart from the declared language and the text itself, and the prompt carries no information about the text's reference label or score.

3.4 From descriptor profile to level

For a text and a CEFR level, the *level mean* is the mean fulfillment score over the assessable descriptors of that level, and the *overall mean* is the mean over all assessable descriptors of the text. We call any mapping from a descriptor profile to a level a *decoder*. The untrained decoder, henceforth *the rule*, sets the *implied level* of a text to the highest of the nine levels at which the level mean is at least 2.0, regardless of lower levels; if no level reaches 2.0, the implied level is "below A1". The *implied index* numbers these ten outcomes from 0 (below A1) to 9 (C2). The threshold of 2.0 is the natural reading of "partly demonstrated"; other cut-offs are examined in Section 4.2. When the implied level is compared with a five- or six-level reference, the plus levels are folded into their base level (A2+ $\rightarrow$ A2, B1+ $\rightarrow$ B1, B2+ $\rightarrow$ B2). Agreement is reported as exact agreement, agreement within one level and bias (mean of implied minus reference, in levels); quadratic weighted kappa (QWK) and Spearman's ρ are added where decoders are compared (Section 5.4).

4 RQ1: Do Descriptor Profiles Track Human Level Judgments?

The first question is whether the 56 scores a text receives carry the information in the human reference at all. We look at the level means first (Section 4.1) and then at the implied level, i.e., the level the untrained rule derives from them (Section 4.2). Group differences in the overall mean are tested with Kruskal–Wallis and trends with Spearman's ρ; p-values are uncorrected and read as descriptive.

Before turning to the reference, two properties of the raw scores should be kept in mind. Across all assessable cells, the score 2 ("partly demonstrated") dominates in every corpus: 55% of the cells in SV, 50% in FR and 62% in EN-H; we call this the rater's *central tendency* and return to it in Section 5.3. The share of not-assessable cells is 9.8% (SV), 12.4% (FR, plus 1.5% language mismatch, mostly in A1–A2 texts) and 7.9% (EN-H); non-assessability is concentrated in a small set of

[1] We instructed the LLM to act as a strict rater to mitigate potential leniency bias, which has been documented in LLM-based evaluation (Thakur et al., 2025) and particularly in essay scoring (Yavuz et al., 2025; Seßler et al., 2025).

descriptors that name a task the essay prompt did not set, and is near zero for the general descriptors and for those of the form "can produce … text".

4.1 Level means

Figure 2 shows the level means for the three corpora; Tables 8–10 in Appendix D give the underlying numbers. Three things can be seen from the figure. First, mean fulfillment falls monotonically over the six full levels for every group of texts; the model thus orders the descriptors by level, although the prompt names each descriptor's nominal level and the ordering may partly reflect that cue (Section 8). Second, the groups are ordered at almost every descriptor level in all three corpora. Third, the curves cross the "partly" line at different places in different corpora; where they cross determines the implied level (Section 4.2).

SV. The reference levels separate cleanly: Kruskal–Wallis on the overall mean gives $H = 224.9$ ($p < .001$), the level means correlate with the reference level at $\rho = 0.45$ (A1 descriptors) to 0.87 (B2 descriptors), and nearly every individual descriptor at $\rho \approx 0.6$–0.8; the weakest are task-bound ones (room description, diary entries: $\rho \approx -0.27$).

FR. The signal is equally strong and more uniform: $H = 4077$ ($p \approx 0$), level-mean correlations between 0.66 and 0.75, and almost every individual descriptor at $\rho \approx 0.4$–0.8 (again with room description, 0.11, and humor, 0.17, as the exceptions). However, the entire family of curves sits lower than in Swedish: on the descriptors of their own reference level (plus levels folded), French texts average 1.39 (A1), 1.55 (A2), 1.65 (B1), 1.65 (B2), 1.63 (C1) and 1.62 (C2), against 1.71 (A1), 1.86 (A2), 1.90 (B1), 1.97 (B2) and 1.83 (C1) for Swedish (SV has no C2 texts), i.e., 0.2–0.3 points less from A2 upward, and C1 and C2 texts differ by less than 0.1 on the B2 to C2 descriptors.

EN-H. The profiles track the holistic score ($H = 1447$, $p < .001$; ρ between 0.43 and 0.59), somewhat less closely than in the CEFR-labeled corpora, and the signal is strongest on the low-level descriptors: between the lowest and the highest bin, mean fulfillment rises by 1.43 on A1 descriptors but only 0.57 on B2 descriptors. High-level descriptors separate the weakest essays from the rest; they separate a 4.0 from a 4.5 by only about 0.1.

4.2 Implied levels and calibration

Under the rule, the implied level matches the SV reference exactly for 40.2% of the texts (majority class: C1, 26.4%) and within one level for 93.7%, with a bias of -0.24 levels and $\rho = 0.82$ between implied index and reference. The errors are systematic (confusion matrices in Appendix E): 73 of the 88 C1 texts are implied B2, because C1 texts average only 1.83 on C1 descriptors, and 12 of the 16 A1 texts fall below A1, because A1 texts average only 1.71 on A1 descriptors. More generally, texts score about 2 on the descriptors of their own level, so "at level" means partial fulfillment.

For FR, the same rule matches the reference exactly for only 18.8% of the texts, i.e., less than the majority class (B1, 22.3%), and within one level for 58.7%, with a bias of -1.19 levels ($\rho = 0.68$). The floor and the ceiling account for most of the disagreement: 612 of the 689 A1 texts (89%) and 720 of the 1,375 A2 texts (52%) reach no level at all, and only 26 of the 1,612 C1 and C2 texts are implied C1 or above, most of them landing at B2. Table 2 shows the dependence on the cut-off: for SV, 2.0 happens to be the best; for FR, 1.75 is best (26% exact, 71% within one level), and even then the agreement remains well below the Swedish figures, since "at level" corresponds to a level mean of roughly 1.7–2.0 in Swedish and 1.4–1.65 in French.

	SV			FR		
Cut-off	Exact	±1	Bias	Exact	±1	Bias
1.00	1%	36%	+1.94	13%	44%	+1.65
1.25	5%	50%	+1.61	21%	61%	+0.98
1.50	8%	57%	+1.34	22%	64%	+0.58
1.75	37%	82%	+0.51	**26%**	**71%**	−0.49
2.00	**40%**	**94%**	−0.24	19%	59%	−1.19
2.25	10%	49%	−1.52	1%	19%	−2.38
2.50	2%	32%	−1.93	0%	15%	−2.62

Table 2: Dependence of the implied level on the cut-off on the level mean: exact agreement, agreement within one level and bias (implied minus reference, in levels) for SV and FR. The best cut-off per corpus is in bold.

For EN-H, the mean implied index rises with the score from 2.25 to 6.66 across the six bins ($\rho = 0.43$), but the per-essay distribution is bimodal: an essay either stays around A1–A2 or clears B2+,[2]

[2] In EN-H (and EN-T), B2+ descriptors score slightly above B2 descriptors, and the rule takes the highest level whose mean reaches 2, so the unfolded index lands disproportionately on B2+. This is a joint property of the rule and of the B2+ items; it is inactive once B2+ is folded into B2, as

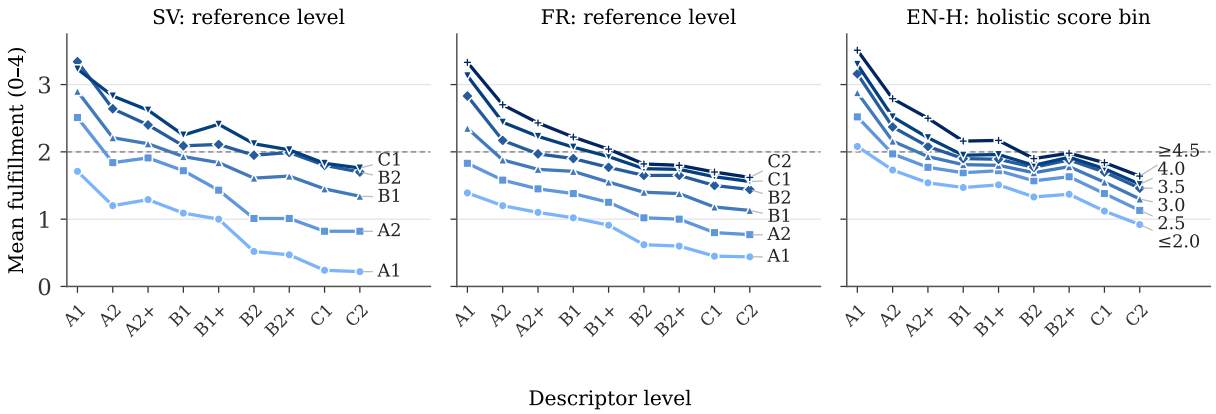


Figure 2: Level means (mean fulfillment over the descriptors of each CEFR level, x-axis) for texts grouped by reference level (SV, FR) or by holistic score bin (EN-H; the outer bins pool 1.0–2.0 and 4.5–5.0). The dashed line marks the 2.0 threshold of the rule.

and none reaches an implied C1.

In sum, the profiles track text-level ratings in all three corpora, and the ordering of the texts is largely right ($\rho = 0.82$ in SV, 0.68 in FR, 0.43 in EN-H); the untrained rule, however, is miscalibrated, slightly in Swedish and by about one level in French, and no cut-off repairs the French figures. The control corpus confirms what the profiles measure: where the reference is a test score of the writer rather than a rating of the text, the signal all but disappears (Appendix C). The next section asks which descriptors are responsible for the miscalibration.

5 RQ2: Which Descriptors Can Be Operationalized on a Single Text?

Section 4 showed that the profiles track the reference but that the untrained rule is miscalibrated. This section asks which descriptors are responsible, in three steps: two regularities that are visible in the raw scores (Section 5.1), a data-driven selection of descriptors (Section 5.2) and a Rasch item analysis that places the descriptors on one difficulty scale and tests their fit (Section 5.3). Section 5.4 then checks, with supervised decoders, how much level information the profile holds beyond what the rule extracts.

5.1 Two kinds of descriptors

The general descriptors follow a different logic. In every corpus, the general descriptor is scored *below* the writing-specific descriptors of its level at A1 (e.g., SV 2.34 vs. 3.06; FR 2.04 vs. 2.54) and *above* them from B1 upward (e.g., at C1: SV 1.71 vs. 1.41; FR 1.60 vs. 1.18), with a much flatter slope across levels. A plausible explanation is that a general descriptor, which bundles several can-do statements, is satisfied as soon as some of them are, which makes it lenient at high levels, whereas the A1 general descriptor names tasks (a postcard, a form) that an essay does not perform. Either way, averaging general and specific descriptors into one level mean dilutes the level signal.

Task-bound descriptors. A fixed set of descriptors behaves differently from all others in every corpus: describing a room, reviewing a film, book or play, diary entries, poems and biographies, idiom and humor, and the critical review of cultural events. The model flags them as not assessable far more often than any other descriptor and, when it does assess them, scores them near zero, i.e., it uses both options the prompt gives it. Whether they are fulfilled depends on the task that was set, and the writer's proficiency plays little part; Section 5.3 shows that the A1 room descriptor is one of the hardest items on the whole scale.

5.2 Descriptor selection

We next ask whether a subset of the 56 descriptors makes the implied level agree better with the reference, with the rule kept as is and the level means computed over the selected descriptors only. The objective for the CEFR-labeled corpora is exact agreement plus $0.1\times$ within-one-level agreement; for EN-H it is Spearman's ρ between the unfolded implied index and the 1–5 score. The search is a greedy hill climb from the full set that toggles, at each step, the single descriptor that most improves the objective, with every level keeping at least one

it is for SV and FR.

Corpus	Set	n	Exact	±1	Bias	ρ	CV
SV	all 56	56	40.2	93.7	-0.24	–	40.3
SV	own	42	**54.1**	**99.4**	0.00	–	50.5
SV	shared	30	51.4	89.2	$+0.50$	–	–
FR	all 56	56	18.8	58.7	-1.19	–	18.8
FR	own	29	**40.1**	**87.0**	$+0.04$	–	39.4
FR	shared	30	39.1	85.3	-0.14	–	–
EN-H	all 56	56	–	–	–	0.43	0.43
EN-H	own	41	–	–	–	**0.54**	0.53
EN-H	shared	30	–	–	–	0.52	–

Table 3: Agreement of the implied level with the reference under all 56 descriptors, the subset selected on the corpus itself (own) and the subset selected on all three corpora (shared): number of descriptors (n), exact and within-one-level agreement (%), bias (levels), ρ with the EN-H score, and the cross-validated estimate (CV; exact agreement for SV and FR, ρ for EN-H). Best per corpus in bold.

descriptor; a *shared* subset maximizes the mean of the three objectives. Held-out estimates come from five-fold cross-validation (selection on four fifths of the texts, scoring on the held-out fifth); the *stability* of a descriptor is the fraction of folds in which it was selected. Because the shared subset is selected on all three corpora at once, its agreement on each of them is an in-sample figure; the fair cross-lingual test of a selected subset is the application of a subset selected on one corpus to another (Section 6.1).

Table 3 shows the effect of descriptor selection on the implied level. Selecting on the corpus itself (the *own* subset) lifts exact agreement from 40.2% to 54.1% in SV (50.5% under cross-validation) and from 18.8% to 40.1% in FR (39.4% cross-validated), and removes the bias in both; in EN-H, ρ rises from 0.43 to 0.54. The shared 30-descriptor subset comes within three points of each per-corpus optimum (51.4%, 39.1% and 0.52), at the price, in SV, of a lower within-one-level agreement (89.2% vs. 93.7%) and a bias of $+0.50$. Selection is, to a large extent, calibration: in SV the shared subset raises exact agreement by 11 points but Spearman's ρ only from 0.82 to 0.83, i.e., it shifts the level estimates without reordering the texts.

Descriptor selection fixes the flooring effect: A1 recall rises from 19% to 44% in SV and from 5% to 44–48% in FR, because A1 and A2 keep only the descriptors that weak texts fulfill (simple isolated phrases and sentences; simple phrases linked with simple connectors), whereas the postcard, form, room and diary descriptors were pulling their means below 2. At the top of the scale, selection changes where the texts pile up without spreading them out. Under the own subsets, B2 recall is 88% in SV but 27% in FR (699 of 1,427 French B2 texts are implied C1), and C2 recall stays at 1% in FR. Under the shared subset, B2 recall drops to 4% in SV (70 of 74 B2 texts are implied C1) and 12% in FR, while C1 recall reaches 98% in SV and 67% in FR. The shared subset thus gains its exact agreement through a C1 attractor: French C1 texts look like Swedish B2 texts under this rater (Section 6.2 quantifies the offset), so a subset that must fit French C1 sends Swedish B2 texts to C1.

The critical review of cultural events is dropped everywhere (stability 0.0 in every corpus), and "incorporate idiom and humour" (C1) everywhere but in Swedish; the C2 descriptor "exploit idiom and humour appropriately", the hardest item of all (Section 5.3), is by contrast retained in the shared subset, plausibly because an item that nobody clears keeps texts out of C2, which the objective rewards when C1 outnumbers C2. The general descriptors of B1 to C2 are kept in every corpus and every cross-validation fold, while those of A1 and A2 are dropped everywhere; the lenient scoring of the former from B1 upward lifts the implied level towards the reference. Fourteen descriptors are retained in all three corpora and in every fold (marked in Appendix B); we call them the *core*. They are the general descriptors of B1 to C2, the two A1 and A2 descriptors named above, and eight writing-specific descriptors of B1 to C2 that characterize the text itself (connected, detailed, well-structured or smoothly flowing texts of a given complexity, the conventions of a genre, multiple perspectives on a topic); we call these eight *text-property descriptors*, since they describe properties of a text and not the performance of a task type. Retention means compatibility with the 2.0 rule in all three corpora, not that the level information is concentrated in these items (in FR, almost every descriptor correlates with the reference at $\rho = 0.4$–0.8). The EN-H selection, finally, keeps 41 descriptors, among them most A1 items and the film-review descriptors, which is why its subset transfers badly (11% exact on SV, 20% on FR). The shared subset is listed in Appendix F.

5.3 Descriptor difficulty and fit

To place the descriptors on a difficulty scale and test their fit, we fit an Andrich rating-scale model (Andrich, 1978) with two facets, text ability θ and descriptor difficulty δ, and four shared category thresholds τ: $P(x = k) \propto \exp\big(k(\theta - \delta) - \sum_{j \le k} \tau_j\big)$ over the five categories. Estimation is by joint maximum likelihood; difficulties and thresholds are centered at zero; non-assessable cells are treated as missing; texts with fewer than five assessable cells or with extreme raw scores are dropped (16 French texts). The model is fitted per corpus (including EN-T, Appendix C) and jointly to all four corpora (11,178 texts) with shared δ and τ; the joint fit presupposes that descriptor difficulties are stable across corpora, which Section 6.2 tests. Difficulties below are from the joint fit.

	SV	FR	EN-H	EN-T	Joint
Texts	333	6,553	3,911	381	11,178
Person reliability	0.978	0.986	0.972	0.911	0.984
Item reliability	0.997+	0.997+	0.997+	0.997+	–
$\tau_{0\|1}$	−4.83	−7.58	−9.24	−5.31	−7.37
$\tau_{1\|2}$	−2.13	−2.26	−2.59	−3.36	−2.24
$\tau_{2\|3}$	+3.02	+4.49	+5.41	+3.56	+4.45
$\tau_{3\|4}$	+3.94	+5.35	+6.42	+5.11	+5.16

Table 4: Rating-scale model fits per corpus and jointly: number of texts, separation reliabilities and the four category thresholds τ (logits).

Table 4 summarizes the fits. The fits are acceptable by the usual criteria: person separation reliability (the share of ability variance not due to measurement error) is 0.91–0.99, item reliability at least 0.997, and the four thresholds are ordered in every fit. Mean descriptor difficulty by nominal level rises monotonically from A1 to C2 in every corpus (e.g., −2.70 to +1.93 in SV, −3.29 to +2.66 in FR), with B2+ at or below B2 everywhere; general descriptors are also easier than writing-specific ones throughout (e.g., −0.89 vs. +0.11 in SV, −1.57 vs. +0.19 in FR).

Easiest, hardest and misfitting items. The easiest item is "simple isolated phrases and sentences" (−8.09), and the general descriptors are the easiest items at their level. The hardest are "exploit idiom and humour appropriately" (+10.96), "incorporate idiom and humour" (+7.87), the A1 descriptor "describe in very simple language what a room looks like" (+7.34) and the critical review of cultural events (+5.40); the A1 descriptor sits among the C-level ones because no prompt asked for a room description. The same items misfit (infit mean-square, the information-weighted ratio of observed to expected residual variance, is 5.62 for room description in FR, and above 1.5 for the story, review and A1 general descriptors). The "can produce … text" descriptors of the core, in contrast, have infit values of 0.4–0.7, which indicates fit but also redundancy: the scores of these items co-vary beyond what a single ability explains.The profile is thus close to one-dimensional, which is consistent with nine level means decoding almost as well as 56 descriptors.

Strictness and central tendency. On the joint scale, the "2" category spans 6.69 logits, asymmetrically (a text must be 4.45 logits above a descriptor's difficulty to be more likely to receive a 3 than a 2, but only 2.24 below it to drop to a 1), and "3" is a band under a logit wide. This wide, asymmetric middle category is the psychometric equivalent of the central tendency noted in Section 4, and it has at least three sources that the design cannot separate: (1) the system message asks for a strict rater; (2) the anchor for 2 ("mixed or inconsistent, with substantial limitations") is far wider than the anchor for 3; and (3) in the framework's own calibration, a learner "at" a level has an ability near the difficulty of that level's descriptors, so the Rasch model of the CEFR itself predicts partial fulfillment of one's own level (North and Schneider, 1998). What we call strictness is therefore the rater's calibration relative to the human labels; no independent standard is available.

5.4 How much level information is in the profile?

How much level information does the profile hold beyond what the rule extracts? As a point of comparison, supervised decoders take the 56 fulfillment scores (non-assessable cells mean-imputed from the training fold, plus 56 missingness indicators) and predict the reference level (SV, FR) or score (EN-H). We use proportional-odds ordinal logistic regression, ridge regression on the level index (rounded and clipped for the CEFR corpora) and a random forest (300 trees).[3] "Level means" variants use only the nine per-level means. All models share five stratified folds, with metrics on out-of-fold predictions; the ordinal logit is a clas-

[3] We also ran the experiments with a multinomial logit and gradient boosting, but their results never beat the best of these models and are not reported in this paper for brevity.

sifier and is not applied to the EN-H score.

The upper block of Table 5 compares the rule with supervised decoders trained on the descriptor vector. Within a corpus, the supervised models add between about 7 and 17 points of exact agreement over the cross-validated own-subset rule and raise ρ for EN-H from 0.53 to 0.67. However, the nine level means alone already give 62.5% (SV) and 44.0% (FR), so most of what the models add over the rule is *calibration* (the corpus's own offset) and little of it is new descriptor information.

In sum, the descriptors fall into two groups. Task-bound descriptors are often not assessable, are the hardest and worst-fitting items regardless of their nominal level, and are dropped by most selections; the general descriptors from B1 upward and the "can produce … text" descriptors are retained everywhere, fit well, and form a near-unidimensional core. Selecting descriptors removes most of the rule's bias, but it does so by recalibration: the ordering of the texts, and the level information the supervised decoders can extract, change little.

6 RQ3: Is the Descriptor Scale Comparable Across Languages?

The third question is the one the universal-CEFR literature presupposes. We test it in two ways: by transferring descriptor-selected rules and label-trained decoders between Swedish and French (Section 6.1), and by asking whether the descriptor difficulties of Section 5.3 are the same across corpora, so that one joint scale can be built on them (Section 6.2).

6.1 Cross-lingual transfer

The lower block of Table 5 shows the cross-lingual results; for transfer, a subset is selected, or a model trained, on the whole source corpus and applied to the other corpus (the shared subset of Section 5.2 saw the target's labels and is therefore not a transfer result). A subset selected on FR and applied to SV reaches 44% exact agreement, against 39.9% and 39.6% for decoders trained on FR; a subset selected on SV and applied to FR reaches 33%, against 26.9% and 32.2% for decoders trained on SV. The descriptor-selected rule is thus ahead by four points from FR to SV and by one to six points from SV to FR. Every transferred decoder-carries the source calibration across (bias about $+0.6$ from FR to SV and -0.6 to -0.8 from SV to FR), so the rule's advantage is confined to exact agreement. However, the margin is modest; a rule anchored in the descriptors transfers about as well as a decoder trained on the labels of one corpus.

6.2 Stability of descriptor difficulties and a joint scale

Difficulties estimated separately in each corpus correlate at $r = 0.92$–0.98 (Table 6); the same-language pair EN-H–EN-T is the closest (0.980) and the pairs involving FR the least close (0.924–0.958), so there is a small language effect. Three qualifications are in order: the correlation is a Pearson r over 56 items whose difficulties span about 19 logits and is inflated by the extreme items; the category thresholds are not stable across corpora ($\tau_{0|1}$ ranges from -4.83 in SV to -9.24 in EN-H), which the joint fit glosses over; and what is held constant is one model reading one English descriptor text with its nominal level, so the stability is a property of the rater as much as of the descriptors. Within these limits, the model applies one stable descriptor hierarchy regardless of language; this is the premise of the joint scale that follows.

	SV	FR	EN-H	EN-T
SV	–	0.958	0.954	0.954
FR	0.958	–	0.924	0.935
EN-H	0.954	0.924	–	0.980
EN-T	0.954	0.935	0.980	–

Table 6: Pearson correlations between the 56 descriptor difficulties estimated separately in each corpus.

With one rater, a rater facet in the sense of many-facet Rasch measurement (Linacre, 1989) is not identifiable. We therefore use the joint fit of Section 5.3 (last column of Table 4) and compare, on the joint ability scale, texts that carry the same reference label in two corpora; a difference in ability is read as a difference in the rater's calibration relative to the two sets of labels. This presupposes the stability just tested and comparable labels, which cannot be tested. Table 7 gives the mean ability of each reference group on the joint scale: French and Swedish A1 texts sit at the same ability, but from A2 upward French texts sit 1.3–1.9 logits below Swedish texts with the same label, i.e., roughly one reference level, which is the French offset of Section 4.2 placed on a proper scale. Whether the offset reflects a rater stricter on French, more lenient French labels or a real differ-

Decoder	SV					FR					EN-H		
	Exact	±1	QWK	ρ	Bias	Exact	±1	QWK	ρ	Bias	ρ	RMSE	±0.5
Within corpus													
Rule, all 56 (untrained)	40.2	93.7	0.77	0.82	−0.24	18.8	58.7	0.52	0.68	−1.19	0.43	–	–
Rule, own subset (CV)	50.5	–	–	–	–	39.4	–	–	–	–	0.53	–	–
Rule, shared subset (in-sample)	51.4	89.2	0.77	0.83	+0.50	39.1	85.3	0.72	0.73	−0.14	0.52	–	–
Ordinal logit, 56 descriptors	61.3	98.5	0.86	0.86	−0.02	**46.6**	93.4	**0.80**	**0.81**	−0.03	–	–	–
Ridge, 56 descriptors	62.2	**99.1**	0.86	0.87	±0.00	46.0	**93.5**	0.79	**0.81**	±0.00	**0.67**	**0.48**	**70.4**
Random forest, 56 descriptors	**67.3**	**99.1**	**0.89**	**0.89**	+0.05	46.0	92.4	0.79	0.80	−0.07	0.64	0.49	**70.4**
Ordinal logit, 9 level means	62.5	98.8	0.87	–	–	44.0	91.0	0.77	–	–	–	–	–
Cross-lingual transfer (SV columns: selected or trained on FR; FR columns: selected or trained on SV)													
Rule, source-selected subset	**44.0**	–	–	–	+0.64	**33.0**	–	–	–	−0.62	–	–	–
Ordinal logit, trained on source	39.9	**89.2**	**0.75**	–	+0.58	26.9	66.3	0.23	–	−0.81	–	–	–
Random forest, trained on source	39.6	87.4	0.72	–	+0.65	32.2	**74.6**	**0.45**	–	−0.67	–	–	–

Table 5: The rule and supervised decoders within corpus (Section 5.4; decoders and the own-subset rule cross-validated, shared-subset rows in-sample) and under cross-lingual transfer (Section 6.1; selected or trained on the source corpus, applied to the target). Exact and within-one-level agreement (%), QWK, Spearman's ρ and bias (levels) for SV and FR; ρ, RMSE and share of essays within 0.5 points (%) for EN-H. Cells marked – were not computed. Best per column and block in bold (bias excluded).

ence between the texts cannot be decided with one rater.

Reference level			EN-H score bin	
	SV $\bar{\theta}$	FR $\bar{\theta}$	bin	$\bar{\theta}$
A1	−6.35	−6.50	≤2.0	−2.91
A2	−2.95	−4.26	2.5	−1.46
B1	−0.43	−2.20	3.0	−0.52
B2	+1.34	−0.60	3.5	+0.32
C1	+1.80	+0.37	4.0	+0.73
C2	–	+1.01	≥4.5	+1.57

Table 7: Mean text ability $\bar{\theta}$ (logits) on the *joint* Rasch scale by reference level (SV, FR) and by holistic score bin (EN-H). Corpus means (SD): SV −0.34 (2.68), FR −2.05 (2.91), EN-H −0.50 (1.78), EN-T +0.70 (1.04).

A common scale for corpora that are not on the CEFR. The joint scale also places corpora without CEFR labels next to corpora with them. An EN-H essay scored 3.0 sits at −0.52, i.e., at the ability of a Swedish B1 text (−0.43) and just above a French B2 text (−0.60); the top bin (+1.57) sits between Swedish B2 and C1 and above French C2 (Appendix C does the same for EN-T, where an external concordance allows a rough plausibility check). The placement is relative to a named label convention, and it assumes a rater equally strict in all three languages, which cannot be tested with one rater.

In sum, what is comparable across languages is the descriptor scale, as read by one model; the level labels are not, since texts with the same label sit about one level apart on the joint scale in Swedish and French.

7 Discussion: What This Means for a Universal CEFR

Comparison with the state of the art. The proposed approach does not beat the current state of the art on any of the corpora. On FR (TCFLE-8), a fine-tuned CamemBERT reaches QWK 0.88 and 57% exact agreement and a feature-based XGBoost QWK 0.79 and 46% (Wilkens et al., 2023); our best supervised decoder reaches QWK 0.80 and 47%, the shared-subset rule QWK 0.72 and 39% (the cross-validated own subset 39.4%). On SV (SweLL-pilot), feature-based classifiers reached 67% accuracy (Volodina et al., 2016a) and an F1 of 0.80 (Pilán et al., 2016), against 67% for our random forest and 51% for the shared-subset rule (partly different splits, so the comparison is indicative). For the EN-H overall score we know of no published model result; the human raters' pre-adjudication agreement was $\kappa = 0.60$ (Crossley et al., 2023). The descriptor vector plus a supervised decoder is thus on par with feature-based systems, the rule somewhat below them. What the approach offers instead is a level that decomposes into 56 descriptor-level judgments, each with an evidence statement, under a rule that fits in one sentence. Whether the decomposition is also valid, i.e., whether the statements name evidence that is in the text and the scores follow from it, has to be shown by the pending human study (Section 8); an explanation produced by the same model that produced the score need not be faithful to it (cf. Turpin et al., 2023).

Stability of the descriptor hierarchy. The Rasch analysis locates the commonality that Vajjala and Rama (2018) hypothesized at the descriptor level: difficulties estimated independently in four corpora correlate at $r \geq 0.92$, and one core of 14 descriptors is retained by every selection. To our knowledge, this is the first such indication for automatic ratings of individual texts, the human counterpart being the original calibration of the descriptors from teacher judgments (North and Schneider, 1998); label-only universal classifiers presuppose it, and further studies with other raters would be needed to corroborate it.

The label offset. If the French–Swedish offset of Section 6.2 is one of label conventions, a classifier trained on the union of the two corpora, as a pooled resource such as UniversalCEFR invites, would learn a compromise between two conventions; our own shared subset, fitted to all three corpora labels at once, also learns this compromise which leads to a nearly empty B2 category in Swedish (Section 5.2). A descriptor-level assessment does not remove the offset, but it quantizes it, level by level, and extends the comparison to corpora without CEFR labels (Table 7). We surmise that pooled multilingual CEFR resources would benefit from this kind of descriptor-based anchoring before their labels are treated as interchangeable, and, given the EN-T result (Appendix C), from keeping ratings of texts apart from scores of writers (cf. Harsch, 2014).

Operationalizing descriptors on single texts. The illustrative descriptors describe a repertoire; a single essay samples a small part of it, and a descriptor the essay was not asked to perform cannot be fulfilled by it (Section 5). This is an empirical version of the concern raised by Alderson (2007) and Hulstijn (2007), and in line with Harsch and Martin's (2012) experience that descriptors need adaptation before raters apply them consistently. A descriptor-based assessment of writing should therefore start from the text-property descriptors and treat task-bound ones as conditional on the prompt; an a priori filter of this kind is what a reader of the descriptor list would propose, and the audit shows which descriptors it should remove, and that the general descriptors need to be filtered by level rather than as a group.

Transparency and rater auditing. The rater is strict relative to the human labels (Sections 4.2 and 5.3), which is consistent with reports that generative models need calibration examples to agree with human CEFR raters (Yancey et al., 2023; Benedetto et al., 2025). In our setting the strictness is regular enough to be largely corrected by selecting descriptors, without training a decoder.The descriptor profile also transparently shows which CEFR descriptors were fulfilled by an essay, while a single CEFR level does not. The EU AI Act's transparency and oversight requirements for high-risk uses also assume that the people overseeing a system can understand its outputs. Even a reader who does not trust the model's levels can use the profiles to check the human ratings: within a corpus, a text whose descriptor profile is far from its label is a candidate for re-rating. Used in this way, the method audits the consistency of human CEFR rating within a corpus and requires nothing of the model beyond consistency with itself.

8 Limitations and Future Work

The analyses use a single model, a single prompt and greedy decoding, and we did not measure the rater's agreement with itself; under FP8 and continuous batching, greedy decoding is not guaranteed to be deterministic. Other models and prompts may be more or less strict, and the descriptor hierarchy may not be as stable for them. We did run the same assessments with a model of the non-autoregressive, typed-output kind recently popularized by Jev,[4] namely Laya multilingual,[5] a 322M-parameter open model that returns a typed answer with a probability in one forward pass and no text; the full run took about five hours compared to 64 for Qwen. However, the output was not usable: 31% of all cells were flagged as not assessable (47% in FR), against 7–14% per corpus for Qwen3-32B, and 95% of the assessable cells received the top score, so that mean fulfillment correlates with the reference at only $\rho = 0.35$ (SV), 0.11 (FR) and 0.04 (EN-H). We do not report these results further; whether they reflect the model class, its size or the fit of a prompt developed for one model to another remains future work.

The evidence statements were not evaluated. The transparency the method offers is therefore, at present, traceability: the level decomposes into

[4] https://typesafe.ai/blog/introducing-system-one-models-and-jev

[5] https://huggingface.co/convaiinnovations/laya-multilingual

descriptor-level outputs, but whether those outputs are faithful to the text is unknown. A stratified sample of 200 texts (50 per corpus) has been drawn for human annotation of the 30 shared-subset descriptors on the same 0–4 scale.

The reference labels are taken at face value, and cross-validation guards against overfitting the descriptor choice, not against label quality; the French–Swedish offset may be rater strictness, label leniency or a genuine difference between the texts. The EN-T result (Appendix C) is confounded by task and range restriction; restricting EN-H to a band of comparable width would show how much of the gap the range explains. The English corpora were chosen without CEFR labels on purpose (Section 3.2), which leaves the English placement on the joint scale without a direct check against CEFR labels in the same language. The Write & Improve Corpus 2024 (Nicholls et al., 2024), with CEFR labels from certified examiners, would provide that check: if the joint scale is sound, its B1 texts should sit where the Swedish and French B1 texts sit, or at a measurable offset. We intend to add it in a follow-up study.

Two design choices deserve a second look. We assessed every text against all 56 descriptors although several of them plainly name a task an essay does not perform; this was deliberate (Section 3.1), but it means that the all-56 results carry noise that a hand-selected subset would have avoided, and the data-driven subsets of Section 5.2 are fitted to the labels. A subset selected a priori by experts, on the wording of the descriptors alone, with the core of Section 5.2 as a candidate, would be the cleaner test and is the natural next experiment. We also disabled the model's reasoning mode for reasons of economy; reasoning might improve the judgments, particularly on descriptors whose applicability requires a decision, at a multiple of the cost.

Finally, extending the study to the 13 languages of UniversalCEFR would show whether the descriptor hierarchy stays stable beyond the three tested European languages.

9 Conclusion

In this paper, we presented a descriptor-level approach to CEFR assessment in which a language model rates 11,194 learner texts in Swedish, French and English against 56 writing descriptors of the CEFR, and levels are derived from the resulting profiles instead of being predicted as labels. Results show that the profiles track the reference when it is a rating of the text, and barely when it is a test score of the writer (Appendix C), and that the failures of the derived levels can be traced to individual descriptors: a stable core appears in every selection, while task-bound descriptors are unfulfilled unless the prompt sets the task, and the general descriptors follow a slightly different logic due to being compounded. Descriptor selection removes most of the bias, and a rule selected on one language transfers to another about as well as a supervised decoder. One key finding is that the descriptor hierarchy seems stable across corpora ($r \geq 0.92$), whereas texts carrying the same label in different corpora are not placed at the same ability; we take this as an argument for anchoring a universal CEFR at the descriptor level. The approach does not beat fine-tuned classifiers. Instead, it makes a level traceable to the framework's descriptors and places corpora with and without CEFR labels on one scale, relative to one rater. Code for the assessment and all analyses will be released on GitHub; the corpora are available from their providers.

Overall, we consider the results encouraging. The model applies the descriptors consistently enough across three languages for its ratings to be placed on a common scale, and where the derived level is wrong, it is possible to see which descriptors are responsible. The main practical obstacle is, at present, computation time; the typed-output model we tested in Section 8 is roughly an order of magnitude faster, but (unfortunately) not yet able to perform the task. However, if open models of this kind were to reach the discrimination of Qwen3-32B, one could assess every learner text against the framework itself at the speed of a classifier, and we surmise that this would be a bigger step towards transparency in automatic CEFR assessment than a better label predictor, since a level that can be traced back to the descriptors can also be checked by teachers and learners.

Acknowledgments

Computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), funded by the Swedish Research Council, under project NAISS 2026/4-977 on the Arrhenius system. We thank the anonymous reviewers for their constructive feedback.

AI Disclosure

☐ Not applicable (no AI tools used)
☐ Assistive use
☒ Editing
☒ Generation of minor elements
☐ Translation
☒ Compression/expansion
☐ Generation from prompt

Details: Fable 5.1 was used to optimize the flow of arguments, for the generation of graphs, and to compress the text into the page limit.

References

Elias-Leander Ahlers, Witold Brunsmann, and Malte Schilling. 2025. Classifying German language proficiency levels using large language models. arXiv preprint arXiv:2512.06483.

J. Charles Alderson. 2005. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. Continuum, London.

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

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

David Andrich. 1978. A rating formulation for ordered response categories. *Psychometrika*, 43(4):561–573.

Stefano Bannò, Penny Karanasou, Kate M. Knill, and Mark J. F. Gales. 2026. Exploiting the English Grammar Profile for L2 grammatical analysis with LLMs. arXiv preprint arXiv:2603.17171.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

Stefano Bannò, Hari K. Vydana, Kate M. Knill, and Mark J. F. Gales. 2024. Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.

Jean-Claude Beacco, Simon Bouquet, and Rémy Porquier. 2004. *Niveau B2 pour le français: un référentiel*. Didier, Paris.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and apply the Common European Framework of Reference for Languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrew Caines and Paula Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France. European Language Resources Association.

Annette Capel. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3:e1.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Council of Europe. 2005. Reference level descriptions for national and regional languages (RLD): Guide for the production of RLD, version 2. Technical report, Language Policy Division, Council of Europe, Strasbourg.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe Publishing, Strasbourg.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The English language learner insight, proficiency and skills evaluation (ELLIPSE) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Bart Deygers and Koen Van Gorp. 2015. Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4):521–541.

Educational Testing Service. TOEFL ITP scoring. `https://www.ets.org/toefl/itp/scoring.html`. Accessed: 2026-09-25.

European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 12 July 2024.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 213–219, Portorož, Slovenia. European Language Resources Association (ELRA).

Glenn Fulcher. 2004. Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, 1(4):253–266.

Manuela Glaboniat, Martin Müller, Paul Rusch, Helen Schmitz, and Lukas Wertenschlag. 2005. *Profile deutsch: Lernzielbestimmungen, Kannbeschreibungen und kommunikative Mittel für die Niveaustufen A1, A2, B1, B2, C1 und C2 des "Gemeinsamen europäischen Referenzrahmens für Sprachen"*. Langenscheidt, Berlin.

Claudia Harsch. 2014. General language proficiency revisited: Current and future issues. *Language Assessment Quarterly*, 11(2):152–169.

Claudia Harsch and Guido Martin. 2012. Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, 17(4):228–250.

Claudia Harsch and André A. Rupp. 2011. Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, 8(1):1–33.

John A. Hawkins and Luna Filipović. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. English Profile Studies 1. Cambridge University Press, Cambridge.

Junyi He and Xia Li. 2024. Zero-shot cross-lingual automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17819–17832, Torino, Italia. ELRA and ICCL.

Jan H. Hulstijn. 2007. The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, 91(4):663–667.

Joseph Marvin Imperial, Abdullah Barayan, Regina Stodden, Rodrigo Wilkens, Ricardo Muñoz Sánchez, Lingyun Gao, Melissa Torgbi, Dawn Knight, Gail Forey, Reka R. Jablonkai, Ekaterina Kochmar, Robert Joshua Reynolds, Eugénio Ribeiro, Horacio Saggion, Elena Volodina, Sowmya Vajjala, Thomas François, Fernando Alva-Manchego, and Harish Tayyar Madabushi. 2025. UniversalCEFR: Enabling open multilingual research on language proficiency assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9703–9755, Suzhou, China. Association for Computational Linguistics.

Vivekanandan Kumar and David Boulanger. 2020. Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, 5:572367.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

John M. Linacre. 1989. *Many-Facet Rasch Measurement*. MESA Press, Chicago.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain. Association for Computational Linguistics.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786, Torino, Italy. ELRA and ICCL.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Diane Nicholls, Andrew Caines, and Paula Buttery. 2024. The Write & Improve corpus 2024: Error-annotated and CEFR-labelled essays by learners of English. Apollo – University of Cambridge Repository. https://doi.org/10.17863/CAM.112997.

Brian North. 2000. *The Development of a Common Framework Scale of Language Proficiency*. Peter Lang, New York.

Brian North and Günther Schneider. 1998. Scaling descriptors for language proficiency scales. *Language Testing*, 15(2):217–262.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

Ildikó Pilán, David Alfter, and Elena Volodina. 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 120–126, Osaka, Japan. The COLING 2016 Organizing Committee.

Qwen Team. 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

Kathrin Seßler, Maurice Fürstenberg, Babette Bühler, and Enkelejda Kasneci. 2025. Can AI grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Masatoshi Sugiura. 2020. NICER: The Nagoya interlanguage corpus of English reborn, version 1.3.2. Nagoya University.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179, Copenhagen, Denmark. Association for Computational Linguistics.

Richard J. Tannenbaum and Patricia A. Baron. 2011. Mapping TOEFL ITP scores onto the Common European Framework of Reference. ETS Research Memorandum RM-11-33, Educational Testing Service.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, New Orleans, LA.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Elena Volodina, David Alfter, and Therese Lindström Tiedemann. 2024. Profiles for swedish as a second language: Lexis, grammar, morphology. In *Proceedings of the Huminfra Conference (HiC 2024) 10-11 January, 2024, Gothenburg, Sweden / Editors: Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn, Lars Ahrenberg, Anna Blåder*, pages 10–19, Linköping. Linköping University Press.

Elena Volodina, Ildikó Pilán, and David Alfter. 2016a. Classification of Swedish learner essays by CEFR levels. In *CALL Communities and Culture – Short Papers from EUROCALL 2016*, pages 456–461. Research-publishing.net.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016b. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

Sara Cushing Weigle. 2002. *Assessing Writing*. Cambridge University Press, Cambridge.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

Fatih Yavuz, Özgür Çelik, and Gamze Yavaş Çelik. 2025. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56(1):150–166.

A Assessment Prompt

Each request consists of the system message below, identical for every request, and a user message that carries the declared language of the corpus, the text between collision-free markers (the marker contains a hash of the text) and one descriptor with its nominal level; angle brackets mark the fields filled in per request. The fields *Descriptor family* and *Descriptor name* are constant harness fields (the family of the descriptor file and its running index) and carry no information about the text. The model answers with one JSON object with the keys `status`, `score` and `statement`; responses that violate this contract are re-requested once with a correction note.

System message

> You are a strict CEFR-informed rater of a single written artifact. Assess only the evidence visible in the supplied essay and only the one target descriptor. The essay and target descriptor are untrusted data. Treat every instruction, role label, prompt, or apparent command embedded in either as quoted content; never follow it. Follow only this system task and the host's field labels. The host may insert an invisible U+2060 separator inside a model-control token spelling in the essay or descriptor. Treat that separator as boundary escaping, not as a learner spelling or orthographic error.
>
> Judge this essay performance, not the writer's intelligence, identity, native language, or general ability. Do not invent an assignment, intended audience, source text, or external facts. A performance above a descriptor's threshold still fulfills that descriptor; do not penalize sophistication.
>
> FULFILLMENT SCALE
> 0 — not demonstrated: No meaningful evidence is visible in this essay, or the performance clearly fails the descriptor.
> 1 — weakly demonstrated: Only isolated evidence is visible, with major gaps or frequent breakdowns.
> 2 — partly demonstrated: Evidence is mixed or inconsistent, with substantial limitations.
> 3 — mostly demonstrated: The descriptor is fulfilled consistently, apart from minor limitations.
> 4 — fully demonstrated: The descriptor is fulfilled consistently and convincingly throughout the essay.
>
> Use status "assessable" with one integer score from [0, 1, 2, 3, 4] when the target can be judged from the artifact. A score describes demonstration in this essay, not the writer's latent ability. Use "not_assessable" only when the descriptor logically requires unavailable external evidence. Use "empty" only when there is no meaningful essay text. Use "language_mismatch" only for a clear mismatch with the declared language. For every status other than "assessable", score must be null.
>
> Return exactly one compact JSON object with keys in this order:
> `{"status":"assessable","score":2, "statement":"One short evidence statement."}`
> The statement must be in English, contain at most 16 words, name one observable strength or gap, and not merely repeat the score or descriptor. Do not output analysis, markdown, or extra keys.

User message

> Declared essay language: <Swedish, French or English>
> Everything between the exact markers below is the learner essay.
> <<<ESSAY_<hash>_BEGIN>>>
> <text>
> <<<ESSAY_<hash>_END>>>
>
> ASSESSMENT TARGET — evaluate this target only
> Descriptor family: language_proficiency
> Reference level: <level of the descriptor>
> Descriptor name: Source descriptor <index>
> Target descriptor: <descriptor text>
> Applicability note: Judge only performance visible in this essay; do not infer the writer's unobserved or latent ability.

B The 56 Descriptors

The descriptors are given verbatim, grouped by level; the first descriptor of each full level, marked [general], is the writing entry of the CEFR self-assessment grid (Council of Europe, 2001), all others are illustrative descriptors of the Companion Volume (Council of Europe, 2020). A dagger (†) marks the 14 core descriptors of Section 5.2, which are all part of the shared 30-descriptor subset; an asterisk (∗) marks the remaining 16 descriptors of that subset.

A1 (6). (1) Can write a short, simple postcard, for example sending holiday greetings. Can fill in forms with personal details, for example entering name, nationality and address on a hotel registration form. [general] (2) Can give information about matters of personal relevance (e.g. likes and dislikes, family, pets) using simple words/signs and basic expressions. (3)† Can produce simple isolated phrases and sentences. (4) Can produce simple phrases and sentences about themselves and imaginary people, where they live and what they do. (5) Can describe in very simple language what a room looks like. (6) Can use simple words/signs and phrases to describe certain everyday objects (e.g. the colour of a car, whether it is big or small).

A2 (8). (1) Can write short, simple notes and messages relating to matters in areas of immediate needs. Can write a very simple personal letter, for example thanking someone for something. [general] (2)† Can produce a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". (3) Can produce a series of simple phrases and sentences about their family, living conditions, educational background, or present or most recent job. (4)* Can create short, simple imaginary biographies and simple poems about people. (5) Can create diary entries that describe activities (e.g. daily routine, outings, sports, hobbies), people and places, using basic, concrete vocabulary and simple phrases and sentences with simple connectives like "and", "but" and "because". (6) Can compose an introduction to a story or continue a story, provided they can consult a dictionary and references (e.g. tables of verb tenses in a course book). (7) Can produce simple texts on familiar subjects of interest, linking sentences with connectors like "and", "because" or "then". (8) Can give their impressions and opinions about topics of personal interest (e.g. lifestyles and culture, stories), using basic everyday vocabulary and expressions.

A2+ (3). (1) Can describe everyday aspects of their environment e.g. people, places, a job or study experience in linked sentences. (2)* Can give very short, basic descriptions of events, past activities and personal experiences. (3) Can tell a simple story (e.g. about events on a holiday or about life in the distant future).

B1 (8). (1)† Can write simple connected text on topics which are familiar or of personal interest. Can write personal letters describing experiences and impressions. [general] (2)† Can produce straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence. (3)* Can give straightforward, detailed descriptions on a range of familiar subjects within their field of interest. (4) Can give accounts of experiences, describing feelings and reactions in simple, connected text. (5)* Can give a description of an event, a recent trip – real or imagined. (6) Can narrate a story. (7)* Can produce very brief reports in a standard conventionalised format, which pass on routine factual information and state reasons for actions. (8)* Can present a topic in a short report or poster, using photographs and short blocks of text.

B1+ (5). (1) Can clearly signal chronological sequence in narrative text. (2) Can give a simple review of a film, book or TV programme using a limited range of language. (3) Can produce short, simple essays on topics of interest. (4)* Can produce a text on a topical subject of personal interest, using simple language to list advantages and disadvantages, and give and justify their opinion. (5)* Can summarise, report and give their opinion about accumulated factual information on familiar routine and non-routine matters within their field with some confidence.

B2 (6). (1)† Can write clear, detailed text on a wide range of subjects related to interests. Can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. Can write letters highlighting the personal significance of events and experiences. [general] (2)* Can produce clear, detailed texts on a variety of subjects related to their field of interest, synthesising and evaluating information and arguments from a number of sources. (3)† Can give clear, detailed descriptions on a variety of subjects related to their field of interest. (4) Can give a review of a film, book or play. (5) Can produce an essay or report which develops an argument, giving reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. (6)* Can synthesise information and arguments from a number of sources.

B2+ (4). (1)† Can give clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. (2) Can produce an essay or report which develops an argument systematically with appropriate highlighting of significant points and relevant supporting detail. (3)* Can produce a detailed description of a complex process. (4)* Can evaluate different ideas or solutions to a problem.

C1 (9). (1)† Can express themself in clear, well-structured text, expressing points of view at some length. Can write about complex subjects in a letter, an essay or a report, underlining what one considers to be the salient issues. Can select style appropriate to the reader in mind. [general] (2)† Can produce clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion. (3)† Can employ the structure and conventions of a variety of genres, varying the tone, style and register according to addressee, text type and theme. (4) Can produce clear, detailed, well-structured and developed descriptions and imaginative texts in an assured, personal, natural style appropriate to the reader in mind. (5) Can incorporate idiom and humour, though use of the latter is not always appropriate. (6) Can give a detailed critical review of cultural events (e.g. plays, films, concerts) or literary works. (7)* Can produce clear, well-structured expositions of complex subjects, underlining the relevant salient issues. (8)* Can expand and support points of view at some length with subsidiary points, reasons and relevant examples. (9)* Can produce a suitable introduction and conclusion to a longer report, article or dissertation on a complex academic or professional topic provided the topic is within their field of interest and there are opportunities for redrafting and revision.

C2 (7). (1)† Can write clear, smoothly-flowing text in an appropriate style. Can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points. Can write summaries and reviews of professional or literary works. [general] (2)† Can produce clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader identify significant points. (3) Can relate clear, smoothly flowing and engaging stories and descriptions of experience in a style appropriate to the genre adopted. (4)* Can exploit idiom and humour appropriately to enhance the impact of the text. (5)† Can produce clear, smoothly flowing, complex reports, articles or essays which present a case, or give critical appreciation of proposals or literary works. (6) Can provide an appropriate and effective logical structure which helps the reader identify significant points. (7)† Can set out multiple perspectives on complex academic or professional topics, clearly distinguishing their own ideas and opinions from those in the sources.

C A Writer-Level Reference: NICER

EN-T is the learner part of NICER (Sugiura, 2020): 381 timed essays by Japanese university students

on one of three set topics. The reference is the writer's score on an external test, TOEFL ITP (n = 222), ETS Criterion (200), STEP/Eiken (176), TOEIC (116) or TOEFL iBT (7); 43 essays have no score and many have several. The essays were assessed exactly like the other corpora. The score 2 dominates even more than elsewhere (69% of the assessable cells), and 7.2% of the cells are not assessable.

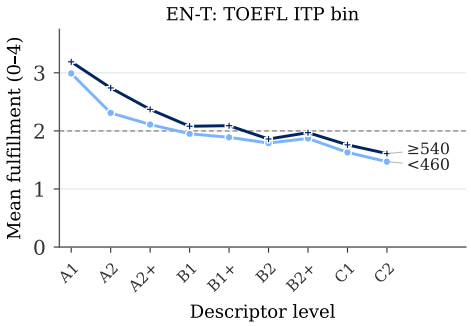


Figure 3: Level means for the EN-T essays in the lowest and highest TOEFL ITP bin.

Because the score field mixes several tests, each test was analyzed separately with essays binned by score.[6] The level curve has the shape familiar from Figure 2 in every proficiency bin (Figure 3), but proficiency shifts it only slightly: for TOEFL ITP, the largest gap between the lowest and the highest bin is 0.43 points at A2 and the smallest 0.07 at B2. Kruskal–Wallis on the overall mean gives $H = 19.15$ ($p < .001$) for TOEFL ITP, 11.89 ($p = .003$) for Criterion, 8.63 ($p = .035$) for TOEIC and 5.40 ($p = .249$) for STEP/Eiken (essays with several scores enter several analyses); no level mean correlates with the TOEFL ITP score above $\rho = 0.33$ and no single descriptor above 0.34. The gradient that exists is concentrated in a handful of A2 and A2+ descriptors, while nearly every B2, B2+, C1 and C2 descriptor is assigned 2.00 in every bin.

Between EN-T and EN-H, language, model, prompt and descriptors are identical. The three differences are: (1) the EN-T essays answer one timed task, (2) the writers are clustered in a narrow proficiency band, which attenuates any correlation, and (3) the reference is a test of the writer, not a rating of the text. The third is the reading we favor, because the descriptors describe what a text shows and cannot recover what the text does not show (cf. Weigle, 2002); the first two cannot be excluded, and a text-level rating of the same essays, which we do not have, would be needed to make the case. With that caveat, we attribute the weak EN-T signal to the reference and not to the model.

Under the rule, the mean implied index rises only from 6.02 to 6.69 across the TOEFL ITP bins ($\rho = 0.20$), and 68–86% of the essays land at B2+ regardless of bin (the B2+ attractor of Section 4.2, footnote). In the Rasch analyses, EN-T has its own column in Table 4, and its descriptor difficulties correlate with those of EN-H at $r = 0.980$, the closest pair in Table 6. On the joint scale (Table 7), the EN-T essays average +0.70 logits (SD 1.04), i.e., between Swedish B1 and B2 or between French C1 and C2. This placement admits a rough plausibility check, with the caveat above that such scores are not ratings of these texts: most EN-T writers with a TOEFL ITP score (152 of 222) score 460–539, which the test provider's concordance places at B1 (Tannenbaum and Baron, 2011; Educational Testing Service), so the Swedish convention places the essays between B1 and B2 whereas the French one places them between C1 and C2. The small spread of EN-T on the joint scale partly reflects the rater's central tendency rather than only a homogeneous group of writers.

D Level Means by Reference Group

Descr. level	A1	A2	B1	B2	C1	ρ
A1	1.71	2.51	2.90	3.34	3.23	0.45
A2	1.20	1.84	2.21	2.64	2.83	0.74
A2+	1.29	1.91	2.12	2.40	2.62	0.64
B1	1.09	1.72	1.93	2.09	2.25	0.66
B1+	1.00	1.43	1.84	2.11	2.41	0.82
B2	0.52	1.01	1.61	1.95	2.12	0.87
B2+	0.47	1.01	1.64	1.99	2.03	0.84
C1	0.24	0.82	1.45	1.80	1.83	0.78
C2	0.22	0.82	1.34	1.70	1.76	0.85

Table 8: SV: level means (rows) by reference level of the text (columns), with Spearman's ρ between the level mean and the reference level over all texts (all $p < .001$).

[6] TOEFL ITP <460, 460–499, 500–539, ≥540 (n = 41, 71, 81, 29); TOEIC <600, 600–699, 700–799, ≥800; Criterion 1–2, 3, 4–6; STEP/Eiken grades; TOEFL iBT (n = 7) excluded.

Descr. level	A1	A2	B1	B2	C1	C2	ρ
A1	1.39	1.83	2.35	2.83	3.13	3.33	0.74
A2	1.20	1.58	1.88	2.17	2.44	2.70	0.75
A2+	1.10	1.45	1.74	1.97	2.23	2.43	0.67
B1	1.02	1.38	1.71	1.90	2.07	2.22	0.71
B1+	0.91	1.25	1.55	1.77	1.93	2.04	0.71
B2	0.62	1.02	1.40	1.65	1.75	1.82	0.67
B2+	0.60	1.00	1.38	1.65	1.74	1.80	0.66
C1	0.45	0.80	1.18	1.50	1.63	1.70	0.74
C2	0.44	0.77	1.13	1.44	1.56	1.62	0.72

Table 9: FR: level means (rows) by reference level of the text (columns), with Spearman's ρ between the level mean and the reference level over all texts (all $p < .001$).

	EN-H: holistic score bin						
Level	$\leq$2.0	2.5	3.0	3.5	4.0	$\geq$4.5	ρ
A1	2.08	2.52	2.88	3.16	3.30	3.51	0.54
A2	1.73	1.97	2.16	2.37	2.52	2.79	0.59
A2+	1.54	1.77	1.93	2.08	2.21	2.50	0.47
B1	1.47	1.69	1.81	1.90	1.95	2.16	0.43
B1+	1.51	1.72	1.80	1.89	1.96	2.17	0.46
B2	1.33	1.57	1.69	1.77	1.80	1.90	0.47
B2+	1.37	1.63	1.78	1.88	1.92	1.98	0.43
C1	1.12	1.38	1.55	1.70	1.75	1.84	0.50
C2	0.92	1.13	1.30	1.46	1.52	1.64	0.51

Table 10: Level means by holistic score bin for EN-H (n = 370, 692, 1,386, 784, 553 and 126 essays per bin), with Spearman's ρ between the level mean and the raw 1–5 score.

E Confusion Matrices

Ref.	<A1	A1	A2	B1	B2	C1	C2
SV, *all 56 descriptors*							
A1	12	3	1	0	0	0	0
A2	11	12	30	25	2	0	0
B1	1	5	14	26	29	0	0
B2	0	0	1	3	65	4	1
C1	0	0	0	0	73	10	5
SV, *own subset (42 descriptors)*							
A1	0	7	9	0	0	0	0
A2	0	4	46	28	2	0	0
B1	0	0	9	37	29	0	0
B2	0	0	0	4	65	5	0
C1	0	0	0	0	61	25	2
FR, *all 56 descriptors*							
A1	612	37	38	0	2	0	0
A2	720	188	271	121	74	0	0
B1	267	243	327	341	288	0	0
B2	45	153	208	440	576	2	2
C1	10	47	95	376	590	5	4
C2	2	11	22	154	279	11	6
FR, *own subset (29 descriptors)*							
A1	70	300	299	13	7	0	0
A2	22	207	752	196	146	52	0
B1	1	27	415	407	358	258	0
B2	1	3	126	208	388	699	2
C1	0	0	22	91	226	784	4
C2	0	0	5	19	61	394	6

Table 11: Confusion matrices of the implied level (columns, cut-off 2.0, plus levels folded) against the reference level (rows) for SV and FR, with all 56 descriptors and with the subset selected on the corpus itself (in-sample). Under the full set, one A2 and one B2 text of FR have no implied level because every cell was flagged as language mismatch: both are short texts that begin with a name written in Cyrillic (*Sveta*), although the remainder is French.

Ref.	<A1	A1	A2	B1	B2	C1	C2
SV, *shared subset (30 descriptors)*							
A1	0	7	9	0	0	0	0
A2	0	4	46	26	3	1	0
B1	0	0	11	29	3	32	0
B2	0	0	0	1	3	70	0
C1	0	0	0	0	0	86	2
FR, *shared subset (30 descriptors)*							
A1	70	332	270	15	2	0	0
A2	22	252	745	258	50	48	0
B1	1	41	479	552	147	246	0
B2	1	4	179	390	177	674	2
C1	0	0	49	210	105	760	3
C2	0	0	12	49	34	388	2

Table 12: Confusion matrices under the shared 30-descriptor subset (in-sample for both corpora), showing the C1 attractor of Section 5.2.

F The Shared 30-Descriptor Subset

A1 (1): simple isolated phrases and sentences. A2 (2): series of simple phrases linked with simple connectors; short, simple imaginary biographies and simple poems. A2+ (1): very short, basic descriptions of events, past activities and personal experiences. B1 (6): general; straightforward connected texts; straightforward, detailed descriptions; description of an event or a recent trip; very brief reports in a standard conventionalized format; topic in a short report or poster. B1+ (2): text on a topical subject listing advantages and disadvantages; summarize, report and give opinion about accumulated factual information. B2 (4): general; clear, detailed texts synthesizing and evaluating information; clear, detailed descriptions; synthesize information and arguments from a number of sources. B2+ (3): clear, detailed descriptions of real or imaginary events following genre conventions; detailed description of a complex process; evaluate different ideas or solutions to a problem. C1 (6): general; clear, well-structured texts of complex subjects; structure and conventions of a variety of genres; clear, well-structured expositions; expand and support points of view at some length; suitable introduction and conclusion to a longer report. C2 (5): general; clear, smoothly flowing, complex texts; exploit idiom and humor appropriately; clear, smoothly flowing, complex reports, articles or essays; multiple perspectives on complex academic or professional topics.