

What a Machine Can Hear, Read and Judge: A Measurability-First Taxonomy of Correctness for Spoken L2 Spanish and French, and a Computable Rubric Specification

Jasper Vande Moortel
Orion-Solution, Belgium
jaspervandemoortel@orion-solution.net

Manuscript version 0.4. Reading edition: single column, one section per page.

Abstract

Automated grading of spoken second-language production is usually organised by linguistic level: pronunciation, grammar, vocabulary, discourse. A grading system needs a different organisation, by the artifact the evidence for a criterion survives in: the audio, the verbatim transcript, or human judgement. We propose a nine-layer taxonomy of what speaking rubrics score for L2 Spanish and French on that principle, with each layer assigned to the artifact that carries its evidence, spelling excluded because it has no spoken correlate and punctuation admitted only as a prosodic proxy. We test the principle on 166 criteria extracted verbatim from 10 published speaking rubrics and descriptor scales (CEFR, DELE, DELF and DALF, IELTS, TOEFL, and two Flemish sources), classified blind by two separate runs of one language model (Cohen's κ of 0.77 on the artifact label) with a third adjudicating disagreements: 56.6% of criteria are transcript-recoverable, 24.7% audio-recoverable, and 18.7% depend on a judgement no pipeline supplies, a share that does not differ measurably between exam grids, band descriptors and Flemish classroom material. We release a rubric specification that binds every criterion to its evidence source, scoring rule, weight and abstention policy, with an engine that re-scores two companion studies' released outputs exactly, and show from a logged search that no interoperability standard carries that binding, CASE and Open Badges 3.0 coming nearest with scored levels and no evidence field. The taxonomy is an argument tested on rubric text rather than on learner data, part of the criteria corpus is reconstructed from secondary quotations where primary documents blocked automated access, and the classifiers share one model; each limit is stated where it applies.

Contents

1	Introduction	3
2	Related work	4
3	A measurability-first taxonomy	6
3.1	The organising principle	6
3.2	Nine layers	6
3.3	Punctuation and spelling	8
3.4	Gravity is two axes, neither in the transcript	8
3.5	Why no existing scheme was adopted	8
4	The criteria corpus	9
5	A computable rubric specification	12
6	Discussion	14
6.1	The companion papers as one measurement per artifact	14
6.2	What a builder and a teacher do with the division	14
6.3	The specification as the contract	14
6.4	Conclusion	14
	Limitations	16
A	Provenance of the criteria corpus	21
B	Classification protocol	22
C	Judgement-dependent criteria	23
D	Artifact disagreements	27
E	Candidate schemes	30
F	Fields of the specification	31
G	Reference-implementation rescoring	32
H	Search log for the standards claim	33

1 Introduction

A teacher who wants a machine to grade a two-minute spoken answer in Spanish or French arrives with a rubric. The rubric says things like “uses the past tenses correctly”, “pronunciation is clear”, “answers the question with ease”. The builder of the grading system has to decide, criterion by criterion, whether the system can score it, and the available vocabulary is the wrong one: applied linguistics organises correctness by construct dimension and learner-corpus research by linguistic level crossed with operation, and neither says where the evidence for a criterion can be found once the utterance has been recorded, which is the only question the builder has.

This paper proposes to organise correctness by that question. Three artifacts exist for a spoken answer: the audio, a verbatim transcript, and a human judgement. Some evidence survives only in the first (a vowel quality, a pause), some survives transcription (a verb ending, a missing element), and some exists in neither and has to be supplied by a rater (whether the answer was appropriate or natural). The taxonomy in Section 3 assigns each of nine layers of correctness to the artifact that carries its evidence, and the assignment is a prediction: it says which modules a grading system will need, which rubric criteria each module can score, and which criteria no module can score however good the underlying models become.

The prediction is tested in Section 4 on a corpus of 166 rubric criteria and descriptor cells taken verbatim from 10 published sources for speaking assessment, classified blind by two independent classifiers and adjudicated by a third. Section 5 turns the taxonomy into a resource: a rubric specification in which every criterion names its evidence source, scoring rule, weight and abstention policy, with a validator and a scoring engine that reproduce the scores of two companion studies from their released outputs. Section 2 shows, from a logged search of the interoperability standards, that no existing standard carries that binding.

The paper makes no new measurement of learner speech. Its empirical content is the classification of published rubric text and the exact re-scoring of two released datasets; the per-layer measurements it rests on come from three companion papers in the same series, which scored transcript grammar with a language model, audio segmentals with a commercial pronunciation scorer, and transcript verb morphology with a conjugation index. Section 6.1 reads those three results as one measurement per artifact.

The contributions are the taxonomy, with the Dutch-L1 confusions the layers must carry and an explicit treatment of punctuation, spelling and gravity; a classified criteria corpus with inter-classifier agreement, released with its protocol, showing that about one criterion in five depends on judgement and that the transcript-versus-judgement boundary, not the audio boundary, is where classifiers disagree; a computable rubric specification with a reference engine and three example rubrics; and a documented absence: no interoperability standard binds a rubric criterion to its evidence and a scoring rule.

2 Related work

How correctness is organised today. Applied linguistics describes L2 performance through complexity, accuracy and fluency, with accuracy defined as adherence to the target language's norms (Housen and Kuiken, 2009), and the same programme refines how complexity and lexis are measured (Norris and Ortega, 2009; Skehan, 2009); that is a construct dimension, not a statement about where the evidence for a violation can be found. Learner-corpus error schemes organise by linguistic level crossed with operation: the Cambridge Learner Corpus assigns one of about eighty codes per surface deviation (Nicholls, 2003); MERLIN and Falko instead annotate a target hypothesis and read the errors off the difference, allowing competing hypotheses where more than one correction is defensible (Boyd et al., 2014; Reznicek et al., 2013); Díaz-Negrillo and Fernández-Domínguez (2006) survey the schemes. All of them presuppose a written text or a finished transcript. The one major spoken learner corpus with a published scheme, LINDSEI, tags errors with the Louvain manual written for written production, whose seven categories (form, grammar, lexico-grammar, lexis, word, punctuation, infelicities) contain no pronunciation category, and codes disfluencies as a separate layer (Gilquin et al., 2010; Gilquin and De Cock, 2013). That separation is the nearest precedent to the division drawn here, but it divides by phenomenon type, not by evidence source: a wrong verb ending has a code and a mispronounced segment none, because the transcript that carries the scheme has already discarded it. The CEFR's 2020 Companion Volume splits phonological control into overall control, sound articulation and prosodic features (Council of Europe, 2020), a modality-aware division that stops at pronunciation.

Error gravity. Munro and Derwing (1995) showed that intelligibility, comprehensibility and accentedness are dissociable; Burt and Kiparsky (1974) separate global errors, which break the relations between clauses, from local ones; Vann et al. (1984) began the empirical gravity literature, whose rankings vary by rater population. The two axes are kept apart here because neither is recoverable from a transcript alone and a system that treats accent as error measures the wrong construct.

Automated scoring and its constructs. The operational tradition already reasons about what a scorer can and cannot see: Williamson et al. (2012) frame the evaluation of automated scoring around construct representation, Xi (2010) sets out the validity argument for automated speaking assessment, and SpeechRater's feature families are an implicit measurability map for English (Evanini and Zechner, 2019). None of that work states the map as a taxonomy a teacher could apply to a rubric criterion before a system exists. The three companion papers in this series supply the per-layer measurements the map rests on: transcript grammar scored by a language model behind four gates; audio segmentals scored by a commercial service that flags planted errors in Spanish and none in French (Microsoft, 2026); and transcript verb morphology, which tagger features recover on most forms and lose on the imperative, conditional and subjunctive (Section 6.1).

Rubrics and their machine forms. Analytic rubrics with explicit criteria raise reliability and make expectations teachable (Jonsson and Svingby, 2007); rubric-referenced grading with language models inherits the rubric's structure and its biases (Kucia et al., 2026). What the rubric

research does not supply is a machine-readable form that binds a criterion to its evidence. QTI 3's rubric block and Common Cartridge's rubric element are presentational markup with no scoring semantics (1EdTech Consortium, 2026g,e); the 2005 1EdTech Rubric Specification says of itself that it provides no "information model to support automated evaluation software" (1EdTech Consortium, 2005); LTI Advantage's Assignment and Grade Services carry one numeric score per line item and no criteria (1EdTech Consortium, 2026a); xAPI has no rubric profile; learning-management systems expose rubrics through proprietary interfaces (Instructure, 2026). Two standards go further. Open Badges 3.0's result descriptions carry named criterion levels with points and a pass threshold (1EdTech Consortium, 2026b), and 1EdTech CASE carries rubric criteria with a numeric weight and criterion levels with a numeric score, a computable weighted multi-criterion rule (1EdTech Consortium, 2026c); the Comprehensive Learner Record reuses Open Badges' fields and OneRoster adds a free-text rubric value to a single-score line item (1EdTech Consortium, 2026d,f). CASE's rule stops at a criterion weight and a level score, Open Badges' at points and a threshold, and neither of the two, nor any other standard checked, has a field that binds evidence to a criterion. Section 5 states what a schema has to carry for spoken production, and Appendix H logs the searches.

Resources for Spanish and French. No spoken, CEFR-linked, error-tagged corpus with pronunciation annotation was found for French. For Spanish one exists: CORELE, a semi-spontaneous oral corpus of forty A2 and B1 learners from more than nine first languages, error-annotated at four levels including pronunciation and consulted through an online interface (Campillos Llanos, 2014); the pre-registered search missed it, its pronunciation tagset is documented at a coarser grain than the segmental and suprasegmental split used here, and whether the annotated data can be downloaded for reuse could not be established because its host was unreachable (Appendix E). The other resources each lack a property: COWS-L2H is written Spanish (Davidson et al., 2020), CEDEL2 is mostly written (Lozano, 2022), FLLOC is spoken French by level with orthographic transcripts and no error or pronunciation coding (Myles and Mitchell, 2026), NICT-JLE is the closest spoken precedent and excludes pronunciation from its tagset (Izumi et al., 2005), and the ERRANT family has no port for either language (Qiu et al., 2025; Masciolini et al., 2025). Flemish secondary practice grades speaking through descriptor levels rather than error counts (Centrum voor Taal en Onderwijs, 2026; Katholiek Onderwijs Vlaanderen, 2026), and the national assessment of French found speaking the weakest skill in the arts and technical tracks (Vlaams Ministerie van Onderwijs en Vorming, 2013); both facts shape what a rubric specification for this setting must express.

3 A measurability-first taxonomy

3.1 The organising principle

A grading system does not care whether an error is morphological or syntactic. It cares whether the evidence for the error survives into the artifact it can inspect. Three artifacts exist for spoken production.

Audio. Evidence recoverable from the waveform and destroyed by transcription: segmental realisation, stress, intonation, rhythm, pausing, rate, linking. A verbatim transcript with disfluency marks preserves pausing and repair and is counted here as audio-derived.

Transcript. Evidence recoverable from the words, whatever produced them: morphology, agreement, syntax, word choice, range, cohesion devices, length, the presence of required elements.

Judgement. A construct that is inherently evaluative or has no reliable automatic proxy: appropriateness, naturalness, listener effort, the interest of content, a comparison with a human standard.

Correctness is used here as a term of art: whatever a speaking rubric assigns points to. In the complexity, accuracy and fluency programme, fluency is a performance dimension distinct from accuracy (Housen and Kuiken, 2009), and task fulfilment is not a linguistic dimension at all; both are kept as layers because published rubrics score them and the artifact question applies to them exactly as it applies to a verb ending. The principle is a prediction, not a preference. It predicts that any system grading speech will have an audio module, a transcript module and a judgement stage with a human in it, that the modules will fail independently, and that a rubric criterion assigned to the wrong artifact cannot be scored however good the models are. Section 4 tests the prediction against published rubrics; Section 6.1 reads the three companion papers as one measurement per artifact.

3.2 Nine layers

Segmental (audio). Phoneme substitution, deletion and insertion, vowel quality, consonant voicing, allophonic realisation; in Spanish the approximant allophones of intervocalic /b d g/ mark non-nativeness without harming intelligibility. The Spanish confusions documented for Dutch-speaking learners are the tap–trill contrast and the palatal nasal (Sarmiento and De Vriendt, 1977); for French, nasal vowels and final obstruent voicing are the typologically expected trouble spots for a Dutch speaker, whose language has neither, but no study documents either error for Dutch-L1 learners specifically, and the taxonomy carries them as expectations, not findings. The companion pronunciation paper measured one commercial scorer’s sensitivity to planted errors of this kind and found it graded in Spanish and absent at word level in French.

Suprasegmental (audio). Lexical stress, phrase intonation, rhythm, pausing, rate, linking. Spanish word stress follows a default rule from the final letter of the word, and the written accent marks the exceptions, so a learner who has not internalised the rule and the exceptions places stress on the wrong syllable; French has no lexical stress and phrase-final prominence, so a Dutch learner who imports word stress onto every content word makes a phrasal error. Spanish and French are

conventionally syllable-timed and Dutch stress-timed, a difference that is computable as durational variability. French liaison is obligatory, forbidden or optional by context, so a liaison error is a segmental event governed by a syntactic rule and needs both artifacts to adjudicate.

Fluency (audio, with a verbatim transcript). Articulation and speech rate, filled and silent pauses (*eh, este, o sea; euh, ben, en fait*), false starts, repetitions, self-corrections, mean length of run. The metrics are language-independent; their baselines are not.

Lexical (transcript). Word choice, false friends (*embarazada* is not “embarrassed”; *actuellement* is not “actually”), collocation, register, range, repetition. Two of the companion grammar paper’s items carry English-mediated false friends: Flemish pupils learn English before Spanish or French, and a second language transfers into third-language lexis more readily than the first, the L2 status factor (Williams and Hammarberg, 1998; Hammarberg, 2001; Bardel and Falk, 2007). The one study with Dutch L1, English L2 and French L3 concerns syntax and finds the first language dominant there (Stadt et al., 2018), so the lexical expectation rests on the general literature, not on a population-matched study.

Morphological (transcript). Conjugation, gender and number agreement, article and pronoun forms: the richest error zone for this population. Two traps for a scorer: Spanish *leísmo* is dialectally standard and must not be penalised; French object clitics, *y* and *en* among them, are omitted by second-language learners far more often than they are misplaced or substituted (Grüter and Crago, 2011; Herschensohn, 2004), with Dutch *er* slowing the acquisition of *en* in particular (Sleeman and Ihsane, 2017), so a detector that looks for wrong forms misses the actual error. The companion tense paper measures how much of this layer a tagger and a conjugation index recover.

Syntactic (transcript). Word order, clitic placement, negation, question formation, subordination and mood triggers, agreement across clauses, missing or extra obligatory elements. French subject clitics are obligatory and Spanish is null-subject, so the same omission is an error in one language and unremarkable in the other; the taxonomy records the rule per language, not the operation.

Semantic (transcript plus judgement). Meaning errors (*conocer/saber*), contradiction, non-sense; whether the words mean what the task needed is a judgement.

Discourse and pragmatic (transcript plus task context). Cohesion, coherence, register, politeness, turn-taking. The *tú/usted* and *tú/vous* choice is a pragmatic failure distinct from the agreement of the pronoun with its verb; a system that checks only agreement scores a perfectly conjugated, socially wrong answer as correct.

Task fulfilment (transcript plus prompt). Content coverage, required elements, length; a fluent answer that ignores the prompt scores low here.

3.3 Punctuation and spelling

Teachers ask about both; the honest answers differ. Punctuation has a proxy: it stands in writing for intonational phrase boundaries, which pitch reset, pre-boundary lengthening and pausing mark acoustically, so the defensible construct is whether the speaker marked boundaries a listener could parse, reported as *phrasing*, an audio-layer criterion. Spelling has no spoken correlate: the orthography in a transcript belongs to the recogniser or the transcriber, so a spelling criterion on speech measures the pipeline, unless the task is to spell aloud.

3.4 Gravity is two axes, neither in the transcript

Intelligibility, comprehensibility and accentedness are dissociable (Munro and Derwing, 1995), and the CEFR moved its phonological target from native-likeness to intelligibility for that reason (Council of Europe, 2020); global errors break relations between clauses, local ones stay inside a clause (Burt and Kiparsky, 1974, cited from the catalogue record; the chapter itself was not read); empirical gravity rankings vary with the rater population (Vann et al., 1984). The taxonomy therefore carries gravity as two optional labels a criterion may add, listener consequence and structural scope, treats both as judgement-layer evidence unless a rater study has anchored them for the population at hand, and the engine reports them without scoring them.

3.5 Why no existing scheme was adopted

Table 7 (Appendix E) compares the candidate resources. None found is spoken, CEFR-linked, error-tagged and pronunciation-aware for French; for Spanish, CORELE describes all four properties and its reusability could not be established; the closest spoken precedent with a published scheme excludes pronunciation from its tagset (Izumi et al., 2005). Two design patterns are borrowed: the Cambridge crossing of level with operation, used inside the transcript layers, and Falko's competing target hypotheses, used wherever a learner utterance admits more than one correction, which the companion grammar paper met on its own items.

4 The criteria corpus

Sources and extraction. The pre-registration fixed the source list before extraction: the CEFR speaking scales (sustained monologue, the accuracy column of the qualitative-aspects table, grammatical accuracy, the three phonological-control subscales, coherence and cohesion) (Council of Europe, 2020; Relang / University of Novi Sad Faculty of Philosophy, 2020); the DELE B1 and B2 oral scales (Instituto Cervantes, 2014a,b); the examiner-facing DELF B1, DELF B2 and DALF C1 grilles (Centre international d'études pédagogiques, 2026; France Éducation internationale, 2026a,b); the IELTS speaking band descriptors (Ielts Partners, 2026); the TOEFL iBT independent speaking rubric (Educational Testing Service, 2024); and two Flemish sources, the rubric examples on the taalbeleid.be guidance page (Centrum voor Taal en Onderwijs, 2026) and the speaking-specific text characteristics of the attainment targets reproduced in the Katholiek Onderwijs Vlaanderen French curriculum (Katholiek Onderwijs Vlaanderen, 2026). One criterion is one descriptor cell or one rubric criterion row, copied verbatim with its source, level and language into the released corpus file. Two CEFR scales could not be obtained, several CEFR cells come from secondary quotations, the taalbeleid rubrics were transcribed from images and one of them concerns Dutch as language of schooling, and the Flemish attainment-target text describes expected performance rather than a grading grid; Appendix A gives the provenance per source and Table 1 the count. The corpus is a census of that fixed list, not a sample; the Wilson intervals describe sampling variability under the population of criteria of these kinds.

Classification. Two classifiers, each a separate run of the same language model (Claude Sonnet 5) given the protocol in Appendix B and one criterion at a time, blind to each other and to the source's identity, assigned each criterion a primary artifact (audio, transcript, judgement) and a layer (the nine of Section 3.2, plus *mixed* when two layers are equally present and *none* for a purely evaluative cell). A third run of the same model received both labels for every disagreement, without knowing which classifier had produced which, and picked one or a third label with a one-line reason. Agreement was $\kappa = 0.77$ (bootstrap 0.66 to 0.85) on the artifact and $\kappa = 0.83$ (0.76 to 0.89) on the layer; 40 of 166 criteria (24.1%) went to adjudication, 21 for the artifact, 24 for the layer and 5 for both. The two were not symmetric: the second called judgement more readily, the adjudicator sided with it in 26 of the 40 disputes against 13 (1 third label), and the judgement share would be 7.8% (Wilson 4.6 to 12.9) under the first classifier alone and 18.1% (13.0 to 24.6) under the second. A post-hoc keyword rule built from the protocol's own trigger words reaches 63.3% agreement with the adjudicated labels ($\kappa = 0.35$), so the classifiers did more than match the protocol's vocabulary.

Where the artifacts fall. Figure 1 and Table 1 give the shares. Over all 166 criteria, 56.6% (Wilson 49.0 to 63.9) are transcript-recoverable, 24.7% (18.8 to 31.8) audio-recoverable and 18.7% (13.5 to 25.3) judgement-dependent. By layer, discourse and pragmatics (22.9%) and mixed cells (22.9%) are the largest classes, followed by lexis (14.5%), task fulfilment (13.9%) and syntax (12.7%); the audio layers are small on their own (segmental 3.0%, suprasegmental 1.8%, fluency 3.0%) because published pronunciation descriptors bundle them: 28 of the 38 mixed cells are single “pronunciation” cells naming segments, stress and rhythm together. Morphology has 4 cells of its own; published rubrics fold it into “grammatical accuracy”.

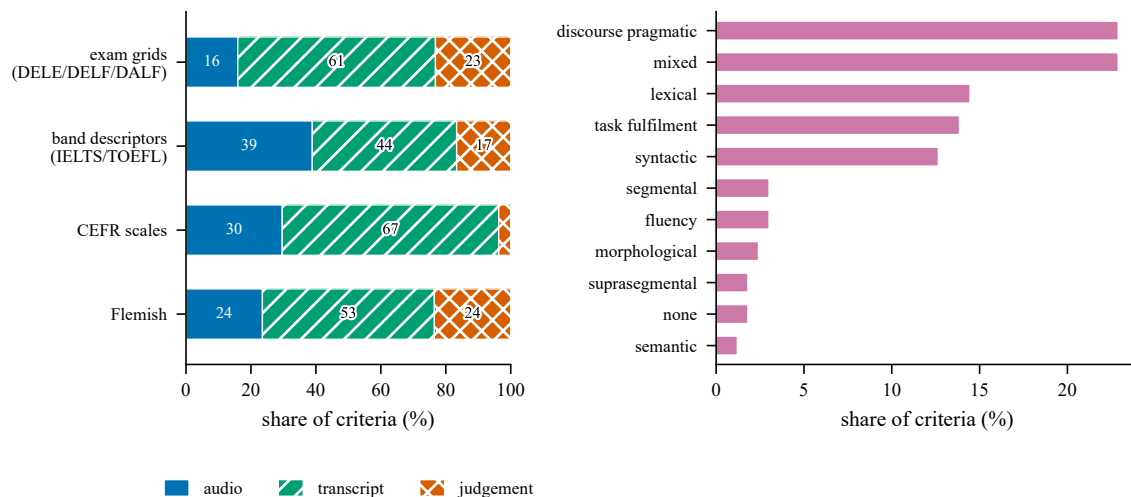


Figure 1: Left: share of criteria per source family recoverable from the audio, the transcript, or judgement only (percentages; n per family in Table 1). Right: share of all 166 criteria per taxonomy layer; *mixed* is a cell naming two layers equally, almost always a bundled pronunciation descriptor.

Where the classifiers disagreed. The audio boundary was never disputed: every one of the 41 criteria that one classifier called audio, the other called audio too. Of the 21 artifact disagreements, 19 were the first classifier reading a cell as transcript-recoverable and the second as judgement, and the adjudicator sided with judgement in 20 of the 21 (Tables 5 and 6). The disputed cells are of one kind: an observable behaviour qualified with ease, naturalness or effectiveness (“con facilidad y destreza . . . sin esfuerzo aparente”; “can interact with ease”), the boundary a rubric author has to decide on and the one the specification (Section 5) makes explicit through the criterion kind.

The judgement fifth. The 31 judgement-dependent criteria (Appendix C) sit in discourse and pragmatics (12), task fulfilment (10), mixed cells (5), no layer (3) and lexis (1); none is segmental, morphological or syntactic. They are the ceiling of automation for these rubrics: a pipeline without a rater must abstain on them or route them to a human, which is what the specification’s abstention policy is for.

The Flemish subset. The pre-registration expected the Flemish sources’ criteria to fall predominantly in the judgement class. They do not. Of the 34 Flemish criteria, 8 (23.5%, Wilson 12.4 to 40.0) are judgement-dependent, against 23.2% for the exam grids, 16.7% for the band descriptors and 3.7% for the CEFR scales; a chi-square test across the four families gives $p = 0.137$ and Fisher’s exact test of Flemish against the rest $p = 0.461$, and the Fisher test gives $p = 0.307$ under the first classifier’s labels alone, $p = 0.626$ under the second’s, and $p = 0.397$ with the off-population presentation rubric removed. Nor does the Flemish material phrase errors differently from the exam grids: its one error statement (errors occur but the message is mostly not compromised) uses the same comprehensibility-anchored, non-numeric language as the DELE, DELF, IELTS and CEFR cells, and no family in the corpus grades errors by count. A count-based criterion is therefore a design choice the specification has to make explicit, not a transcription of any family’s practice; the pre-registered claim is reported as not supported.

Table 1: Share of the 166 rubric criteria classified audio-, transcript-, and judgement-recoverable, by source and source family (Wilson 95% intervals, whole percentage points). Cohen’s κ between the two blind classifiers: artifact $\kappa = 0.77$ [0.66, 0.85] (3×3 , $n = 166$), layer $\kappa = 0.83$ [0.76, 0.89] ($n = 166$), both bootstrap 95% intervals over 2000 resamples of the criteria; 40/166 criteria (24%) were adjudicated by a third classifier.

family	source	n	audio	transcript	judgement
exam grids (DELE/DELF/DALF)	DELE B1	20	20 [8, 42]	70 [48, 85]	10 [3, 30]
	DELE B2	20	20 [8, 42]	55 [34, 74]	25 [11, 47]
	DELF B1	11	9 [2, 38]	45 [21, 72]	45 [21, 72]
	DELF B2	8	12 [2, 47]	88 [53, 98]	0 [0, 32]
	DALF C1	10	10 [2, 40]	50 [24, 76]	40 [17, 69]
<i>exam grids (DELE/DELF/DALF) total</i>		69	16 [9, 26]	61 [49, 72]	23 [15, 34]
band descriptors (IELTS/TOEFL)	IELTS	24	42 [24, 61]	50 [31, 69]	8 [2, 26]
	TOEFL	12	33 [14, 61]	33 [14, 61]	33 [14, 61]
<i>band descriptors (IELTS/TOEFL) total</i>		36	39 [25, 55]	44 [30, 60]	17 [8, 32]
CEFR scales	CEFR scales	27	30 [16, 48]	67 [48, 81]	4 [1, 18]
Flemish	taalbeleid	19	32 [15, 54]	47 [27, 68]	21 [9, 43]
	KOV	15	13 [4, 38]	60 [36, 80]	27 [11, 52]
<i>Flemish total</i>		34	24 [12, 40]	53 [37, 69]	24 [12, 40]
Overall		166	25 [19, 32]	57 [49, 64]	19 [13, 25]

5 A computable rubric specification

A rubric that a system can score must say, for every criterion, where the evidence comes from and how the evidence becomes a score. The specification released with this paper (spec/rubric.schema.json, JSON Schema 2020-12, with a validator and a scoring engine) makes both explicit. Table 8 (Appendix F) lists the fields; the design decisions follow.

Three kinds of criterion. Every criterion is one of three kinds, and the kind fixes its scoring rule. A *count* criterion counts typed entries in an error list produced upstream and maps the count to a band through a table; the companion grammar paper’s rubric is one such table (9 layers are available as the criterion’s layer). A *threshold* criterion compares a measured signal (a pronunciation score, a rate, a length) with a value or a piecewise table, or combines such comparisons by disjunction or conjunction; the conjunction (a pass bar on two scores at once) was missing from the first release, a refutation attempt built a pipeline-scorable criterion the schema could not express, and it was added, so the sufficiency claim is limited to the criteria in the corpus and the two companion rubrics. A *band* criterion is a judged level anchored to exemplar performances, with the rater named as human or model. The distinction matters because the first two kinds are auditable to the item and the third is not; a rubric that hides a band inside a count, by asking a model for a score and calling it a count, is what the companion grammar paper measured extract-then-score against.

Evidence source and artifact. Each criterion names its evidence source (a pipeline output and field) and its artifact, declared independently of the current implementation, so a criterion can be audio-recoverable in principle and unimplemented in practice. A rubric can therefore be checked against the taxonomy before any system exists; a judgement criterion with `rater: model` is a flag the engine reports, not an error.

Weight, abstention, readings, constraints. Every criterion carries a weight and its own abstention policy: the condition under which the engine declines to score it (evidence missing, a parse failure, a confidence below a stated value) and the route the item then takes (teacher, re-run, or a zero with a reason); one global advisory flag is not enough when the evidence for one criterion can be missing while the others are sound. Two fixed readings, errors per hundred words and a CEFR estimate with its basis, keep different teachers’ rubrics comparable in one dimension the teacher did not choose. Assignment constraints (required words with lemma matching, required tenses, topic adherence, anti-gaming) are separate from criteria, because failing a constraint and scoring low on a criterion are different events for the teacher and the learner.

Validation on the companion papers. The engine re-scores the companion grammar paper’s 24 items from their released error lists with the grammar rubric expressed in the schema and reproduces the recorded band on 24 of 24 (23 of 24 against the planted-error gold, the one difference being the corpus-convention item that paper documents), and re-scores the companion pronunciation paper’s 80 broken clips with its word-level flag rule and reproduces the recall count in 8 of 8 cells (Table 9, Appendix G). Agreement is expected, since the engine implements the same rules; the value is that the rules now sit in one machine-checkable form, and a third example rubric in

the release (a Flemish French speaking task with one criterion of each kind) shows the form a teacher's rubric takes.

6 Discussion

6.1 The companion papers as one measurement per artifact

The taxonomy predicts three modules and a human stage. The three companion papers built one module each, and each headline result is a measurability result for its artifact. The grammar paper scored transcript-layer errors with a language model and found that the usable configurations pass four gates at once and that a model which returns a band without the error list behind it cannot be audited to the item: transcript evidence is recoverable, and the recovery has to be checked per criterion, which is what a count criterion with an evidence source does. The pronunciation paper planted segmental errors in synthesised Spanish and French and found that a commercial scorer flags them in Spanish and scores them without flagging them in French: audio evidence is recoverable in principle and its recovery is language-dependent in practice, which is why the specification declares the artifact independently of the implementation. The tense paper found that a dependency tagger returns the teacher's tense label for most forms and loses the imperative, the conditional and the subjunctive, while an exhaustive conjugation index covers every form and disambiguates few: the transcript carries the evidence, and whether it is recovered depends on asking the artifact the right question. None of the three touched the judgement fifth, and the corpus says why: no evidence source exists for it.

6.2 What a builder and a teacher do with the division

For a builder, the taxonomy is a module map and a failure map: audio and transcript modules fail independently, so criteria should be scored from their own artifact and abstain on their own condition. For a teacher, it is a filter to run before a system exists: a judgement-dependent criterion can be kept with a human rater declared or rewritten into a count or a threshold, but cannot be handed to a pipeline as it stands; about a fifth of the criteria in the rubrics teachers copy from are of that kind.

6.3 The specification as the contract

The rubric specification is where the two meet: it requires the criterion decision, kind, artifact, evidence source, rule, weight, abstention, to be written down rather than left implicit. A rubric in the schema can be checked against the taxonomy before any audio is recorded, and the check reports a judgement criterion with a model as rater as a flag, not an error. The engine's exact reproduction of the two companion papers' scores is not a validation of those papers; it shows that their scoring rules and a teacher's rubric with one criterion of each kind fit one machine-checkable form. Whether teachers will write rubrics in it is an empirical question this paper does not answer.

6.4 Conclusion

Correctness in spoken L2 production is recoverable from three artifacts, and a rubric criterion belongs to one of them. Published speaking rubrics for Spanish and French are mostly transcript-recoverable, a quarter audio-recoverable and about a fifth judgement-dependent, with the disputed

boundary between transcript and judgement. No interoperability standard binds a criterion to its evidence; the released specification does. The taxonomy's next test is a rubric written in the specification and scored on real learner speech.

Limitations

The corpus is rubric text, not learner speech: the shares are shares of what published documents ask for, and a rubric a teacher writes for one class may differ. Its provenance is uneven and is listed per source in Appendix A: two pre-registered CEFR scales were not obtained, several CEFR cells come from secondary quotations rather than the Council of Europe's document, the DELF and DALF grilles were read from institutional mirrors of the examiner documents, the taalbeleid rubrics were transcribed from images and one of them concerns Dutch as language of schooling, and the Flemish curriculum entries are attainment-target descriptions rather than a grading grid; the Flemish subset is therefore the weakest part of the corpus, and its negative result (Section 4) would also look different if the TCF, TEF and SIELE grids that Flemish schools meet alongside DELF and DELE, excluded by the pre-registration, had been in the exam-grid family. The criterion texts are quoted verbatim from copyrighted assessment documents for research and criticism; the release licenses the annotations, not the texts (Appendix A). The three classifiers were separate runs of one language model applying a protocol the authors wrote, so the agreement figures measure the protocol's clarity as much as the taxonomy's, a shared bias of the model would not show as disagreement, and the judgement share moves by more than ten points depending on which single classifier is believed (Section 4); a human classification with the same protocol has not been run, and the transcript-versus-judgement boundary, where nearly all disagreement sits, is the one a human panel would most need to settle. The taxonomy's nine layers and three artifacts are a proposal; the corpus tests whether criteria can be sorted into them consistently. Whether that sorting predicts grading accuracy is a separate question, one only a system scoring real learner speech against a rubric in the specification could answer. The engine validates against two papers whose rules it implements, so exact agreement is expected and the validation is of the form, not of the scores, and the schema's sufficiency is a claim about the criteria in this corpus, one of which a refutation attempt could not express until conjunctive threshold rules were added. The English-language sources (IELTS, TOEFL) are included because they are the descriptors most often copied into classroom rubrics for other languages, not because the taxonomy was tested on English speech. No teacher has yet written a rubric in the specification.

References

- 1EdTech Consortium. 2005. Iedtech rubric specification v1.0. https://www.imsglobal.org/ep/epv1p0/imsrubric_specv1p0.html. Accessed September 10, 2026.
- 1EdTech Consortium. 2026a. Iedtech learning tools interoperability advantage: Assignment and grade services (AGS) v2.0. <https://www.imsglobal.org/spec/lti-ags/v2p0>. Accessed September 10, 2026.
- 1EdTech Consortium. 2026b. Iedtech open badges 3.0 specification: Appendix b, ResultDescription and RubricCriterionLevel. <https://www.imsglobal.org/spec/ob/v3p0/>. Accessed September 10, 2026.
- 1EdTech Consortium. 2026c. Case service v1.1 information model (CFRubric, CFRubricCriterion, CFRubricCriterionLevel). https://www.imsglobal.org/sites/default/files/spec/case/v1p1/information_model/caseservicev1p1_infomodelv1p0.html. Accessed September 10, 2026.
- 1EdTech Consortium. 2026d. Comprehensive learner record standard v2.0. <https://www.imsglobal.org/spec/clr/v2p0/>. Accessed September 10, 2026.
- 1EdTech Consortium. 2026e. Ims common cartridge v1.3 implementation guide: the rubric element. https://www.imsglobal.org/cc/ccv1p3/imsc_Implementation-v1p3.html. Accessed September 10, 2026.
- 1EdTech Consortium. 2026f. Oneroster v1.2. <https://www.imsglobal.org/spec/oneroster/v1p2/>. Accessed September 10, 2026.
- 1EdTech Consortium. 2026g. Qti v3 best practices and implementation guide (rubricBlock). <https://www.imsglobal.org/spec/qti/v3p0/impl>. Accessed September 10, 2026.
- Camilla Bardel and Ylva Falk. 2007. The role of the second language in third language acquisition: the case of germanic syntax. *Second Language Research*, 23(4):459–484.
- 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. <https://aclanthology.org/L14-1488/>. Accessed September 10, 2026.
- Marina K. Burt and Carol Kiparsky. 1974. Global and local mistakes. <https://archive.org/details/tails/newfrontiersinse0000unse>. Accessed September 10, 2026.
- Leonardo Campillos Llanos. 2014. A spanish learner oral corpus for computer-aided error analysis. *Corpora*, 9(2):207–238.
- Centre international d’études pédagogiques. 2026. Grille d’évaluation, production orale, DELF B1. http://deniscousineau.pbworks.com/f/grille+orale-b1_doc4.pdf. Accessed September 10, 2026.
- KU Leuven Centrum voor Taal en Onderwijs. 2026. Rubrics (aan de slag met taalbeleid). <https://www.taalbeleid.be/nl/aan-de-slag/rubrics>. Accessed September 10, 2026.

- Council of Europe. 2020. Common european framework of reference for languages: Learning, teaching, assessment — companion volume. <https://www.coe.int/en/web/common-european-framework-reference-languages>. Accessed September 10, 2026.
- Sam Davidson, Aaron Yamada, Paloma Fernández-Mira, Agustina Carando, Claudia H. Sánchez-Gutiérrez, and Kenji Sagae. 2020. Developing NLP tools with a new corpus of learner spanish. <https://aclanthology.org/2020.lrec-1.894/>. Accessed September 10, 2026.
- Ana Díaz-Negrillo and Jesús Fernández-Domínguez. 2006. Error tagging systems for learner corpora. <https://dialnet.unirioja.es/descarga/articulo/2198610.pdf>. Accessed September 10, 2026.
- Educational Testing Service. 2024. Toefl iBT independent speaking rubric. <https://www.ets.org/content/dam/ets-india/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf>. Accessed September 10, 2026.
- Keelan Evanini and Klaus Zechner. 2019. *Overview of Automated Speech Scoring*, pages 3–20. Routledge.
- France Éducation international. 2026a. Grille d'évaluation de la production orale B2 (DELFL). https://moodle.insa-toulouse.fr/pluginfile.php/31016/mod_resource/content/1/grille-correction_DELF-B2.pdf. Accessed September 10, 2026.
- France Éducation international. 2026b. Grille d'évaluation de la production orale C1 (DALF). https://d11.cuni.cz/pluginfile.php/576537/mod_folder/content/0/dalf_c1_grille_po.pdf. Accessed September 10, 2026.
- Gaëtanelle Gilquin and Sylvie De Cock, editors. 2013. *Errors and Disfluencies in Spoken Corpora*. John Benjamins Publishing Company.
- Gaëtanelle Gilquin, Sylvie De Cock, and Sylviane Granger. 2010. Lindsei: Louvain international database of spoken english interlanguage. <https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html>. Accessed September 10, 2026.
- Theres Grüter and Martha Crago. 2011. Object clitics and their omission in child L2 french: The contributions of processing limitations and L1 transfer. *Bilingualism: Language and Cognition*, 15(3):531–549.
- Björn Hammarberg. 2001. *Roles of L1 and L2 in L3 production and acquisition*, pages 21–41. Multilingual Matters.
- Julia Herschensohn. 2004. *Functional categories and the acquisition of object clitics in L2 French*, pages 207–242. John Benjamins Publishing Company.
- A. Housen and F. Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.
- Ielts Partners. 2026. Ielts speaking band descriptors (public version). <https://ielts.org/cdn/ielts-guides/ielts-speaking-band-descriptors.pdf>. Accessed September 10, 2026.

- Instituto Cervantes. 2014a. Guía del examen DELE B1. https://exámenes.cervantes.es/sites/default/files/guia_examen_dele_b1.pdf. Accessed September 10, 2026.
- Instituto Cervantes. 2014b. Guía del examen DELE B2. https://exámenes.cervantes.es/sites/default/files/guia_examen_dele_b2_0.pdf. Accessed September 10, 2026.
- Instructure. 2026. Canvas LMS REST API: Rubrics. <https://www.canvas.instructure.com/doc/api/rubrics.html>. Accessed September 10, 2026.
- Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2005. Error annotation for corpus of japanese learner english. <https://aclanthology.org/I05-6009.pdf>. Accessed September 10, 2026.
- Anders Jonsson and Gunilla Svingby. 2007. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2):130–144.
- Katholiek Onderwijs Vlaanderen. 2026. Taalonderwijs frans: aan de slag in de klas (evaluatie). <https://pro.katholiekonderwijs.vlaanderen/taalonderwijs-frans/aan-de-slag-in-de-klas>. Accessed September 10, 2026.
- Filip J. Kucia, Anirban Chakraborty, and Anna Wróblewska. 2026. Llm essay scoring under holistic and analytic rubrics: Prompt effects and bias. arXiv:2604.00259. *arXiv preprint*.
- Cristóbal Lozano. 2022. Cedel2: Design, compilation and web interface of an online corpus for L2 spanish acquisition research. <https://journals.sagepub.com/doi/full/10.1177/02676583211050522>. Accessed September 10, 2026.
- Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction. <https://aclanthology.org/2025.nlp4call-1.1/>. Accessed September 10, 2026.
- Microsoft. 2026. Use pronunciation assessment (Azure AI Speech documentation, model-version note). <https://learn.microsoft.com/azure/ai-services/speech-service/how-to-pronunciation-assessment>. Accessed September 10, 2026.
- Murray J. Munro and Tracey M. Derwing. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1):73–97.
- Florence Myles and Rosamond Mitchell. 2026. Flloc: French learner language oral corpora. <https://web.archive.org/web/2023/https://www.flloc.soton.ac.uk/>. Accessed September 10, 2026.
- Diane Nicholls. 2003. The cambridge learner corpus: Error coding and analysis for lexicography and ELT. <https://ucrel.lancs.ac.uk/publications/CL2003/papers/nicholls.pdf>. Accessed September 10, 2026.
- J. M. Norris and L. Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4):555–578.

- Mengyang Qiu, Tran Minh Nguyen, Zihao Huang, Zelong Li, Yang Gu, Qingyu Gao, Siliang Liu, and Jungyeul Park. 2025. Multilingual grammatical error annotation: Combining Language-Agnostic framework with Language-Specific flexibility. arXiv:2506.07719. *arXiv preprint*.
- Relang / University of Novi Sad Faculty of Philosophy. 2020. Speaking assessment criteria and scales (RELANG 2020 workshop materials, reproducing CEFR speaking scales). <https://www.ff.uns.ac.rs/uploads/files/Fakultet/Centri/Centar%20za%20jezike/RELANG2020/Materijali%20za%20radionice/Speaking%20assessment%20criteria%20and%20scales.pdf>. Accessed September 10, 2026.
- Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. *Competing target hypotheses in the Falko corpus: A flexible multi-layer corpus architecture*, pages 101–124. John Benjamins Publishing Company.
- Sarmiento and De Vriendt. 1977. Corrección fonética de los neerlandófonos que estudian español. https://cvc.cervantes.es/ensenanza/biblioteca_ele/aepe/pdf/boletin_17_10_77/boletin_17_10_77_09.pdf. Accessed September 10, 2026.
- P. Skehan. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4):510–532.
- Petra Sleeman and Tabea Ihsane. 2017. *The L2 acquisition of the French quantitative pronoun <i>en</i> by L1 learners of Dutch: Vulnerable domains and cross-linguistic influence*, pages 303–330. John Benjamins Publishing Company.
- Rosalinde Stadt, Aafke Hulk, and Petra Sleeman. 2018. The influence of L1 dutch and L2 english on L3 french: A longitudinal study. *Journal of the European Second Language Association*, 2(1):63.
- Roberta J. Vann, Daisy E. Meyer, and Frederick O. Lorenz. 1984. Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly*, 18(3):427.
- Vlaams Ministerie van Onderwijs en Vorming. 2013. Peiling frans in de derde graad aso, kso en tso van het secundair onderwijs: luisteren en spreken. https://docs.vlaamsparlament.be/docs/biblio/opensdigbib/monografie/2013/728_peiling_frans.pdf. Accessed September 10, 2026.
- S. Williams and B. Hammarberg. 1998. Language switches in L3 production: Implications for a polyglot speaking model. *Applied Linguistics*, 19(3):295–333.
- David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.
- Xiaoming Xi. 2010. Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3):291–300.

A Provenance of the criteria corpus

The corpus file (`corpora/rubric_criteria.jsonl`) carries one record per criterion with its verbatim text, an English gloss where the source is not English, and its source, level and language; the released sources log records the fetch date, the URL and the document for every source, and Table 1 gives the count per source. Provenance a reader should weigh: the CEFR sustained-monologue scale and the accuracy column of the qualitative-aspects table were read from a university workshop document that reproduces them (Relang / University of Novi Sad Faculty of Philosophy, 2020), because the Council of Europe’s server refuses automated access; the phonological-control, sound-articulation and prosodic-features cells were recovered from secondary quotations and cover only part of each scale; the grammatical-accuracy and coherence scales of the Companion Volume (Council of Europe, 2020) came by the same secondary route; the vocabulary-range and overall-oral-production scales were not obtained and are absent. The DELE guides are the official Instituto Cervantes documents (Instituto Cervantes, 2014a,b). The DELF B1, DELF B2 and DALF C1 grilles are the examiner-facing point-scored documents, read from institutional mirrors because the issuing body’s hosts refuse automated access (Centre international d’études pédagogiques, 2026; France Éducation internationale, 2026a,b). The IELTS and TOEFL documents are the official public versions (Ielts Partners, 2026; Educational Testing Service, 2024). The taalbeleid.be rubrics are two example rubrics on a guidance page, a telephone-conversation rubric and a presentation rubric, transcribed from the page’s images; the second is for Dutch as language of schooling and neither carries a CEFR level (Centrum voor Taal en Onderwijs, 2026). The Katholiek Onderwijs Vlaanderen entries are the text characteristics of the attainment targets for production and interaction, including the additions for spoken production, as reproduced in the French curriculum for the second stage; every entry was checked to occur verbatim in the document’s extracted text (Katholiek Onderwijs Vlaanderen, 2026). SIELE, TCF and TEF, the other exams Flemish learners meet, and the ACTFL proficiency guidelines were outside the pre-registered list and are absent; the exam-grid family would be larger and the Flemish comparison different with them in. The criterion texts are quoted verbatim from copyrighted assessment documents for research and criticism, with attribution per row; the release licenses the annotations, identifiers, glosses and structure added here under CC BY 4.0 and not the texts, whose copyright stays with their publishers.

B Classification protocol

The protocol given to the classifiers, fixed before any criterion was read, asked one question: from which artifact is the evidence for this criterion primarily recoverable by an automated system? It defined the three answers as in Section 3.1, with the added rules that a verbatim transcript with disfluency marks counts as audio-derived for pausing and repair, that a mixed criterion takes the artifact of its dominant clause with the secondary one noted, and that a criterion naming a proficiency level with no observable behaviour is judgement. The second label was the layer, with *mixed* for two equally present layers and *none* for a purely evaluative cell. Classifiers were not told the other classifier's labels, the paper's hypothesis, or the source's identity beyond what the criterion text carries; the adjudicator saw the source only after labelling. Each classifier was a separate run of Claude Sonnet 5 with its own context, recorded in the packet; the raw label files of all three and the adjudication key are released, and a post-hoc keyword baseline built from the protocol's trigger words is reported in Section 4 as a floor.

C Judgement-dependent criteria

Tables 2 to 4 list every criterion whose adjudicated artifact is judgement, with its source and layer, verbatim and with an English gloss where the source is not English. Reading down the list, the recurring markers are ease, naturalness, effectiveness, appropriateness, and a comparison with a human standard; none of them names a behaviour a module could count or threshold without a rater.

Table 2: Every criterion classified judgement-dependent (no automatic proxy), with its source and taxonomy layer. $n = 31$, part 1 of 3.

id	source	layer	criterion (verbatim, English gloss in parentheses where the source is not English)
cefr_sustmono_c2	CEFR scales	discourse_pragmatic	Can give clear, smoothly-flowing, elaborate and often memorable descriptions.
dalf_c1_r1	DALF C1	task_fulfilment	Peut analyser avec un regard critique les textes sources, extraire et utiliser des informations importantes. (0 / 0.5 / 1 / 1.5) (Can critically analyse source texts, extract and use important information. (0-1.5 points))
dalf_c1_r4	DALF C1	discourse_pragmatic	Peut faire une présentation claire et organisée avec aisance, spontanéité et pertinence pour parvenir à une conclusion appropriée. (0/0.5 / 1 / 1.5 / 2 / 2.5) (Can give a clear, organised presentation with ease, spontaneity and relevance, reaching an appropriate conclusion. (0-2.5 points))
dalf_c1_r5	DALF C1	discourse_pragmatic	Peut facilement préciser et défendre sa position en répondant aux questions, commentaires et contre-arguments. (0 / 0.5 / 1 / 1.5 / 2) (Can easily clarify and defend his/her position by responding to questions, comments and counter-arguments. (0-2 points))
dalf_c1_r6	DALF C1	discourse_pragmatic	Peut faciliter le développement de la discussion en recentrant et / ou élargissant le débat. (0 / 0.5 / 1 / 1.5 / 2) (Can facilitate the development of the discussion by refocusing and/or widening the debate. (0-2 points))
dele_b1_holistica_0	DELE B1	task_fulfilment	Apenas proporciona información, por lo que no cumple con el objetivo comunicativo de las tareas. Las limitaciones lingüísticas provocan dificultades en la formulación de lo que quiere decir. Requiere que el interlocutor repita lo que ha dicho o lo diga con otras palabras y a una velocidad lenta y que le ayude a formular lo que intenta decir. (Barely provides information, so does not meet the tasks' communicative goal. Linguistic limitations cause difficulty formulating what he/she wants to say. Requires the interlocutor to repeat/rephrase slowly and help formulate the intended message.)
dele_b1_holistica_1	DELE B1	task_fulfilment	Aunque hace descripciones y presentaciones sencillas, no aporta la información suficiente para cumplir con el objetivo comunicativo de las tareas. Aunque su repertorio lingüístico limitado le permite transmitir información sobre temas personales y del entorno más cercano en situaciones sencillas y cotidianas, tiene que adaptar el mensaje y buscar las palabras y comete repetidamente errores básicos. Participa en conversaciones e intercambia información siempre que el interlocutor le ayude. (Although makes simple descriptions and presentations, does not provide enough information to meet the tasks' communicative goal. Limited repertoire allows transmitting personal/everyday information but requires adapting the message, searching for words, and repeatedly making basic errors. Participates only with the interlocutor's help.)
dele_b2_coherencia_3	DELE B2	discourse_pragmatic	Muestra un discurso coherente y cohesionado, con un uso apropiado y variado de estructuras organizativas, conectores y otros mecanismos de cohesión. Conversa con facilidad y destreza, utilizando sin esfuerzo aparente los recursos lingüísticos apropiados y siguiendo con naturalidad los turnos de palabra. (Shows coherent, cohesive discourse with appropriate varied use of organisational structures, connectors and other cohesion devices. Converses with ease and skill, using appropriate resources apparently effortlessly and following turn-taking naturally.)
dele_b2_holistica_0	DELE B2	task_fulfilment	Apenas es capaz de expresar sus opiniones o describir situaciones, por lo que no cumple con el objetivo comunicativo de las tareas. Las limitaciones lingüísticas provocan dificultades en la formulación de lo que quiere decir. Contesta a preguntas y responde a afirmaciones sencillas e indica cuándo comprende. Necesita la colaboración de su interlocutor porque no comprende lo suficiente. (Barely able to express opinions or describe situations, not meeting the communicative goal. Linguistic limitations cause difficulty formulating the intended message. Answers simple questions/statements and indicates comprehension. Needs the interlocutor's collaboration due to insufficient comprehension.)

Table 3: Every criterion classified judgement-dependent (no automatic proxy), with its source and taxonomy layer. $n = 31$, part 2 of 3.

id	source	layer	criterion (verbatim, English gloss in parentheses where the source is not English)
dele_b2_holistica_1	DELE B2	task_fulfilment	Aunque aporta algunas ideas o impresiones, son poco claras o incompletas, por lo que no cumple con el objetivo comunicativo de las tareas. Se expresa con esfuerzo y son evidentes sus limitaciones lingüísticas: errores gramaticales, incorrecciones e imprecisiones léxicas. Mantiene conversaciones sencillas sobre temas cotidianos. Puede necesitar alguna aclaración y repetir parte de lo que su interlocutor ha dicho para confirmar la comprensión mutua. (Provides some ideas/impresions but they are unclear or incomplete, not meeting the communicative goal. Speaks with effort; linguistic limitations evident (grammar errors, lexical inaccuracies). Holds simple conversations, may need clarification/repetition.)
dele_b2_holistica_2	DELE B2	task_fulfilment	Expresa sus opiniones y describe las situaciones planteadas con claridad. Cumple con el objetivo comunicativo de las tareas. Tiene un repertorio lingüístico que le permite expresarse con claridad, dando pocas muestras de restringir lo que dice. Demuestra un control gramatical y léxico relativamente alto, aunque comete errores e imprecisiones que en ocasiones corrige. Mantiene la conversación de forma adecuada, confirmando su comprensión y colaborando con su interlocutor. (Expresses opinions and describes situations clearly. Meets the communicative goal of the tasks. Has a repertoire allowing clear expression with little sign of restricting content. Relatively high grammatical/lexical control despite errors, sometimes self-corrected. Keeps the conversation going, confirming comprehension and collaborating.)
dele_b2_holistica_3	DELE B2	mixed	Expresa sus opiniones, describe las situaciones planteadas con claridad y desarrolla sus ideas aportando detalles relevantes y resaltando los aspectos significativos. Tiene un repertorio lingüístico que le permite expresarse con claridad y formular lo que quiere decir. Demuestra un alto control gramatical y léxico. Comete escasos errores e imprecisiones y en ocasiones los corrige. Conversa con facilidad y destreza, utilizando sin esfuerzo aparente los recursos lingüísticos apropiados. Hace alusiones a lo dicho por su interlocutor de forma natural. (Expresses opinions, describes situations clearly and develops ideas with relevant detail and emphasis. Has a linguistic repertoire allowing clear expression. Shows high grammatical/lexical control, few errors, sometimes self-corrected. Converses with ease and skill, alluding naturally to the interlocutor's words.)
delf_b1_r2	DELFB1	discourse_pragmatic	Peut aborder sans préparation un échange sur un sujet familier avec une certaine assurance (0 / 0.5 / 1) (Can undertake, without preparation, an exchange on a familiar topic with some assurance. (0-1 point))
delf_b1_r3	DELFB1	discourse_pragmatic	Peut faire face sans préparation à des situations même un peu inhabituelles de la vie courante (respect de la situation et des codes sociolinguistiques) (0 / 0.5 / 1) (Can deal, without preparation, with somewhat unusual everyday situations (respecting the situation and sociolinguistic codes). (0-1 point))
delf_b1_r4	DELFB1	discourse_pragmatic	Peut adapter les actes de parole à la situation (0 / 0.5 / 1 / 1.5 / 2) (Can adapt speech acts to the situation. (0-2 points))
delf_b1_r6	DELFB1	task_fulfilment	Peut présenter d'une manière simple et directe le sujet à développer (0 / 0.5 / 1) (Can present the topic to be developed in a simple and direct way. (0-1 point))
delf_b1_r8	DELFB1	discourse_pragmatic	Peut relier une série d'éléments en un discours assez clair pour être suivi sans difficulté la plupart du temps (0 / 0.5 / 1 / 1.5) (Can link a series of elements into discourse clear enough to be followed without difficulty most of the time. (0-1.5 points))
ielts_pron_5	IELTS	none	Displays all the positive features of band 4, and some, but not all, of the positive features of band 6.
ielts_pron_7	IELTS	none	Displays all the positive features of band 6, and some, but not all, of the positive features of band 8.

Table 4: Every criterion classified judgement-dependent (no automatic proxy), with its source and taxonomy layer. $n = 31$, part 3 of 3.

id	source	layer	criterion (verbatim, English gloss in parentheses where the source is not English)
kov_fra_da_et37_ coherence	KOV	discourse_pragmatic	Redelijk herkenbare samenhang (Reasonably recognisable coherence.)
kov_fra_da_et37_ modeling_bodylang	KOV	none	Gepaste, niet-storende lichaamstaal (Appropriate, non-distracting body language.)
kov_fra_da_et37_ register	KOV	discourse_pragmatic	Gebruik van gepaste beleefdheidsconventies bij alledaagse taalhandelingen en van gepast register: neutraal, familiair (Use of appropriate politeness conventions in everyday speech acts and of appropriate register: neutral, familiar.)
kov_fra_da_et38_ modeling_interaction	KOV	discourse_pragmatic	Kan redelijk gemakkelijk interactief zijn in gestructureerde situaties en korte gesprekken, mits er om herhaling of herformulering kan gevraagd worden; kan een eenvoudig rechtstreeks gesprek beginnen, gaande houden en beëindigen; kan op eenvoudige wijze om herhaling vragen (zoals vragen om belangrijke woorden of zinnen te verduidelijken a.d.h.v. standaardzinnnetjes) (Can interact with reasonable ease in structured situations and short conversations, provided repetition or rephrasing can be requested; can begin, maintain and end a simple face-to-face conversation; can ask for repetition in a simple way (such as asking for key words or phrases to be clarified using standard phrases).)
toefl_topic_1	TOEFL	task_fulfilment	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
toefl_topic_2	TOEFL	mixed	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
toefl_topic_3	TOEFL	mixed	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall development is somewhat limited, usually lacks elaboration or specificity. Relationships between ideas may at times not be immediately clear.
toefl_topic_4	TOEFL	mixed	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or there is a clear progression of ideas).
taalbeleid_pres_an_0	taalbeleid	mixed	Je maakt te veel storende taalfouten. (You make too many disruptive language errors.)
taalbeleid_pres_ vaktaal_0	taalbeleid	lexical	Je gebruikt de vaktermen niet correct of gebruikt vage woorden. Dit komt niet geloofwaardig over. Het lijkt alsof je de inhoud van de presentatie zelf niet beheerst. (You use the technical terms incorrectly or use vague words. This does not come across as credible. It seems as if you do not master the content of the presentation yourself.)
taalbeleid_telefoon_ boodschap_1	taalbeleid	task_fulfilment	Je slaagde er in om een afspraak vast te leggen, maar dit kon op een efficiëntere manier. (You managed to fix an appointment, but it could have been done more efficiently.)
taalbeleid_telefoon_ boodschap_2	taalbeleid	task_fulfilment	Je slaagde er op een efficiënte manier in om een afspraak vast te leggen. (You managed to fix an appointment in an efficient way.)

D Artifact disagreements

Tables 5 and 6 list every criterion on which the two classifiers disagreed about the artifact, with both labels, the adjudicated label and the adjudicator's reason. Nearly every row is the same dispute: an observable behaviour qualified by ease or effectiveness, read as transcript by one classifier and as judgement by the other.

Table 5: Every criterion where the two blind classifiers disagreed on artifact, with each classifier’s artifact label, the adjudicated final label, and the third classifier’s stated reason. $n = 21$, part 1 of 2.

id	source	label A	label B	final	adjudicator’s reason
dalf_c1_r5	DALF C1	transcript	judgement	judgement	‘Facilement’ (easily) matches the protocol’s judgement example ‘can interact with ease’; the underlying question-answering moves are secondary transcript evidence.
dele_b1_holistica_0	DELE B1	transcript	judgement	judgement	‘Apenas proporciona información... no cumple con el objetivo comunicativo’ is an adequacy/effectiveness judgement per the protocol, consistent with the same pattern in sibling criteria (b006, b021, b046); the described limitations are secondary transcript evidence.
dele_b1_holistica_1	DELE B1	transcript	judgement	judgement	‘No aporta información suficiente para cumplir el objetivo comunicativo’ is an adequacy/effectiveness judgement per the protocol; the lexical adaptation and errors are secondary transcript evidence.
dele_b1_holistica_2	DELE B1	transcript	judgement	transcript	‘Aporta la información requerida’ names a specific required-content checklist rather than a clarity/sufficiency judgement, unlike sibling criteria (b006, b021) that explicitly hinge on clarity or sufficiency.
dele_b2_coherencia_3	DELE B2	transcript	judgement	judgement	‘Con facilidad y destreza... sin esfuerzo aparente... con naturalidad’ directly matches the protocol’s judgement examples ‘can interact with ease’ and ‘naturalness’, and this clause dominates the cohesion-device inventory named earlier.
dele_b2_holistica_0	DELE B2	transcript	judgement	judgement	‘No cumple con el objetivo comunicativo’ is an effectiveness judgement per the protocol; the interlocutor-collaboration detail is secondary transcript evidence.
dele_b2_holistica_1	DELE B2	transcript	judgement	judgement	‘No cumple con el objetivo comunicativo’ over unclear/incomplete ideas maps to the protocol’s ‘effectiveness’ judgement example; the named grammar/lexical errors are secondary transcript evidence.
dele_b2_holistica_2	DELE B2	transcript	judgement	judgement	‘Cumple con el objetivo comunicativo... con claridad’ is an effectiveness judgement per the protocol; grammatical/lexical control is secondary transcript evidence.
dele_b2_holistica_3	DELE B2	transcript	judgement	judgement	‘Con facilidad y destreza... sin esfuerzo aparente... de forma natural’ directly matches the protocol’s ease/naturalness judgement examples and dominates the criterion, consistent with the same pattern in b067; grammatical/lexical control is secondary transcript evidence.
delf_b1_r2	DELFB1	transcript	judgement	judgement	With the criterion reduced to a single short clause, ‘avec une certaine assurance’ is the actual scored distinction and matches the protocol’s ease/manner judgement examples; the exchange itself is secondary transcript evidence.
delf_b1_r6	DELFB1	transcript	judgement	judgement	The scored distinction is the manner of presentation (‘simple and direct’), a style construct with no automatic proxy, not merely whether a topic was raised.
delf_b1_r8	DELFB1	judgement	transcript	judgement	‘Clear enough to be followed without difficulty’ is the protocol’s explicit judgement example ‘listener effort’; the underlying element-linking is secondary transcript evidence.
kov_fra_da_et37_coherence	KOV	transcript	judgement	judgement	This terse cell names no observable feature, only a degree of an unqualified construct (‘reasonably recognisable coherence’), matching the protocol’s explicit carve-out for level-only descriptors with no observable behaviour.
kov_fra_da_et37_register	KOV	judgement	transcript	judgement	‘Gepaste beleefdheidsconventies’ and ‘gepast register’ are near-verbatim instances of the protocol’s own judgement examples ‘appropriateness’ and ‘register fit’.
toefl_topic_1	TOEFL	transcript	judgement	judgement	Relevance and substance of content match the protocol’s judgement example ‘persuasiveness or interest of content’; reliance on repeating the prompt is secondary transcript evidence.

Table 6: Every criterion where the two blind classifiers disagreed on artifact, with each classifier’s artifact label, the adjudicated final label, and the third classifier’s stated reason. $n = 21$, part 2 of 2.

id	source	label A	label B	final	adjudicator’s reason
toefl_topic_2	TOEFL	transcript	judgement	judgement	‘Limited... development’, ‘vaguely expressed’, and ‘connections... unclear’ are evaluative quality judgements with no automatic proxy, spanning task_fulfilment and discourse layers equally; the raw idea count is secondary transcript evidence.
toefl_topic_3	TOEFL	transcript	judgement	judgement	‘Mostly coherent’, ‘somewhat limited’ development and relationships ‘not immediately clear’ are evaluative quality judgements with no automatic proxy, consistent with the same pattern in sibling criteria (b042, b052); raw presence of ideas is secondary transcript evidence.
toefl_topic_4	TOEFL	transcript	judgement	judgement	‘Well developed and coherent’ with idea relationships ‘clear’ names no automatic proxy and dominates over the merely transcript-checkable ‘sustained and sufficient’ length clause.
taalbeleid_pres_an_0	taalbeleid	transcript	judgement	judgement	‘Te veel storende taalfouten’ is a severity/impact judgement (‘too many’, ‘disruptive’) with no automatic proxy, so judgement is right, but the unspecified error type spans morphological, syntactic and lexical evidence equally, so layer should be mixed rather than either label’s specific/arbitrary layer choice.
taalbeleid_telefoon_boodschap_1	taalbeleid	transcript	judgement	judgement	The criterion’s point is the contrast ‘managed, but not efficiently’; ‘efficiently’ is the scored evaluative distinction with no automatic proxy, while the appointment outcome itself is secondary transcript evidence.
taalbeleid_telefoon_boodschap_2	taalbeleid	transcript	judgement	judgement	‘In an efficient way’ is the scored evaluative distinction with no automatic proxy, consistent with the same contrast in the companion criterion (b044); the appointment outcome itself is secondary transcript evidence.

Table 7: Candidate annotation resources for spoken L2 Spanish and French against the properties a measurability-first scheme needs. None has all of them; the closest spoken precedent excludes pronunciation from its tagset.

scheme	language(s)	spoken	CEFR-linked	pronunciation	organised by	reusable here
Cambridge Learner Corpus (Nicholls, 2003)	English	no	loosely	no	level $\times$ operation	design pattern
MERLIN (Boyd et al., 2014)	German, Italian, Czech	no	yes	no	target hypothesis	method
Falko (Reznicek et al., 2013)	German	no	no	no	competing target hypotheses	method
LINDSEI (Gilquin et al., 2010; Gilquin and De Cock, 2013)	English	yes	by task	no	disfluency vs error	nearest precedent
NICT-JLE (Izumi et al., 2005)	English (Japanese L1)	yes	test bands	excluded	level $\times$ operation	closest spoken precedent
CORELE (Campillos Llanos, 2014)	Spanish (A2–B1)	yes	yes	yes, coarser grain	error level $\times$ operation	closest Spanish precedent; reuse unconfirmed data
CEDEL2 (Lozano, 2022)	Spanish	mostly no	no	no	none (raw)	data
FLLOC (Myles and Mitchell, 2026)	French	yes	by level	not found	none (raw)	data
COWS-L2H (Davidson et al., 2020)	Spanish	no	no	no	corrections	data

E Candidate schemes

Table 7 compares the error schemes and resources considered as a basis for the taxonomy on the properties the setting needs: spoken, CEFR-linked, error-tagged, pronunciation-aware, Spanish or French, and organised by evidence source. The table is hand-maintained from each scheme’s documentation and is a reading, not a measurement; the CORELE row rests on the corpus paper’s abstract and the author’s thesis record, since the corpus host could not be reached.

Table 8: Fields of the computable rubric specification (spec/rubric.schema.json, JSON Schema 2020-12). Full field-by-field rationale is in the released spec/README.md.

field	type	meaning
id, title, version, author, licence	string	Rubric identity and provenance; version is semver.
language	array[string]	BCP-47-ish locale tags the rubric applies to, e.g. ["es", "fr"].
task	string	Free-text description of what is being graded.
criteria[].id, name	string	Stable identifier and display name for one scored dimension.
criteria[].layer	enum (9)	Which taxonomy layer the criterion belongs to (segmental . . . task_fulfilment).
criteria[].artifact	enum (3)	What the criterion is recoverable from: audio, transcript, or judgement.
criteria[].kind	enum (3)	count, threshold, or band; selects the shape of rule below.
criteria[].evidence	object	source (named pipeline output), field, unit: where the criterion's input comes from.
criteria[].rule (count)	object	error_types_counted plus an ascending band_table mapping error count to score.
criteria[].rule (threshold)	object	A comparator/value pair, an any_of or all_of list of such pairs, or a piecewise table; abstains if any referenced field is missing.
criteria[].rule (band)	object	levels[] with label, score, descriptor, anchors; rater is human or model.
criteria[].weight	positive number	This criterion's share of the weighted-sum aggregation.
criteria[].abstain	object	condition (free text) and route (teacher, rerun, or zero).
readings.errors_per_100_words	object	Fixed CAF-style secondary reading; enabled flag plus evidence pointers.
readings.cefr_estimate	object	Fixed secondary reading mapping a score basis onto CEFR bands.
constraints[]	object	required_words, required_tenses, topic_adherence, or anti_gaming; separate from the criteria.
aggregation	object	method (weighted_sum), per_criterion_rounding, total_scale, missing_weight_policy.

F Fields of the specification

Table 8 lists the fields a rubric author fills in, one row per field, with nested rule shapes summarised as one row per criterion kind rather than expanded property by property; the released schema carries the full structure, including the conjunctive threshold rule and the optional gravity labels added after the review, and the README the rationale per field. The table was moved here from the body to hold the venue's page budget.

Table 9: Reference-implementation rescoring of the companion grammar and pronunciation papers from their released derived outputs (Section 5). The engine reproduces every recorded score exactly for both; the planted-error gold row is included for orientation and is not part of the validation claim.

comparison	unit	n	agreement	note
Grammar rubric, model score (extract-then-score)	items	24	24/24	exact agreement with the model's recorded score
Grammar rubric, planted-error gold	items	24	23/24	agreement with the planted-error gold band (not an engine-fidelity claim)
Pronunciation flag rule	cells	8	8/8	exact agreement with <code>pa_discrimination.json</code> recall counts
Pronunciation flag rule, per clip	clips	80	80/80	engine flag equals the recorded per-clip flag (10 broken clips per cell, 8 cells)

G Reference-implementation rescoring

Table 9 gives the engine's rescoring of the companion grammar paper's items and the companion pronunciation paper's broken clips from their released derived outputs, per item and per clip. The derivation script, the schema, the two rubrics that express the companion papers' rules and the third, illustrative teacher rubric (a Flemish secondary French speaking task with one criterion of each kind, required words, a required tense and an anchored band) are in the release; the engine's tests exercise validation failures as well as scoring.

H Search log for the standards claim

The absence claim in Section 2 rests on a logged search of the interoperability standards a rubric could live in, run on the dates recorded in the released literature notes, each standard read from its own specification or implementation guide, with a second pass after review that added the 1EdTech standards a reviewer named. QTI 3: the `rubricBlock` element carries rubric text to a named audience and has no scoring semantics; response processing scores items, not criteria (1EdTech Consortium, 2026g). Common Cartridge: the `rubric` element in the assessment object has `material` as its sole child and is display-only guidance (1EdTech Consortium, 2026e). The 1EdTech Rubric Specification of 2005, found by search rather than assumed, declares in its own introduction that it provides no “information model to support automated evaluation software” (1EdTech Consortium, 2005). LTI Advantage Assignment and Grade Services: line items, scores and results carry one numeric grade per activity and no rubric object (1EdTech Consortium, 2026a). xAPI: the official profile registry lists no rubric profile, and `cmi5` is scoped to launch, authorisation, reporting and course structure; the registry was read from its public mirror because the live server refuses automated access. Open Badges 3.0: result descriptions carry named criterion levels with points and a pass threshold, with no field binding evidence to a criterion and no rule beyond a threshold (1EdTech Consortium, 2026b). CASE: rubric criteria carry a numeric weight and criterion levels a numeric score, a computable weighted rule, and the criterion’s item reference points at a curriculum standard, not at evidence (1EdTech Consortium, 2026c). The Comprehensive Learner Record reuses Open Badges’ result fields (1EdTech Consortium, 2026d); OneRoster’s gradebook adds a free-text rubric value to a single-score line item (1EdTech Consortium, 2026f); SCORM’s sequencing rolls up weighted objectives without named criterion levels; Caliper is event telemetry and its own issue tracker records that it cannot represent rubric-based assessment. Canvas: a real JSON rubric schema with criteria, ratings and points, proprietary to one platform (Instructure, 2026). A title search of OpenAlex for Caliper learning analytics returned nothing on topic; Cross-ref queries for rubric-reliability research returned the review by Jonsson and Svingby (2007), and the rubric-scoring work by Kucia et al. (2026) was reused from the companion grammar paper’s pool. The full query strings, hit counts and screening decisions are in the released literature notes.