

Where Taggers Lose the Tense: Imperative, Conditional and Subjunctive Forms in Spanish and French, and What an Exhaustive Conjugation Index Adds

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

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

Abstract

A grader of spoken learner Spanish or French needs, for every verb form in a transcript, the tense and mood label the pupil was taught, and needs to know when a form has none. We measure how well Universal Dependencies taggers (spaCy, Stanza) and an exhaustive reverse conjugation index supply that label on a pre-registered grid of 1005 well-formed forms of twelve verbs per language across fifteen Spanish and thirteen French teacher labels, on the same sentences synthesised and recognised by a commercial speech pipeline, and on 83 rule-generated ill-formed learner forms. Taggers do better than expected: Stanza names the label on 93.0% of Spanish and 94.0% of French forms, and the losses are specific: the Spanish imperative, most of whose forms are homographs of indicative and subjunctive forms and whose remaining forms fail just the same; the conditional and subjunctive in both languages; and, for spaCy's French models, the future. The index's reading set always contains the label but is unique for a minority of forms; its value is coverage and a closed-world check, since it returns nothing on the non-word forms that every tagger assigns a tense to, and a post-hoc combination of the two beats either. All stimuli are synthetic and well-formed, so the tagger figures are upper bounds; corpus, index, logs and code are released.

Contents

1	Introduction	3
2	Related work	4
3	Method	5
4	Results	7
5	Discussion	9
	Limitations	10
A	Teacher inventory and verb set	14
B	Label mapping	15
C	Per-label accuracy and confusion matrices	16
D	Ill-formed forms	20
E	Recognised speech and paired tests	22
F	Lexicon audit (T5)	26
G	Resources	27
H	Search log	28

1 Introduction

A teacher grading spoken Spanish or French asks a question no dependency tagger was built to answer: which tense is this? Not *Mood=Ind*, *Tense=Pres*, but the label the pupil was taught, *condicional*, *passé composé*, *subjonctif présent*, from an inventory of about a dozen labels per language. A grading pipeline that reads a transcript needs that label for every verb form, and needs to know when it does not have one, because a learner’s verb may be a form that does not exist.

The obvious source is the morphology a Universal Dependencies tagger returns, whose per-feature accuracy nobody publishes (Section 2). This paper measures the question systematically: a grid of every well-formed finite form of twelve verbs per language across the teacher’s inventory, under a clean and a transcript-like condition; three taggers and an exhaustive reverse conjugation index; the same sentences synthesised and transcribed by a commercial recogniser; and rule-generated ill-formed learner verbs on which the right answer is no answer.

The result is not that taggers cannot tell the tense: Stanza is right on more than nine forms in ten in both languages, and the losses are specific to the imperative (partly homography, partly a form the models have not learned), the conditional, the subjunctive and, for spaCy’s French models, the future. The index always lists the correct label but usually lists others too; what it adds is coverage, a closed-world check on non-words, and a candidate set that makes a post-hoc combination of the two better than either. Section 3 gives the design, Section 4 the measurements and Section 5 what a grader should build.

2 Related work

Universal Dependencies annotates the French and Spanish future as a tense value, and the conditional and subjunctive as mood values, in both the feature guidelines and the treebank documentation (Universal Dependencies, 2026b,a; Nivre et al., 2020); a tagger that returns present indicative for *je mangerai* or *coma* is wrong by the annotation’s own terms. The training data differ by toolkit: spaCy’s French models train on Sequoia, where the imperative and the conditional are rare, and Stanza’s on GSD, whose morphological features were assigned by a program and not fully hand-verified; Spanish AnCora’s features were annotated by hand and then converted automatically to the UD scheme. spaCy and Stanza publish morphological-feature accuracy as one aggregate score per language (Honnibal et al., 2020; Qi et al., 2020); the CoNLL 2018 and SIGMORPHON 2019 shared tasks and the strongest taggers of that generation likewise score the whole feature bundle as one tag, on the French and Spanish treebanks among others (Straka, 2018; Bohnet et al., 2018; McCarthy et al., 2019), so the accuracy of the one feature a grader needs is not in the published record. Two user reports on spaCy’s smallest French model (the medium and large models are tested here) describe the failure: one counts about a third of future forms, under a tenth of conditional forms and about an eighth of imperatives tagged correctly, the other reports singular imperatives tagged as nouns (spaCy GitHub discussion 13717 (user report), 2024; spaCy GitHub issue 8131 (user report), 2021).

Grammatical-error correction treats verb tense as a standard category and reports per-type scores for English (Bryant et al., 2017, 2019); spoken error detection has been studied for English by the Cambridge group and by FBK (Knill et al., 2019; Bannò and Matassoni, 2022; Bannò et al., 2023; Qian et al., 2024), and the corrected corpus of learner Spanish is written (Yamada et al., 2020). No study found measures tense identification on learner or transcribed Spanish or French: the nearest work classifies learner level from automatically extracted features on the same Spanish corpus (Miaschi et al., 2020), and none measures the drop in tagger morphology on lowercased, unpunctuated text for either language (Appendix H). Exhaustive lexicons and finite-state analysers are the rule-based alternative to a tagger’s guess: Lefff for French (Sagot, 2010), Apertium’s analysers (Forcada et al., 2011), and for Spanish verb generation an evaluation of lexicons against a classifier (Ferrés et al., 2017). The index here is verbecc (Tolbert, 2024), an open conjugator, inverted from generation to recognition.

3 Method

Task and grid. Given a verb form in a transcript, name its tense and mood in the teacher’s inventory: 15 Spanish and 13 French labels at the granularity of a teaching reference grammar, fixed before any system was run (Appendix A). It is broader than any single curriculum stage teaches: the one Flemish plan checked, the second-stage French plan of the Catholic network, requires seven of the thirteen French labels productively and no subjunctive, conditional or anterior form (Katholiek Onderwijs Vlaanderen, 2022) (Appendix A). Claims are numbered T1 to T6 after the pre-registration. 12 verbs per language, three regular, three stem- or spelling-changing and six irregular, are conjugated in every label for three persons with verbecc 2.0.3 (Tolbert, 2024), the periphrases built by rule: 540 Spanish and 465 French forms (*pouvoir* has no imperative). Each form sits in one neutral carrier per language with an overt subject pronoun and a bare object (*Yo hablo la pizza; Je parle le repas*), odd for several verbs by design; imperatives take no pronoun and subjunctives the matrix *es posible que / il est possible que*, the teacher’s own usage. Two conditions: *clean* (capitalised, punctuated) and *transcript-like* (lowercased, punctuation stripped, apostrophes kept). The datasheet records every departure from the pre-registered text, including two corrections after the first run: the Spanish carrier noun *comida* is itself a participle and was replaced by *pizza*, and the *passé récent* forms lacked elision before a vowel, a defect the lexicon audit found; every number below is from the run after both.

Systems and metrics. spaCy 3.8.16 (md and lg for each language) (Honnibal et al., 2020) and Stanza 1.14.0 (Qi et al., 2020), default configurations. A tagger’s features on the gold form’s tokens (Mood, Tense, VerbForm, and the auxiliary’s features for compound and periphrastic forms) are mapped to the teacher’s label by a table fixed before the run (Appendix B); an uncovered combination, or a form the tokeniser splits differently from the gold, counts as wrong. The index enumerates every form of every verb in the conjugator’s lexicon and inverts the table; lookup takes unigrams, bigrams and periphrasis patterns and returns the set of labels the form can carry, which is often larger than one because *hable* or *parle* is spelled the same across persons and moods; a two-step context rule keeps subjunctive readings only under the matrix trigger and drops imperative readings after a subject pronoun, and the first remaining reading is the resolved label. Metrics: label accuracy per system, language, condition and label, with label-cluster bootstrap intervals (a system gets a whole tense right or wrong, so forms are not independent; Wilson intervals stay in the derived data) and McNemar’s exact test for the paired comparisons; for the index, set containment (pre-registered, exact by construction), resolved accuracy (the top-one measure a tagger can be compared with, added once the ambiguity was visible) and the unique-reading share on the form’s own span, since the subjunctive matrix adds a verb form of its own to a whole-sentence lookup.

Ill-formed forms, recognised speech, audit. 35 Spanish and 48 French non-existent forms are built by rule from the same verbs (regularised stems, wrong thematic vowel, dropped accents, infinitive for finite, *-ai/-ais* confusions) plus one real form per verb in the wrong slot for the two documented Dutch-L1 confusions per language; the measure is whether the index returns no reading and whether a tagger assigns a tense. A stratified subsample of the grid, ten forms per label per language with a fixed seed, was synthesised with one Azure neural voice per language and transcribed with Azure speech-to-text, one call each, under a cost cap and a futility rule that

never triggered; the four systems then ran on the recognised text, and index misses are split into forms the recogniser changed and forms it left intact. Sixty grid forms per language were checked against Wiktionary's conjugation tables and ten per language against the Real Academia Española and Le Robert (T5, Appendix F).

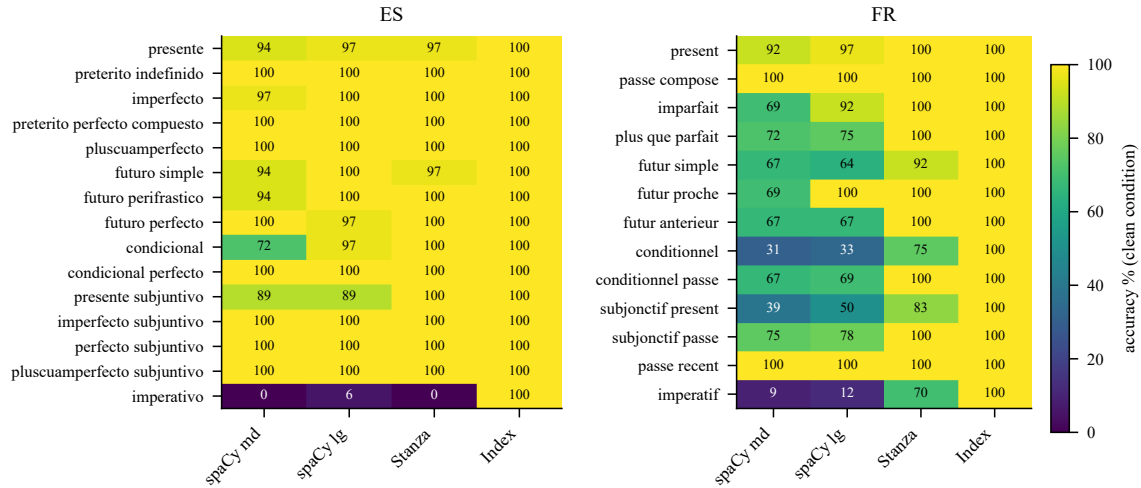


Figure 1: Label accuracy in percent per teacher label (rows) and system (columns) on the clean grid, Spanish left and French right. The Spanish imperative row is near zero for every tagger; the index column is set containment, exact by construction.

4 Results

Where the taggers lose the tense (T1). Figure 1 gives label accuracy per teacher label and system, and Table 1 (Appendix C) every system, language and condition with label-cluster intervals. On the clean grid Stanza names the teacher’s label on 93.0% of Spanish forms (label-cluster 79.4 to 100.0) and 94.0% of French forms (87.9 to 98.7); spaCy reaches 89.4% (md) and 92.4% (lg) in Spanish and 66.2% and 72.5% in French. After Benjamini–Hochberg correction over the 42 paired tests (Appendix E), only one single-system condition comparison survives: lowercasing costs Stanza 94.0 to 87.1 points in French ($q0.000$, 33 forms lost against 1 gained), where it removes the sentence-initial cue, and the two combinations inherit that drop; the rest are within noise. The losses are concentrated. Every tagger scores the Spanish imperative at or near zero (0.0%, 5.6% with Wilson 1.5 to 18.1, 0.0%) for two reasons the raw rows separate: the regular second-singular and the first-plural forms coincide with present indicative and subjunctive forms and come back as those, and the forms with no homograph (*hablad*, *ten*, *haz*, *di*) fail just the same, so the imperative is above all a form the models have not learned; in French, Stanza recovers the plural imperatives and loses the second-singular homograph. For the spaCy models in both languages and for Stanza in French, the conditional and the present subjunctive are the next worst cells (the subjunctive carrier’s matrix clause is a confound the design does not separate), spaCy’s French models at or below half on both and losing a third of the *futur simple*; compound pasts, periphrases and the imperfect are near ceiling, and a lost form comes back as present indicative (Appendix C).

What the index returns (T2). The reading set contains the gold label for every well-formed form (100.0% and 100.0%), as expected by construction. On the form’s own span the set is unique for 40.2% of Spanish and 23.2% of French forms (43.1% and 27.5% once the periphrases, never unique, are set aside), and the resolved label is right on 52.8% and 38.1%, of which 36.5% and 44.6% come from the two context rules the carriers make deterministic. Two post-hoc combina-

tions both beat Stanza alone ($q0.000$ and $q0.000$ for the first). Tagger first (Stanza's label when the index lists it, else the index's resolved label) reaches 98.0% in Spanish and 97.6% in French, no verb carrying it (per-verb minimum 95.6% and 92.3%). Index first (the index's label when one of its rules resolves the form, else Stanza's) reaches 100.0% and 100.0%, exact because both rules are deterministic on these carriers: an upper bound for this design, not a result about running speech.

Ill-formed forms (T4). The index returns no reading for 68.6% of the Spanish and 37.5% of the French ill-formed forms, while the taggers assign a tense to 45.7 to 62.9% of the Spanish and 72.9 to 77.1% of the French. A check against Wiktionary's conjugation tables shows why the languages differ: 62.9% of the Spanish and 81.2% of the French ill-formed forms are real forms of some verb (the *-ail/-ais* confusions yield future and conditional forms; 12 and 12 rows are infinitives, which the index holds without a tense label). On the 13 Spanish and 9 French forms Wiktionary does not know, the index returns nothing for 92.3% and 66.7%; the French exceptions are the three auxiliary swaps (*suis mangé*), which it reads compositionally as compound tenses, so lookup fails closed on non-words except where the error is the auxiliary (Appendix D).

Recognised speech (T3). The recogniser kept the gold form in 98.7% of the Spanish and 86.2% of the French sentences (148 of 150, 112 of 130) (apostrophe-aware match; Appendix E); the French losses are worst in the *subjonctif passé*, where the form survived in 6 of ten: recognition errors, not tense errors. On the recognised text Stanza scores 92.0% and 82.3%, spaCy lg 90.0% and 67.7%; paired against the same forms' transcript-like text, only spaCy lg's French drop survives correction ($q0.006$). The index's set still contains the label on 98.7% and 96.9% of sentences; every one of its 2 Spanish and 4 French misses is a sentence whose form the recogniser changed, and no intact form had a wrong set (0 in both languages). The combinations score 98.0% and 93.1% (tagger first) and 98.7% and 93.8% (index first). The run cost EUR 0.2119.

Resources (T6). A lookup takes 0.04 ms against 5.95 ms for spaCy and 58.09 ms for Stanza; build time and size are in Appendix G.

5 Discussion

The pre-registered expectation was that taggers would be wrong on a large share of forms. They are not: Stanza gives the teacher's label on more than nine forms in ten in both languages, and the failures are a short list a grader can name in advance. Two entries belong partly to the languages: most Spanish imperative forms coincide with indicative or subjunctive forms, so nothing that reads a form without its clause can recover them, and the ones that do not coincide fail because the models have not learned them; the French conditional and subjunctive share endings with the imperfect and the present while the future shares its stem with the conditional, so a tagger trained on treebanks where those forms are rare learns the frequent reading. The index's contribution is not accuracy: its reading set always contains the answer, its resolved label is right on about half the Spanish and under two fifths of the French forms, and what it adds is a closed-world check and a candidate set a tagger's label can be validated against. The post-hoc combinations do that and beat every single system on the grid and on recognised speech; they have no fitted parameter, the index-first ordering is exact on the grid only because its two context rules are deterministic on these carriers, and their value on learner speech is untested.

For a grader, then: look the form up first and fail closed when the lexicon has nothing; take the tagger's label when the lexicon allows it and the lexicon's resolved label when it does not; treat the Spanish imperative and the French conditional and subjunctive as labels that need the clause; and measure the recogniser's verb-form error rate on the target population first, because it rewrote a seventh of the French forms here.

Limitations

Every form in the grid is well-formed and synthetic, and every recognised transcript comes from synthesised speech in one voice per language: the tagger figures are upper bounds on what a tagger would do on learner production, and the recogniser effect is a lower bound, because learner speech carries accent, hesitation and errors that synthesis does not. The carrier sentences are semantically odd for several verbs by design, and the imperative and subjunctive carriers differ in shape from the rest (no subject pronoun; a matrix clause), so part of what the taggers lose on those labels is the carrier, not the form; a tagger trained on running text has never seen *Habla la pizza* as an imperative and has every statistical reason to read *habla* as third-person present. The index and the grid's gold labels come from the same conjugator; the lexicon audit (Appendix F) checks a random sample of forms against independent references, but a systematic defect in a rare paradigm that the sample missed would be shared by both, and the audit itself found one defect in the grid that the sample had not covered. Set-containment accuracy for the index is exact by construction and says nothing about disambiguation; its resolved accuracy depends on a two-token context rule that was written for these carriers and would resolve less in running speech. The teacher inventory is a choice: fifteen Spanish and thirteen French labels at the granularity of a teaching reference grammar, broader than the one curriculum stage checked (Appendix A), which merges some distinctions a grammarian would keep and keeps some a tagger cannot express, so a tagger is measured on a task it was not trained for; the paper's claim is exactly that this is the task a grader needs. Rule-generated from documented error types rather than drawn from learner data, the ill-formed set's second provenance places real forms in wrong slots, which no lexicon can reject. The taggers were run in their default configurations; a tagger fine-tuned on tense-labelled data would be a different system, and so would a larger context window for the index.

References

- Stefano Bannò and Marco Matassoni. 2022. Cross-corpora experiments of automatic proficiency assessment and error detection for spoken english. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 82–91. Association for Computational Linguistics.
- Stefano Bannò, Michela Rais, and Marco Matassoni. 2023. Grammatical error correction for L2 speech using publicly available data. https://www.isca-archive.org/slate_2023/banno23b_slate.html. Accessed September 10, 2026.
- Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a Meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652. Association for Computational Linguistics.
- Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75. Association for Computational Linguistics.
- Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805. Association for Computational Linguistics.
- Daniel Ferrés, Ahmed AbuRa’ed, and Horacio Saggion. 2017. Spanish morphological generation with wide-coverage lexicons and decision trees. <https://repositori.upf.edu/items/25d2a462-7983-4304-8ce4-d7bd34008a03>. Accessed September 10, 2026.
- Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. <https://doi.org/10.5281/zenodo.1212303>. Accessed September 10, 2026.
- Katholiek Onderwijs Vlaanderen. 2022. Leerplan frans b — 2de graad D/A-finaliteit (versie januari 2022). [https://pro.katholiekonderwijs.vlaanderen/download/content/96fba272-f8ec-4ec1-a969-db50a61f39fb/attachments/Leerplan%20II-Fra-da%20\(1ste%20leerjaar%202de%20graad%20vanaf%202022-2023\).pdf](https://pro.katholiekonderwijs.vlaanderen/download/content/96fba272-f8ec-4ec1-a969-db50a61f39fb/attachments/Leerplan%20II-Fra-da%20(1ste%20leerjaar%202de%20graad%20vanaf%202022-2023).pdf). Accessed September 10, 2026.
- K. M. Knill, M. J. F. Gales, P. P. Manakul, and A. P. Caines. 2019. Automatic grammatical error detection of non-native spoken learner english. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8127–8131. IEEE.

- Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and Cross-Lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244. Association for Computational Linguistics.
- Alessio Miaschi, Sam Davidson, Dominique Brunato, Felice Dell’Orletta, Kenji Sagae, Claudia Helena Sanchez-Gutierrez, and Giulia Venturi. 2020. Tracking the evolution of written language competence in L2 spanish learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. <https://aclanthology.org/2020.1rec-1.497/>. Accessed September 10, 2026.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.
- Mengjie Qian, Kate Knill, Stefano Banno, Siyuan Tang, Penny Karanasou, Mark J. F. Gales, and Diane Nicholls. 2024. The speak & improve challenge 2025: Tasks and baseline systems. arXiv:2412.11985. *arXiv preprint*.
- Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. <https://aclanthology.org/L10-1487/>. Accessed September 10, 2026.
- spaCy GitHub discussion 13717 (user report). 2024. French morphologizer mis-tags future, conditional and imperative forms (fr_core_news_sm). <https://github.com/explosion/spaCy/discussions/13717>. Accessed September 10, 2026.
- spaCy GitHub issue 8131 (user report). 2021. Singular present imperatives tagged as NOUN by fr_core_news_sm. <https://github.com/explosion/spaCy/issues/8131>. Accessed September 10, 2026.
- Milan Straka. 2018. Udpipes 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207. Association for Computational Linguistics.
- Brett Tolbert. 2024. verbecc: Verb conjugator for catalan, spanish, french, italian, portuguese and romanian. <https://github.com/bretttolbert/verbecc>. Accessed September 10, 2026.
- Universal Dependencies. 2026a. Ud_French-GSD and UD_Spanish-AnCora treebank documentation. https://universaldependencies.org/treebanks/fr_gsd/index.html. Accessed September 10, 2026.

Universal Dependencies. 2026b. Universal dependencies feature guidelines: Mood and tense. <https://universaldependencies.org/u/feat/Mood.html>. Accessed September 10, 2026.

Aaron Yamada, Sam Davidson, Paloma Fernández-Mira, Agustina Carando, Kenji Sagae, and Claudia Sánchez-Gutiérrez. 2020. Cows-l2h: A corpus of spanish learner writing. *Research in Corpus Linguistics*, 8:17–32.

A Teacher inventory and verb set

The Spanish inventory (15 labels): *presente*, *pretérito indefinido*, *pretérito imperfecto*, *pretérito perfecto compuesto*, *pluscuamperfecto*, *futuro simple*, *futuro perifrástico* (*ir a* + infinitive), *futuro perfecto*, *condicional simple*, *condicional perfecto*, *presente de subjuntivo*, *imperfecto de subjuntivo*, *pretérito perfecto de subjuntivo*, *pluscuamperfecto de subjuntivo*, *imperativo*. The French inventory (13 labels): *présent*, *passé composé*, *imparfait*, *plus-que-parfait*, *futur simple*, *futur proche* (*aller* + infinitive), *futur antérieur*, *conditionnel présent*, *conditionnel passé*, *subjonctif présent*, *subjonctif passé*, *passé récent* (*venir de* + infinitive), *impératif*. The inventory is that of a teaching reference grammar; the one Flemish curriculum checked, the second-stage French plan of the Catholic network, lists productively the *présent*, *passé récent*, *passé composé*, *imparfait*, *futur proche*, *futur simple* and *impératif*, with a fixed conditional-of-politeness formula and no subjunctive or anterior form (Katholiek Onderwijs Vlaanderen, 2022); the third-stage French plan and the Spanish plans were not checked. Where the conjugator lists doublets, the *-ra* imperfect subjunctive and *hubiera* pluperfect subjunctive are used; where it lists several pronouns for one person, the first (*tú*, *él*, *ellos*; *il*, *ils*, masculine participle agreement). The verb set and every carrier sentence are in the released grid file, and the datasheet lists each departure from the pre-registered text with its reason.

B Label mapping

A tagger's output is mapped to the teacher's label by a table fixed before the run and released with the code. A single finite token maps from its Mood and Tense features (indicative present, past, imperfect, future; conditional mood; subjunctive present and imperfect; imperative mood). A two-token span maps from the auxiliary's Mood and Tense and a participle (*haber*, *avoir* or *être* plus a past participle gives the compound tense that the auxiliary's tense selects) or an infinitive (*ir a* and *aller* plus an infinitive give the periphrastic future); a three-token span with *venir de* or *venir d'* plus an infinitive gives the *passé récent*. Any other combination, and any span the tokeniser cuts differently from the gold form, is *unmapped* and counted wrong. The mapping is deliberately literal: a tagger that returns present indicative for a conditional form is scored wrong, because Universal Dependencies annotates the conditional as a mood value (Section 2).

Table 1: Label accuracy on the E3 grid (label-cluster bootstrap intervals). “Index, any”: gold label among the readings; “resolved”: the context rule’s label; “Combined”: Stanza’s label when the index lists it, else the index’s; “index first”: the index’s label when its rule resolves, else Stanza’s.

system	condition	ES %	FR %
spaCy md	clean	89 [76, 98]	66 [52, 79]
spaCy md	transcript	91 [79, 98]	66 [52, 79]
spaCy lg	clean	92 [80, 100]	72 [58, 86]
spaCy lg	transcript	93 [81, 100]	73 [58, 87]
Stanza	clean	93 [79, 100]	94 [88, 99]
Stanza	transcript	92 [78, 100]	87 [79, 95]
Index, any	clean	100 [100, 100]	100 [100, 100]
Index, any	transcript	100 [100, 100]	100 [100, 100]
Index, resolved	clean	53 [27, 79]	38 [18, 60]
Index, resolved	transcript	53 [27, 79]	38 [18, 60]
Combined	clean	98 [94, 100]	98 [94, 100]
Combined	transcript	97 [92, 100]	92 [85, 99]
Combined, index first	clean	100 [100, 100]	100 [100, 100]
Combined, index first	transcript	99 [97, 100]	96 [90, 100]

C Per-label accuracy and confusion matrices

Table 1 gives label accuracy per system, language and condition with label-cluster intervals, and Tables 2 to 7 each tagger’s predicted label against the gold label on the clean condition, with *unmapped* as a column.

Reading down the imperative rows shows both causes: the Spanish forms come back as the present indicative or subjunctive forms they coincide with, or unmapped, for all three persons including the unique second plural; the French *impératif* splits between present indicative, unmapped and, for Stanza, the right label in the plural.

Table 2: Confusion matrix, spaCy md, ES, clean condition: rows are gold label, columns are the label mapped from the tagger’s own morphology via experiments/tense_labelmap.py (“unmapped” when no rule matched).

gold	condicional	condicional_perfecto	futuro_perfecto	futuro_perifrastico	futuro_simple	imperfecto	imperfecto_subjuntivo	perfecto_subjuntivo	pluscuamperfecto	pluscuamperfecto_subjuntivo	presente	presente_subjuntivo	preterito_indefinido	preterito_perfecto_compuesto	unmapped
condicional	26	0	0	0	0	10	0	0	0	0	0	0	0	0	0
condicional_perfecto	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0
futuro_perfecto	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0
futuro_perifrastico	0	0	0	34	0	0	0	0	0	0	0	0	0	0	2
futuro_simple	0	0	0	0	34	0	0	0	0	0	1	0	0	0	1
imperativo	0	0	0	0	2	0	0	0	0	0	21	6	3	0	4
imperfecto	1	0	0	0	0	35	0	0	0	0	0	0	0	0	0
imperfecto_subjuntivo	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0
perfecto_subjuntivo	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0
pluscuamperfecto	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0
pluscuamperfecto_subjuntivo	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0
presente	0	0	0	0	0	0	0	0	0	0	34	0	1	0	1
presente_subjuntivo	0	0	0	0	0	0	0	0	0	0	4	32	0	0	0
preterito_indefinido	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
preterito_perfecto_compuesto	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0

Table 3: Confusion matrix, spaCy md, FR, clean condition: rows are gold label, columns are the label mapped from the tagger’s own morphology via experiments/tense_labelmap.py (“unmapped” when no rule matched).

gold	conditionnel	conditionnel_passe	futur_anterieur	futur_proche	futur_simple	imparfait	imperatif	passe_compose	passe_recent	passe_simple	plus_que_parfait	present	subjonctif_passe	subjonctif_present	unmapped
conditionnel	11	0	0	0	1	13	0	0	0	0	0	11	0	0	0
conditionnel_passe	0	24	0	0	0	0	0	12	0	0	0	0	0	0	0
futur_anterieur	0	0	24	0	0	0	0	12	0	0	0	0	0	0	0
futur_proche	0	0	0	25	0	0	0	0	0	0	0	0	0	0	11
futur_simple	0	0	0	0	24	0	0	0	0	0	0	10	0	2	0
imparfait	2	0	0	0	0	25	0	0	0	0	0	9	0	0	0
imperatif	0	0	0	0	2	1	3	0	0	0	0	13	0	0	14
passe_compose	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0
passe_recent	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0
plus_que_parfait	0	0	0	0	0	0	0	10	0	0	26	0	0	0	0
present	0	0	0	0	1	0	0	0	0	1	0	33	0	1	0
subjonctif_passe	0	0	0	0	0	0	0	9	0	0	0	0	27	0	0
subjonctif_present	0	0	0	0	0	0	0	0	0	0	0	22	0	14	0

Table 4: Confusion matrix, spaCy lg, ES, clean condition: rows are gold label, columns are the label mapped from the tagger’s own morphology via experiments/tense_labelmap.py (“unmapped” when no rule matched).

gold	condicional	condicional_perfecto	futuro_perfecto	futuro_perifrastico	futuro_simple	imperativo	imperfecto	imperfecto_subjuntivo	perfecto_subjuntivo	pluscuamperfecto	pluscuamperfecto_subjuntivo	presente	presente_subjuntivo	preterito_indefinido	preterito_perfecto_compuesto	unmapped
condicional	35	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
condicional_perfecto	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0
futuro_perfecto	0	0	35	0	0	0	0	0	0	0	0	0	0	0	0	1
futuro_perifrastico	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0
futuro_simple	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0
imperativo	0	0	0	0	0	2	0	0	0	0	0	21	3	2	0	8
imperfecto	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0
imperfecto_subjuntivo	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0
perfecto_subjuntivo	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0
pluscuamperfecto	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0
pluscuamperfecto_subjuntivo	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0
presente	0	0	0	0	0	0	0	0	0	0	0	35	0	0	0	1
presente_subjuntivo	0	0	0	0	0	1	0	0	0	0	0	3	32	0	0	0
preterito_indefinido	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
preterito_perfecto_compuesto	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0

Table 5: Confusion matrix, spaCy lg, FR, clean condition: rows are gold label, columns are the label mapped from the tagger’s own morphology via experiments/tense_labelmap.py (“unmapped” when no rule matched).

gold	conditionnel	conditionnel_passe	futur_anterieur	futur_proche	futur_simple	imparfait	imperatif	passe_anterieur	passe_compose	passe_recent	passe_simple	plus_que_parfait	present	subjonctif_passe	subjonctif_present	unmapped
conditionnel	12	0	0	0	0	14	0	0	0	0	1	0	9	0	0	0
conditionnel_passe	0	25	0	0	0	0	0	0	9	0	0	2	0	0	0	0
futur_anterieur	0	0	24	0	0	0	0	0	12	0	0	0	0	0	0	0
futur_proche	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0
futur_simple	0	0	0	0	23	0	0	0	0	0	2	0	11	0	0	0
imparfait	0	0	0	0	0	33	0	0	0	0	0	0	3	0	0	0
imperatif	1	0	0	0	1	1	4	0	0	0	0	0	9	0	0	17
passe_compose	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0
passe_recent	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0
plus_que_parfait	0	0	0	0	0	0	0	0	9	0	0	27	0	0	0	0
present	0	0	0	0	0	1	0	0	0	0	0	0	35	0	0	0
subjonctif_passe	0	0	0	0	0	0	0	1	7	0	0	0	0	28	0	0
subjonctif_present	0	0	0	0	0	0	0	0	0	0	0	0	18	0	18	0

Table 6: Confusion matrix, Stanza, ES, clean condition: rows are gold label, columns are the label mapped from the tagger’s own morphology via experiments/tense_labelmap.py (“unmapped” when no rule matched).

gold	condicional	condicional_perfecto	futuro_perfecto	futuro_perifrastico	futuro_simple	imperfecto	imperfecto_subjuntivo	perfecto_subjuntivo	pluscuamperfecto	pluscuamperfecto_subjuntivo	presente	presente_subjuntivo	preterito_indefinido	preterito_perfecto_compuesto	unmapped
condicional	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0
condicional_perfecto	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0
futuro_perfecto	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0
futuro_perifrastico	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0
futuro_simple	0	0	0	0	35	0	0	0	0	0	1	0	0	0	0
imperativo	0	0	0	0	0	0	0	0	0	0	29	0	0	0	7
imperfecto	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0
imperfecto_subjuntivo	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0
perfecto_subjuntivo	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0
pluscuamperfecto	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0
pluscuamperfecto_subjuntivo	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0
presente	0	0	0	0	0	0	0	0	0	0	35	0	0	0	1
presente_subjuntivo	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0
preterito_indefinido	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0
preterito_perfecto_compuesto	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0

Table 7: Confusion matrix, Stanza, FR, clean condition: rows are gold label, columns are the label mapped from the tagger’s own morphology via experiments/tense_labelmap.py (“unmapped” when no rule matched).

gold	conditionnel	conditionnel_passe	futur_anterieur	futur_proche	futur_simple	imparfait	imperatif	passe_compose	passe_recent	plus_que_parfait	present	subjonctif_passe	subjonctif_present	unmapped
conditionnel	27	0	0	0	0	5	0	0	0	0	4	0	0	0
conditionnel_passe	0	36	0	0	0	0	0	0	0	0	0	0	0	0
futur_anterieur	0	0	36	0	0	0	0	0	0	0	0	0	0	0
futur_proche	0	0	0	36	0	0	0	0	0	0	0	0	0	0
futur_simple	0	0	0	0	33	0	0	0	0	0	3	0	0	0
imparfait	0	0	0	0	0	36	0	0	0	0	0	0	0	0
imperatif	0	0	0	0	0	0	23	0	0	0	6	0	0	4
passe_compose	0	0	0	0	0	0	0	36	0	0	0	0	0	0
passe_recent	0	0	0	0	0	0	0	0	36	0	0	0	0	0
plus_que_parfait	0	0	0	0	0	0	0	0	0	36	0	0	0	0
present	0	0	0	0	0	0	0	0	0	0	36	0	0	0
subjonctif_passe	0	0	0	0	0	0	0	0	0	0	0	36	0	0
subjonctif_present	0	0	0	0	0	0	0	0	0	0	6	0	30	0

D Ill-formed forms

Table 8 gives, per language, condition and system, the index’s no-reading share and each tagger’s assigned-a-tense share on the 83 ill-formed forms, and Table 9 breaks them down by error type with the share of each type that is a real form of some verb, judged two ways: by the index’s own reading and, independently, by whether the form appears in the English Wiktionary conjugation table of any of the twelve verbs (every page fetched and cached, the per-form record released as a raw file). The forms come in two provenances, recorded per row in the released file: rule-generated from the pre-registered error types on all twelve verbs, and one real form placed in the wrong tense slot for six verbs per language (the regular and stem- or spelling-changing ones), three per confusion category taken from the contrastive notes on Dutch-L1 learners (subjunctive avoidance and preterite for imperfect in Spanish; subjunctive avoidance and *avoir* for *être* in French); the categories are documented, the individual forms are constructed, and the auxiliary swap on verbs that take *avoir* is a construction, not an attested error. Real forms in a wrong slot are words, and no lexicon can reject them; the tables are read with that in mind.

Table 8: Ill-formed learner verbs (rule-generated Dutch-L1 conjugation errors): the index’s no-reading share vs. each tagger’s assigned-a-tense share (Wilson 95% intervals). Lookup fails closed on non-words, the auxiliary swaps excepted; tagging fails open.

lang	condition	system	outcome	% [CI]
ES	clean	spaCy md	assigned-a-tense	46 [30, 62]
ES	clean	spaCy lg	assigned-a-tense	40 [26, 56]
ES	clean	Stanza	assigned-a-tense	63 [46, 77]
ES	clean	Index, any	no-reading	69 [52, 81]
ES	transcript	spaCy md	assigned-a-tense	49 [33, 64]
ES	transcript	spaCy lg	assigned-a-tense	43 [28, 59]
ES	transcript	Stanza	assigned-a-tense	63 [46, 77]
ES	transcript	Index, any	no-reading	69 [52, 81]
FR	clean	spaCy md	assigned-a-tense	73 [59, 83]
FR	clean	spaCy lg	assigned-a-tense	73 [59, 83]
FR	clean	Stanza	assigned-a-tense	77 [63, 87]
FR	clean	Index, any	no-reading	38 [25, 52]
FR	transcript	spaCy md	assigned-a-tense	73 [59, 83]
FR	transcript	spaCy lg	assigned-a-tense	73 [59, 83]
FR	transcript	Stanza	assigned-a-tense	75 [61, 85]
FR	transcript	Index, any	no-reading	38 [25, 52]

Table 9: Ill-formed learner-error breakdown by language and error type (Wilson 95% intervals): whether the error form happens to spell a real conjugated form of some verb (“real form (index)”), judged by the index’s own form-span reading; the same check done independently against English Wiktionary conjugation tables (“real form (Wikt.)”, round 2, any cell of the form’s own verb or any of the paper’s other 11 verbs of that language); the index’s no-reading share; and each tagger’s assigned-a-tense share, all on the clean condition. “source” is plan-rule (a deterministic learner-error rule) or notes-category (drawn from a documented Dutch-L1 error category).

lang	error type	source	n	real form (index)	real form (Wikt.)	index none	spaCy md	spaCy lg	Stanza
ES	dropped_accent	plan-rule	6	17 [3, 56]	17 [3, 56]	83 [44, 97]	100 [61, 100]	67 [30, 90]	100 [61, 100]
ES	infinitive_for_finite	plan-rule	12	0 [0, 24]	100 [76, 100]	100 [76, 100]	0 [0, 24]	0 [0, 24]	0 [0, 24]
ES	preterite_imperfect_confusion	notes-category	3	100 [44, 100]	100 [44, 100]	0 [0, 56]	100 [44, 100]	100 [44, 100]	100 [44, 100]
ES	regularised_stem	plan-rule	8	12 [2, 47]	0 [0, 32]	88 [53, 98]	25 [7, 59]	25 [7, 59]	100 [68, 100]
ES	subjunctive_avoidance	notes-category	3	100 [44, 100]	100 [44, 100]	0 [0, 56]	67 [21, 94]	67 [21, 94]	67 [21, 94]
ES	wrong_thematic_vowel	plan-rule	3	100 [44, 100]	100 [44, 100]	0 [0, 56]	100 [44, 100]	100 [44, 100]	100 [44, 100]
FR	ai_ais_confusion	plan-rule	24	100 [86, 100]	100 [86, 100]	0 [0, 14]	100 [86, 100]	100 [86, 100]	100 [86, 100]
FR	auxiliary_choice	notes-category	3	100 [44, 100]	0 [0, 56]	0 [0, 56]	100 [44, 100]	100 [44, 100]	100 [44, 100]
FR	infinitive_for_finite	plan-rule	12	0 [0, 24]	100 [76, 100]	100 [76, 100]	0 [0, 24]	0 [0, 24]	8 [1, 35]
FR	regularised_stem	plan-rule	6	0 [0, 39]	0 [0, 39]	100 [61, 100]	83 [44, 97]	83 [44, 97]	100 [61, 100]
FR	subjunctive_avoidance	notes-category	3	100 [44, 100]	100 [44, 100]	0 [0, 56]	100 [44, 100]	100 [44, 100]	100 [44, 100]

E Recognised speech and paired tests

Table 10 gives every system's label accuracy on the recognised transcripts against its accuracy on the same forms' transcript-like text, Table 11 the verb-form survival rate per language and the split of index misses, and Table 12 McNemar's exact test on the two paired comparisons (clean against transcript-like on every grid form; transcript-like against live on the 280 live forms). The subsample, the recognised text and the recogniser's raw response for every call are in the released log, and the synthesised audio is released with the packet.

Table 10: Grid transcript-like accuracy vs. live accuracy on Azure-recognised speech (E3-ASR, Wilson 95% intervals on the live column), by language and system, and the drop between them in percentage points (positive = live is worse). “Index, any reading” and “Index, resolved” are the same two index metrics as Table 1, evaluated on the recognised text.

lang	system	transcript %	live % [CI]	drop (pp)
ES	spaCy md	91	87 [81, 92]	+3.2
ES	spaCy lg	93	90 [84, 94]	+2.8
ES	Stanza	92	92 [87, 95]	-0.3
ES	Index, any	100	99 [95, 100]	+1.3
ES	Index, resolved	53	52 [44, 60]	+0.8
ES	Combined	97	98 [94, 99]	-1.1
ES	Combined, index first	99	99 [95, 100]	+0.2
FR	spaCy md	66	62 [54, 70]	+3.9
FR	spaCy lg	73	68 [59, 75]	+5.6
FR	Stanza	87	82 [75, 88]	+4.8
FR	Index, any	100	97 [92, 99]	+3.1
FR	Index, resolved	38	35 [28, 44]	+2.7
FR	Combined	92	93 [87, 96]	-0.8
FR	Combined, index first	96	94 [88, 97]	+2.1

Table 11: Form-intact rate (Wilson 95% intervals): whether the gold verb form survives Azure speech recognition intact, and the index-miss split for the set-containment (“Index, any”) metric – misses where the form was changed by the recogniser (expected) vs. misses where the form was intact but the index’s reading was still wrong (expected zero). Corrected (round 2): the survival check now tokenises after apostrophes as well as whitespace, so French elision (“j’aurais” vs. the bare gold form “aurais”) no longer counts a verbatim-recognised sentence as a miss; the originally logged, uncorrected rate is in the derived data (form_intact_rate_logged).

lang	metric	% [CI]	count
ES	gold form intact	99 [95, 100]	–
ES	index miss, form changed	–	2
ES	index miss, form intact	–	0
FR	gold form intact	86 [79, 91]	–
FR	index miss, form changed	–	4
FR	index miss, form intact	–	0

Table 12: McNemar’s exact test on per-form correctness: clean vs. transcript-like (paired by form, within the E3 grid), grid transcript-like vs. live (paired by language, verb, label, person and form, on the 280 live forms, E3-ASR), and the combined system vs. Stanza alone (paired by form, on the clean grid and on the live rows). b/c are the two discordant-pair counts. q is the Benjamini–Hochberg FDR-adjusted p -value across the full family of 42 McNemar tests reported in this paper (both comparison types together, $\alpha = 0.05$).

system	lang	comparison	b	c	p	q
spaCy md	ES	clean vs. transcript	0	6	0.031	0.077
spaCy md	FR	clean vs. transcript	3	3	1.000	1.000
spaCy lg	ES	clean vs. transcript	2	4	0.688	1.000
spaCy lg	FR	clean vs. transcript	4	8	0.388	0.678
Stanza	ES	clean vs. transcript	8	1	0.039	0.086
Stanza	FR	clean vs. transcript	33	1	<0.001	<0.001
Index, any	ES	clean vs. transcript	0	0	1.000	1.000
Index, any	FR	clean vs. transcript	0	0	1.000	1.000
Index, resolved	ES	clean vs. transcript	0	0	1.000	1.000
Index, resolved	FR	clean vs. transcript	0	0	1.000	1.000
Combined	ES	clean vs. transcript	6	0	0.031	0.077
Combined	FR	clean vs. transcript	25	0	<0.001	<0.001
Combined, index first	ES	clean vs. transcript	6	0	0.031	0.077
Combined, index first	FR	clean vs. transcript	19	0	<0.001	<0.001
spaCy md	ES	grid vs. live	5	0	0.062	0.131
spaCy md	FR	grid vs. live	10	2	0.039	0.086
spaCy lg	ES	grid vs. live	4	1	0.375	0.678
spaCy lg	FR	grid vs. live	10	0	0.002	0.006
Stanza	ES	grid vs. live	2	3	1.000	1.000
Stanza	FR	grid vs. live	14	6	0.115	0.231
Index, any	ES	grid vs. live	2	0	0.500	0.808
Index, any	FR	grid vs. live	4	0	0.125	0.239
Index, resolved	ES	grid vs. live	2	0	0.500	0.808
Index, resolved	FR	grid vs. live	4	3	1.000	1.000
Index, resolved (form span)	ES	grid vs. live	0	0	1.000	1.000
Index, resolved (form span)	FR	grid vs. live	0	0	1.000	1.000
Combined	ES	grid vs. live	2	2	1.000	1.000
Combined	FR	grid vs. live	6	5	1.000	1.000
Combined, index first	ES	grid vs. live	2	2	1.000	1.000
Combined, index first	FR	grid vs. live	7	4	0.549	0.854
Combined vs. Stanza	ES	clean grid	27	0	<0.001	<0.001
Combined vs. Stanza	FR	clean grid	17	0	<0.001	<0.001
Combined vs. Stanza	ES	live	9	0	0.004	0.012
Combined vs. Stanza	FR	live	14	0	<0.001	<0.001
Index first vs. Combined	ES	clean grid	11	0	<0.001	0.004
Index first vs. Combined	FR	clean grid	11	0	<0.001	0.004
Index first vs. Combined	ES	live	1	0	1.000	1.000
Index first vs. Combined	FR	live	1	0	1.000	1.000
Index first vs. Stanza	ES	clean grid	38	0	<0.001	<0.001
Index first vs. Stanza	FR	clean grid	28	0	<0.001	<0.001
Index first vs. Stanza	ES	live	10	0	0.002	0.006
Index first vs. Stanza	FR	live	15	0	<0.001	<0.001

F Lexicon audit (T5)

Sixty grid forms per language, drawn with the pre-registered seed, were checked by script against the English Wiktionary conjugation tables of each verb (compound forms composed from the auxiliary's table and the participle), and ten per language were checked by hand against the Real Academia Española's conjugation tables and Le Robert's conjugation pages. Every form matched. The pre-registration named the RAE tables and the Bescherelle pages as the primary references; the RAE's conjugation route is a script-rendered page that a page reader cannot fetch and Bescherelle has no citable static page, so the reference was changed before any result was read, and the audit file records the source URL and the reference form for each check. The audit found the passé récent elision defect in the first grid out of sample.

Table 13: Resources: index build, size, and lookup latency (nearest-rank over the E3 grid, both languages pooled) vs. tagger latency. No GPU, no model download at run time (spaCy/Stanza models are pinned and pre-downloaded; see environment/).

metric	value
index build time (s)	1046.9
index size on disk, gzip (MB)	10.9
index memory, Python peak (MB)	1069.8
index lookup p50 / p95 (ms)	0.043 / 1.480
spaCy md tagging p50 / p95 (ms)	5.95 / 7.02
spaCy lg tagging p50 / p95 (ms)	5.58 / 6.63
Stanza tagging p50 / p95 (ms)	58.09 / 72.21
verbecc / spaCy / Stanza versions	2.0.3 / 3.8.16 / 1.14.0

G Resources

Table 13 gives the index build time, its size on disk and in memory, and lookup and tagging latency. The memory figure is Python-object accounting at load, not resident set size; the index is loaded once per process. The verbecc build is not deterministic across machines (its conjugation-class predictor is a learned model), so the exact index every number in the paper used is released as two compressed files with their hashes in the manifest.

H Search log

The absence claim in Section 2 rests on the searches in the released literature notes, run on the dates recorded there: title-and-abstract searches of OpenAlex for morphological tagging accuracy on Universal Dependencies treebanks and for the Stanza toolkit; web searches for the Universal Dependencies annotation of the French and Spanish conditional and future, for spaCy’s French morphologizer accuracy per feature, for the CoNLL 2018 shared task’s per-language feature scores, for reverse-conjugation libraries, for learner tense-error detection in French and Spanish, and for morphological lexicons and analysers for both languages; direct reads of the Universal Dependencies feature pages and treebank documentation, the two spaCy GitHub threads, Stanza’s performance page, the CoNLL 2018 results page and the Ferrés et al. repository page; a Crossref title query for the Spanish morphological generation paper; and, after review, a second pass through the SIGMORPHON 2019 and CoNLL-SIGMORPHON 2018 shared-task overview and system papers (which score the whole feature bundle as one tag on the French and Spanish treebanks), UDPipe 2’s composite metrics, fourteen OpenAlex queries for learner-corpus tense or mood detection in French or Spanish (FLLOC, CEDEL2, COWS-L2H), which returned level classification and agreement-error detection but no tense identification result, and direct reads of the FLLOC and CEDEL2 project sites and publication lists (manual acquisition studies throughout, no automatic tense identification) and of the spoken French treebank’s documentation (no morphological features). Nothing found reports tense- or mood-specific tagger accuracy for either language, and nothing measures tense identification on learner or transcribed input.