Beyond Acoustics: Isolating Dialectal and Sociolinguistic Bias in Spanish ASR

Johnatan E. Bonilla
Humboldt-Universität zu Berlin
j.bonilla@hu-berlin.de

Abstract

Large-scale ASR systems such as Whisper achieve competitive aggregate Word Error Rate (WER) on multilingual benchmarks, but this aggregate conceals systematic disparities across speaker populations. We evaluate Whisper large-v3 on 276 recordings from the *Corpus Oral y Sonoro del Español Rural* (COSER), a dialectological archive of elderly rural speakers across all Spanish provinces. WER is computed separately for Informants and Interviewers within each recording, revealing that mixed-role evaluation underestimates Informant WER in the majority of provinces, with the largest corrections in southern areas. Negative Binomial regression with cluster-robust standard errors shows that Andalusia and Extremadura generate significantly more Informant errors than the Castilian heartland (Andalusia IRR = 1.20, $p < 0.001$; Extremadura IRR = 1.24, $p = 0.020$), while no geographic predictor reaches significance for Interviewers sharing the same recording environment. Male Informants generate 12.5% more errors than females after geographic adjustment ($p < 0.001$), consistent with differential vernacular retention in traditional rural communities. The geographic pattern aligns with established dialectological classifications of Peninsular Spanish. These results demonstrate that role-disaggregated evaluation is a necessary methodological prerequisite for fairness audits of ASR systems applied to sociolinguistically diverse corpora: aggregate benchmarks systematically suppress disparities that are borne disproportionately by the most underrepresented speaker populations, and their use in isolation constitutes both an allocative harm and a measurement failure.

1 Introduction

The adoption of large-scale ASR systems such as Whisper (Radford et al., 2023) for transcription of spoken-language archives has grown rapidly, yet aggregate performance figures carry an implicit assumption of demographic neutrality that does not survive empirical scrutiny. Word Error Rate (WER)—the proportion of reference words incorrectly transcribed, computed as the minimum edit distance between hypothesis and reference divided by total reference words—is the dominant evaluation metric for ASR, but its aggregate form masks population-level disparities. Systematic disparities have been documented across racial groups in English (Koenecke et al., 2020), across gender and dialect in automatic captioning (Tatman, 2017), across stigmatised regional varieties of British English (Markl, 2022), and at the intersection of non-standard phonology and gender (Harris et al., 2024). For elderly speakers, Vipperla et al. (2010) found WER increases of approximately 10 percentage points relative to younger adults. For Spanish, and for non-English languages more broadly, the systematic study of ASR bias remains sparse (Kantharuban et al., 2023).

The most relevant antecedent is San Martín et al. (2024), who evaluated Whisper and SeamlessM4T on the *Corpus Oral y Sonoro del Español Rural* (COSER; Fernández-Ordóñez 2005), over 1,700 hours of sociolinguistic interviews with elderly rural residents across all Spanish provinces. Their principal finding—Whisper large-v3 achieves a mean WER of 0.292 largely stable across dialect regions—leads them to conclude that the model is a viable transcription aid for rural corpus work. We identify two design limitations. First, WER is computed without separating the rural *Informant* from the *Interviewer*; without role-segregated evaluation, any disparity attributable to rurality is absorbed into a global average. Second, no multivariate analysis simultaneously controls for acoustic quality, geography, speaker age, and sex.

Our study provides a role-disaggregated evaluation of a large-scale ASR model on COSER. By computing WER separately for Informants and In-

Proceedings of the 7th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 123–132

terviewers within the same recordings, we isolate performance differences associated with rural vernacular speech while holding recording conditions approximately constant. We then analyse these differences using a multivariate Negative Binomial regression that controls jointly for audio quality, geography, speaker sex, and age, providing the first geographically disaggregated account of ASR performance disparity in European rural Spanish.[1]

2 Background

2.1 The COSER

The *Corpus Oral y Sonoro del Español Rural* (COSER; Fernández-Ordóñez, 2005; Fernández-Ordóñez and Pato, 2020) comprises 1,947 hours of semi-directed sociolinguistic interviews covering 1,325 rural localities across all 52 Spanish provinces, with 2,574 registered Informants (mean age: 73; 52.4% female). Participants are elderly residents of low formal education who were born and have lived continuously in small rural communities—the sociolinguistic profile plausibly absent from the urban, broadcast, and web-sourced data that dominate ASR training corpora. Informants account for approximately 81.8% of total speaking time (SD = 7.56; San Martín et al., 2024). The Interviewer, generally younger and educated, speaks a variety close to the spoken standard that grounds Whisper's language model.

Transcriptions follow a semi-conventional orthographic norm (Fernández-Ordóñez and Pato, 2020) that encodes surface-level phonological reduction and morphological dialectal variants in surface-faithful orthography, but normalises the most salient phonetic features of southern varieties to standard orthography. This norm produces two competing and partially opposing effects on WER measurement.

Mechanism 1: WER inflation from encoded dialectal forms. Segment deletions and morphological variants are transcribed as produced: forms such as *comprao* (standard: *comprado*), *pa* (standard: *para*), *na* (standard: *nada*), *tá* (standard: *está*), and *to* (standard: *todo*) appear as reference tokens. Morphological variants—*marcharsen*, *traíba*, *tuviendo*—are preserved verbatim. Whisper, whose language model is grounded in standard written Spanish, systematically restores these forms to their standard equivalents: it hypothesises *comprado* where the reference reads *comprao*, *para* where the reference reads *pa*. Each such normalisation generates a substitution or deletion error in the WER computation, despite the fact that Whisper may have correctly identified the acoustic signal. The density of these reduced forms is substantially higher in southern and rural speech—where syllable-final consonant deletion and intervocalic /-d-/ deletion are most advanced—than in the Castilian heartland, producing a systematic WER gradient that is partly orthographic in origin rather than purely acoustic.

Mechanism 2: artificial WER suppression from normalised forms. Conversely, the most salient phonological features of southern varieties—*seseo* (merger of /s/ and the interdental fricative /θ/), *ceceo*, *yeísmo* (loss of the palatal lateral /ʎ/), and glotalisation of coda consonants—are explicitly *not* transcribed, being normalised to standard orthography. The reference always reads *caza* and *casa* with the same sibilant, and *pollo* regardless of whether the speaker produces a palatal lateral or a palatal fricative. When Whisper's output also defaults to standard orthography—whether because it correctly perceived the acoustic signal or because its language model overrides a non-standard input—the two transcriptions agree and no error is registered, masking what may be a genuine recognition failure at the acoustic level. The net effect is that WER *underestimates* Whisper's actual difficulty with southern phonology on normalised features while *overestimating* it on the features that COSER does encode. The WER disparities reported in § 4 therefore represent a conservative lower bound on the true performance gap for southern varieties.

2.2 Whisper's Training Distribution

Whisper's Spanish training data derives from 680,000 hours of weakly supervised web audio (Radford et al., 2023), filtered by language identification but not by speaker demographics. Although the exact composition is not disclosed, indirect evidence suggests systematic under-representation of non-standard varieties: Conneau et al. (2022) showed that multilingual ASR models trained on web-crawled data consistently underperform on low-resource language variants relative to high-resource standard registers; and Pratap et al. (2024) documented that even Massively Multilingual Speech models exhibit performance gaps on re-

[1] Scripts, per-speaker WER tables, and model code: https://github.com/johnatanebonilla/socio-asr-bias/.

Statistic	Value
Recordings analysed	276
Provinces covered	50
Total segments (after filtering)	1,321
Informant segments	530
Interviewer segments	791
Informant segments w/ sex metadata	505 (90.0%)
Informant segments w/ age metadata	447 (79.7%)
Informant mean age (COSER)	73 years
Informant % female (COSER)	52.4%
Recording-level mean WER	0.302

Table 1: Summary of the analysed dataset.

gional varieties absent from their training distributions. These findings concern different architectures (FLEURS, MMS) rather than Whisper directly, but the shared mechanism—web-crawled data over-representing standard registers—makes the inference plausible for Whisper as well. For Spanish specifically, the bulk of online audio plausibly consists of broadcast media, podcasts, and video content produced in urban, educated registers approximating the written norm—precisely the variety closest to the Interviewer's speech. The COSER Informants occupy the opposite pole of this distribution: elderly, rural, low-education speakers whose phonological and morphosyntactic surface forms diverge maximally from the written standard (§ 5.2). The performance gap between Informants and Interviewers, measured within the same recording environment, operationalises the distance between Whisper's training distribution and the target speech.

3 Methodology

3.1 Data and ASR Model

We use 276 COSER recordings with audio and verified transcriptions, spanning 50 provinces.[2] We evaluate Whisper large-v3 (Radford et al., 2023); recording-level mean WER is 0.302, closely replicating San Martín et al.'s result (0.292); the 1 pp difference is attributable to our larger sample (276 vs. 226 recordings), which includes more recently released files from peripheral and island provinces.

3.2 Speaker Segmentation and WER Computation

The COSER XML release includes speaker-turn timestamps, which would in principle allow direct acoustic segmentation by role: each timestamped interval could be extracted, transcribed independently by Whisper, and attributed to the corresponding speaker tag. To assess whether this approach was viable, we evaluated timestamp reliability by comparing, for each XML segment, the reference text falling within the declared boundaries against the word-level timestamps produced by Whisper's own decoder when processing the full recording—a measure of whether the XML boundary corresponds to the acoustic content Whisper actually finds there. Levenshtein similarity between the two text sequences showed substantial and systematic inconsistency: even after text normalisation, only 5,673 of 26,379 segments (21.5%) achieved a similarity score above 0.9, while over 5,000 segments fell below 0.5, indicating that the declared boundaries frequently do not correspond to the actual acoustic content of the recording at those positions. Because timestamp-based segmentation would therefore introduce uncontrolled boundary errors into the attribution procedure, we discarded the XML timestamps for acoustic segmentation entirely. Role attribution is instead achieved through the transcription-level tag system described immediately below, with the full WER computation procedure detailed in § 3.3.

COSER transcriptions encode speaker role through a structured tag system at the segment level. Tags of the form `I`n (`I1`, `I2`, ...) identify Informants and map directly to sociodemographic metadata entries (sex, age, birth year); tags `E`n identify Interviewers—university-trained fieldworkers for whom no demographics are recorded. Tags `IE`n and `II`n mark simultaneous speech involving at least one Informant and one Interviewer, or two Informants, respectively. All overlap segments are excluded from both WER computation and demographic attribution, since overlapping speech cannot be unambiguously attributed to a single speaker's acoustic footprint. Although this strict filtering reduces the analysed volume, it ensures that the WER measured for each role reflects exclusively that speaker's uninterrupted output.

Table 1 summarises the corpus as analysed. After filtering, 530 Informant and 791 Interviewer segments are retained across 276 recordings. The asymmetry—fewer Informant segments despite Informants contributing 81.8% of speaking time—reflects the interview structure: Informants produce long, uninterrupted narrative turns while Interviewers contribute many short question segments. Sex metadata is available for 505 segments (90.0%)

[2] Downloaded from `corpusrural.es`, 2025.

and age for 447 (79.7%); coverage is lower for higher-numbered Informants (I3–I5), who joined interviews opportunistically, as noted in the table caption.

3.3 WER Computation and Error Attribution

WER is computed with `jiwer` following San Martín et al. (2024): bracketed annotations are removed, text is lowercased, and punctuation stripped. Per-speaker attribution traverses the `jiwer.process_words` alignment: substitutions and deletions are attributed to the speaker of the aligned reference word; insertions to the nearest preceding reference word. Concretely, Whisper transcribes the complete audio file as a single linear text. The full reference word sequence is constructed by concatenating normalised text from all non-overlapping segments in order, maintaining a parallel array of per-word role labels. The `jiwer` alignment between Whisper's output and the concatenated reference is then traversed word by word, and each error is assigned to the role label of the corresponding reference position. This approach requires that Whisper's output preserves the temporal order of speech, which is satisfied by its autoregressive decoding.

To illustrate, consider a reference sequence of four words with roles I1: *fue*, I1: *pa*, E1: *para*, I1: *comprarlo*, against which Whisper hypothesises *fue para comprarlo*. The alignment produces: a match on *fue*, a deletion on *pa* (charged to I1), a match on *para* (E1, no error), and a match on *comprarlo*. The single deletion—Whisper's normalisation of the dialectally reduced form *pa* to its standard equivalent—is attributed exclusively to the Informant counter, not pooled into a global figure.

All values reported are *micro-averages* (total errors / total words) unless otherwise noted; this ensures that short segments with high WER do not inflate descriptive statistics relative to the predominant Informant contributions. Because COSER metadata are not uniformly available for all speakers, sociodemographic analyses are conducted on the subset of segments with available annotations; differences in sample size across models therefore reflect metadata coverage rather than sampling decisions.

3.4 Audio Quality

Three complementary objective metrics characterise acoustic conditions per recording, independently of Whisper's output. **SNR** (Signal-to-Noise Ratio, dB) measures the decibel difference between speech power and background noise power; it is a low-level signal measure that does not capture perceptual characteristics such as reverberation or loudness adequacy. **UTMOS** (Saeki et al., 2022) is a neural non-intrusive Mean Opinion Score predictor trained on naturalness judgements from human listeners (VoiceMOS Challenge 2022); it produces a scalar quality estimate on a 1–5 scale without requiring a clean reference signal, making it applicable to field recordings. **NISQA-MOS** (Mittag et al., 2021) is a multi-dimensional perceptual quality model that decomposes overall MOS into four sub-scores: Noisiness (NOI), Coloration (COL), Discontinuity (DIS), and Loudness (LOUD), each on a 1–5 scale.

The choice of quality covariate for multivariate modelling is determined empirically by Pearson and Spearman correlations between each metric and recording-level WER ($N = 276$). Contrary to intuition, raw SNR does not significantly predict WER ($r = -0.081$, $p = 0.156$), while UTMOS ($r = -0.235$, $p < 0.001$) and NISQA-Loudness ($r = -0.233$, $p < 0.001$) do. Noisiness, notably, does not predict WER ($p = 0.456$). This pattern indicates that Whisper is not sensitive to background noise per se but to signal level and perceptual naturalness—a distinction with direct implications for the SNR×dialect interaction reported in § 4.4. SNR is retained in the multivariate model given its wider dynamic range (SD = 10.27 dB) and interpretability. As a robustness check, we re-estimated all models replacing SNR with UTMOS; all geographic coefficients retained sign, magnitude, and significance, confirming that the choice of quality covariate does not drive the reported effects.[3]

3.5 Multivariate Modelling

We model raw error counts using **Negative Binomial GLMs** (NB2, log link), treating the number of transcription errors per speaker segment as the outcome and including $\log(N_{\text{words}})$ as an offset to account for differences in segment length. This formulation is equivalent to modelling error rate on the log scale while respecting the count nature of the data.

Informant predictors entered simultaneously are:

[3] Andalusia Informant IRR shifts from 1.201 (SNR model) to 1.198 (UTMOS model); Sex IRR from 1.125 to 1.122. Full UTMOS models available in the repository.

Autonomous Community (16 dummies, Castile and León as reference), centred SNR (continuous), sex (binary), and age cohort (categorical: 50–70 ref., 71–85, 86+, residual). Cluster-robust standard errors (sandwich estimator) grouped by recording (261 clusters) account for within-session correlation among segments from the same file, functionally equivalent to random intercepts per recording without distributional assumptions on the random effects (Abadie et al., 2022).

Separate but structurally identical models are estimated for Informants and Interviewers, both with cluster-robust standard errors. The parallelism of these two models is the central inferential strategy: a geographic coefficient that is positive and significant for Informants but absent for Interviewers—who occupy the same physical recording environment—is consistent with linguistic variety as the primary source rather than recording conditions. Coefficients are reported as incidence rate ratios (IRR $= e^{\hat{\beta}}$) with 95% confidence intervals and both standard and cluster-robust p-values.

4 Results

Global micro-averaged WER is 0.309 for Informants and 0.294 for Interviewers—a 1.5 pp gap that appears modest in aggregate but conceals pronounced geographic and sociodemographic structure, as the following subsections demonstrate.

4.1 Province-Level Geographic Distribution

Figure 1 presents Informant and Interviewer WER across all 50 provinces. Table 2 reports the 10 highest and 5 lowest Informant WER provinces.

The nine highest-WER provinces with positive Informant–Interviewer gaps are all southern (Andalusian: Almería, Sevilla, Cádiz, Málaga, Córdoba), western peripheral (Galician: Orense, Lugo), or Extremaduran (Cáceres). The exception is Soria, a northern province with small sample size. Albacete shows equally high Interviewer WER, suggesting shared acoustic difficulty rather than a linguistic effect. Orense shows an exceptionally high Interviewer WER (0.384), nearly matching its Informant WER, which may reflect shared recording-quality issues in the Galician sessions rather than a purely linguistic effect. The five lowest-WER provinces—all northern—show inverted gaps where Interviewers produce *higher* WER than Informants, consistent with rural varieties close to the Castilian standard.

Province	Inf.	Int.	Gap	N
Almería	.418	.295	+.123	4
Orense	.415	.384	+.032	5
Sevilla	.409	.315	+.095	4
Cádiz	.407	.278	+.129	5
Málaga	.401	.295	+.106	4
Lugo	.382	.255	+.127	4
Albacete	.381	.392	−.011	4
Cáceres	.380	.359	+.022	5
Córdoba	.372	.290	+.082	5
Soria	.371	.331	+.040	5
Álava	.236	.276	−.040	6
Gerona	.231	.233	−.002	6
Segovia	.227	.279	−.052	6
Valladolid	.223	.300	−.077	7
Vizcaya	.201	.282	−.082	7

Table 2: Provinces with the 10 highest and 5 lowest Informant micro-WER. Gap = Inf. − Int. Of the 50 provinces, 14 show gap $> +0.05$ (all southern or western); 8 show gap < -0.05 (all northern). Provincial estimates with $N = 4$ should be interpreted with caution given the small number of recordings.

4.2 Autonomous Community Aggregation

Table 3 and Figure 2 aggregate the provincial data to the 17 Autonomous Communities (*Comunidades Autónomas*), which serve as geographic predictors in the multivariate model.

Interviewer WER does not track the Informant gradient. Andalusia, the community with the largest sample ($N = 38$), shows Informant WER of 0.370 against Interviewer WER of 0.298—a gap of +0.072. Murcia (0.367 vs. 0.292, gap +0.075) and Extremadura (0.355 vs. 0.318, gap +0.037) follow the same pattern. Galicia shows a similar pattern (0.340 vs. 0.303, gap +0.038). Cantabria shows a large gap (+0.072) but with only $N = 6$ recordings. Castile and León (0.285), Navarre (0.258), and Basque Country (0.255) show inverted or near-zero gaps. In Castile and León the Interviewer WER (0.292) actually exceeds the Informant WER (0.285). The Canary Islands represent an interesting case: Informant WER (0.311) is elevated relative to the best WER peninsular communities but the gap (+0.064) is moderate, a pattern we discuss in § 5.2. Madrid ($N = 4$) and La Rioja ($N = 5$) show extreme inversions driven by short recording samples and should not be interpreted inferentially; note that Murcia ($N = 5$) similarly has a small sample, and its high IRR in the regression model should be interpreted with corresponding caution. Because both speakers share the same

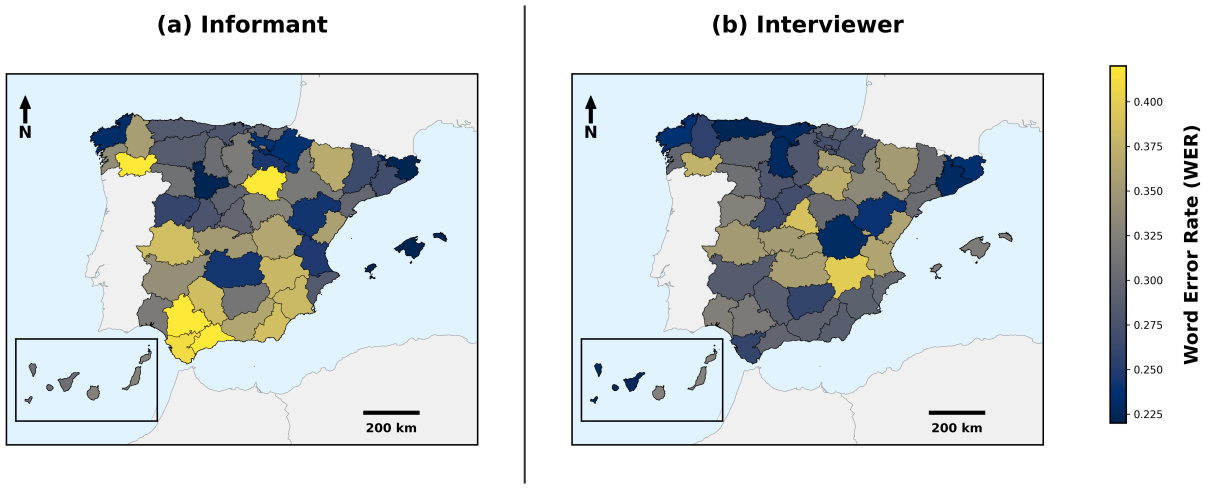

Figure 1: Province-level Informant (left) and Interviewer (right) micro-WER.

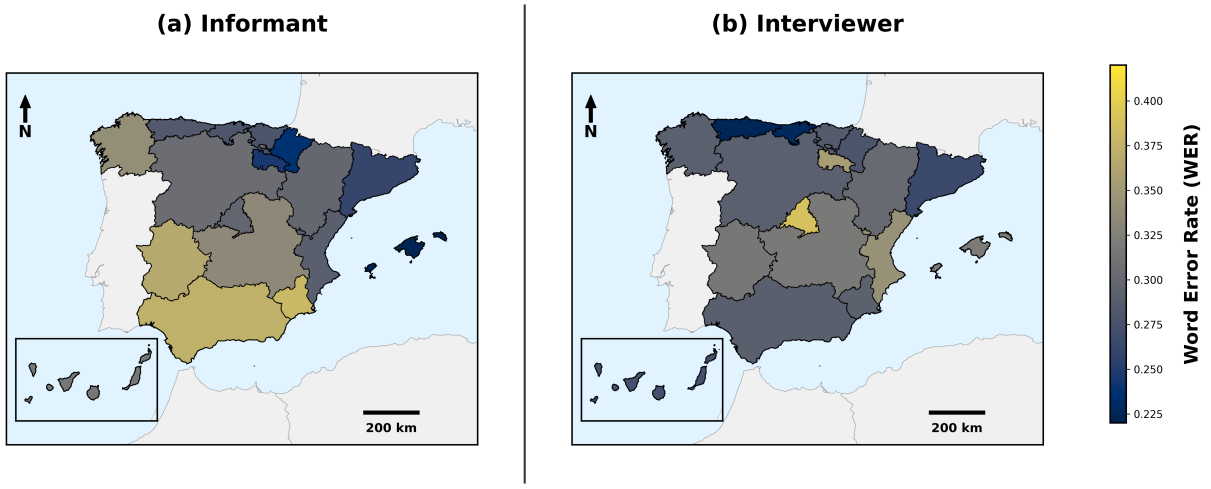

Figure 2: Autonomous Community-level Informant (left) and Interviewer (right) micro-WER.

recording environment within each recording, the systematic dissociation between the Informant gradient and the flat Interviewer pattern is consistent with linguistic variety as the primary source of the disparity.

4.3 Sociodemographic Effects

Among 505 Informant segments with sex metadata, males show consistently higher WER than females. The male micro-average is 0.320 versus 0.299 for females—a 2.1 percentage point gap. The difference is statistically significant (Mann-Whitney $U = 36{,}266$, $p = 0.003$, rank-biserial $r = 0.12$). The rank-biserial indicates a small effect size; the amplification to IRR = 1.125 (12.5%) in the multivariate model reflects the redistribution of variance after geographic adjustment.

Among the 447 Informant segments with age metadata, 422 correspond to speakers aged 50 or above. Three cohorts were defined: 50–70 ($n = 139$), 71–85 ($n = 237$), and 86+ ($n = 46$). No significant age difference emerges (Kruskal-Wallis $H = 1.37$, $p = 0.503$; means 0.310, 0.311, 0.303). The male–female gap is stable across all three cohorts at approximately +0.02, indicating that sex and age are orthogonal predictors in this population.

4.4 Multivariate Analysis

Table 4 reports the Negative Binomial GLM with both standard and cluster-robust p-values.

The Informant model reveals two robustly significant communities. Andalusia generates 20.1% more errors than Castile and León (IRR $= 1.201$, $p_{\text{clust}} < 0.001$), and Extremadura generates 24.0% more errors (IRR $= 1.240$, $p_{\text{clust}} = 0.020$). Galicia (IRR $= 1.199$, $p_{\text{clust}} = 0.117$) and Murcia (IRR $= 1.294$, $p_{\text{clust}} = 0.123$) show consistent positive

Aut. Community	Inf.	Int.	Gap	N
Andalusia	**.370**	.298	+.072	38
Murcia	.367	.292	+.075	5
Extremadura	**.355**	.318	+.037	10
Galicia	.340	.303	+.038	20
Castile-La Mancha	.314	.319	−.005	27
Canary Islands	.311	.246	+.064	11
Aragon	.307	.298	+.009	18
Cantabria	.303	.231	+.072	6
Castile & León	.285	.292	−.007	55
Valencian C.	.284	.307	−.022	16
Madrid	.282	.404	−.122	4
Asturias	.273	.232	+.041	7
Catalonia	.272	.272	.000	22
Navarre	.258	.290	−.032	7
Basque Country	.255	.281	−.026	20
La Rioja	.253	.373	−.120	5
Balearic Isl.	.251	.304	−.054	5
Global	*.309*	*.294*	*+.015*	*276*

Table 3: Micro-WER by Autonomous Community.

effects of similar magnitude that do not survive the more conservative cluster correction, likely reflecting the limited number of recording clusters in those communities ($N = 20$ and $N = 5$ respectively).

No geographic predictor is positive and significant in the Interviewer model: the largest Interviewer IRR is Extremadura (1.156), which does not reach significance ($p = 0.133$). The two significant Interviewer coefficients are both *negative*—Cantabria (0.739) and Asturias (0.805)—indicating that Interviewers in those communities generate *fewer* errors than the Castile and León reference.

Male Informants generate 12.5% more errors than females (IRR $= 1.125$, $p_{\text{clust}} < 0.001$), a robust effect that survives all geographic specifications. Neither age cohort reaches significance (71–85: IRR $= 0.971$, $p = 0.488$; 86+: IRR $= 0.961$, $p = 0.573$), confirming the bivariate null result. SNR is significant for Informants ($p_{\text{clust}} = 0.011$) but not for Interviewers ($p = 0.718$), indicating that audio quality affects recognition of vernacular speech more than standard speech.

The pseudo-R^2 values (Informant: 0.048, Interviewer: 0.013) indicate that geography, audio quality, sex, and age together explain a modest share of total WER variance; unmeasured variables such as speech rate, lexical density, and individual articulatory characteristics likely account for a substantial portion of the remaining variance. Nevertheless, the 3.6× ratio between models confirms that the measured predictors structure Informant performance far more than Interviewer performance—precisely the asymmetry predicted by a linguistic account of the disparity.

	IRR	95% CI	p_{std}	p_{clust}
Informant model (N=530 segments, 261 clusters)				
Andalusia	1.201	[1.07, 1.35]	.002	.0004***
Extremadura	1.240	[1.04, 1.48]	.019	.020*
Galicia	1.199	[1.05, 1.37]	.009	.117
Murcia	1.294	[1.01, 1.66]	.041	.123
Canary Isl.	1.087	[0.90, 1.31]	.373	.414
SNR (/dB)	0.995	[0.99, 1.00]	.005	.011*
Sex (male)	1.125	[1.05, 1.20]	.001	.0005***
Age 71–85	0.971	[0.90, 1.05]	.468	.488
Age 86+	0.961	[0.84, 1.09]	.541	.573
Interviewer model (N=791 segments, 261 clusters)				
Andalusia	1.022	[0.92, 1.13]	.677	.702
Extremadura	1.156	[0.96, 1.40]	.133	.158
Galicia	1.080	[0.95, 1.23]	.250	.289
Murcia	1.001	[0.81, 1.23]	.995	.996
SNR (/dB)	1.001	[1.00, 1.00]	.718	.734

α: Inf. 0.133, Int. 0.149. Pseudo R^2: Inf. 0.048, Int. 0.013 (3.6×). Ref.: Castile and León (geog.), 50–70 (age).
11 remaining CCAA non-significant ($p_{\text{clust}} > 0.05$) in both. Interviewer significant: Cantabria 0.739 (p=.009), Asturias 0.805 (p=.037)—both negative.
***$p < 0.001$; *$p < 0.05$. Cluster-robust SEs in both models: sandwich estimator grouped by recording (261 clusters).

Table 4: Negative Binomial GLM results for the Informant (top) and Interviewer (bottom) models.

5 Discussion

5.1 The Geographic Disparity is Linguistic

The within-recording Informant/Interviewer contrast is the core empirical contribution of this study. Two speakers sharing the same recording environment produce divergent WER values that track geography systematically. This within-recording contrast substantially reduces the plausibility of acoustic quality as the primary explanation for the geographic gradient, although we note that microphone distance and angle may vary between speakers within a session (§ 6).

The finding extends San Martín et al. (2024), who documented geographic variation in overall WER but could not isolate its source because their evaluation conflated Informant and Interviewer speech. Our role-disaggregated analysis reveals that the geographic gradient is entirely concentrated in the Informant channel: the Interviewer channel is geographically flat. The pseudo-R^2 ratio (3.6× for Informants vs. Interviewers) quantifies this asymmetry.

This pattern converges with findings across typologically distinct contexts. Koenecke et al. (2020) showed roughly double the error rate for African American English relative to white American English across five commercial ASR systems, attributing the gap to training-data composition. Markl (2022) extended this analysis to stigmatised British English varieties, arguing that performance gaps constitute both allocative and representational harms: speakers of non-standard varieties receive worse service from ASR and are implicitly positioned as deviations from a norm. Harris et al. (2024) showed that the interaction of gender and dialect is the primary driver of ASR error in non-standard American English. Our contribution extends this framework to a Romance language context where the relevant axis is not race but the rural–urban, vernacular–standard continuum structuring Peninsular Spanish dialectology.

	Group	Inf.	Int.	IRR
García Mouton (phonetic)				
Inf.	Northern	.290	.290	ref.
	Southern	.340	.301	1.131***
Int.	Northern	—	.290	ref.
	Southern	—	.301	$1.026^{\text{n.s.}}$
Fernández-Ordóñez (morphosyntactic)				
Inf.	North-Central	.296	.299	ref.
	Western	.330	.299	1.169**
	Southern	.346	.289	1.108*
	Eastern	.277	.290	$0.974^{\text{n.s.}}$
Int.	North-Central	—	.299	ref.
	Western	—	.299	$1.071^{\text{n.s.}}$
	Southern	—	.289	$0.981^{\text{n.s.}}$
	Eastern	—	.290	$1.027^{\text{n.s.}}$

Table 5: Auxiliary NB GLMs with dialectological classifications. Micro-WER (Inf./Int.) and Informant/Interviewer IRRs. All Informant models include SNR, sex, and age. ***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; n.s. not significant.

5.2 Alignment with Peninsular Spanish Dialectology

The communities and provinces with the highest Informant WER correspond to what established dialectological frameworks identify as the varieties most distant from the Castilian standard. We briefly describe the two classifications used and report auxiliary Negative Binomial models that replace the CCAA dummies with these classifications (Table 5).

García Mouton (1994) organises Peninsular Spanish along a phonetic axis, distinguishing a conservative *Northern* area characterised by maintenance of the *distinción* (contrast between alveolar /s/ and interdental fricative) and strong articulation of coda consonants, from an innovative *Southern* area—encompassing Andalusia, Canary Islands, Murcia, Extremadura, and the Valencian and Albacete transition zones—where three convergent phonological changes produce surface forms maximally distant from orthographic norms: yeísmo (loss of the palatal lateral), deletion of intervocalic /-d-/, and the progressive assimilation, neutralisation, and loss of coronal consonants in syllable coda. As García Mouton notes, the epicentre of these changes is western Andalusia, from which they radiate in successive stages northward and to the Canary Islands.

Fernández-Ordóñez (2016) departs from phonetic criteria by grounding the classification in grammatical evidence from the ALPI and the COSER itself. Her division identifies a *Western* area (Cantabria, both Castiles, Asturias, and Galician-Portuguese contact provinces) defined by the mass/count pronominal system (*neutro de materia*), leísmo, laísmo, and loísmo; an *Eastern* area characterised by inflected infinitives (*-sen*) and subjunctive-to-conditional displacement; and a *Southern* area where the etymological pronominal system predominates. The Western area—shaped by contact with Galician-Portuguese—cuts across the phonological north–south divide, capturing a morphosyntactic dimension that purely phonetic classifications miss.

Under García Mouton, Southern Informants generate 13.1% more errors than Northern Informants (IRR $= 1.131$, $p < 0.001$); the Interviewer contrast is 2.6% and non-significant ($p = 0.435$). Under Fernández-Ordóñez, both the Southern (IRR $= 1.108$, $p = 0.029$) and the Western area (IRR $= 1.169$, $p = 0.002$) show significant Informant effects.

A notable exception is the Canary Islands (Informant WER $= 0.311$, IRR $= 1.087$, $p_{\text{clust}} = 0.414$), classified as Southern yet showing no significant disparity. Canarian Spanish occupies a well-documented position as an interdialect between Peninsular and Latin American varieties (García Mouton, 1994): it shares with Caribbean Spanish generalised seseo and /-s/ aspiration—features plausibly well represented in Whisper's web-sourced training data given the large volume of Latin American audio online. This hypothesis—

that Canarian speech enjoys proximity to Whisper's training distribution that mainland Southern varieties do not share—is consistent with the observed pattern but remains speculative without access to the training composition.

5.3 The Sex Effect

The male–female disparity is robust across all specifications (Mann-Whitney $p = 0.003$; NB $p_{\text{clust}} < 0.001$ in the CCAA model; $p = 0.001$ under both García Mouton and Fernández-Ordóñez). Male Informants generate 12–13% more errors after geographic, acoustic, and age adjustment. The direction reverses the English pattern (Tatman, 2017; Feng et al., 2024)—a difference explained by the operative axis of variation. In English audits, the disparity is attributed to over-representation of male broadcast speech; in rural Spanish, the operative axis is variety proximity to the standard. In traditional rural communities, women adopt prestige phonological features at higher rates while men retain local vernacular phonology (Labov, 2001), producing speech further from Whisper's standard-oriented language model. This is consistent with Harris et al. (2024), who show that the gender–dialect interaction is the primary driver of ASR error in non-standard American English. We acknowledge that this interpretation is post-hoc: the same WER difference could partially reflect sex-linked differences in speech rate or articulatory precision, and direct evidence of the mediating mechanism (differential adoption of prestige features) would require a phonological error analysis that the current design does not provide.

6 Conclusions

This study demonstrates that aggregate WER conceals systematic sociolinguistic disparities in Whisper's performance on rural Spanish. Role-segregated evaluation reveals that mixed-role benchmarks underestimate the Informant WER in the majority of provinces, with the largest corrections in southern communities where dialectal divergence from the standard is greatest. Negative Binomial regression with cluster-robust standard errors identifies Andalusia and Extremadura as robustly generating 20–24% more Informant errors than the Castilian heartland, while no geographic predictor reaches significance for Interviewers sharing the same recording environment. Male Informants generate 12.5% more errors than females—a pattern consistent with differential vernacular retention and opposite to English audits. The disparity aligns with established dialectological classifications: both García Mouton's phonetic axis and Fernández-Ordóñez's morphosyntactic framework predict the geographic gradient, with the latter revealing a Western (Galician-Portuguese contact) effect invisible to administrative boundaries.

Limitations and Future Work

The within-recording design assumes approximately shared acoustic conditions for Informant and Interviewer. In practice, microphone distance and angle may vary: the Informant is typically seated facing the recorder while the Interviewer may move, consult notes, or sit at a different distance. This potential asymmetry cannot be measured from the audio alone and represents a residual confound.

Speaker attribution relies on transcription-level segmentation rather than time-aligned diarisation. In segments with very high WER—where alignment between Whisper's output and the reference degrades—error attribution to roles becomes less precise. This limitation affects Southern Informants disproportionately, since they exhibit the highest WER. Time-aligned diarisation of the COSER audio would strengthen the attribution procedure and enable fine-grained acoustic analyses at the speaker level.

All evaluations use Whisper large-v3; generalisation to wav2vec 2.0, MMS, or fine-tuned models requires further work. Galicia and Murcia show consistent positive effects that do not survive cluster correction (both $p \approx 0.12$); for Murcia ($N = 5$ clusters), the asymptotic properties of the sandwich estimator may not hold, and the cluster-robust p-value should be interpreted cautiously (Abadie et al., 2022).

Acknowledgments

This research was carried out within the Collaborative Research Centre SFB/CRC 1412 *Register: Language Users' Knowledge of Situational-Functional Variation*, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number 416591334.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2022. When should you ad-

just standard errors for clustering?*. *The Quarterly Journal of Economics*, 138(1):1–35.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. 2024. Towards inclusive automatic speech recognition. *Computer Speech and Language*, 84:101567.

Inés Fernández-Ordóñez. 2005. COSER: Corpus oral y sonoro del español rural. Universidad Autónoma de Madrid.

Inés Fernández-Ordóñez. 2016. Dialectos del español peninsular. In Javier Gutiérrez Rexach, editor, *Enciclopedia lingüística hispánica*, volume 2, pages 387–404. Routledge, London and New York.

Inés Fernández-Ordóñez and Enrique Pato. 2020. El COSER (Corpus Oral y Sonoro del Español Rural) y su contribución al estudio de la variación gramatical del español. In Ángel J. Gallego and Francesc Roca, editors, *Dialectología digital del español*, number 80 in Verba. Anexo, pages 71–100. Universidade de Santiago de Compostela, Santiago de Compostela.

Pilar García Mouton. 1994. *Lenguas y dialectos de España*. Arco Libros, Madrid.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15166–15184, Miami, Florida, USA. Association for Computational Linguistics.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

William Labov. 2001. *Principles of Linguistic Change, Volume 2: Social Factors*. Blackwell, Oxford.

Nina Markl. 2022. Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 521–534, Seoul, Republic of Korea. ACM.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep CNN-Self-Attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2021)*, pages 2127–2131, Brno, Czechia.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pages 4521–4525.

Mirari San Martín, Jónathan Heras, Gadea Mata, and Sara Gómez. 2024. Is ASR the right tool for the construction of spoken corpus linguistics in European Spanish? *Procesamiento del Lenguaje Natural*, 73:165–176.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Ravichander Vipperla, Steve Renals, and Joe Frankel. 2010. Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP J. Audio Speech Music. Process.*, 2010.