ESPANLI: un benchmark adversarial de NLI para el español rioplatense mediante competencias humano–modelo

Mauro Santelli[2,3] Joaquín Toranzo Calderón[1,2,3]
Ramiro Caso[2,3] Bruno Muntaabski[2]

[1] Universidad Tecnológica Nacional, FRBA, Grupo de IA y Robótica
[2] Universidad de Buenos Aires, FFyL
[3] Instituto de Investigaciones Filosóficas (SADAF–CONICET)

mesantelli@uba.ar jtoranzocalderon@frba.utn.edu.ar

ASAID — 55 JAIIO (2026)

La tarea de NLI

Dadas una premisa P y una hipótesis H, decidir si H se infiere de P (Dagan y Glickman, 2004).
Clasificación canónica: *implicación* (entailment), *contradicción*, *neutralidad*.

Ejemplo

P: Recién vencido el plazo, el equipo entregó el informe.

- **I** *El informe se entregó tarde.* — implicada
- **C** *El informe se entregó dentro del plazo.* — contradictoria
- **N** *El director rechazó el informe.* — neutral

Tarea asimétrica sobre unidades léxico-sintácticas: captura patrones probabilísticos de inferencia humana que la lógica formal no alcanza.

Motivación: una paradoja

- Los LLMs superan el 90 % en los benchmarks estándar de NLI (SNLI, MultiNLI).
- Pero fallan sistemáticamente ante instancias adversariales (Nie et al., 2020).
- Explotan atajos superficiales que no requieren comprensión inferencial:
  - superposición léxica (*lexical overlap*) (McCoy et al., 2019)
  - sesgo de hipótesis aislada: la etiqueta se predice sin ver P (Poliak et al., 2018)
  - inserción de negaciones $\Rightarrow$ contradicción
- ANLI y HANS respondieron con estrategias adversariales, pero siguen concentradas en el inglés.

El problema del castellano

- Estrategia dominante: traducir benchmarks del inglés, p. ej. XNLI (Conneau et al., 2018); excepción: InferES (Kovatchev y Taulé, 2022).
- La traducción distorsiona superposición léxica, polisemia y condiciones de verdad.
- Además, transfiere un sesgo pragmático anglocéntrico: el modelo falla por conocimiento cultural, no por competencia inferencial.
- La variante rioplatense (léxico, modos verbales locales) queda fuera.
- *Somos 600M* subraya la necesidad de recursos nativos (Grandury, 2024); la cobertura adversarial sigue siendo escasa.

ESPANLI: benchmark nativo y adversarial

Doble carencia a cubrir:

1. fragilidad adversarial de los benchmarks saturados.
2. sub-representación del castellano, en particular del rioplatense.

Propuesta

Benchmark de NLI construido desde cero para el español, con estrategia adversarial humano-en-el-bucle: corpus nativo y culturalmente anclado, en lugar de traducir benchmarks del inglés.

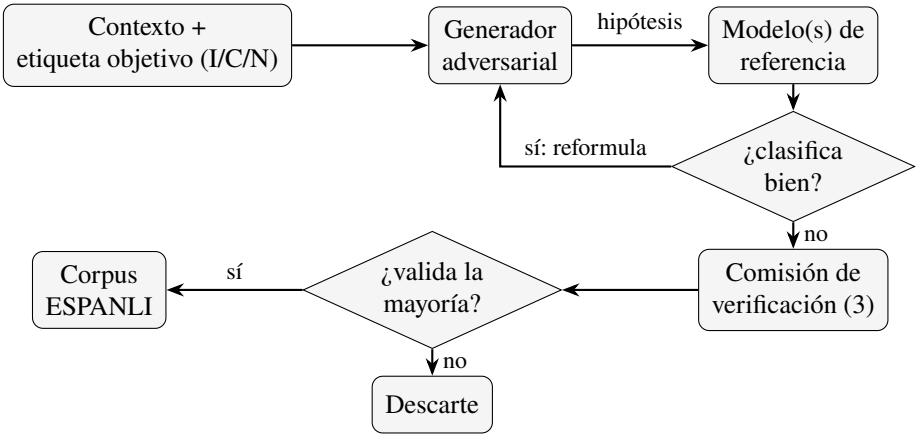
Principal diferencia metodológica respecto de los conjuntos existentes.

Corpus: tesis doctorales de repositorios públicos

- Fuente: tesis doctorales en repositorios públicos de universidades argentinas.
- Cobertura inicial: Filosofía y Humanidades; luego Matemática, Ingeniería y Derecho.
- Estructura: tuplas *(contexto, hipótesis, etiqueta)*; hipótesis adversarial.
- Selección de contextos: polisemia, marcas argumentales, complejidad textual.
- Fenómenos rioplatenses presentes aun en registro formal.

Estimación preliminar pesimista (FFyL–UBA): ∼1500 tesis (∼1200 digitales) → ∼1M de párrafos; rondas de curación de 5000 párrafos.

Metodología: competencias adversariales humano–modelo



Los anotadores actúan como adversarios del modelo (cf. ANLI: Nie et al., 2020).

Componentes del protocolo

Generadores	seleccionados por formación académica; gamificación en competencias regulares; rechazo justificado tras agotar las iteraciones permitidas.
Modelos de referencia	estado del arte al momento de cada competencia.
Verificadores	tres personas formadas en análisis de argumentos.
Consignas	específicas por etiqueta (I, C, N); su redacción es un parámetro experimental.

Iteraciones periódicas del benchmark para anular la saturación entre ediciones (sin objetivo de entrenamiento, a diferencia de ANLI).

Contribuciones previstas y trabajo futuro

1. Subcorpus académico adversarial derivado de tesis doctorales.
2. Plataforma web interactiva de generación abierta de instancias.
3. Protocolo de gamificación con estudiantes universitarios, supervisado por especialistas.
4. Sub-tareas de NLI de nivel experto nativas en castellano (filosófica, formal, jurídica).

Las pruebas piloto ajustarán: iteraciones por instancia, modelos de referencia, consignas por etiqueta, selección de contextos.

Evaluación de modelos sobre el corpus e incorporación de fuentes no académicas: publicaciones separadas.

Conclusión

- ESPANLI apunta a una doble brecha: fragilidad adversarial y sub-representación del castellano rioplatense y experto.
- Combina curado nativo de corpus y competencias humano–modelo con verificación.
- El objetivo no es sólo medir desempeño, sino caracterizar *qué* resuelven los modelos cuando aciertan, *qué* "habilidades" poseen y *cómo* fallan cuando no.

Gracias — ¿preguntas?