



Bajo el dominio de la ley

Desidentificación y extracción de eventos en la Base de
Jurisprudencia Nacional del Uruguay

Diego Garat Baridón

Doctorado en Informática

PEDECIBA

Facultad de Ingeniería, Universidad de la República

Montevideo – Uruguay

Junio de 2026



Bajo el dominio de la ley

Desidentificación y extracción de eventos en la Base de
Jurisprudencia Nacional del Uruguay

Diego Garat Baridón

Tesis de Doctorado presentada al PEDECIBA y a la Facultad de Ingeniería de la Universidad de la República, como parte de los requisitos necesarios para la obtención del título de Doctor en Informática.

Directores:

Dr. Guillermo Moncecchi

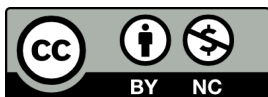
Dra. Dina Wonsever

Director académico:

Dr. Guillermo Moncecchi

Montevideo – Uruguay

Junio de 2026



© 2026 Diego Garat Baridón

Esta obra se encuentra bajo una Licencia Creative Commons
Reconocimiento - No comercial 4.0 ([BY-NC](#)).

Garat Baridón, Diego

Bajo el dominio de la ley / Diego Garat Baridón. –
Montevideo : Universidad de la República. Facultad de
Ingeniería, 2026.

XXXV, 210 p.: il.

Directores: Guillermo Moncecchi y Dina Wonsever

Director académico: Guillermo Moncecchi

Tesis de Doctorado en Informática (PEDECIBA) –
Universidad de la República. Facultad de Ingeniería.

1. PROCESAMIENTO DEL LENGUAJE NATURAL.
2. INFORMÁTICA JURÍDICA. 3. ANONIMIZACIÓN
DE DATOS. 4. SEMÁNTICA DE MARCOS.
I. Moncecchi, Guillermo. II. Wonsever, Dina.
III. Universidad de la República. Facultad de Ingeniería.
IV. Título.

INTEGRANTES DEL TRIBUNAL DE DEFENSA DE TESIS

Dra. Libertad Tansini (presidenta)

Dra. Nuria Bel (revisora)

Dr. Alfonso Ureña López (revisor)

Dr. Gustavo Betarte

Mag. Gabriel Valentín

Montevideo – Uruguay

Junio de 2026

K: Imagine a tribunal working like a laboratory. Do you think that one day it'll be like that?

Scientist: You don't really think we're going to start telling fortunes, do you? Let's be serious.

K: You're not going to tell me that your machine couldn't do it?

Scientist: She wouldn't want to.

THE TRIAL (escena borrada)

ORSON WELLES

Agradecimientos

Tengo la convicción de que ningún logro personal sería posible sin el apoyo y la colaboración de otros. La presente tesis no es la excepción.

En primer lugar, quisiera agradecer a mis dos supervisores. A Dina, que desde el primer día me acompañó con sus sugerencias y su paciencia; este trabajo nunca habría comenzado sin ella. A Guillermo, quien tuvo la generosidad de aceptar ser el director académico con el proyecto ya comenzado, aportando su experiencia y conocimiento para enriquecerlo.

En sus distintas etapas, esta investigación contó con el respaldo de varias instituciones, a las que también quiero expresar mi agradecimiento: a Pyxis, a la ANII y al Ministerio de Industria, Energía y Minería por su financiamiento a través del Fondo Sectorial de Investigación a partir de Datos, y al Programa de Desarrollo de las Ciencias Básicas (PEDECIBA) y la Comisión Sectorial de Investigación Científica (CSIC) de la Universidad de la República.

Agradezco también al Poder Judicial por autorizar el acceso a la Base de Jurisprudencia Nacional, que constituye el corpus de esta tesis; en particular, a la Dra. María Luisa Tosi y a la División Tecnología por su colaboración en el marco de este trabajo.

Finalmente, quisiera agradecer a los integrantes del tribunal y a los revisores de esta tesis por el tiempo dedicado a su evaluación y por las observaciones y sugerencias realizadas, que contribuyeron a mejorar su calidad y claridad.

A quienes estuvieron. A quienes siempre estarán.

Resumen

El procesamiento de lenguaje natural (PLN) en el dominio jurídico es un área de investigación clave, debido a que gran parte de la información se encuentra en leyes, sentencias y reglamentos, con desafíos derivados de su complejidad lingüística, semántica y discursiva.

El objetivo de esta investigación es mejorar la calidad y accesibilidad de la información en la Base de Jurisprudencia Nacional (BJN), una colección de más de 100 000 sentencias del Poder Judicial de Uruguay.

Como primer objetivo, se aborda la desidentificación automática de sentencias. Esta se centra en nombres propios y se basa en un modelo de reconocimiento de entidades nombradas combinado con agrupamiento aglomerativo para la resolución de correferencias, entrenado y evaluado sobre el corpus B*N, creado a partir de sentencias de la BJN con este fin. Los resultados se integran en un prototipo funcional para desidentificación asistida por humanos.

Como segundo objetivo, se estudia la extracción de eventos relevantes. Esta se basa en la Teoría de Marcos Semánticos de Fillmore y en el aprendizaje en contexto con grandes modelos de lenguaje. Con ejemplos de FrameNet en inglés se obtienen resultados competitivos respecto de enfoques previos, mientras que la combinación con el corpus ViolenciaR en español, derivado de sentencias de la BJN y anotado manualmente, permite observar el impacto del uso de recursos multilingües y de la complejidad del dominio jurídico sobre el rendimiento del sistema.

Este trabajo introduce recursos lingüísticos inéditos y metodologías innovadoras, constituyendo una contribución significativa al desarrollo del PLN en el dominio jurídico, particularmente en la BJN.

Palabras clave: Base de Jurisprudencia Nacional de Uruguay, desidentificación, extracción de eventos, análisis de marcos semánticos, aprendizaje en contexto, FrameNet

Abstract

Natural Language Processing (NLP) in the legal domain is a key research area, as much of the information is contained in laws, court rulings, and regulations, which pose challenges due to their linguistic, semantic, and discursive complexity.

The aim of this research is to improve the quality and accessibility of information in the National Jurisprudence Database (BJN), a collection of over 100 000 court rulings from the Judiciary of Uruguay.

As a first objective, the automatic de-identification of court rulings is addressed. This focuses on proper names and is based on a named entity recognition model combined with agglomerative clustering for coreference resolution, trained and evaluated on the B*N corpus, created from rulings in the BJN for this purpose. The results are integrated into a functional prototype for human-assisted de-identification.

As a second objective, the extraction of relevant events is studied. This is grounded in Fillmore’s Frame Semantic Theory and in-context learning with large language models. Using FrameNet examples in English yields results competitive with prior approaches, while combining them with the manually annotated ViolenciaR corpus in Spanish, derived from BJN rulings, allows observation of the impact of multilingual resources and legal-domain complexity on system performance.

This work introduces new linguistic resources and innovative methodologies, constituting a significant contribution to the development of NLP in the legal domain, particularly with respect to the BJN.

Keywords: National Jurisprudence Database of Uruguay; automatic de-identification; event extraction; frame semantic parsing; in-context learning; FrameNet

Lista de figuras

- 1.1 Fragmento extraído de la Sentencia 15/2022, del Tribunal de Apelaciones Penal de 1º Tº de Uruguay. Se trata de un enunciado complejo cuya oración principal se prolonga hasta el punto final, incorporando citas directas de testigos. En negrita se señalan las siglas que sustituyen nombres; en naranja y verde se resaltan los testimonios del tío y tía de la víctima, respectivamente. 3
- 1.2 Fragmento extraído de la Sentencia 15/2022, del Tribunal de Apelaciones Penal de 1º Tº de Uruguay. Este fragmento presenta un estilo técnico-formal, con oraciones complejas correctamente puntuadas y bien estructuradas. 4
- 1.3 Resultado del reconocimiento de entidades por tres herramientas sobre un texto extraído la Base de Jurisprudencia Nacional de Uruguay. En celeste se señalan los segmentos indetificados como personas, en verde, aquellos identificados como organizaciones, mientras que en violeta, segmentos etiquetados como otros. 6
- 1.4 Ejemplo de una sentencia en el portal de la BJN. La información publicada incluye: (A) información general de la sentencia: número de identificación, fecha, tipo de resolución, etc.; (B) tema; (C) tribunal; (D) clasificación; (E) resumen; y (F) texto de la sentencia. 8
- 2.1 Fragmentos extraídos de la Sentencia 15/2022, del Tribunal de Apelaciones Penal de 1º Tº de Uruguay. El nombre del acusado se encuentra oculto (resaltado en rojo), mientras que los nombres de jueces, fiscales, abogados (en verde) y referentes (en celeste) permanecen incambiados. 17

2.2	Captura de la herramienta BRAT utilizada para la corrección manual de las etiquetas en el corpus B*N. Las etiquetas asignadas son visibles sobre los nombres, los cuales se encuentran ofuscados por razones de confidencialidad.	26
2.3	Archivo de anotación en BRAT. Cada línea contiene una entidad T_i , su etiqueta, sus posiciones de comienzo y fin en el texto original, y el texto del segmento en sí mismo. En este caso, hay cinco menciones a personas marcadas en el texto, de las cuales T_1 , T_2 y T_4 refieren a la misma persona, mientras las otras dos refieren a otras dos personas distintas.	27
2.4	Definiciones de precisión (P), cobertura (R) y medida F_1 ; VP , FP y FN representan verdaderos positivos, falsos positivos y falsos negativos, respectivamente (Jurafsky y Martin, 2025). . .	33
2.5	Valores de F_1 al variar la tasa de <i>dropout</i> para los distintos experimentos.	34
2.6	Experimentos PER y PER-500 con tasa de <i>dropout</i> de 0.35 para 120 (arriba) y 500 épocas (abajo). Los mejores valores de F_1 son 0.899 en la ronda 80 y 0.902 en la ronda 420, respectivamente. .	35
2.7	Pseudocódigo del algoritmo Agrupamiento Aglomerativo con criterio de parada por umbral de distancia μ	37
2.8	Dendrograma resultante de aplicar el Agrupamiento Aglomerativo hasta llegar a un único agrupamiento. Se muestran las fusiones de agrupamientos para cinco nombres de entrada y dos posibles umbrales, μ_1 y μ_2 , que generan dos y tres agrupamientos finales, respectivamente.	39
2.9	Definiciones de las distancias basadas en las funciones de similitud <i>Bag</i> , <i>Jaccard</i> , <i>Overlap</i> y <i>Sørensen</i> , donde $Bag(X)$ representa la bolsa de palabras de la entrada X	41
2.10	Definiciones de homogeneidad (H), cobertura (C), V_1 y ARI . Las clases reales y las estimadas se representan con T y K , respectivamente; N es el número total de elementos.	42
2.11	ARI para las distancias <i>Overlap</i> , <i>Jaccard</i> , <i>Sørensen</i> y <i>Bag</i> por valor de umbral en el conjunto de entrenamiento.	43
2.12	ARI , Acc_{DOC} , C y H en el conjunto de entrenamiento para <i>Overlap</i> (arriba) y <i>Jaccard</i> (abajo) por valor de umbral.	44

3.1	Ejemplo extraído de FrameNet para el marco «Asesinato» (<i>Killing</i>), con unidad léxica «matar» (<i>kill</i>). Se marcan los elementos: «Asesino», «Víctima» —resaltados en azul y rojo, respectivamente—, y el «Objetivo», en negrita.	50
3.2	Anotación según PropBank para el Ejemplo 3.1. Los roles «Arg0» y «Arg1» corresponden a los roles semánticos de «Asesino» y «Víctima» de FrameNet. El marco «kill.01» indica al verbo «kill» en su primer sentido: causar la muerte, asesinar. . .	51
3.3	Ejemplos de entrada y salida para las tres tareas FSP, FI y FSRL con el marco «Asesinato».	56
3.4	FSP en contexto: los marcos de FrameNet, junto con sus ejemplos anotados, se combinan de acuerdo a una plantilla para construir una guía específica para la tarea. Esta guía, junto con un texto, se proporcionan a un LLM como entrada para generar la salida correspondiente.	57
3.5	Ejemplos de sección «Objetivo» para las tareas de (a) Identificación de Marcos y (b) Extracción de Argumentos.	60
3.6	Ejemplo de la sección «Eventos» para (a) detección de marcos y (b) extracción de argumentos. En el primer caso se detallan posibles valores para los objetivos, mientras que en el segundo se provee la lista de posibles elementos asociados a un marco. . .	61
3.7	Directrices para la tarea de Identificación de Marco.	62
3.8	Sección de ejemplos para la tarea FSRL.	62
3.9	Resultados micro F_1 según el número de ejemplos para la tarea FSP (Experimento 1).	68
3.10	P (arriba) y R (abajo) para Claude Haiku 3.5 y GPT-4o Mini según número de ejemplos para la tarea FSP (Experimento 1). .	69
3.11	F_1 para los argumentos «Objetivo» y «Víctima» según el número de <i>shots</i> para el Experimento 1.	71
3.12	F_1 para los argumentos «Asesino» y «Causa» según el número de <i>shots</i> para el Experimento 1.	72
3.13	Mejores valores de F_1 para los modelos Sonnet 3.5 y GPT-4o por argumento para el Experimento 1.	74
3.14	F_1 para la detección del par marco–objetivo según número de ejemplos para el Experimento 2.	75

3.15	F_1 para el Marco (arriba) y Objetivo (abajo) según número de ejemplos para el Experimento 2.	76
3.16	F_1 para el Experimento 3a según número de ejemplos con pares marco-objetivo de referencia.	77
3.17	P (arriba) y R (abajo) para el Experimento 3a según número de ejemplos con pares marco-objetivo de referencia.	78
3.18	F_1 , con 150 ejemplos, para FI (Experimento 2) y FSRL con pares marco-objetivo de referencia (Experimento 3a) y extraídos en la etapa previa (Experimento 3b).	79
3.19	P (arriba) y R (abajo), con 150 ejemplos, para FI (Experimento 2) y FSRL con pares marco-objetivo de referencia (Experimento 3a) y extraídos en la etapa previa (Experimento 3b). . . .	80
4.1	Mapa de áreas de los verbos violentos seleccionados en la Base Nacional de Jurisprudencia, donde se detalla el porcentaje de cada verbo sobre el total de ocurrencias de la lista. Los 20 verbos más frecuentes se muestran individualmente; la categoría «Otros» agrupa a los verbos restantes.	95
4.2	Ejemplo extraído de FrameNet para el marco «Compromiso» (<i>Commitment</i>), con unidad léxica «prometer» (<i>promise</i>). Se marcan los elementos «hablante» (<i>speaker</i>), «destinatario» (<i>addressee</i>) y «mensaje» (<i>message</i>), resaltados en rojo, verde y azul, respectivamente.	96
4.3	Marcado de eventos de violencia con BRAT. Los verbos están marcados automáticamente; los complementos y las relaciones son anotadas de forma manual.	98
4.4	Porcentaje de eventos de violencia y no violencia por marco. Cada barra representa un marco y está normalizada al 100 %, mostrando la proporción de casos violentos y no violentos, en rojo y azul respectivamente.	100
4.5	P (arriba) y R (abajo) en función de la proporción relativa de ejemplos provenientes de la BJN y FrameNet en el conjunto de entrenamiento para Claude Sonnet 3.7, GPT-4o y DeepSeek Reasoner (0: 100 % FrameNet; 1: 100 % BJN.	106

4.6	F_1 en función de la proporción relativa de ejemplos provenientes de la BJN y FrameNet en el conjunto de entrenamiento para Claude Sonnet 3.7, GPT-4o y DeepSeek Reasoner (0: 100 % FrameNet; 1: 100 % BJN).	107
4.7	F_1 para <i>zero-shot</i> y <i>500-shot</i> con 0 % y 100 % de ejemplos provenientes de la BJN para cada uno de los tres modelos utilizados.	107
4.8	Valores de F_1 para los ocho atributos con mejor desempeño: a la izquierda, mejor valor absoluto, independientemente de la proporción de ejemplos; a la derecha, valores correspondientes a la distribución con 100 % de ejemplos provenientes de la BJN.	111
4.9	F_1 para los argumentos «Objetivo» y «Parte del cuerpo» al variar la proporción de <i>shots</i> de la BJN y FrameNet. En el eje de las abscisas se indica la proporción de ejemplos de la BJN, donde 0 corresponde a 100 % FrameNet y 1 a 100 % BJN.	112
4.10	F_1 para los argumentos «Destinatario» y «Víctima» al variar la proporción de <i>shots</i> de la BJN y FrameNet. En el eje de las abscisas se indica la proporción de ejemplos de la BJN, donde 0 corresponde a 100 % FrameNet y 1 a 100 % BJN.	113
A.1	Plantilla Jinja para el <i>prompt</i> de la tarea FSP.	157
A.2	<i>Prompt</i> para la tarea FSP, escenario sin información, ejemplificado con el marco «Asesinato».	158
A.3	<i>Prompt</i> para la tarea FSP, escenario solo definición, ejemplificado con el marco «Asesinato».	158
A.4	<i>Prompt</i> para la tarea FSP, escenario <i>zero-shot</i> , ejemplificado con el marco «Asesinato».	159
A.5	<i>Prompt</i> para la tarea FSP, escenario <i>3-shot</i> , ejemplificado con el marco «Asesinato».	160
A.6	Plantilla Jinja para el <i>prompt</i> de la tarea FI.	161
A.7	<i>Prompt</i> para la tarea FI, escenario sin información, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	162
A.8	<i>Prompt</i> para la tarea FI, escenario solo definición, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	163
A.9	<i>Prompt</i> para la tarea FI, escenario <i>zero-shot</i> , ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	164

A.10	<i>Prompt</i> para la tarea FI, escenario 3-shot, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	165
A.11	Plantilla Jinja para el <i>prompt</i> de la tarea FSRL.	166
A.12	<i>Prompt</i> para la tarea FSRL, escenario sin información, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	167
A.13	<i>Prompt</i> para la tarea FSRL, escenario solo definición, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	168
A.14	<i>Prompt</i> para la tarea FSRL, escenario zero-shot, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	169
A.15	<i>Prompt</i> para la tarea FSRL, escenario 3-shot, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».	170
C.1	Fragmento de la definición del marco «Amenaza» en inglés, utilizada como insumo para la creación de <i>prompts</i> .	177
C.2	Fragmento en español del marco «Amenaza», basado en su definición original en inglés.	178
E.1	Plantilla Jinja para el <i>prompt</i> de la tarea FSP sobre el corpus ViolenciaR.	184
E.2	<i>Prompt</i> para la tarea FSP sobre el corpus ViolenciaR, escenario zero-shot, ejemplificado para los marcos «Robo», «Violación» y «Amenaza».	185
E.3	Sección «Ejemplos» para la tarea FSP sobre el corpus ViolenciaR, escenario <i>n-shot</i> , para $n = 10$, con una distribución 20–80 % de ejemplos BJN–FrameNet.	186
I.1	Fragmento de la definición del marco «Abuso» (<i>Abusing</i>) en inglés.	191
I.2	Fragmento de la definición del marco «Abuso» (<i>Abusing</i>) en español.	192
I.3	Fragmento de la definición del marco «Arma de fuego» (<i>Use Firearm</i>) en inglés.	193
I.4	Fragmento de la definición del marco «Arma de fuego» (<i>Use Firearm</i>) en español.	194
I.5	Fragmento de la definición del marco «Asesinato» (<i>Killing</i>) en inglés.	195
I.6	Fragmento de la definición del marco «Asesinato» (<i>Killing</i>) en español.	196

I.7	Fragmento de la definición del marco «Compromiso» (<i>Commitment</i>) en inglés.	197
I.8	Fragmento de la definición del marco «Compromiso» (<i>Commitment</i>) en español.	198
I.9	Fragmento de la definición del marco «Daño» (<i>Cause Harm</i>) en inglés.	199
I.10	Fragmento de la definición del marco «Daño» (<i>Cause Harm</i>) en español.	200
I.11	Fragmento de la definición del marco «Disparo» (<i>Shoot Projectiles</i>) en inglés.	201
I.12	Fragmento de la definición del marco «Disparo» (<i>Shoot Projectiles</i>) en español.	202
I.13	Fragmento de la definición del marco «Impacto» (<i>Hit Target</i>) en inglés.	203
I.14	Fragmento de la definición del marco «Impacto» (<i>Hit Target</i>) en español.	204
I.15	Fragmento de la definición del marco «Robbery» (<i>Robo</i>) en inglés.	205
I.16	Fragmento de la definición del marco «Robbery» (<i>Robo</i>) en español.	206
I.17	Fragmento de la definición del marco «Violación» (<i>Rape</i>) en inglés.	207
I.18	Fragmento de la definición del marco «Violación» (<i>Rape</i>) en español.	208
I.19	Fragmento de la definición del marco «Violencia» (<i>Violence</i>) en inglés.	209
I.20	Fragmento de la definición del marco «Violencia» (<i>Violence</i>) en español.	210

Lista de tablas

2.1	Segmentación en oraciones y <i>tokens</i> de 17 350 documentos de la BJA realizada por CoreNLP, Freeing, NLTK y SpaCy.	28
2.2	Ejemplos de segmentación de oraciones por NLTK, CoreNLP y Freeing. Cada grupo de segmentos (44–45, 56–58 y 108–111) pertenecen a una misma oración del texto original, a pesar de que los segmentadores las dividan incorrectamente.	29
2.3	Parámetros y valores experimentados para el reentrenamiento del módulo NER de SpaCy.	32
2.4	Resumen de los experimentos para el reentrenamiento del módulo NER de SpaCy con el corpus B*N.	33
2.5	Mejores resultados sobre el conjunto de evaluación para cada experimento según F_1 . Se agregan los resultados en el mismo conjunto del modelo sin reentrenar junto al resultado reportado por los autores de SpaCy sobre su corpus de evaluación, WikiNER. Los mejores resultados se indican en negrita, mientras que los segundos mejores resultados se muestran subrayados. . .	36
2.6	Parámetros del experimento del entrenamiento de Agrupamiento Aglomerativo.	40
2.7	Resultados de las cuatro mejores soluciones distancia- μ según <i>ARI</i> sobre el conjunto de evaluación.	45
3.1	Descripciones de los grandes modelos de lenguaje evaluados según la información disponible públicamente.	65
3.2	Configuraciones de los experimentos de las distintas tareas con ICL. Claude, GPT y Deepseek se refieren a ambos modelos evaluados por familia: Haiku 3.5 y Sonnet 3.5, GPT-4o y GPT-4o Mini, y Chat y Reasoner, respectivamente.	65

3.3	Mejores valores de F_1 obtenidos en los experimentos, indicando entre paréntesis la cantidad de <i>shots</i> . Se muestran resultados para FSP (Exp. 1), FI (Exp. 2), FSRL con pares marco-objetivo (Exp. 3a) y FSRL con resultados de FI como entrada (Exp. 3b). Los valores más altos se indican en negrita y los segundos mejores se encuentran subrayados.	67
3.4	F_1 por argumento para distintos modelos en el Experimento 1. Los números entre paréntesis indican la cantidad de <i>shots</i> donde se alcanza el rendimiento máximo.	73
3.5	F_1 para el Experimento 2. Se reportan los valores para el par marco-objetivo, y para cada componente por separado. Los números entre paréntesis indican la cantidad de ejemplos en el que se alcanza el rendimiento máximo.	74
3.6	Resultados para los experimentos 1 y 3b para la tarea de FSP. Para el Experimento 1 se detalla entre paréntesis la cantidad de ejemplos en que se obtuvo el resultado; para el Experimento 3b esa cantidad es de 150 ejemplos.	81
3.7	F_1 por modelo y estrategia en los experimentos 1, 2 y 3a con ablación: (a) sin información de los marcos ni ejemplos en contexto; (b) únicamente con la definición del marco; (c) con información completa de los marcos, pero sin ejemplos en contexto; (d) con información completa de los marcos junto con ejemplos en contexto. En el caso (d), el número entre paréntesis indica la cantidad de ejemplos utilizados con el mejor resultado. Los valores más altos se muestran en negrita y los segundos mejores, subrayados.	82
4.1	Lista de los 50 verbos más frecuentes del total de 8764 lemas detectados en 5000 documentos de la BJN según SpaCy, ordenados por cantidad de ocurrencias.	92
4.2	Lista de los 33 verbos, ordenados por frecuencia, resultantes del proceso iterativo de búsqueda de la segunda etapa.	93
4.3	Lista de los 59 verbos resultantes de la tercera etapa, ordenados según su conteo en los 5000 documentos de la BJN. La columna posición indica el lugar por frecuencia en la lista original de verbos.	94

4.4	Distribución de los eventos anotados por marco semántico. Se presenta el conteo de casos correspondientes a violencia y a no violencia, así como el total por marco. Entre paréntesis se indica el porcentaje que representa cada valor sobre el total de eventos.	99
4.5	Distribución de los eventos positivos y negativos por marco semántico en los 517 y 200 textos de entrenamiento y evaluación, respectivamente. Entre paréntesis, el porcentaje en el total del conjunto respectivo.	103
4.6	Número de argumentos en los conjuntos de 500 ejemplos para las distribuciones 100 % FrameNet y 100 % BJN, así como en el conjunto de evaluación. Entre paréntesis se indica el porcentaje que representa cada valor sobre el total de argumentos de cada conjunto.	104
4.7	Configuraciones utilizadas para la experimentación de ICL-FSP en la BJN. La evaluación se realiza sobre un conjunto de 200 ejemplos provenientes exclusivamente de ViolenciaR.	104
4.8	Mejores resultados de F_1 , junto con sus valores de precisión y cobertura, para los modelos evaluados. En todos los modelos, el mejor F_1 se obtiene utilizando 500 ejemplos provenientes exclusivamente de la BJN, sin incorporar ejemplos de FrameNet. Los valores más altos se destacan en negrita y los segundos mejores se encuentran subrayados.	105
4.9	Mejor F_1 por argumento y modelo. El resultado para cada argumento se muestra en negrita y el segundo mejor se subraya. Entre paréntesis, se indican las proporciones de <i>shots</i> de la BJN con las que se produce ese resultado.	109
4.10	Valores de F_1 por argumento y modelo para la configuración con 100 % de ejemplos provenientes de la BJN. El mejor resultado por argumento se indica en negrita y el segundo mejor se subraya.	110
A.1	Escenarios experimentales y figuras asociadas a los <i>prompts</i> generados para cada tarea	156
B.1	Lista de los 20 verbos de violencia de la lista inicial con su correspondiente categoría en ADESSE.	172

B.2	Definición de las categorías en ADESSE de los verbos semi- lla. Se listan los verbos de violencia asociados a cada categoría. Los verbos en negrita pertenecen a la lista inicial; mientras que aquellos subrayados no figuran en el corpus.	173
B.3	Lista de los 59 verbos de violencia ordenados según su conteo en la BJN.	174
D.1	Lista de sustitutos de fantasía utilizados para la desidentifica- ción según su categoría.	180

Lista de siglas

BiLSTM Red neuronal recurrente bidireccional de memoria a corto y largo plazo (*Bidirectional Long Short-Term Memory*)

BJN Base de Jurisprudencia Nacional

CNN Red neuronal convolucional (*Convolutional Neural Network*)

CRF Campo aleatorio condicional (*Conditional Random Field*)

FI Identificación de marco (*Frame Identification*)

FSP Análisis de marcos semánticos (*Frame Semantic Parsing*)

FSRL Etiquetado de roles semánticos (*Frame Semantic Role Labeling*)

ICL Aprendizaje en contexto (*In-Context Learning*)

IE Extracción de información (*Information Extraction*)

LLM Gran modelo de lenguaje (*Large Language Model*)

NER Reconocimiento de entidades nombradas (*Named Entity Recognition*)

PLN Procesamiento del lenguaje natural

POS Categoría gramatical (*Part-of-Speech*)

SCJ Suprema Corte de Justicia

SVM Máquina de vectores de soporte (*Support Vector Machine*)

Tabla de contenidos

Agradecimientos	XI
Resumen	XV
Abstract	XVII
Lista de figuras	XIX
Lista de tablas	XXVII
Lista de siglas	XXXI
1 Introducción	1
1.1 Marco de investigación	5
1.2 Objetivos	7
1.3 Contribuciones	10
1.4 Organización del documento	11
2 Desidentificación	13
2.1 Desidentificación en la BJN	15
2.2 Trabajos Relacionados	21
2.3 Corpus B*N	25
2.4 Reconocimiento de Entidades	27
2.4.1 NER en el Corpus B*N	30
2.4.2 Experimentación	31
2.4.3 Resultados	33
2.5 Resolución de Correferencias	37
2.5.1 Agrupamiento de Nombres	38
2.5.2 Experimentación	40

2.5.3	Resultados	42
2.6	Síntesis	45
3	Análisis de marcos semánticos	47
3.1	FrameNet	49
3.2	Trabajos Relacionados	52
3.3	FSP en Contexto	55
3.3.1	Generación de <i>Prompts</i>	59
3.3.2	Experimentación	63
3.3.3	Resultados	67
3.3.4	Ablación	82
3.4	Síntesis	83
4	Eventos de violencia	85
4.1	Trabajos Relacionados	87
4.2	Verbos de violencia en la BJN	90
4.3	Corpus ViolenciaR	94
4.4	ICL-FSP en la BJN	100
4.4.1	<i>Prompt</i> para ICL en la BJN	101
4.4.2	Experimentación	102
4.4.3	Resultados	105
4.4.4	Ejemplos de extracciones	114
4.5	Síntesis	118
5	Conclusiones	121
	Referencias bibliográficas	127
	Glosario	147
	Apéndices	153
	Apéndice A ICL para Análisis de Marcos Semánticos	155
	Apéndice B Verbos de violencia.	171
	Apéndice C Marco de Amenaza.	175
	Apéndice D Desidentificación de ViolenciaR	179
	Apéndice E <i>Prompt</i> para FSP en ViolenciaR	183
	Anexos	187
	Anexo I Marcos de FrameNet.	189

Capítulo 1

Introducción

El área de la inteligencia artificial aplicada al derecho constituye un campo de investigación de gran interés, debido a su relevancia e impacto social. Ya en la década de 1970, D’Amato (1977) se pregunta si las computadoras pueden o deben reemplazar a los jueces. En aquel trabajo, tomando como ejemplo al psicoterapeuta artificial ELIZA/DOCTOR desarrollado por Weizenbaum (1966) en la década anterior, se discute la posibilidad de sustituir a los jueces por algoritmos. Medio siglo después, debates similares resurgen en torno a la utilización de grandes modelos de lenguaje (*Large Language Models*, LLM) y otras técnicas de inteligencia artificial en el ámbito jurídico (Bex, 2024; Blair-Stanek et al. 2023; Greenstein, 2022; Katz et al. 2024; Martínez, 2024; Sachoulidou, 2023; Schweitzer y Conrads, 2024).

La inteligencia artificial aplicada al derecho (*AI & Law*) comienza a establecerse como un campo de investigación propio en la década de 1980, con la aparición de congresos académicos y publicaciones especializadas. Entre estas iniciativas destacan la «International Conference on Artificial Intelligence and Law» (Conferencia Internacional en Inteligencia Artificial y Derecho, ICAIL) en 1987, «JURIX», en 1988, y la revista «Artificial Intelligence and Law» (Inteligencia Artificial y Derecho), cinco años más tarde; más recientemente se han sumado talleres especializados que complementan las conferencias principales, como el de «Automated Semantic Analysis of Information in Legal Text» (Análisis Semántico Automatizado de Información en Textos Legales, ASAIL).

Los trabajos en esa década y la siguiente se centran en el desarrollo de sistemas expertos que simulaban el razonamiento legal, donde la representación

del conocimiento jurídico y la lógica formal eran los enfoques predominantes (Governatori et al. 2022). Entre los proyectos más influyentes de esta etapa se encuentra «HYPO» (Ashley, 1990), que aplican el razonamiento basado en casos previos para resolver disputas en derecho de propiedad intelectual. Este sistema da origen a sucesores como «CATO» (Aleven y Ashley, 1997), orientado a la enseñanza de argumentación jurídica, y motivó aplicaciones en otras ramas, como el derecho impositivo.

Tras la etapa inicial centrada en sistemas expertos, la creciente disponibilidad de bases de datos documentales condujo a plantearse nuevas tareas: no solamente es necesario procesar y mejorar estos datos, sino que también surge la posibilidad de abordar nuevos problemas mediante técnicas de aprendizaje automático, como sucede también en otros dominios relacionados, particularmente en el área del procesamiento del lenguaje natural (PLN). A comienzos de este siglo, aparecen trabajos relacionados con la generación de ontologías para el dominio legal, la predicción de resultados judiciales y el procesamiento de texto legal, incluyendo minería de argumentos y clasificación de documentos legales, entre otros (Sartor et al. 2022).

En los últimos años, el uso de aprendizaje automático y, en particular, de redes neuronales profundas y LLM se ha vuelto la norma, cubriendo tareas como búsqueda de respuestas, reconocimiento de entidades, resumen y clasificación de documentos y predicción de fallos judiciales (Ariai et al. 2025; Siino et al. 2025; Villata et al. 2022).

Entre las bases de datos legales utilizadas se encuentran el corpus de Jurisprudencia Europea¹, el de la Corte Europea de Derechos Humanos², el de Jurisprudencia de la Corte Suprema de los Estados Unidos³ así como bases de datos nacionales, como las de Alemania, Italia, Japón e India, entre otros países (Ariai et al. 2025; Siino et al. 2025). Además de trabajos para la lengua inglesa, existen estudios en distintos idiomas, entre ellos el japonés (Yamada et al. 2019), el alemán (Waltl et al. 2019), el italiano (Ragazzi et al. 2024; Tagarelli y Simeri, 2022) y el coreano (J. Lee, 2025), así como investigaciones sobre transferencia de conocimiento entre lenguas (Moro et al. 2023).

Más allá de los recursos disponibles, el lenguaje jurídico presenta, desde la perspectiva del PLN, desafíos particulares que dificultan la aplicación directa

¹Disponible en <https://eur-lex.europa.eu/>.

²Disponible en <https://hudoc.echr.coe.int>.

³Disponible en <http://scdb.wustl.edu>.

Es obvio que todos los homicidios de mujeres no son femicidios, pero como relevara la primera instancia, surge probado que el vínculo entre los concubinos **AA** y **HH** revestía características como las previstas en la norma transcrita, y que la muerte de la víctima es totalmente compatible con la agravante muy especial cuyo cómputo, infundadamente la Defensa cuestiona, minimizando la situación previa como la de una pareja “normal” en pie de igualdad: “...ellos se peleaban discutían como pareja, yo vi y presencié las discusiones, **AA** era militar lo corrieron por un asunto de drogas, fue después que vino del Congo. Él compró lo de la casa (heladera, cocina) y lo vendía para comprar drogas, vendía la ropa de las niñas que le compraba **JJ**, les vendió un par de botas... ella nos pedía dinero yo le daba, a veces era para comer y darle algo a las niñas... ese día a las 22:15 pasó eso y ella estaba muerta, la niña más grande **NN** estaba cortada, ella estaba muerta sentada en un banco y él estaba en el piso y estaba vivo” (**BB**, tío de la víctima, audiencia de 26/7, pista 9.9); “ella no salía casi nada y nosotros nos ocupábamos de las niñas... era ama de casa, a él lo sacaron por drogadicto del cuartel, la ayudamos en todo, la ropa de las niñas la comprábamos nosotros y a veces **JJ** con la asignación, **KK** no le compraba ropa a las niñas, yo le veía drogado, acelerado, altanero siempre hablando de malos modos, los ojos rojos. Él un día le quiso pegar le levantó la mano y yo me metí... el día de los hechos le dimos de comer a las niñas, miramos televisión, las niñas tenían sueño al rato se fueron. Nos acostamos y como a las 22.00 horas nos golpean la puerta, eran las nenas decían “mató a mamá, mató a mamá, todas ensangrentadas”, las abracé y las acosté, subí y todo era un charco de sangre ella en una silla, él tirado en el suelo y llamamos a mi hermana y al servicio 911, ella tenía puntadas en el cuello en la nuca. Ellos discutían todos los días, el me pedía plata siempre, las cosas que él se llevó no volvieron nunca, nosotros le dimos una heladera y cocina vieja que teníamos... la niña pequeña dice hoy en día si su papá la iba a matar como a su mamá” (declaración de la tía de la víctima en la misma audiencia, pista 15).

Figura 1.1: Fragmento extraído de la Sentencia 15/2022, del Tribunal de Apelaciones Penal de 1º Tº de Uruguay. Se trata de un enunciado complejo cuya oración principal se prolonga hasta el punto final, incorporando citas directas de testigos. En negrita se señalan las siglas que sustituyen nombres; en naranja y verde se resaltan los testimonios del tío y tía de la víctima, respectivamente.

Como agravantes genéricas computó alevosía, grave sevicia, y nocturnidad. Desestimó la agravante especial reclamada para el Homicidio (“Si concurre la agravante prevista en el artículo 312 numeral 8 no corresponde el cómputo de la agravante prevista en el artículo 311 inciso 1 del Código penal”, dijo) lo que no fue objeto de agravio por parte de Fiscalía, y en base a ello disminuyó la pena requerida en la acusación. (...) Pero in dubio pro reo, que el ataque haya ocurrido de noche, parece contingente, esto es, no querida ni buscada en concreto, y esta agravante “exige un gran discernimiento en el Juez; estas circunstancias ontológicamente, resultan una veces indiferentes, otras atenuantes y algunas agravantes del hecho” (Irureta, cfm. Código Penal anotado por Reta y Grezzi, 2ª edición, p. 182)

Figura 1.2: Fragmento extraído de la Sentencia 15/2022, del Tribunal de Apelaciones Penal de 1º Tº de Uruguay. Este fragmento presenta un estilo técnico-formal, con oraciones complejas correctamente puntuadas y bien estructuradas.

de herramientas estándar de procesamiento de lenguaje: se trata de textos extensos, con un estilo técnico-formal, uso intensivo de referencias cruzadas y una terminología específica que a menudo difiere significativamente de los textos de uso general con los que suelen entrenarse estas herramientas.

Por ejemplo, consideremos el texto que se presenta en la Figura 1.1, extraído de la Sentencia 15/2022, del Tribunal de Apelaciones Penal de 1º Tº de Uruguay. Este texto presenta algunas particularidades. Primeramente, es un enunciado complejo —el primero de un párrafo— que contiene citas directas de testigos y es de una extensión considerable. No se mencionan nombres de personas en el texto: para preservar su identidad, estas fueron sustituidas por las siglas AA, HH, JJ, etc.

Las citas entrecomilladas corresponden a las transcripciones de dos testimonios: el del tío, referido como BB, y el de la tía de la víctima (sin siglas de referencia), que en la figura se marcan en amarillo y verde, respectivamente. Estas transcripciones, al reproducir literalmente el carácter oral del enunciado, no comparten el mismo registro formal que el resto del texto de la sentencia, como se observa, por ejemplo, en el fragmento de la Figura 1.2 extraído del mismo documento. Presentan, además, falta de puntuación y cierta agramaticalidad: «todo era un charco de sangre ella en una silla». Además, estas transcripciones contienen oraciones y hasta citas a lo dicho por terceros entrecomilladas como, por ejemplo, en «las nenas decían “mató a mamá, mató a mamá (...)”».

Estas características particulares alejan a este tipo de textos de aquellos

periodísticos o científicos, que son los tipos de textos con los que las herramientas de PLN suelen ser entrenadas. Los problemas detectados aparecen en las etapas iniciales del análisis estándar de texto: los analizadores cometen errores importantes, desde la segmentación en componentes básicos (tokenización) y segmentación en oraciones hasta la detección de entidades nombradas (Garat y Wonsever, 2019). En la Figura 1.3 se muestran algunos ejemplos de errores en esta última tarea: los nombres transpuestos no son detectados como personas, «Dra.» es etiquetada como organización, «Martínez Mágico» como organización pero nunca como persona, etc.

En definitiva, la investigación en inteligencia artificial aplicada al derecho constituye un área de interés académico en creciente desarrollo. En particular, el procesamiento de textos jurídicos plantea retos lingüísticos y estructurales especialmente exigentes —oraciones extensas, estilo técnico-formal, terminología especializada, así como fragmentos transcritos del habla oral y citas literales— que dificultan la aplicación directa de modelos entrenados en dominios generales.

1.1. Marco de investigación

La Base de Jurisprudencia Nacional (BJN) es una base documental digital que comienza a funcionar en el año 2008 como parte de un plan de modernización del Poder Judicial de Uruguay. Su objetivo es facilitar el acceso libre y gratuito a la jurisprudencia nacional mediante la sistematización de los fallos judiciales en un sistema informático, no solo para los integrantes del propio Poder Judicial, sino también para todos los actores del ámbito jurídico en general. Aunque formalmente la jurisprudencia y las costumbres no son una fuente de derecho en Uruguay, constituyen, de todas formas, precedentes importantes para todos los actores de un proceso judicial.

Desde su creación se han incorporado sistemáticamente a la BJN fallos de diversos tribunales, incluyendo la Suprema Corte de Justicia (SCJ), los Tribunales de Apelaciones y los Juzgados Letrados de Primera Instancia. En la actualidad, esta base de datos jurídica digital cuenta con más de 100 000 documentos, disponibles públicamente en su portal de Internet⁴. En la Figura 1.4 se muestra un ejemplo de una sentencia disponible en este portal.

⁴Disponible en: <http://bjn.poderjudicial.gub.uy/>

Entidades: Persona, Organización, Otros.

CoreNLP

En segunda instancia para sentencia interlocutoria estos autos “PÉREZ CABRERA, María José c/ MÁRTINEZ MÁGICO, Adolfo Ramón.” Créditos salariales e indemnización por despido. (F.356-252/2005) venidos a conocimiento de este Tribunal en mérito al recurso de apelación interpuesto contra la resolución No.3072 de fecha 4 de agosto de 2006 dictada por la Sra. Juez Letrado de Primera Instancia de Salto de 5º turno Dra. Mónica Bórtoli Porro.

Freeling

En segunda instancia para sentencia interlocutoria estos autos “PÉREZ CABRERA, María José c/ MÁRTINEZ MÁGICO, Adolfo Ramón.” Créditos salariales e indemnización por despido. (F.356-252/2005) venidos a conocimiento de este Tribunal en mérito al recurso de apelación interpuesto contra la resolución No.3072 de fecha 4 de agosto de 2006 dictada por la Sra. Juez Letrado de Primera Instancia de Salto de 5º turno Dra. Mónica Bórtoli Porro.

SpaCy

En segunda instancia para sentencia interlocutoria estos autos “PÉREZ CABRERA, María José c/ MÁRTINEZ MÁGICO, Adolfo Ramón.” Créditos salariales e indemnización por despido. (F.356-252/2005) venidos a conocimiento de este Tribunal en mérito al recurso de apelación interpuesto contra la resolución No.3072 de fecha 4 de agosto de 2006 dictada por la Sra. Juez Letrado de Primera Instancia de Salto de 5º turno Dra. Mónica Bórtoli Porro.

Figura 1.3: Resultado del reconocimiento de entidades por tres herramientas sobre un texto extraído la Base de Jurisprudencia Nacional de Uruguay. En celeste se señalan los segmentos indetificados como personas, en verde, aquellos identificados como organizaciones, mientras que en violeta, segmentos etiquetados como otros.

Antes de su publicación, las sentencias judiciales atraviesan un proceso de curaduría manual, a través del cual se enriquecen con información sobre su prioridad, se clasifican dentro de una taxonomía jurídica, se les agrega un resumen, etc. Además, el propio texto de cada sentencia es modificado para ocultar datos personales considerados sensibles —nombres de menores, testigos protegidos, delincuentes primarios, etc.— conforme a los mandatos legales de protección de datos personales.

Con el fin de impulsar el desarrollo de tecnologías de procesamiento automático que permitan explotar de mejor forma la información contenida en las sentencias, así como automatizar o semiautomatizar las tareas requeridas para su publicación, en el año 2018 se firma un convenio entre la Facultad de Ingeniería y el Poder Judicial. En este marco, y dentro del Grupo de Procesamiento de Lenguaje Natural, del Instituto de Computación, se define y desarrolla la presente tesis de doctorado. Además, de forma complementaria a este trabajo, se desarrollan dos tesis de grado (Cruces y Ferraro, 2023; Fulloni et al. 2020) y un proyecto del Fondo Sectorial de Investigación a partir de Datos de la ANII-MIEM⁵.

1.2. Objetivos

Este trabajo se centra en el desarrollo de técnicas de procesamiento de lenguaje natural aplicadas a tareas relevantes en el ámbito jurídico, tomando como caso de estudio la Base de Jurisprudencia Nacional del Poder Judicial de Uruguay. En este marco, se proponen dos objetivos principales: (a) la desidentificación automática de sentencias judiciales y (b) la extracción de eventos relevantes en el ámbito jurídico.

La desidentificación forma parte de los procesos de curaduría previos a la publicación de las sentencias en el portal de la BJN, y su resolución fue señalada de interés por parte de los propios integrantes del Poder Judicial. El objetivo es ocultar toda aquella información que permita identificar a los involucrados en un proceso judicial, protegiendo así su identidad y privacidad en aquellos casos en que lo exigen las normativas legales vigentes sobre protección de datos personales. Esta procedimiento implica, por ejemplo, la sustitución de nombres de personas, matrículas de vehículos o incluso profesiones, procurando

⁵Curación automática de sentencias de la Base de Jurisprudencia Nacional (FSDA_1.-2017-143138), desarrollado entre 2018 y 2020.

A	Número	Sede	Importancia	Tipo
	15/2022	Tribunal Apelaciones Penal 1° T°	ALTA	DEFINITIVA
	Fecha	Ficha	Procedimiento	
	02/03/2022	573-3933/2018	PROCESO PENAL ORDINARIO	
B	Materias			
	DERECHO PENAL			
C	Firmantes			
	Nombre	Cargo		
	Dr. Alberto Domingo REYES OEHNINGER	Ministro Trib.Apela.		
	Dra. Graciela Susana GATTI SANTANA	Ministro Trib.Apela.		
	Dr. Sergio TORRES COLLAZO	Ministro Trib.Apela.		
	Redactores			
	Nombre	Cargo		
	Dr. Alberto Domingo REYES OEHNINGER	Ministro Trib.Apela.		
D	Abstract			
	Camino	Descriptores Abstract		
	DERECHO PROCESAL PENAL->NCPP			
	DERECHO PENAL->DELITOS CONTRA LA PERSONALIDAD FISICA Y MORAL DEL ->HOMICIDIO MUY ESPECIALMENTE AGRAVADO, ART. 312 DEL->FEMICIDIO			
	Descriptores			
E	Resumen			
	Femicidio			
	Texto de la Sentencia			
F	VISTOS para definitiva de segunda instancia, en autos: "AA. HOMICIDIO ESPECIALMENTE AGRAVADO Y MUY ESPECIALMENTE AGRAVADO EN CONCURRENCIA FORMAL CON UN DELITO DE LESIONES PERSONALES EN CALIDAD DE AUTOR" (IUE: 573-3933/2018) venidos del Juzgado Letrado de Primera Instancia en lo Penal de 34° Turno, en virtud del recurso interpuesto por la Defensa Pública a cargo de la Dra. Virginia Ripoll y del Dr. Diego Moreira, contra la Sent. 195/2021 de 13.09.2021 dictada por la Dra. Blanca Rieiro, con intervención de la Fiscalía de Delitos Sexuales, Violencia Doméstica y Violencia Basada en			

Figura 1.4: Ejemplo de una sentencia en el portal de la BJN. La información publicada incluye: (A) información general de la sentencia: número de identificación, fecha, tipo de resolución, etc.; (B) tema; (C) tribunal; (D) clasificación; (E) resumen; y (F) texto de la sentencia.

mantener el sentido del relato de los hechos narrados en el documento, como se observa en el texto de la Figura 1.1. Actualmente la desidentificación se realiza de forma manual, lo que demanda un esfuerzo y un tiempo considerables, y no está exenta de errores humanos: la automatización o semiautomatización de este proceso permitiría agilizar la publicación de las sentencias y minimizar estos errores.

Con este propósito, la investigación se orienta al estudio de técnicas de PLN que permitan identificar y ocultar automáticamente los nombres de los involucrados en las sentencias de la BJN, que constituyen la categoría de información más sensible y relevante. Se busca generar un prototipo que, además, sea integrable a los sistemas informáticos del Poder Judicial. La tarea se divide en dos fases: una de reconocimiento de entidades nombradas (NER) y otra de resolución de correferencias entre los nombres identificados. Para la primera fase, debido al bajo desempeño en este dominio de los modelos generales existentes, se ajustan modelos específicamente a los textos de la BJN; para la segunda, se experimenta con técnicas de aprendizaje no supervisado. Ambas se desarrollan sobre un corpus de sentencias anotadas manualmente, creado como parte de este trabajo.

El segundo objetivo, la extracción de eventos relevantes, apunta a identificar automáticamente los eventos significativos que ocurren en las sentencias judiciales, junto con sus participantes y circunstancias. La extracción de eventos permite mejorar la accesibilidad y la búsqueda de información en la BJN, ayudando a los usuarios a encontrar rápidamente los casos y decisiones que son más relevantes a sus necesidades. Por ejemplo, identificados los eventos relacionados con homicidios, se podría buscar y analizar todos los casos que involucren víctimas mujeres, facilitando estudios estadísticos o jurídicos sobre fenómenos como el femicidio. A su vez, los resultados de esta tarea pueden contribuir a la creación de resúmenes automáticos y a la generación de otros metadatos que faciliten la organización y recuperación de la información jurídica.

El marco teórico elegido para representar los eventos es la Teoría de Marcos Semánticos (*Frame Semantics Theory*), propuesta por Fillmore (1976), la cual postula que las palabras evocan situaciones prototípicas que involucran eventos, participantes, ubicaciones y tiempos. En especial, dado el carácter del corpus de trabajo, la atención se centra en la extracción de eventos de violencia —asesinatos, robos o violaciones, entre otros— presentes en las sentencias

judiciales de la BJN.

Para la extracción de los eventos, se exploran técnicas de aprendizaje en contexto (*In-Context Learning*, ICL), que consisten en presentar a un LLM preentrenado ejemplos de entrada-salida —en este caso un texto y los eventos identificados en él— como parte de su entrada, junto con el texto a procesar. Para la obtención de ejemplos, se consideran los eventos del proyecto FrameNet, una base de datos léxica y semántica desarrollada por el Instituto de Ciencia de la Computación de la Universidad de Berkeley (C. F. Baker et al. 1998), que define más de 1220 marcos semánticos y contiene más de 200 000 ejemplos anotados manualmente en inglés contemporáneo.

Dado que prácticamente no se encuentran antecedentes de explotación de este recurso con ICL para la extracción de eventos, se realiza inicialmente una prueba de viabilidad utilizando únicamente ejemplos de FrameNet. Posteriormente, se aborda el desafío de transferir este conocimiento al español y aplicarlo a los textos jurídicos de la BJN, combinando los ejemplos de FrameNet con un corpus de sentencias de la BJN anotadas manualmente en el marco de esta investigación.

En definitiva, el objetivo general de esta investigación es mejorar la calidad, accesibilidad y aprovechamiento de la información contenida en la BJN mediante la aplicación de técnicas de procesamiento de lenguaje natural. Para alcanzarlo, se definen dos objetivos específicos: (a) la desidentificación automática de nombres de personas y (b) la extracción de eventos de violencia en sentencias judiciales. Con ello, se busca contribuir tanto a la eficiencia institucional del Poder Judicial como a la investigación académica y social en torno al derecho.

1.3. Contribuciones

Las principales contribuciones de esta investigación abarcan la construcción de recursos lingüísticos especializados, la aplicación de técnicas novedosas, así como el desarrollo de modelos y prototipos sobre la Base de Jurisprudencia Nacional. Estos trabajos son, hasta donde se sabe, los primeros en investigar el uso de técnicas de procesamiento de lenguaje natural aplicadas a este corpus judicial, contribuyendo así al desarrollo de tecnologías lingüísticas para el ámbito jurídico en español y, concretamente, en Uruguay.

En primer lugar se desarrollan dos corpus anotados manualmente: uno pa-

ra la desidentificación de nombres propios en sentencias judiciales —el corpus B*N—, y otro destinado a la extracción de eventos de violencia basado en la Teoría de Marcos Semánticos —el corpus ViolenciaR—. Estos recursos constituyen aportes para el estudio del procesamiento automático del lenguaje jurídico en español y en particular para la BJN. Cabe destacar que el trabajo de anotación es aceptado como ponencia en «Workshop on Legal Annotation, JURIX» (Garat y Wonsever, 2023).

En segundo lugar, sobre el primer corpus se desarrollan y evalúan modelos automáticos para la desidentificación de sentencias judiciales, combinando técnicas de reconocimiento de entidades nombradas y métodos de agrupamiento para la resolución de correferencias. El resultado de este trabajo es publicado en la revista arbitrada «Information» (Garat y Wonsever, 2022), y es integrado en un prototipo funcional en el proyecto del Fondo Sectorial de Investigación a partir de Datos de la ANII-MIEM.

En tercer lugar, se propone la explotación de FrameNet para su aplicación en aprendizaje en contexto, con el objetivo de extraer eventos de violencia en sentencias judiciales; este es un enfoque para el cual no se registran antecedentes en la literatura. Se explora, además, la transferencia de conocimiento hacia el español utilizando como apoyo al segundo corpus. Los resultados de esta investigación son aceptados para su presentación en la conferencia arbitrada «Foundation and Large Language Models» (Garat et al. 2025).

En conjunto, estas contribuciones aportan recursos y modelos específicos para la BJN, junto con metodologías innovadoras que abren nuevas líneas de investigación en tareas complejas como la desidentificación automática y la extracción de eventos, contribuyendo así tanto al desarrollo tecnológico como al estudio académico del PLN aplicado al derecho en español.

1.4. Organización del documento

El resto del documento se organiza de la siguiente forma. El Capítulo 2 aborda la tarea de desidentificación en la BJN. El Capítulo 3 presenta la extracción de eventos mediante ICL basada en FrameNet sobre textos en inglés, mientras que su aplicación en la BJN, en un escenario de transferencia entre lenguas, se desarrolla en el Capítulo 4. Finalmente, las conclusiones y las líneas de trabajo futuras se exponen en el Capítulo 5.

Capítulo 2

Desidentificación

En un número creciente de países, el manejo de datos que contienen información sensible de carácter personal se encuentra regulado por leyes de protección de datos, y su publicación puede acarrear sanciones legales (Centers for Medicare & Medicaid Services, United States of America, [1996](#); Parlamento Europeo y Consejo de la Unión Europea, [2016](#); Poder Legislativo, República Oriental del Uruguay, [2008](#)). Por otra parte, la difusión de documentos con este tipo de información es muchas veces relevante o incluso imprescindible, ya sea con fines de investigación científica, ya sea en virtud del derecho a la información establecido por ley. Este es el caso, por ejemplo, de la publicación con fines científicos de las historias clínicas en el dominio de la salud, en donde los resultados de estudios, patologías detectadas y procedimientos médicos son considerados información confidencial; o en el dominio judicial, en donde por ley se protegen datos sobre menores de edad, infractores primarios, etc.

Con el objetivo de permitir la publicación de información que contiene datos sensibles sin comprometer la privacidad de las personas, se aplican procesos de *desidentificación*, los cuales remueven, ocultan u ofuscan aquellos datos que permiten la individualización de personas o instituciones. El procedimiento de desidentificación es muchas veces manual: una tarea costosa, tanto en tiempo como en recursos humanos, y sin garantías de dar un resultado completamente confiable (Glaser et al. [2021](#); Oksanen et al. [2019](#); Sweeney, [1996](#)).

El Poder Judicial del Uruguay, en su Acordada Nro. 7564 (Suprema Corte de Justicia, República Oriental del Uruguay, [2006](#)), establece directrices sobre cuáles son los datos personales a protegerse en sus bases de datos de carácter documental o jurisprudencial, lo que comprende a la Base de Jurisprudencia

Nacional (BJN). En esta Acordada se indica que deben ocultarse, entre otros, los datos filiatorios junto con aquellos otros identificatorios de bienes inmuebles, profesión u ocupación, domicilios de «las partes del proceso, terceros, testigos y toda otra persona que actúe en éste en calidad de auxiliar de la justicia». Como se observa en la Figura 1.1, en esa sentencia de la BJN los nombres de las personas son reemplazados por iniciales ficticias: AA, BB, HH, etc. La desidentificación en la BJN es una actividad demandante, tanto en esfuerzo como en tiempo, en la cual pueden cometerse errores involuntarios.

En esta tesis se plantea, entonces, el estudio de técnicas de procesamiento de lenguaje natural (PLN) con el objetivo de realizar esta tarea de forma automática o semiautomática, facilitando así, y con una menor tasa de error, la edición de las sentencias para su publicación. En particular, se estudia la desidentificación de nombres de personas, que es el dato sensible más común e importante de las sentencias judiciales, y son aquellos que, en general, presentan mayores dificultades al momento de su desidentificación automática debido a la necesidad de resolver las cadenas de correferencias para mantener el sentido original del relato.

En la solución propuesta, esta tarea se divide en dos fases consecutivas: una primera fase en la que se detectan las menciones a desidentificar, y una segunda fase de agrupamiento de menciones que refieren a una misma persona, para sustituirlas luego por una misma sigla.

La primera fase corresponde, entonces, a una tarea de reconocimiento de entidades nombradas (*Named Entity Recognition*, NER). NER es una tarea central del Procesamiento del lenguaje natural que tiene por objetivo identificar y clasificar menciones textuales que se refieren a entidades del mundo real —tales como personas, organizaciones o ubicaciones— en categorías semánticas predefinidas. Esta tarea puede entenderse como un problema de etiquetado secuencial, en el cual, a los distintos segmentos del texto, se les asigna una etiqueta que indica si forman parte o no de una entidad nombrada y, en caso afirmativo, su categoría (Grishman y Sundheim, 1996; Jurafsky y Martin, 2025). Por el objetivo propuesto, este estudio se centra específicamente en los nombres de personas a desidentificar en la BJN.

La segunda fase corresponde a una tarea de resolución de correferencias, una tarea en donde el objetivo es identificar y agrupar todas las menciones que refieren a una misma entidad dentro del texto (Jurafsky y Martin, 2025; Karttunen, 1969). Así, en la oración «Juan llamó a Pedro por teléfono y lo

amenazó», «Juan» y «lo» refieren a una misma entidad. Para este trabajo sobre la BJN, la tarea se limita a la resolución de correferencias entre nombres propios de personas, debido a que, por ejemplo, los pronombres no son relevantes para la desidentificación, dado que no revelen información personal de forma directa.

La construcción de modelos para ambas fases requiere un conjunto de ejemplos en donde los nombres y sus agrupaciones estén debidamente identificadas. En consecuencia, se crea un corpus de sentencias anotadas manualmente con la finalidad de construir un modelo NER para su uso específico en la BJN y para la experimentación con técnicas de aprendizaje no supervisado para la resolución de correferencias. Como resultado se obtiene un prototipo de desidentificación integrable a los sistemas informáticos existentes en el Poder Judicial.

El resto de este capítulo se organiza de la siguiente manera: la Sección 2.1 presenta cómo se realiza el proceso de desidentificación en la BJN. En la Sección 2.2 se describen los trabajos relacionados sobre desidentificación, en particular, en texto libre. El trabajo de creación del corpus B*N de sentencias desidentificadas se detalla en la Sección 2.3. Las secciones 2.4 y 2.5 describen las soluciones propuestas para NER y resolución de correferencias que conforman el proceso de desidentificación en la BJN. Finalmente, la Sección 2.6 contiene una síntesis de este capítulo.

2.1. Desidentificación en la BJN

La desidentificación de datos personales en la Base de Jurisprudencia Nacional, según lo establecido por la Suprema Corte de Justicia en su Acordada Nro. 7564, busca «un equilibrio en la protección en el goce de los diversos derechos fundamentales reconocidos por la Constitución Nacional que resultan involucrados con motivo de la creación y difusión de bases de datos, en particular, el derecho a la información y la tutela del derecho a la intimidad». Aunque este trabajo se centra en la BJN, esta Acordada abarca también a todos aquellos datos personales y sensibles «asentados en bancos o bases de datos de carácter documental o jurisprudencial en todos los ámbitos del Poder Judicial y cualquiera sea el soporte que los contenga —papel o magnético—» (Suprema Corte de Justicia, República Oriental del Uruguay, 2006).

La Acordada Nro. 7564 menciona diversos tipos de datos sensibles a desidentificar, entre los que se encuentran:

- a. Datos filiatorios, como nombres y apellidos, documentos de identidad, nacionalidad, profesiones u ocupaciones, domicilios, números de teléfono, etc.
- b. Datos sensibles —salvo que se vinculen esencialmente a la causa— como el origen racial o étnico, las preferencias políticas, la afiliación sindical, la religión o la orientación sexual.
- c. Datos identificatorios de bienes inmuebles o muebles, como números de matrículas, números de padrón, direcciones, etc.

Se observa que algunos de estos datos que, en principio, no aparentan relevancia para la identificación de personas, cobran importancia bajo ciertas combinaciones o en ciertos contextos. Una profesión tiene, en ciertas circunstancias, la misma capacidad de identificación que un documento de identidad: no es igual ser un médico en una gran ciudad —uno entre miles— que en una pequeña localidad —posiblemente el único—. ¹

Es importante destacar que la Acordada alcanza a «las partes del proceso, terceros, testigos y toda otra persona que actúe en éste en calidad de auxiliar de la justicia», dejando por fuera a los nombres de jueces, fiscales y abogados intervinientes en el caso, así como a referentes del derecho y autores de estudios jurisprudenciales citados en la sentencia, que no son considerados datos sensibles.

Esto se evidencia en los fragmentos de la Sentencia 15/2022 del Tribunal de Apelaciones Penal de 1º Tº presentados en la Figura 2.1. En ellos se observa que, mientras el nombre del acusado es sustituido por siglas, los nombres de los jueces, fiscales, abogados y referentes se mantienen incambiados en el texto. Por su parte, la Figura 1.1 presenta otro pasaje de esta misma sentencia, en el cual se aprecia como también los nombres de víctimas y testigos son reemplazados por siglas.

La necesidad de desidentificación también depende del tipo de causa; a modo de ejemplo, en un juicio de divorcio los actores principales no se desidentifican, aunque sí se ocultan los nombres de los menores involucrados.

Al momento de desidentificar entidades sensibles en un texto, la solución más simple consiste, en el caso más general, en detectar todas aquellas menciones de estas entidades y eliminarlas completamente, sustituyéndolas por

¹Algunos ejemplos de reidentificación se pueden encontrar en el trabajo de Csányi et al. (2021).

VISTOS para definitiva de segunda instancia, en autos: “AA. HOMICIDIO ESPECIALMENTE AGRAVADO Y MUY ESPECIALMENTE AGRAVADO EN CONCURRENCIA FORMAL CON UN DELITO DE LESIONES PERSONALES EN CALIDAD DE AUTOR” (IUE: 573-3933/2018) venidos del Juzgado Letrado de Primera Instancia en lo Penal de 34º Turno, en virtud del recurso interpuesto por la Defensa Pública a cargo de la Dra. Virginia Ripoll y del Dr. Diego Moreira, contra la Sent. 195/2021 de 13.09.2021 dictada por la Dra. Blanca Rieiro, con intervención de la Fiscalía de Delitos Sexuales, Violencia Doméstica y Violencia Basada en Género de 2o T. a cargo del Dr. Raúl Iglesias y la Dra. Soledad Albarado, y de la Defensa de las víctimas Dra. Arabie Buroni.

(...)

Pero in dubio pro reo, que el ataque haya ocurrido de noche, parece contingente, esto es, no querida ni buscada en concreto, y esta agravante “exige un gran discernimiento en el Juez; estas circunstancias ontológicamente, resultan una veces indiferentes, otras atenuantes y algunas agravantes del hecho” (Irureta, cfm. Código Penal anotado por Reta y Grezzi, 2ª edición, p. 182).

En igual sentido: “Como se ve, queda a cargo del buen criterio del juez juzgar en qué casos pudo tener consecuencias graves, porque hay delitos en que importa muy poco que sea de noche o de día, incluso porque el lugar puede estar iluminado por luz artificial. Así es como se ha determinado por parte de nuestra jurisprudencia, que no cabe computar la agravante de nocturnidad porque la hora en que se produjo el episodio delictivo no tuvo particular incidencia en la producción del resultado. Tampoco debe computarse la nocturnidad cuando las facilidades de orden natural no fueron deliberadamente buscadas o aprovechadas por el agente para cometer el delito... quedando la apreciación librada a la discrecionalidad del juez” (Cairolí, ob. cit., pp. 452/3 y Sents. 152/98 y 35/04).

Figura 2.1: Fragmentos extraídos de la Sentencia 15/2022, del Tribunal de Apelaciones Penal de 1º Tº de Uruguay. El nombre del acusado se encuentra oculto (resaltado en rojo), mientras que los nombres de jueces, fiscales, abogados (en verde) y referentes (en celeste) permanecen incambiados.

alguna marca del estilo de «■■■■». Este procedimiento constituye, en definitiva, la variante digital de tachar las entidades sensibles en un documento impreso. Aunque efectivo al momento de ocultar todas las menciones, el texto resultante presenta el inconveniente de que los distintos actores del relato son indistinguibles entre sí, dificultando la correcta interpretación de la sentencia judicial.

A modo de ejemplo, se considera el siguiente fragmento de la sentencia 339/2015 del Tribunal de Apelaciones Penal de 2º Turno, que trata sobre un caso de abuso sexual.² En este fragmento se mencionan tres personas: al Sr. Juan Pérez, acusado del abuso, a la Sra. María Rodríguez, denunciante, y a la Sra. Juana Fernández, una tercera persona involucrada en el caso. En el Ejemplo 2.1 se observa el texto original, en donde las menciones a los nombres propios se destacan con distintos colores.

- (2.1) Que ninguno de los funcionarios de la dependencia, incluso el Comisario a cargo constataron u observaron algún tipo de conducta fuera de lugar del Sr. Juan Pérez. Que la denuncia fue realizada por la Sra. María Rodríguez que es quien lideraba la Cárcel y ordenaba a las demás (...) Que lo único admitido por el Sr. Pérez es que compró un chip a la Sra. Juana Fernández, pero no por eso se lo puede acusar de abuso y mucho menos de violación.

Al sustituir las menciones de personas por marcas, se obtiene como resultado el texto del Ejemplo 2.2. Este texto ya no conserva la información sobre la cantidad de actores y se pierde completamente la referencia a los distintos roles: ¿hay dos, tres o cuatro actores en este relato?, ¿es la señora denunciante la misma que le vendió el chip al señor acusado de abuso y violación?

- (2.2) Que ninguno de los funcionarios de la dependencia, incluso el Comisario a cargo constataron u observaron algún tipo de conducta fuera de lugar del Sr. ■■■■. Que la denuncia fue realizada por la Sra. ■■■■ que es quien lideraba la Cárcel y ordenaba a las demás (...) Que lo único admitido por el Sr. ■■■■ es que compró un chip a la Sra. ■■■■, pero no por eso se lo puede acusar de abuso y mucho menos de violación.

Para evitar la pérdida de la narrativa original, se sustituyen los nombres por otros muy comunes o de fantasía, o bien se asocia una etiqueta genérica

²Los nombres y apellidos de esta sentencia se sustituyen por otros de fantasía.

de forma unívoca a cada uno de los actores.³

Este último método es el utilizado en la BJN, según lo establecido en la Acordada Nro. 7564 previamente mencionada. En particular, en su artículo 6°, se establece que: «Para la remoción de los datos filiatorios se procederá a su sustitución por pares de letras tomadas del alfabeto (v. gr. AA, BB), y en caso de agotarse éstas, se recurrirá al empleo de tres de ellas (v. gr. AAA) o cuantas resulte necesario; manteniéndose en el caso de familiares la letra que representa el apellido (v. gr. BA, CA). Debiendo cuidarse en cada caso de no emplear letras que coincidan con las iniciales respectivas».⁴

Así, en el texto del Ejemplo 2.1, las menciones «Sr. Juan Pérez» y «Sr. Pérez» refieren a una misma persona, y a ambas se les debe asignar la misma etiqueta, como se muestra en el Ejemplo 2.3 donde se sustituyen por «AA».

- (2.3) Que ninguno de los funcionarios de la dependencia, incluso el Comisario a cargo constataron u observaron algún tipo de conducta fuera de lugar del Sr. AA. Que la denuncia fue realizada por la Sra. BB que es quien lideraba la Cárcel y ordenaba a las demás (...) Que lo único admitido por el Sr. AA es que compró un chip a la Sra. CC, pero no por eso se lo puede acusar de abuso y mucho menos de violación.

Como se aprecia en ese ejemplo, el resultado de esta sustitución mantiene el hilo de la narración original: aquí queda claro que hay tres actores distintos, AA, BB y CC; el señor AA está acusado de abuso y violación por parte de la

³Cuando el problema se extiende más allá de los nombres de personas y se consideran otras entidades nombradas —por ejemplo, nombres de organizaciones o instituciones—, otra opción es la microagregación, generalizando de acuerdo a alguna taxonomía previamente definida (Mamede et al. 2016). Así, «Universidad de la República» y «Liceo N°35» se sustituyen por «Universidad» y «Liceo», respectivamente, o directamente se reemplazan ambas por «Institución Educativa». Al igual que en el caso de los nombres propios, debe evitarse la pérdida de la semántica original del texto: en el último caso si «Universidad de la República» y «Liceo N°35» figuran en el mismo texto, la sustitución debería ser de la forma «Institución Educativa 1» e «Institución Educativa 2».

En este tipo de sustitución hay que tener en cuenta que, sin perder la coherencia o la semántica de la narrativa original, el grado de generalización del hiperónimo sea suficiente para lograr la desidentificación deseada. Se busca, entonces, un hiperónimo lo más específico posible pero que no permita la identificación de la entidad original, ya sea porque sigue siendo única en su especie, ya sea porque representa a pocas entidades y es posible la identificación mediante el cruzamiento con otros datos presentes en el texto.

Esta forma de sustitución no se restringe únicamente a texto libre, sino que también se aplica sobre atributos categoriales en información estructurada, por ejemplo, para generalizar sobre un tipo de enfermedad (Domingo-Ferrer et al. 2013).

⁴Durante el etiquetado del corpus se encuentra que, en general, no se mantiene constante la letra que representa al apellido entre familiares y las siglas son correlativas como en los otros casos.

señora BB, es AA quien compró un chip a la señora CC, quien a su vez no es la persona denunciante.

En contrapartida, el procedimiento de desidentificación se vuelve ahora más complejo: ya no alcanza con detectar y suprimir todos los nombres, sino que debe mantenerse la coherencia en la asignación de las nuevas etiquetas a los actores originales, lo que implica la resolución de correferencias, al menos entre las distintas variantes de nombres propios.

Esta tarea presenta diversas dificultades, que se acentúan cuando existen múltiples variantes de un mismo nombre, personas relacionadas, etc. Por ejemplo, en los fragmentos del Ejemplo 2.4, extraídos de la sentencia 99/2009 del Tribunal de Apelaciones de Familia de 2º Turno, se debe tener en cuenta que, para el proceso de sustitución, «GRAÑA COSTA, Pedro», «Pedro» y «Pedro Graña Costa» (fragmentos a, b, e y f) refieren a una misma persona, al igual que «Cornelio Graña» y «Cornelio» (fragmentos b y f) refieren a otra. Se observa, en particular, que en el fragmento b se infiere que el Pedro mencionado es de apellido Graña por el plural señores y la distribución de la conjunción en «Pedro y Cornelio», mientras que en el fragmento f se debe asumir que Pedro y Cornelio son los mismos hermanos Graña por ser los únicos mencionados con esos nombres de pila en todo el texto de la sentencia.

- (2.4) a. BRUNO TREZZA, Juan Líber c/ GRAÑA COSTA, Pedro y otros.
- b. Sres. Pedro y Cornelio Graña, deduce recursos de apelación.
- c. No puede considerarse que Pedro Graña ha omitido contestar la demanda,
- d. el Sr. Pedro Graña por haber precluido la oportunidad procesal para promoverla
- e. en que Pedro Graña Costa denunció que su domicilio era en India Muerta (fs. 30)
- f. Se intimó la aceptación de Pedro a fs. 32 vta. y a Cornelio a fs. 36/37 de los autos sucesorios.

Un posible resultado de la desidentificación para el ejemplo anterior se muestra en los fragmentos del Ejemplo 2.5.

- (2.5) a. AA c/ BB y otros.

- b. Sres. BB y CC, deduce recursos de apelación.
- c. No puede considerarse que BB ha omitido contestar la demanda,
- d. el Sr. BB por haber precluido la oportunidad procesal para promoverla
- e. en que BB denunció que su domicilio era en India Muerta (fs. 30)
- f. Se intimó la aceptación de BB a fs. 32 vta. y a CC a fs. 36/37 de los autos sucesorios.

En síntesis, si bien la Acordada Nro. 7564 establece la desidentificación de diversos tipos de datos sensibles, este trabajo se centra específicamente en los nombres de personas, debido a su importancia y frecuencia. Para ello, es necesario distinguir entre los nombres propios que deben ser ocultados —acusados, víctimas, testigos, etc.— y aquellos que deben permanecer incambiados en el texto —jueces, fiscales, abogados y referentes del ámbito jurídico—. El proceso de desidentificación no consiste solamente en la detección de las entidades nombradas, sino también en la resolución de las correferencias entre sus distintas menciones a fin de garantizar una sustitución que no altere el sentido del relato.

2.2. Trabajos Relacionados

Existen distintas estrategias para la desidentificación de datos sensibles, cuya elección depende tanto del tipo de información como de la forma en que se encuentre estructurada.

Cuando los datos se almacenan en bases de datos relacionales, se suelen aplicar algoritmos que *k-anonimizan* la información: los valores de los campos que permiten identificar a los individuos —por sí mismos o en combinación con otros— son reemplazados o eliminados de forma que cada registro resulta indistinguible de al menos otras $(k - 1)$ tuplas, reduciendo la probabilidad de identificación individual a $\frac{1}{k}$. Los problemas en estos procesos suelen estar en la calidad de representación de los datos resultantes versus los originales, la seguridad de la desidentificación lograda y la eficiencia computacional de los métodos aplicados (Bayardo y Agrawal, 2005; Domingo-Ferrer et al. 2013; Heatherly et al. 2016; Karle y Vora, 2017; Patil et al. 2017; X. Zhang et al. 2014).

Así, por ejemplo, el sistema HIDE (Gardner et al. 2008) procesa registros médicos en una base relacional, eliminando el número de identificación de paciente y ofuscando la edad y el código postal transformando esos valores a rangos, lo que permite llevar a la información al nivel de k-anonimización deseado.

El problema es, en cierto modo, diferente cuando la información aparece en texto libre o, incluso, en imágenes (Bayardo y Agrawal, 2005; Dernoncourt et al. 2016; Domingo-Ferrer et al. 2013; Friedrich et al. 2019; Gardner et al. 2008; Heatherly et al. 2016; Karle y Vora, 2017; Y. Li et al. 2018; Newhauser et al. 2014; Patil et al. 2017; X. Zhang et al. 2014). De hecho, nombres sensibles de personas, empresas u hospitales pueden estar contenidos en una narración que no siempre es gramaticalmente correcta —como también sucede con las sentencias de la BJN—, lo que incrementa la complejidad del problema. En el caso de textos legales, además, la información sensible contenida en el texto puede involucrar a varias personas, a diferencia de los textos médicos, que suelen centrarse en un solo paciente (Csányi et al. 2021).

La desidentificación de textos es investigada desde la década de 1990, especialmente en el ámbito de los registros médicos, donde la confidencialidad se encuentra regulada por ley (Centers for Medicare & Medicaid Services, United States of America, 1996; Parlamento Europeo y Consejo de la Unión Europea, 2016). Según la literatura relevada, las soluciones se enfocan en resolver el problema de reconocimiento de información sensible para, posteriormente, ofuscarla o directamente suprimirla. De esta forma, el problema de desidentificación está estrechamente ligado a la tarea de Reconocimiento de Entidades.

Los distintos enfoques siguen de algún modo la progresión general de otras soluciones dentro del área del PLN (Dernoncourt et al. 2016): desde los primeros sistemas desarrollados en los años noventa, basados meramente en diccionarios, expresiones regulares y reglas definidas manualmente por expertos del dominio, que alcanzan alta precisión pero baja cobertura (Gupta et al. 2004; Sweeney, 1996), pasando por las soluciones de los años 2000 con métodos de aprendizaje automático como los campos aleatorios condicionales (*Conditional Random Fields*, CRF) o las máquinas de vectores de soporte (*Support Vector Machines*, SVM) basadas en ingeniería de atributos (Aramaki et al. 2006; Gardner et al. 2008; Krishnan y Manning, 2006; Uzuner et al. 2008), hasta las propuestas de la última década, que utilizan redes neuronales profundas (*Deep Neural Networks*) y grandes modelos de lenguaje (*Large Language Mo-*

deels, LLMs) (Chalapathy et al. 2016; Jia et al. 2019; Lample et al. 2016; J. Y. Lee et al. 2016; J. Li et al. 2016; Z. Liu et al. 2017; Seow et al. 2025; Sharma et al. 2021). Muchas de las soluciones son híbridas: se combinan distintos métodos de aprendizaje automático, para aumentar la cobertura, con reglas que mejoran la precisión del resultado obtenido con los métodos anteriores (Carvalho Dias, 2016; Z. Liu et al. 2015; Phuong y Chau, 2016; Unanue et al. 2017; Yadav et al. 2016).

En particular, la desidentificación de textos del dominio legal es explorada en algunos trabajos recientes, que reportan experimentos en el área o bien discuten los problemas y tecnologías involucrados (Csányi et al. 2021; Glaser et al. 2021; Oksanen et al. 2019). Por ejemplo, Anoppi es un prototipo creado para la pseudonimización semiautomática de sentencias judiciales finlandesas (Oksanen et al. 2019). Utilizando tanto métodos estadísticos como de reglas para el reconocimiento de entidades nombradas y análisis morfológico —particularmente, el módulo NER de CoreNLP—, un analizador sintáctico por dependencias y expresiones regulares creadas por los autores. El servicio aún se encuentra en desarrollo al momento de la publicación, y no se dispone de información detallada sobre los procesos involucrados. Glaser et al. (2021) proponen tratar por separado el reconocimiento de entidades nombradas y la determinación de si una entidad debe o no ser desidentificada. Para lo primero, utilizan Spacy como reconocedor de entidades nombradas, y para lo segundo, un clasificador binario basado en BERT (Devlin et al. 2019).

Cabe señalar que, a diferencia del presente trabajo, ninguno de los artículos mencionados aborda el problema de asignar automáticamente una misma etiqueta a todas las menciones que refieren a una misma entidad, lo cual, como se menciona en la sección anterior, resulta crucial para la publicación en la BJN.

El problema de desidentificación de nombres en la BJN, sí es el objeto del trabajo de Cruces y Ferraro (2023). Ese trabajo, basado en el de esta tesis, propone el uso de técnicas de aprendizaje profundo para NER y la resolución de correferencias. El trabajo se desarrolla sobre el corpus B*N, presentado en la Sección 2.3, pero extendido con nuevas entidades. El mejor resultado en NER lo obtienen reentrenando al modelo BETO (Cañete et al. 2020) obteniendo una medida F_1 de 0.911. Para la resolución de correferencias, utilizan el mismo enfoque de agrupamiento (*clustering*) que se presenta en la Sección 2.5, pero entrenan una red neuronal siamesa con el fin de calcular la similitud entre

menciones, logrando un 0.882 de medida ARI.⁵

Los resultados para NER varían según el idioma y el corpus de entrenamiento y aplicación. Yang et al. (2024) reportan 0.945 de medida F_1 para un modelo CRF-BiLSTM, en el que se combinan características a nivel de carácter y de palabras; mientras que Baevski et al. (2019) logran con el entrenamiento de una CNN más un ajuste fino un F_1 de 0.935, ambos para el inglés sobre el mismo corpus de CoNLL-2003. Para el español, Y. Jiang et al. (2021) aplican aprendizaje por refuerzo para aprender la mejor combinación de vectores para usar como entrada en la tarea, y reporta 0.956 de F_1 . El proyecto Maria (Gutiérrez-Fandiño et al. 2022) se entrenan varios modelos basados en RoBERTa (Y. Liu et al. 2019), obteniendo un F_1 de 0.885, superando al modelo BETO (Cañete et al. 2020) preentrenado, que alcanza un F_1 de 0.884.

El uso de técnicas de aprendizaje automático y, más específicamente, de agrupamiento para la resolución de correferencias es explorado ya desde fines del siglo pasado (Cardie y Wagstaff, 1999; Ma et al. 2016; Mitra et al. 2003; Song et al. 2012; Soon et al. 2001; J. Zheng et al. 2011). Al igual que en otras tareas, la llegada de los modelos de lenguaje basados en transformadores (*Transformers*, Vaswani et al. 2017) ha impulsado el desarrollo de nuevos enfoques para la resolución de correferencias (Bohnet et al. 2023; Joshi et al. 2020; Martinelli et al. 2024; Wu et al. 2020).

SpanBERT es un modelo de lenguaje que extiende BERT para predecir segmentos de textos (Joshi et al. 2020). Este modelo alcanza en el momento de su publicación el mejor valor reportado para OntoNotes (Pradhan et al. 2012), con una medida F_1 de 0.796. Este trabajo sirve como base a otros modelos posteriores, como el de Wu et al. (2020), que proponen un mecanismo pregunta-respuesta, para capturar las relaciones entre menciones, logrando un F_1 de 0.831 sobre OntoNotes.

En el trabajo de Bohnet et al. (2023) se entrena un modelo *seq2seq*, que, en múltiples llamadas, va tomando como entrada una oración junto con una secuencia que codifica el contexto previo, y genera como salida transiciones que alteran los agrupamientos. Este método obtiene un resultado de 0.833 de F_1 promedio para inglés sobre el mismo corpus.

Martinelli et al. (2024) presentan un sistema de resolución de correferencias en etapas basado en mT5 (Xue et al. 2021). Luego de extraer menciones del

⁵El lector interesado puede encontrar una definición de F_1 y ARI en las figuras 2.4 y 2.10, respectivamente.

texto, entrenan redes neuronales para decidir si dos entidades pertenecen a la misma referencia. La mejor variante, *Maverick_{mes}*, utiliza seis modelos distintos basados en categorías gramaticales —por ejemplo, para el caso en que una es un pronombre y otro un nombre propio— y reporta un F_1 promedio de 0.836 sobre OntoNotes. Los autores destacan que *Maverick* puede obtener mejores resultados y con una mayor eficiencia que modelos generativos o de solo codificación (*encoder-only*), como los citados previamente.

Todos los modelos mencionados abordan la resolución de correferencias en su forma más general, considerando distintos tipos de menciones, como los pronombres. En contraste, el presente trabajo se centra exclusivamente en la resolución de correferencias entre nombres propios de personas, lo que permite adoptar métodos más simples y eficientes en términos computacionales, adecuados para su integración en un prototipo de desidentificación.

2.3. Corpus B*N

Para el presente trabajo se dispone aproximadamente un total de 79 000 sentencias como corpus de trabajo, que representan una fracción de la Base de Jurisprudencia Nacional.

Cada sentencia tiene una extensión promedio de tres páginas —unos 10 000 caracteres—. Los textos presentan elementos de enumeración, con paréntesis, numerales y abreviaturas de diversos tipos, e incluyen con frecuencia referencias a otras sentencias, leyes o artículos legales. Todos estos elementos hacen un uso intensivo de los signos de puntuación —especialmente del punto—, lo que dificulta la tarea de segmentación del texto, como se observa en los ejemplos de la Figura 1.3. En los textos aparecen distintos tipos de entidades nombradas: personas, ubicaciones geográficas y organizaciones. Como se menciona anteriormente, en este trabajo se toman en cuenta las referencias a personas y, de ellas, el subconjunto de las que deben ser desidentificadas.

Aproximadamente 17 000 sentencias se encuentran desidentificadas manualmente; esto es, se dispone de alrededor de 17 000 documentos para los cuales se cuenta tanto con una versión original como con una versión desidentificada. Esto constituye, *a priori*, un corpus paralelo de documentos pre y post desidentificación, susceptible de ser aprovechado para entrenar modelos de aprendizaje automático o evaluar herramientas existentes, al estilo del corpus *Acquis Communautaire* (Steinberger et al. 2006) utilizado en traducción automática.

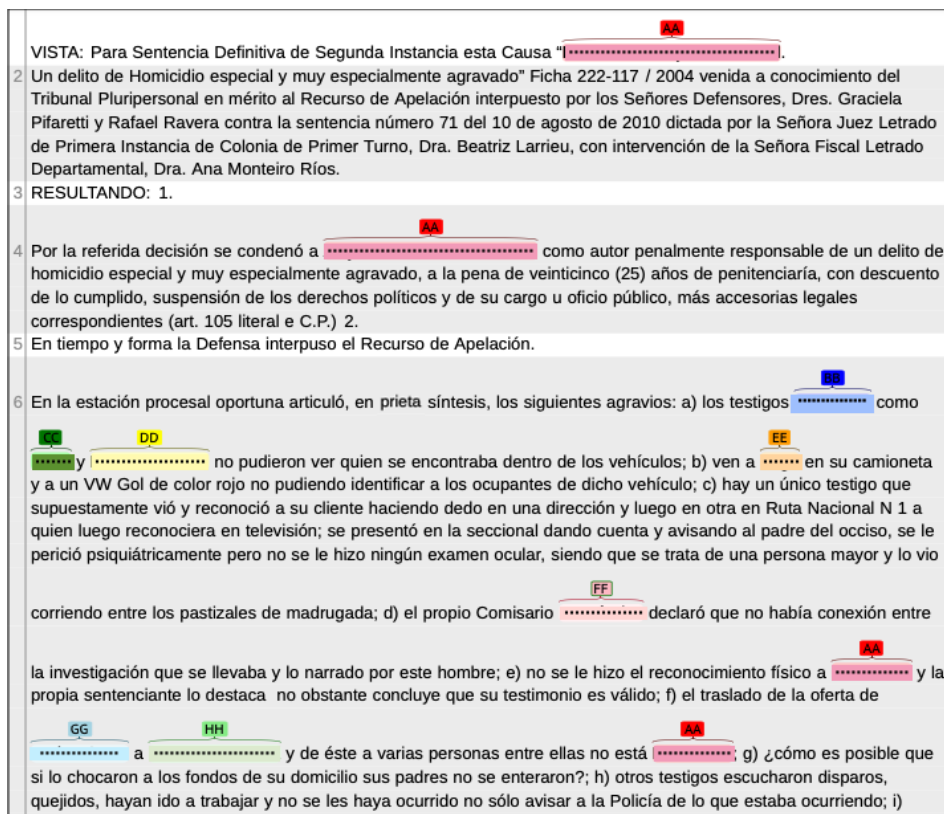


Figura 2.2: Captura de la herramienta BRAT utilizada para la corrección manual de las etiquetas en el corpus B*N. Las etiquetas asignadas son visibles sobre los nombres, los cuales se encuentran ofuscados por razones de confidencialidad.

Lamentablemente, se observa que muchos de estos documentos presentan fallas en su proceso de desidentificación: existen nombres que no fueron marcados como tales en el texto, y etiquetas asignadas de manera inconsistente. A esto se suma que los documentos son parcialmente reescritos durante el proceso de edición, con secciones completamente modificadas o incluso eliminadas.

Los problemas anteriores imposibilitan la construcción automática de ejemplos a partir solamente de las diferencias entre la versión original y la desidentificada. Por ello, se implementa un proceso que, mediante el alineamiento de documentos, preetiqueta las entidades nombradas en los textos originales con la etiqueta del documento desidentificado.

Luego, se procede a la revisión y corrección manual de 1000 de estos documentos preetiquetados, con la ayuda de la herramienta BRAT (Stenetorp et al. 2012). En la Figura 2.2 se puede observar la interfaz de la herramienta durante el proceso de corrección.

T1	AA	81	102	FERNANDEZ, JUAN PABLO
T2	AA	728	742	Juan FERNANDEZ
T3	BB	2262	2272	María José
T4	AA	2491	2500	Fernández
T5	CC	2634	2649	Jorge Hernández

Figura 2.3: Archivo de anotación en BRAT. Cada línea contiene una entidad T_i , su etiqueta, sus posiciones de comienzo y fin en el texto original, y el texto del segmento en sí mismo. En este caso, hay cinco menciones a personas marcadas en el texto, de las cuales T_1 , T_2 y T_4 refieren a la misma persona, mientras las otras dos refieren a otras dos personas distintas.

Entre las correcciones realizadas se encuentran: (a) el ajuste del largo de los segmentos etiquetados; (b) la eliminación de etiquetas asignadas a segmentos incorrectos del texto; (c) la adición de segmentos con sus etiquetas faltantes; y (d) la corrección de las etiquetas asignadas erróneamente. Las correcciones (a) y (b) son debidas generalmente a errores en el proceso de alineamiento, mientras que las correcciones (c) y (d) se deben a fallas en el proceso manual de desidentificación del corpus original.

Luego de la revisión de 1000 documentos se obtiene en un conjunto de 10 102 entidades etiquetadas, distribuidas en 997 documentos.⁶ El corpus resultante, denominado B*N, se compone entonces de pares de textos originales en archivos planos —uno por sentencia—, junto con sus correspondientes archivos de anotaciones en formato BRAT.

Un ejemplo de archivo de anotaciones se presenta en la Figura 2.3. En este ejemplo, entre las posiciones 728 y 742 del texto, se encuentra la mención «T2», etiquetada como «AA», correspondiente a «Juan FERNANDEZ». Las posiciones permiten ubicar el segmento dentro del texto original y, a partir de ello, entrenar y evaluar modelos de reconocimiento de entidades nombradas. Por otra parte, las etiquetas asociadas permiten identificar referencias a las distintas personas, y sirven como insumo para la resolución de correferencias.

2.4. Reconocimiento de Entidades

Como se menciona en la Sección 2.2, el reconocimiento de entidades nombradas suele ser un primer paso en los sistemas de desidentificación de texto

⁶Tres documentos fueron descartados por no contener nombres a desidentificar.

	Oraciones	Tokens	Orac./Doc	Token/Orac.
CoreNLP	1 020 555	29 815 769	58.82	29.22
Freeling	494 306	28 989 140	32.27	58.65
NLTK	973 208	28 711 789	56.09	29.50
SpaCy	995 055	30 294 823	57.35	30.45

Tabla 2.1: Segmentación en oraciones y *tokens* de 17 350 documentos de la BJN realizada por CoreNLP, Freeling, NLTK y SpaCy.

libre. Con el objetivo de evaluar distintas herramientas PLN disponibles —y en particular sus reconocedores de entidades nombradas— se seleccionan y aplican algunas de estas herramientas sobre el conjunto total de documentos: CoreNLP (Manning et al. 2014), Freeling (Padró y Stanilovsky, 2012), SpaCy (Honnibal y Montani, 2017), NLTK (Bird et al. 2009).

Desafortunadamente, todas las herramientas probadas se comportan de manera deficiente, causado probablemente por la diferencia significativa en el estilo y vocabulario de las sentencias judiciales con respecto a los documentos en los que estos analizadores fueron originalmente entrenados —artículos periodísticos y páginas de Wikipedia, entre otros—.

Un caso representativo es, desde una perspectiva cuantitativa, los resultados de la segmentación en oraciones y en *tokens*⁷ que se presentan en la Tabla 2.1. Freeling parece constituir un caso atípico, al menos desde el punto de vista de la segmentación, ya que detecta aproximadamente la mitad de las oraciones que los demás analizadores.

Tras una inspección detallada de los resultados, se observa que todos ellos presentan un bajo rendimiento en esta tarea. La Tabla 2.2 muestra ejemplos de un mismo documento en donde las oraciones son segmentadas incorrectamente por NLTK, Freeling y CoreNLP. Se observa que los puntos utilizados en abreviaturas u ordinales son manejados incorrectamente por las distintas herramientas, como en «7o.» (oración nro. 44), «Sr.» (oración nro. 56) o «pág.» (oración nro. 108). Desde el punto de vista cuantitativo, el número de oraciones detectadas difiere notablemente entre analizadores: NLTK detecta 136 oraciones, mientras que CoreNLP encuentra 87 y Freeling tan solo 47.

⁷La segmentación en *tokens* es frecuentemente una etapa inicial en el procesamiento de texto. Consiste en convertir secuencias de caracteres en elementos —*tokens*— de un vocabulario (Jurafsky y Martin, 2025).

CoreNLP

Entrada:

Como lo plantea el Sr. Fiscal de Corte a fs.183 v., si bien atendiéndonos. . .

Salida:

- [56] Como lo plantea el Sr.
- [57] Fiscal de Corte a fs.
- [58] 183 v., si bien atendiéndonos. . .

Freeling

Entrada:

En el mismo sentido se pronunció el Tribunal en lo Civil de 7o. Turno, cuando en Sentencia No. 134/98 sostuvo que. . .

Salida:

- [44] En el mismo sentido se pronunció el Tribunal en lo Civil de 7o.
- [45] Turno, cuando en Sentencia No. 134/98 sostuvo que. . .

NLTK

Entrada:

Nicolás Coviello, Doctrina General del Derecho Civil, pág. 78, Cf. Eduardo García Maynez, Introducción al estudio del derecho, pág. 329. . .

Salida:

- [108] Nicolás Coviello, Doctrina General del Derecho Civil, pág.
- [109] 78, Cf.
- [110] Eduardo García Maynez, Introducción al estudio del derecho, pág.
- [111] 329. . .

Tabla 2.2: Ejemplos de segmentación de oraciones por NLTK, CoreNLP y Freeling. Cada grupo de segmentos (44–45, 56–58 y 108–111) pertenecen a una misma oración del texto original, a pesar de que los segmentadores las dividan incorrectamente.

El bajo rendimiento no afecta solamente a la segmentación, sino también al reconocimiento de entidades nombradas. La Figura 1.3 muestra un fragmento del análisis de los módulos NER de CoreNLP, Freeling y SpaCy sobre un documento de la BJN. En este ejemplo, existen dos nombres a ser desidentificados —«Pérez Cabrera, María José» y «Martínez Mágico, Adolfo Ramón»—, pero ninguno de los analizadores logra marcarlos como entidades de persona. De hecho, solo se los reconoce parcialmente y, en algunos casos, se les asigna un tipo de entidad incorrecto (organización). El resto de las entidades detectadas tampoco son completamente correctas, presentando errores en su ubicación, extensión de los segmentos o en su categorización.

2.4.1. NER en el Corpus B*N

Todas las herramientas probadas, por tanto, muestran un muy bajo rendimiento sobre el corpus de trabajo, debido a, como se menciona anteriormente, la marcada diferencia en el estilo y vocabulario respecto de los textos con los que fueron originalmente entrenadas. En particular, la evaluación de SpaCy, utilizando su mejor modelo para español disponible al momento de realizar la experimentación, arroja un puntaje de 0.400 de F_1 sobre el conjunto de evaluación. En contraste, los autores de SpaCy reportan un F_1 de 0.899 para el reconocimiento de entidades de persona en su corpus de trabajo original.

Vistos estos resultados, se plantea el objetivo de crear un modelo de reconocimiento de entidades adaptado al dominio legal. Para ello, se procede a entrenar un reconocedor de NER, utilizando el corpus B*N presentado en la Sección 2.3. Los 997 documentos se dividen aleatoriamente en dos conjuntos: 797 documentos se asignan al conjunto de entrenamiento —empleado para ajustar los distintos modelos—, lo que corresponde a 7748 instancias anotadas, y los 200 documentos restantes conforman el conjunto de evaluación, con 2220 instancias.

En lugar de construir una herramienta propia desde cero, se opta por reentrenar el módulo NER de SpaCy. Este modelo se basa en una Red Neuronal Convolutiva (*Convolutional Neural Network*, CNN; Kim, 2014; Lecun et al. 1998) y permite cargar un modelo previamente entrenado y ajustar sus pesos con nuevos ejemplos, habilitando así el aprendizaje por transferencia a partir de modelos ya entrenados. Además, el reentrenamiento se puede efectuar por módulo sin la necesidad de modificar o reentrenar el resto de los módulos pre-

vios en la cadena de procesamiento, como lo es, en este caso, el etiquetado de categorías gramaticales (*POS Tagging*).

Como se menciona en la Sección 2.1, no todos los nombres de personas deben ser desidentificados, por ejemplo, los nombres de jueces, abogados o referentes. Dado que estas menciones no están etiquetadas en el corpus, el modelo de algún modo puede «desaprender» a reconocer ciertos nombres que, en una configuración clásica de NER orientada a la detección exhaustiva de entidades, deberían ser marcados como «Persona» («PER»), aunque no tengan interés para la tarea de desidentificación. Una alternativa más completa sería distinguir también esas entidades, como hacen Cruces y Ferraro (2023), pero debido a limitaciones de recursos no se implementa en esta etapa experimental.

Cabe destacar que, en el enfoque propuesto —y a diferencia de, por ejemplo, Glaser et al. (2021)— el modelo aprende no solo a reconocer las entidades «Persona», sino también, de forma conjunta, cuáles de todas las entidades deben ser desidentificadas: en el corpus la etiqueta a ajustar se asigna únicamente a las entidades que deben ser desidentificadas, mientras que las demás permanecen sin etiquetar.

2.4.2. Experimentación

Todos los experimentos comienzan con el modelo más grande disponible para español (*es_core_news_md*), entrenado sobre los corpus AnCora y WikiNER (Pan et al. 2017; Taulé et al. 2008), y que contiene 20 000 vectores únicos de 50 dimensiones. Este modelo puede realizar la división en *tokens*, el etiquetado de categorías gramaticales, análisis de dependencias y el reconocimiento de entidades nombradas de tipo persona (PER), ubicación (LOC), organización (ORG) y miscelánea (MISC). La medida F_1 reportada para la etiqueta PER es de 0.897 para ese modelo, evaluado sobre el corpus WikiNER.

Se diseñan diferentes experimentos, variando principalmente tres factores: (a) la tasa de *dropout*; (b) reiniciar o mantener los pesos del módulo NER antes del reentrenamiento; y (c) la etiqueta para marcar las entidades a desidentificar. El resto de los parámetros se mantienen con sus valores por defecto, correspondientes a aquellos que obtuvieron los mejores resultados sobre el conjunto de entrenamiento original. En la Tabla 2.3 se resumen los parámetros y valores con los que se experimentan.

La tasa de *dropout* regula la probabilidad de que características individuales

Parámetro	Valor
Etiqueta	Reentrenamiento de PER o una nueva etiqueta PER-X.
Reiniciar NER	Sí/No.
<i>Dropout</i>	0.10, 0.20, 0.30, 0.35, 0.40, 0.50, 0.60.
Parada	Un máximo de 500 rondas o 20 rondas sin mejora de rendimiento.

Tabla 2.3: Parámetros y valores experimentados para el reentrenamiento del módulo NER de SpaCy.

y representaciones internas sean «apagadas» durante el entrenamiento; esta técnica ayuda a evitar que el modelo memorice los ejemplos y sobreajuste al conjunto de entrenamiento (Srivastava et al. 2014). Aunque este valor puede variar en el rango $[0, 1]$, usualmente se establece en 0.5 o un valor menor. Se prueban valores que van de 0.1 a 0.6, y, una vez fijado, el valor se mantiene constante en todas las rondas de un mismo entrenamiento.

Por otra parte, SpaCy permite iniciar el entrenamiento a partir de un modelo previamente entrenado, o reiniciar los valores del modelo y comenzar desde cero. En el primer caso, se transfiere el conocimiento aprendido de un entrenamiento previo —y sobre un corpus diferente—, y se ajusta el modelo a los nuevos ejemplos provenientes del corpus actual. Esta configuración permite analizar el impacto del conocimiento previo en el ajuste al nuevo dominio, frente a un modelo NER entrenado desde cero.

Finalmente, se consideran dos opciones para etiquetar las entidades a desidentificar en los ejemplos del corpus B*N: (a) la etiqueta clásica para persona, «PER», preentrenada en el modelo de SpaCy; o (b) un nuevo subtipo de esta etiqueta, «PER-X». En este último caso, «PER» sigue siendo una posible etiqueta del módulo NER, aunque no se encuentren ejemplos presentes en el nuevo corpus de entrenamiento.

Cada entrenamiento es, entonces, una combinación de etiqueta, reinicio del modelo y tasa de *dropout*. Además, en cada entrenamiento, el modelo se ajusta por rondas⁸, hasta alcanzar veinte rondas consecutivas sin que se supere al mejor desempeño. Adicionalmente, como experimento suplementario, se en-

⁸En cada ronda se utiliza el conjunto de entrenamiento completo para ajustar al modelo.

Experimento	Etiqueta	Reiniciar NER	<i>Dropout</i>	Criterio de Parada
PER-X	PER	No	0.10, 0.20,	20 rondas sin mejora de rendimiento
PER-0		Sí	0.30, 0.35,	
PER		No	0.40, 0.50 y 0.60	
PER-500			0.35 y 0.40	

Tabla 2.4: Resumen de los experimentos para el reentrenamiento del módulo NER de SpaCy con el corpus B*N.

$$P = \frac{VP}{VP + FP} \quad R = \frac{VP}{VP + FN} \quad F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Figura 2.4: Definiciones de precisión (P), cobertura (R) y medida F_1 ; VP , FP y FN representan verdaderos positivos, falsos positivos y falsos negativos, respectivamente (Jurafsky y Martin, 2025).

trenan las dos mejores configuraciones por un máximo de 500 épocas. Estas dos configuraciones corresponden al experimento PER con tasas de *dropout* de 0.35 y 0.40, como se detalla más adelante.

La combinación de todos estos factores da lugar a los experimentos resumidos en la Tabla 2.4.

El desempeño de los modelos se evalúa sobre el conjunto de evaluación mediante las métricas de precisión (P), cobertura (*recall*, R) y medida F_1 , definidas en la Figura 2.4. En el contexto de múltiples documentos, se utiliza la variante *micro*, que se obtiene calculando los valores globales de TP , FP y FN sobre el conjunto total de instancias.⁹ Además, se reporta el acierto por documento (*Document Accuracy*, Acc_{DOC}), que representa la proporción de documentos en los que todas sus entidades son correctamente reconocidas.

2.4.3. Resultados

En la Figura 2.5 se presentan los valores de F_1 obtenidos para los distintos experimentos al variar la tasa de *dropout*. Los resultados para los modelos con

⁹Alternativamente, podría utilizarse la variante *macro*, que promedia los valores de F_1 obtenidos para cada documento. Sin embargo, dado que los documentos de la BBN presentan una gran variabilidad en la cantidad de entidades a desidentificar, esta métrica podría verse sesgada por aquellas sentencias con pocas entidades a desidentificar.

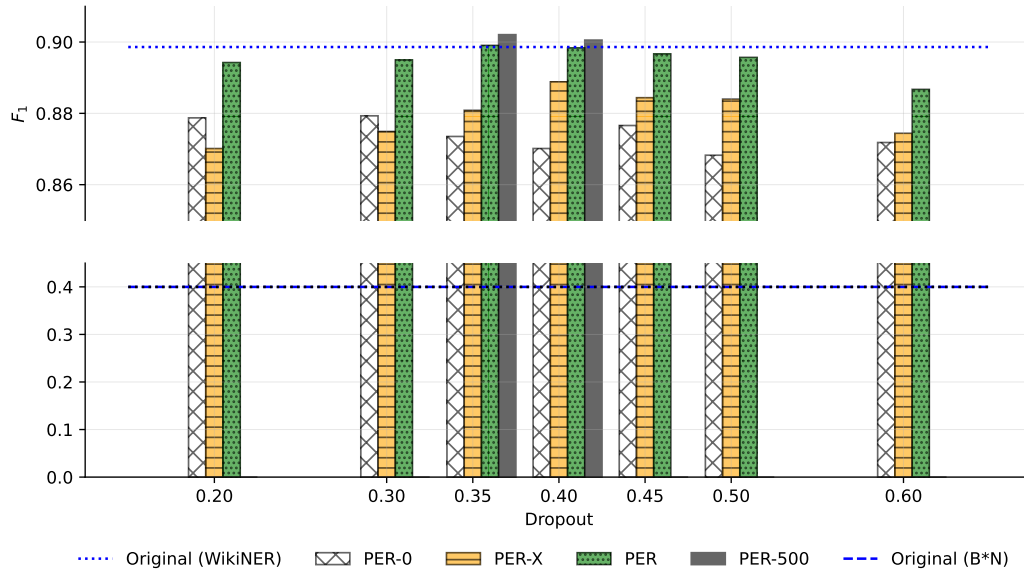


Figura 2.5: Valores de F_1 al variar la tasa de *dropout* para los distintos experimentos.

los mejores valores de F_1 se muestran en la Tabla 2.5.

Los resultados obtenidos permiten observar el efecto diferencial de la tasa de *dropout* y de la elección de etiqueta en el rendimiento de cada experimento.

Sin considerar al experimento PER-500, se observa que los resultados del experimento PER siempre superan a los resultados de PER-X y PER-0, sin importar la tasa de *dropout* utilizada. Los mejores valores de F_1 se obtienen con tasas de 0.35 y 0.40: 0.899 y 0.898 en 80 y 75 iteraciones, respectivamente. Estos valores representan una mejora de casi 0.5 puntos sobre el F_1 de 0.400 obtenido con el modelo antes del reentrenamiento.

En el experimento final, PER-500, se repiten los experimentos con las dos mejores configuraciones, pero modificando el criterio de parada a un número fijo de 500 épocas: estos corresponden al experimento PER con tasas de *dropout* de 0.35 y 0.40. Los modelos resultantes alcanzan un F_1 de hasta 0.902 y 0.901 en 420 y 185 épocas, respectivamente.

La Figura 2.6 muestra el entrenamiento para la tasa de *dropout* 0.35, donde se observa que los modelos se ajustan muy rápidamente y la ganancia comienza a ser marginal después de aproximadamente 100 iteraciones.

El mejor modelo obtenido alcanza un F_1 de 0.902, un valor ligeramente superior al 0.899 reportado originalmente por los autores de SpaCy para su

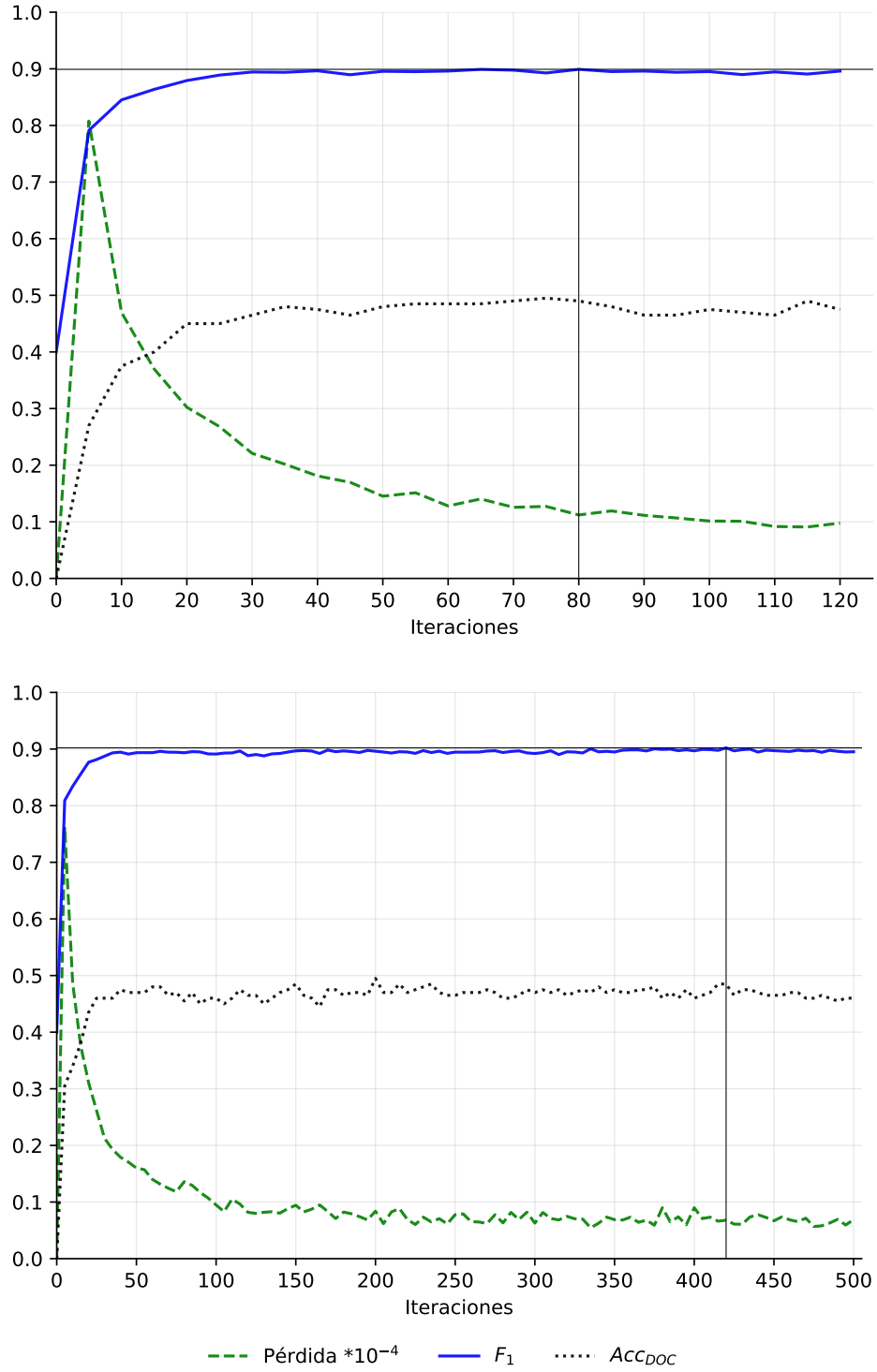


Figura 2.6: Experimentos PER y PER-500 con tasa de *dropout* de 0.35 para 120 (arriba) y 500 épocas (abajo). Los mejores valores de F_1 son 0.899 en la ronda 80 y 0.902 en la ronda 420, respectivamente.

Experimento	<i>Dropout</i>	Iteración	<i>Acc_{DOC}</i>	<i>P</i>	<i>R</i>	<i>F₁</i>
PER-500	0.35	420	<u>0.485</u>	0.899	0.905	0.902
PER	0.35	80	0.490	0.879	0.920	<u>0.899</u>
Original WikiNER	—	—	—	<u>0.899</u>	0.896	0.897
PER-X	0.40	205	0.450	0.885	0.892	0.889
PER-0	0.30	60	0.390	0.852	<u>0.909</u>	0.879
Original B*N	—	—	0.000	0.300	0.599	0.400

Tabla 2.5: Mejores resultados sobre el conjunto de evaluación para cada experimento según F_1 . Se agregan los resultados en el mismo conjunto del modelo sin reentrenar junto al resultado reportado por los autores de SpaCy sobre su corpus de evaluación, WikiNER. Los mejores resultados se indican en negrita, mientras que los segundos mejores resultados se muestran subrayados.

módulo NER sobre el corpus WikiNER, pero levemente por encima de los 0.5 puntos sobre el F_1 de 0.400 que logra ese mismo modelo en el corpus B*N antes de ser reentrenado.

Dada la especificidad de la tarea —reconocer solo un subconjunto de entidades de tipo PER, en un dominio específico y en español— es difícil comparar este modelo con otros sistemas de NER, salvo para los resultados reportados por Cruces y Ferraro (2023) sobre el mismo corpus de la BJN. Los resultados se encuentran 0.054 puntos por debajo de los 0.956 de medida F_1 reportado por Y. Jiang et al. (2021) y a 0.011 puntos de los 0.911 de Cruces y Ferraro (2023), pero 0.017 puntos por encima de los 0.885 de Gutiérrez-Fandiño et al. (2022).

En conclusión, el reentrenamiento del módulo NER de SpaCy mejora significativamente el rendimiento con el corpus B*N, especialmente en el etiquetado de entidades a desidentificar. El mejor modelo alcanza un F_1 de 0.902 —considerado satisfactorio para la tarea de desidentificación propuesta, dada la complejidad del dominio y el número de ejemplos anotados—, superando ampliamente los 0.400 obtenidos por el modelo sin reentrenar. Estos resultados confirman la importancia de un ajuste específico sobre el corpus objetivo para un desempeño adecuado en tareas como NER sobre textos jurídicos.

Entrada: Conjunto de datos $X = \{x_1, x_2, \dots, x_n\}$; función de distancia $d(x_i, x_j)$; función de enlace $linkage(K_i, K_j)$; umbral μ .

Salida: Conjunto de agrupamientos $K = \{K_1, \dots, K_j\}$

1: Inicializar cada punto de datos como un agrupamiento individual:

$$K \leftarrow \{\{x_1\}, \{x_2\}, \dots, \{x_n\}\}$$

2: Calcular la matriz de distancias inicial D entre todos los agrupamientos:

$$D(K_i, K_j) = linkage(K_i, K_j), \quad \forall K_i, K_j \in K$$

3: **Mientras** $\exists K_i, K_j \in K, i \neq j, D(K_i, K_j) \leq \mu$ **hacer**

4: Encontrar un par (K_p, K_q) , $p \neq q$, a mínima distancia.

5: Fusionar los agrupamientos: $K_m = K_p \cup K_q$

6: Actualizar el conjunto de agrupamientos K :

$$K \leftarrow (K \setminus \{K_p, K_q\}) \cup \{K_m\}$$

7: Recalcular la matriz de distancias D para el nuevo K .

8: **Fin Mientras**

9: **Retornar** K

Figura 2.7: Pseudocódigo del algoritmo Agrupamiento Aglomerativo con criterio de parada por umbral de distancia μ .

2.5. Resolución de Correferencias

Como segundo paso en el proceso de desidentificación, las entidades deben ser reemplazadas por etiquetas de manera que todas las referencias a una misma persona reciban la misma y única etiqueta. Esta tarea es una versión simplificada de la resolución de correferencias, en la que solo se consideran nombres propios, mientras que otros fenómenos, por ejemplo pronombres, son ignorados debido a su escaso impacto en la reidentificación de las personas involucradas.

Dado que únicamente se consideran nombres propios —la mayoría de las entidades son variaciones de un mismo nombre: «Pérez Rodríguez, Pedro», «Pedro Pérez» y «Pedro», en el Ejemplo 2.4 de la Sección 2.1—, y la cantidad de entidades a agrupar en cada documento es relativamente pequeña y desconocida, se decide experimentar con un algoritmo no supervisado, agrupamiento aglomerativo (*Agglomerative Clustering*; Everitt et al. 2011).

2.5.1. Agrupamiento de Nombres

El agrupamiento aglomerativo es un algoritmo de agrupamiento jerárquico que fusiona de manera iterativa los grupos más cercanos. El algoritmo utiliza una estrategia *bottom-up*: comienza considerando cada elemento de una entrada X como un conjunto individual y, en cada iteración, fusiona aquellos dos agrupamientos que presentan la menor distancia entre sí, reduciendo en uno la cantidad total de agrupamientos.

El proceso continúa hasta que se obtiene un único conjunto —esto es, todos los elementos pertenecen a una misma clase—, o hasta alcanzar un criterio de parada previamente establecido. Estos criterios son: (a) una cantidad k de agrupamientos, o (b) todos los agrupamientos están a una distancia superior a un umbral μ entre sí.

En la Figura 2.7 se presenta un esquema del algoritmo de agrupamiento aglomerativo con este último criterio de parada.

Para determinar los agrupamientos a fusionar, el algoritmo requiere que se utilicen una función de distancia d y una función de enlace *linkage*. La función d calcula la distancia entre dos instancias del conjunto de entrada X . Por otro lado, la función de enlace *linkage* calcula la distancia entre dos agrupamientos K_i y K_j a partir de las distancias entre las instancias que los componen.

Existen varias formas de definir esta función, como la distancia mínima, máxima o promedio entre las instancias de ambos agrupamientos. Por ejemplo, la distancia promedio *linkage_{avg}* se define en el Ejemplo 2.6.

(2.6)

$$linkage_{avg}(K_i, K_j) = \frac{1}{|K_i| \cdot |K_j|} \sum_{x_p \in K_i} \sum_{x_q \in K_j} d(x_p, x_q)$$

El resultado del algoritmo es interpretable como un dendrograma, un árbol jerárquico que muestra las fusiones realizadas por el algoritmo. En la Figura 2.8 se presenta un ejemplo de dendrograma para un conjunto de cinco instancias. El umbral de distancia μ se representa como una línea horizontal punteada; los agrupamientos que se encuentran por debajo de esta línea se consideran parte del mismo grupo final.

En el ejemplo planteado, para un valor de umbral μ_1 de 0.8, se generan dos agrupamientos finales: «Rodríguez» y «Rodríguez, María» en un conjunto y «Juan Pedro», «Pérez, Juan» y «Juan» en otro. Por otra parte, para un

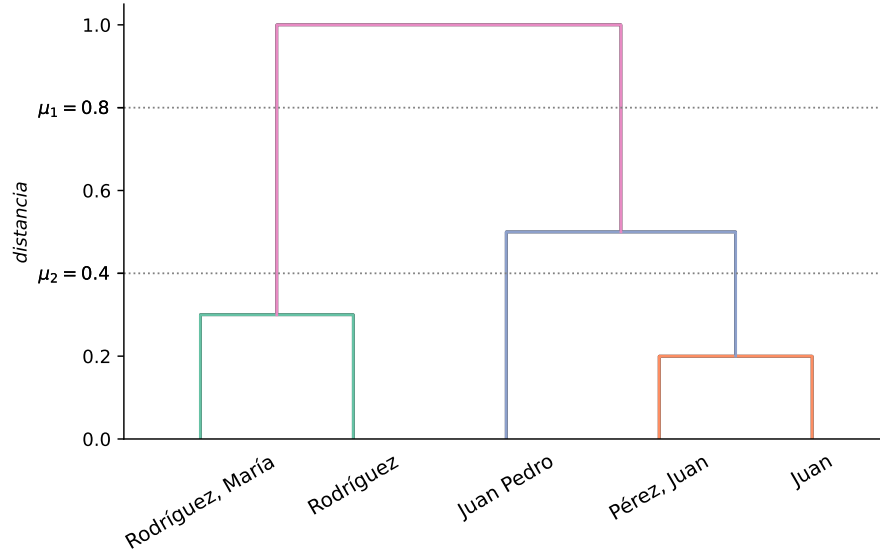


Figura 2.8: Dendrograma resultante de aplicar el Agrupamiento Aglomerativo hasta llegar a un único agrupamiento. Se muestran las fusiones de agrupamientos para cinco nombres de entrada y dos posibles umbrales, μ_1 y μ_2 , que generan dos y tres agrupamientos finales, respectivamente.

umbral μ_2 de 0.4 se obtienen tres agrupamientos, por la separación del segundo en «Juan Pedro» por un lado y «Pérez, Juan» y «Juan» por otro.

Para la tarea de desidentificación, se espera que el algoritmo agrupe correctamente las distintas menciones a una misma persona, mientras que las menciones a personas diferentes queden en agrupamientos separados. Umbrales altos tienden a generar pocos agrupamientos —lo que se traduce en que, posiblemente, distintas personas queden desidentificadas bajo una misma etiqueta—, mientras que umbrales bajos producen soluciones más fragmentadas —y que menciones a una misma persona terminen siendo sustituidas por distintas etiquetas—.

Se experimenta, entonces, con diferentes configuraciones de umbrales y funciones de distancia para evaluar el rendimiento del algoritmo, utilizando el corpus B*N. Se toma la misma división en conjuntos de entrenamiento y evaluación del reentrenamiento del módulo NER de SpaCy, presentado en la Sección 2.4.

El corpus de entrenamiento, en este caso, permite determinar cuáles combinaciones de funciones de distancia y umbral generan los mejores agrupamientos de salida. Luego, el conjunto de evaluación se utiliza para estimar el rendimien-

Parámetro	Valor
Distancia	Overlap, Jaccard, Sørensen, Bag, LCSSTR, Hamming, Goto-H, Bag, Entropy-NCD, Strcmp95, Jaro-Winkler, Smith-Waterman, LCSSEQ, Postfix, Needleman-Wunsch, Ratcliff-Obershelp, Levenshtein, Damerau-Levenshtein
Umbral	Desde 0.05 hasta 0.95 con paso de 0.05
Vínculo	<i>linkage_{avg}</i> (otras opciones fueron descartadas en tests preliminares)

Tabla 2.6: Parámetros del experimento del entrenamiento de Agrupamiento Aglomerativo.

to de las configuraciones de función y umbral encontradas.

2.5.2. Experimentación

Para la experimentación, se utiliza la implementación del algoritmo provista por la biblioteca Scikit-Learn (Pedregosa et al. 2011). Además de una función de enlace, esta implementación recibe una matriz de distancias entre todas las entidades involucradas, en lugar tomar directamente una función de distancia. Se prueba además con distintos valores de umbrales para generar los agrupamientos de salida. En la Tabla 2.6 se resumen los parámetros experimentados.

En cuanto a la función de enlace, en pruebas preliminares se observa que sistemáticamente se obtienen los mejores resultados con el promedio de distancias entre agrupamientos —la función *linkage_{avg}*, definida en el Ejemplo 2.6—. Se fija este parámetro en ese valor para el resto de los experimentos.¹⁰

Por otra parte, en cuanto a la función de distancia d para el cálculo de la matriz de distancias, se consideran las funciones implementadas en el proyecto TextDistance.¹¹ Se prueban 18 distancias o pseudo-distancias de esta biblioteca. En la Figura 2.9 se presentan las definiciones de las cuatro distancias con mejores resultados, como se detalla más adelante en la Sección 2.5.3.¹²

¹⁰Otras opciones descartadas incluyen la distancia máxima o mínima entre observaciones de los agrupamientos o la varianza en los agrupamientos fusionados.

¹¹La biblioteca se encuentra disponible en: <https://pypi.org/project/textdistance/>

¹²En las primeras pruebas se utiliza una función de distancia propia, basada en heurísticas

$$d_{\text{Bag}}(A, B) = 1 - \max(|\text{Bag}(A) - \text{Bag}(B)|, |\text{Bag}(B) - \text{Bag}(A)|)$$

$$d_{\text{Jaccard}}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

$$d_{\text{Overlap}}(A, B) = 1 - \frac{|A \cap B|}{\min(|A|, |B|)}$$

$$d_{\text{Sørensen}}(A, B) = 1 - \frac{2|A \cap B|}{|A| + |B|}$$

Figura 2.9: Definiciones de las distancias basadas en las funciones de similitud *Bag*, *Jaccard*, *Overlap* y *Sørensen*, donde $\text{Bag}(X)$ representa la bolsa de palabras de la entrada X .

Los nombres se preprocesan para mejorar los resultados. Durante este proceso, todas las letras se transforman a minúsculas, se eliminan acentos especiales y los nombres transpuestos se transforman a su forma canónica. Por ejemplo, las menciones «BRUNO TREZZA, Juan Líber» y «Cornelio Graña» del Ejemplo 2.4 se transforman en «juan liber bruno trezza» y «cornelio graña», respectivamente. Las secuencias de nombres se dividen en *tokens* utilizando espacios en blanco como separadores.

Aunque existen varias métricas de desempeño propuestas para el problema general de resolución de correferencias, no hay consenso sobre cuál es la más adecuada. Entre las métricas más antiguas se encuentran MUC y B^3 (Bagga y Baldwin, 1998; Vilain et al. 1995), aunque, debido a sus debilidades, se proponen varias alternativas como CEAF, BLANC y LEA (Luo, 2005; Luo et al. 2014; Márquez et al. 2013; Moosavi y Strube, 2016).

Dado que en la experimentación se aplica un algoritmo de *clustering*, en este trabajo se opta, en cambio, por métricas usuales para evaluar el rendimiento en estos casos: homogeneidad (H), completitud (C), medida V (V_1) y el índice de Rand Ajustado (*Adjusted Rand Index*, ARI), utilizando esta última como métrica objetivo (Hubert y Arabie, 1985; Rosenberg y Hirschberg, 2007). La Figura 2.10 presenta las definiciones de estas cuatro métricas.¹³ Además, al

para determinar el nombre y apellido y comparar sus coincidencias por separado, que se descarta rápidamente por sus bajos rendimientos en comparación con las distancias estándar.

¹³Homogeneidad, completitud y V_1 son análogas a las medidas de precisión, cobertura y F_1 empleadas en NER, pero adaptadas al contexto de *clustering*. La métrica H evalúa en qué medida cada grupo contiene únicamente miembros de una misma clase, mientras que C mide si todos los miembros de una clase están asignados al mismo grupo. La medida V_1 combina H y C en su media armónica, al igual que la medida F_1 lo hace con P y R . Por su parte,

$$H = 1 - \frac{E(T|K)}{E(T)}, \quad C = 1 - \frac{E(K|T)}{E(K)}, \quad V_1 = 2 \cdot \frac{H \cdot C}{H + C}$$

$$ARI = \frac{\sum_{T_i \in T} \sum_{K_j \in K} \binom{|T_i \cap K_j|}{2} - \frac{\sum_{T_i \in T} \binom{|T_i|}{2} \sum_{K_j \in K} \binom{|K_j|}{2}}{\binom{N}{2}}}{\frac{1}{2} \left[\sum_{T_i \in T} \binom{|T_i|}{2} + \sum_{K_j \in K} \binom{|K_j|}{2} \right] - \frac{\sum_{T_i \in T} \binom{|T_i|}{2} \sum_{K_j \in K} \binom{|K_j|}{2}}{\binom{N}{2}}}$$

En donde:

$$E(A) = - \sum_{A_i \in A} \frac{|A_i|}{N} \log \frac{|A_i|}{N}, \quad E(A|B) = - \sum_{B_j \in B} \sum_{A_i \in A} \frac{|A_i \cap B_j|}{N} \log \frac{|A_i \cap B_j|}{|B_j|}$$

Figura 2.10: Definiciones de homogeneidad (H), cobertura (C), V_1 y ARI . Las clases reales y las estimadas se representan con T y K , respectivamente; N es el número total de elementos.

igual que en la tarea NER, se utiliza la métrica Doc_{ACC} , que representa la proporción de sentencias en las que todas sus entidades son correctamente agrupadas.

El algoritmo se aplica para cada sentencia de manera independiente, que es el nivel de granularidad de la evaluación. Las métricas generales en este caso son *macro*, y se calculan como el promedio de los valores obtenidos para las distintas métricas en cada sentencia.

2.5.3. Resultados

El desempeño del algoritmo de *clustering* varía según la distancia seleccionada y el umbral de fusión aplicado. Los mejores valores de ARI en el conjunto de entrenamiento se obtienen para las distancias *Overlap*, *Jaccard*, *Sørensen* y *Bag* en los umbrales 0.45, 0.75, 0.60 y 0.85 respectivamente.

En la Figura 2.11 se muestra el desempeño en el conjunto de entrenamiento al variar el umbral de fusión según la métrica ARI para estas cuatro distancias. En la Figura 2.12 se muestran en detalle las métricas para las dos mejores, *Overlap* y *Jaccard*.

la métrica ARI cuantifica la similitud entre dos particiones, ajustando el valor en función de la probabilidad de coincidencias aleatorias; su rango varía entre -1 (total desacuerdo) y 1 (perfecto acuerdo).

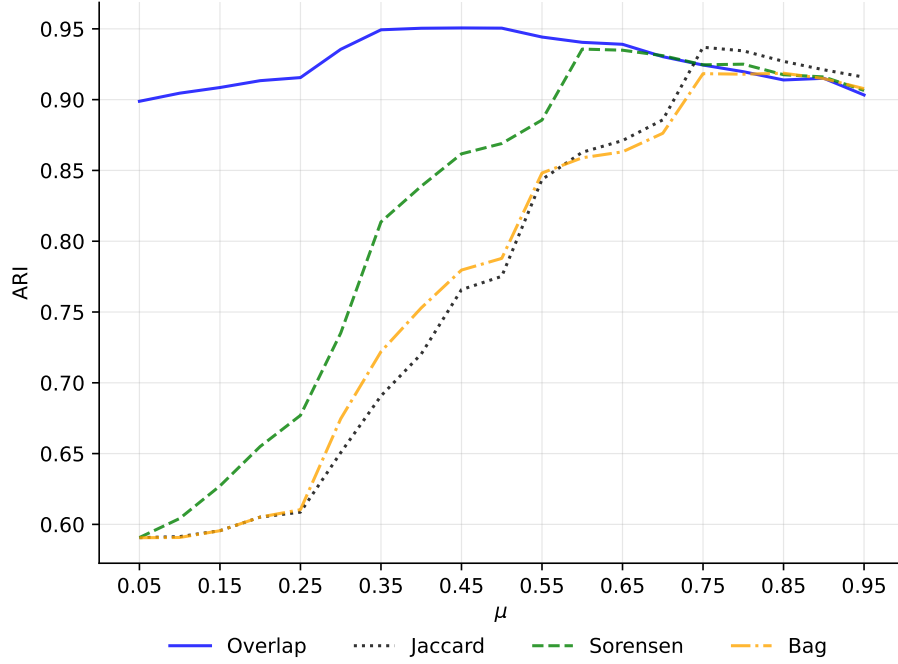


Figura 2.11: *ARI* para las distancias *Overlap*, *Jaccard*, *Sørensen* y *Bag* por valor de umbral en el conjunto de entrenamiento.

Tras determinar el mejor umbral para cada distancia, se evalúan sus desempeños sobre las 2220 referencias de nombres del conjunto de evaluación. En la Tabla 2.7 se muestran los cuatro mejores resultados según *ARI*.

La combinación de la distancia *Overlap* con un umbral de fusión de 0.45 obtiene los mejores resultados en todas las métricas de desempeño —salvo en completitud, en donde queda segunda— con un *ARI* promedio de 0.960, resolviendo correctamente las correferencias en 162 de los 200 documentos del conjunto de evaluación. Esto sugiere que la distancia *Overlap* con un umbral moderado permite agrupar de manera consistente distintas menciones de la misma persona sin fusionar identidades diferentes.

Comparando con (Cruces y Ferraro, 2023) —que trabaja con el mismo enfoque sobre la BJN, pero entrenando una red siamesa para calcular la distancia entre nombres—, *Overlap*-0.45 supera ampliamente los 0.882 de *ARI* y los 0.669 de acierto por documento de esa solución. En general, estos resultados muestran que *Overlap*-0.45 proporciona un desempeño consistente, con una performance superior y un costo de entrenamiento y ejecución menor que soluciones más complejas, ofreciendo así una alternativa simple y eficiente para

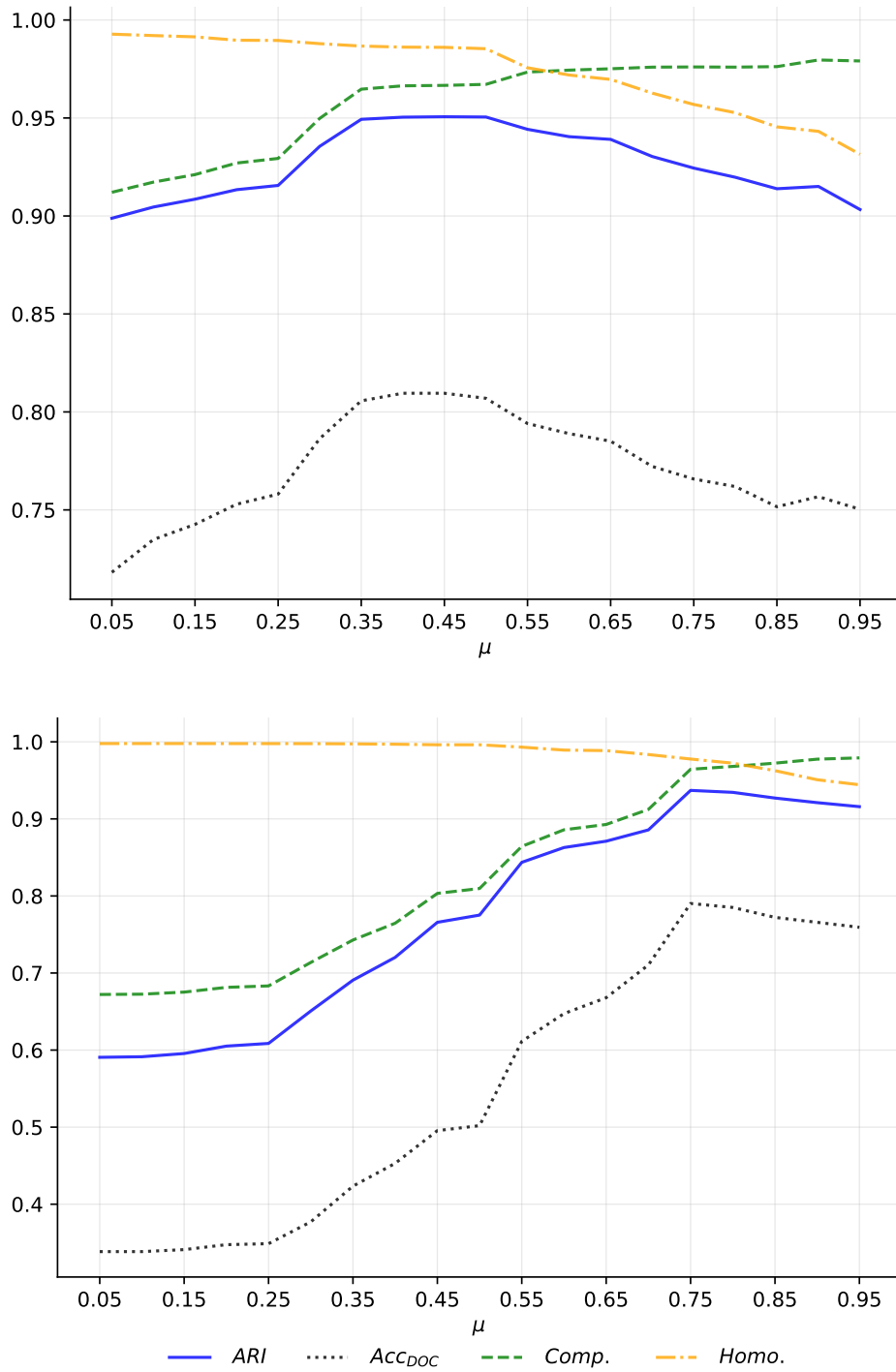


Figura 2.12: ARI , Acc_{DOC} , C y H en el conjunto de entrenamiento para *Overlap* (arriba) y *Jaccard* (abajo) por valor de umbral.

Distancia	μ	Acc_{DOC}	H	C	V_1	ARI
Overlap	0.45	0.810	0.986	<u>0.967</u>	0.970	0.951
Jaccard	0.75	<u>0.790</u>	0.978	0.964	<u>0.961</u>	<u>0.937</u>
Sørensen	0.60	0.786	<u>0.978</u>	0.963	0.960	0.936
Bag	0.85	0.758	0.953	0.972	0.949	0.919

Tabla 2.7: Resultados de las cuatro mejores soluciones distancia- μ según ARI sobre el conjunto de evaluación.

la resolución de esta tarea.

2.6. Síntesis

Este capítulo aborda el primero de los objetivos planteados en esta tesis: la desidentificación automática en sentencias de la Base de Jurisprudencia Nacional del Uruguay. En particular, la investigación se enfoca en la tarea de desidentificar los nombres propios de las personas involucradas en las sentencias, sin afectar la coherencia narrativa de los hechos. El resultado de este trabajo se encuentra publicado en la revista arbitrada «Information» (Garat y Wonsever, 2022).

Para el desarrollo y evaluación de los modelos automáticos que resuelvan esta tarea, se construye un corpus a partir de 997 sentencias judiciales de la BJN. Este corpus, denominado B*N, contiene 10 102 menciones a nombres propios, anotadas manualmente. Este corpus constituye en sí mismo una contribución relevante al ser el primer recurso anotado basado en la BJN.

La solución propuesta se organiza en dos grandes etapas: (a) el reconocimiento de entidades nombradas para identificar las menciones de nombres a desidentificar en el texto, y (b) la resolución de correferencias para agrupar las distintas menciones que refieren a una misma persona.

En la primera etapa, se reentrena el módulo NER de la biblioteca SpaCy utilizando el corpus B*N. Se experimenta con distintas configuraciones de tasa de *dropout* y criterios de parada, logrando un modelo con un F_1 de 0.902 en el corpus B*N, lo que representa una mejora de 0.50 en el desempeño del modelo sin reentrenar y superando el valor reportado para ese mismo modelo en su corpus de entrenamiento.

Para resolver las correferencias, se experimenta con un algoritmo de agrupa-

miento aglomerativo, y a partir del corpus B*N se estiman la mejor distancia y su umbral de fusión. Los mejores resultados se obtienen con la distancia *Overlap*, con un umbral de fusión de 0.45, alcanzando un *ARI* de 0.960 y una precisión por documento de 0.81.

Dada la sensibilidad de los datos a publicar, solo una solución completamente precisa permitiría convertir el proceso de desidentificación en uno totalmente automático. Sin embargo, los resultados obtenidos son prometedores y permiten considerar la posibilidad de un proceso semiautomático, en el que un operador humano revise y corrija las salidas del sistema automático. Cabe mencionar que el resultado se integra en un prototipo funcional para el proyecto «Curación Automática de sentencias de la Base de Jurisprudencia Nacional» del Fondo Sectorial de Investigación a partir de Datos de la ANII-MIEM.

En conjunto, este capítulo demuestra la viabilidad técnica y metodológica de abordar la desidentificación automática de nombres en textos judiciales, con soluciones fácilmente integrables gracias a su base en herramientas de código abierto y bajo costo de ejecución. Los resultados obtenidos sientan las bases para futuras investigaciones y desarrollos orientados a mejorar el procesamiento automático y la desidentificación en la Base de Jurisprudencia Nacional.

Capítulo 3

Análisis de marcos semánticos

La extracción de eventos y asignación de roles semánticos es una tarea de larga data en el campo del Procesamiento del lenguaje natural (PLN) y la Extracción de información (*Information Extraction*, IE). Consiste en identificar eventos específicos en un texto, junto con sus participantes y circunstancias asociadas, respondiendo preguntas como ¿quién hizo qué a quién, cuándo y dónde? (Jurafsky y Martin, 2025).

En esta tesis se plantea, en el contexto de la Base de Jurisprudencia Nacional (BJN), la extracción automática de eventos a partir de sentencias judiciales. El resultado de esta tarea puede proporcionar información valiosa para diversas aplicaciones, como el análisis estadístico de delitos, la identificación de patrones legales o la mejora de los sistemas de búsqueda, entre otras.

El marco teórico elegido para representar los eventos es la Teoría de Marcos Semánticos (*Frame Semantics Theory*; Fillmore, 1976). Su autor introduce el concepto de *framing* como «*the appeal, in perceiving, thinking, and communicating, to structured ways of interpreting experiences*».¹ Fillmore propone que las palabras están ligadas a estructuras cognitivas denominadas «marcos» (*frames*), que representan situaciones prototípicas que involucran eventos, participantes, ubicaciones y tiempos. Ciertas palabras, entonces, «evocan» marcos en la mente del lector o hablante, cuyos componentes sintácticos suelen desempeñar roles semánticos específicos dentro de esos marcos (C. F. Baker et al. 1998).

Un ejemplo ampliamente utilizado es el de una transacción comercial, en la que típicamente se presenta una situación de compra y venta con sus elementos

¹El *framing* consiste en recurrir —al percibir, pensar y comunicar— a estructuras que organizan la interpretación de las experiencias.

lingüísticos asociados: «comprar», que es un verbo que presenta la transacción desde el punto de vista del comprador, y «vender», que lo hace desde la perspectiva del vendedor. Además, intervienen una cantidad de dinero, un tipo de pago y bienes que cambian de propietario.

El análisis de marcos semánticos (*Frame Semantic Parsing*, FSP) adopta la Semántica de Marcos para la extracción de eventos bajo esta teoría: busca la identificación y extracción de los marcos, junto con sus elementos, en un texto dado (C. F. Baker et al. 1998; Gildea y Jurafsky, 2002; Ruppenhofer et al. 2016).

En particular, este trabajo explora la aplicación de técnicas de aprendizaje en contexto (*In-Context Learning*, ICL) para la resolución del Análisis de Marcos. ICL propone utilizar un gran modelo de lenguaje (*Large Language Model*, LLM) para resolver tareas particulares, sin necesidad de modificar su arquitectura ni ajustar sus parámetros (Brown et al. 2020; Radford et al. 2019; Wei et al. 2024). En su lugar, se guía al modelo proporcionándole, como parte de su entrada —el *prompt*—, una descripción de la tarea y ejemplos de entrada con su salida correspondiente. En el caso de FSP, los ejemplos son textos con sus marcos asociados.

Dado que, al momento de comenzar este trabajo, no se encuentran antecedentes del uso de ICL para la tarea de FSP, el primer paso es investigar su viabilidad y rendimiento. Con este fin, se decide aprovechar la base de marcos del Proyecto FrameNet, que define más de 1220 marcos semánticos y contiene más de 200 000 ejemplos anotados manualmente (C. F. Baker et al. 1998).

En síntesis, en esta primera fase del trabajo se investiga la efectividad de aplicar ICL a la tarea de FSP. En particular, se explora si un LLM puede aprovechar el conocimiento obtenido durante su entrenamiento para llevar a cabo esta tarea y cómo su rendimiento se ve influido por la cantidad de ejemplos proporcionados. Las pruebas se basan exclusivamente en las definiciones y anotaciones del proyecto FrameNet, y se evalúan utilizando modelos Claude (Anthropic), GPT (OpenAI) y DeepSeek (Anthropic, 2024; DeepSeek-AI et al. 2025; OpenAI, 2024). En una segunda fase, presentada en el Capítulo 4, se desarrolla la extracción de eventos de violencia en la BJN basada en los resultados obtenidos en esta etapa.

La organización de este capítulo es la siguiente: en la Sección 3.1 se presenta el Proyecto FrameNet. La Sección 3.2 revisa los trabajos relacionados con FSP y el uso de ICL en PLN. La Sección 3.3 describe la propuesta de aplicar

ICL a FSP, detallando las tareas abordadas para la generación automática de *prompts* a partir de FrameNet. Finalmente, en la Sección 3.4 se presenta una síntesis del trabajo desarrollado en este capítulo.

3.1. FrameNet

El proyecto FrameNet², desarrollado en el *International Computer Science Institute* en Berkeley, define más de 1220 marcos semánticos para el inglés contemporáneo (C. F. Baker et al. 1998; Ruppenhofer et al. 2016). En la base de marcos de este proyecto existen más de 200 000 ejemplos anotados manualmente basados en la Teoría de Marcos Semánticos.

Aunque existen proyectos basados en FrameNet para otras lenguas, e incluso un proyecto de *Multilingual FrameNet* que vincula marcos entre diferentes lenguas, sus alcances son limitados (C. F. Baker y Lorenzi, 2020).

A diferencia de otros recursos semánticos como PropBank (Palmer et al. 2005) —que se centra en los roles semánticos de los verbos individuales, utilizando etiquetas numéricas como «Arg0», «Arg1», etc., para identificarlos—, FrameNet organiza los marcos semánticos en torno a situaciones prototípicas y sus elementos, lo que le permite capturar relaciones semánticas más generales y conceptuales. Esto resulta fundamental para el objetivo de extracción de eventos en la BJN y es una de las razones por las que se selecciona FrameNet como marco teórico para este trabajo.

En el proyecto FrameNet, cada marco se encuentra descrito por una definición y por sus posibles elementos (*Frame Element*, FE). Los elementos del marco se dividen en *centrales* y *periféricos*: los FE centrales son esenciales para el significado del marco, mientras que los otros son generalmente circunstanciales.

Además, existe un atributo especial llamado «Objetivo» (*Target*), que señala la presencia misma del marco en la oración. El «Objetivo» está representado por una unidad léxica (*Lexical Unit*, LU), que es la palabra o expresión que *evoca* el marco en el texto, y puede ser tanto un verbo, un adjetivo, un grupo nominal, etc. Hay más de 13 600 unidades léxicas definidas en FrameNet, cada una asociada a uno o más marcos semánticos.

Por ejemplo, el marco «Asesinato»³ (*Killing*) se define como «A *Killer* or

²Disponible en: <https://framenet.icsi.berkeley.edu/>

³Como se detalla más adelante, este es uno de los marcos utilizados durante la ex-

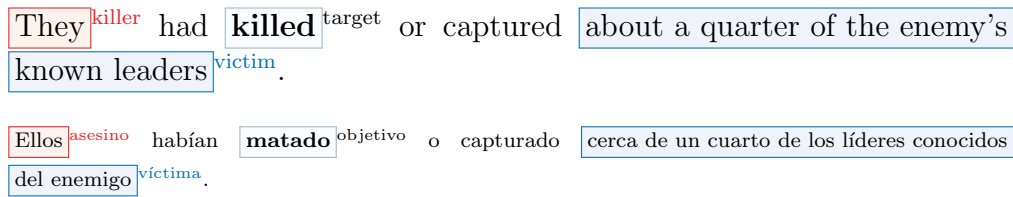


Figura 3.1: Ejemplo extraído de FrameNet para el marco «Asesinato» (*Killing*), con unidad léxica «matar» (*kill*). Se marcan los elementos: «Asesino», «Víctima» —resaltados en azul y rojo, respectivamente—, y el «Objetivo», en negrita.

*Cause causes the death of the Victim*⁴. Los elementos centrales para este marco son los siguientes:

- Asesino (*Killer*): La persona o entidad sintiente que causa la muerte de la víctima.
- Causa (*Cause*): Una entidad inanimada o un proceso que provoca la muerte.
- Víctima (*Victim*): La entidad viviente que muere como resultado del acto de matar.
- Instrumento (*Instrument*): El dispositivo utilizado por el agresor.
- Medio (*Means*): El método o acción que el agente homicida emplea para provocar la muerte.

Dentro de los elementos del marco periféricos del marco «Asesinato» se incluyen entidades como el beneficiario, el tiempo y el lugar.

Algunas posibles unidades léxicas asociadas al «Objetivo» de este marco son los verbos «matar» (*kill*), «masacrar» (*slaughter*) y «asesinar» (*murder*), los sustantivos «decapitación» (*decapitation*), «homicidio» (*homicide*) y «asesinato en masa» (*mass murder*), y el adjetivo «mortal» (*deadly*), entre otros.

La Figura 3.1 presenta un ejemplo extraído de FrameNet correspondiente al marco «Asesinato», evocado por la unidad léxica «asesinado» (*killed*). De todos los posibles elementos de este marco, solo figuran «Asesino» y «Víctima». A

perimentación y es de particular interés considerando la extracción, en una fase posterior, de eventos de violencia en la BJN. Una definición parcial de este marco en español e inglés se presenta en el Anexo I. La versión completa en inglés se encuentra disponible en: <https://framenet.icsi.berkeley.edu/frames/Killing>.

⁴Un asesino o causa provoca la muerte de la víctima.

They^{Arg0} had killed^{kill.01} or captured about a quarter of the enemy's known leaders^{Arg1}.

Figura 3.2: Anotación según PropBank para el Ejemplo 3.1. Los roles «Arg0» y «Arg1» corresponden a los roles semánticos de «Asesino» y «Víctima» de FrameNet. El marco «kill.01» indica al verbo «kill» en su primer sentido: causar la muerte, asesinar.

modo de comparación, la Figura 3.2 ilustra la anotación correspondiente según PropBank para el mismo ejemplo.

Los ejemplos en la base de datos de FrameNet están clasificados por unidad léxica, y cada ejemplo consiste en una única oración. Aunque una oración puede contener varias ocurrencias de un mismo marco o de marcos diferentes, en cada ejemplo de FrameNet se marca exclusivamente una de ellas. En la Figura 3.1 se observa que, por ejemplo, a pesar de que «capturado» (*captured*) evoca a otro marco, no se encuentra identificado en ese caso.⁵

Además, FrameNet organiza los marcos en una estructura relacional, donde algunos marcos heredan propiedades de otros o están conectados mediante relaciones semánticas, lo que permite capturar conexiones conceptuales más complejas entre eventos. Por ejemplo, el marco «Asesinato» hereda del marco «Acción Transitiva» (*Transitive_action*), es ancestro del marco «Ejecución» (*Execution*), y se encuentra relacionado con los marcos «Supervivencia» (*Surviving*) y «Muerte» (*Death*).

En resumen, FrameNet proporciona un marco estructurado para representar situaciones prototípicas, lo que permite identificar de manera sistemática los participantes y las circunstancias de los eventos. El volumen de ejemplos anotados en FrameNet resulta particularmente útil para el desarrollo de soluciones basadas en aprendizaje en contexto, en las que se construyen demostraciones que sirven de guía a los modelos de lenguaje en la resolución de la tarea. Gracias a estas características, FrameNet resulta un recurso adecuado para la extracción automática de información con ICL, con el objetivo final de su aplicación en la BJN.

⁵En ese ejemplo, el marco faltante es «Tomar cautivo» (*Taking captive*). Disponible en: https://framenet.icsi.berkeley.edu/frames/Taking_captive

3.2. Trabajos Relacionados

El análisis de marcos semánticos (*Frame Semantic Parsing*, FSP) tiene como objetivo detectar y extraer las estructuras definidas por Fillmore (1976) en su Teoría de Marcos Semánticos.

El proceso de FSP suele dividirse en dos subtareas: la identificación de marcos (*Frame Identification*, FI) y el etiquetado de roles semánticos (*Frame Semantic Role Labeling*, FSRL). La tarea FI se ocupa de detectar la presencia y el tipo de marco, mientras que FSRL se centra en, dado el marco y su objetivo, identificar los predicados y sus argumentos, etiquetando cada argumento con el roles semánticos correspondiente. A su vez, la detección e identificación se presentan a veces como dos subtareas distintas.

Desde comienzos de la década de 2000, la tarea de etiquetado de roles semánticos ha recibido una atención considerable y ha sido el foco de varias tareas compartidas en conferencias (C. Baker et al. 2007; Carreras y Màrquez, 2005; Litkowski, 2004). Durante las dos últimas décadas, se han evaluado diversos enfoques, que van desde sistemas basados en reglas y máquinas de vectores de soporte (*Support Vector Machines*, SVM) hasta técnicas de comprensión automática de textos (*Machine Reading Comprehension*), redes neuronales profundas y modelos de lenguaje preentrenados como BERT (Ai y Tu, 2024; Bastianelli et al. 2020; Das et al. 2014; Ferraro y Suominen, 2020; FitzGerald et al. 2015; Gildea y Jurafsky, 2002; Roth y Lapata, 2015; C. Zheng et al. 2023).

Un ejemplo de solución para FSRL puede encontrarse en el trabajo de Bastianelli et al. (2020): en primer lugar, se identifican los segmentos en el árbol de constituyentes de la entrada a partir de una red neuronal convolucional de grafos (*Graph Convolutional Network*, GCN) entrenada a tal fin; en segundo lugar, se les asigna un rol utilizando dos campos aleatorios condicionales (*Conditional Random Fields*, CRF), uno para identificar cuáles son los argumentos y otro para etiquetarlos con su correspondiente rol. Con este enfoque, se alcanza un F_1 de 0.756 en el conjunto de datos de FrameNet 1.5, utilizando los marcos y objetivos ya marcados en el conjunto de referencia (*gold standard*).⁶

El trabajo *AGED* aprovecha la información sobre los marcos en FrameNet: las definiciones de los marcos se concatenan con los textos de entrada y se utilizan como ejemplos para entrenar un modelo de lenguaje. En particular, las

⁶Se recuerda que la definición de F_1 se encuentra en la Figura 2.4.

definiciones de las distintas entidades se utilizan como espacios que deben ser completados con los argumentos extraídos (C. Zheng et al. 2023). Este enfoque obtiene una métrica F_1 de 0.769 en el conjunto de datos de FrameNet 1.7. Ai y Tu (2024) siguen el modelo AGED pero, para cada marco, agregan un modelo CRF para modelar la relación entre los argumentos, alcanzando un F_1 de 0.771 en el conjunto de datos de FrameNet 1.7.

Por otra parte, el aprendizaje en contexto es un paradigma que permite a los grandes modelos de lenguaje aprender a resolver una tarea a partir de: (a) la definición en lenguaje natural del problema a resolver, y (b) ejemplos de entrada y salida proporcionados como demostraciones. La principal ventaja de ICL es que no requiere ajuste de parámetros ni cambios de arquitectura, lo cual reduce el tamaño de los conjuntos de entrenamiento necesarios. Sin embargo, esto suele conllevar un menor rendimiento en comparación con los modelos ajustados específicamente para una tarea (Brown et al. 2020; Dong et al. 2024).

El uso de ICL para la extracción de relaciones (*Relation Extraction*, RE), una tarea emparentada con FSRL, ha sido explorado en algunos trabajos recientes. Esta tarea consiste en identificar y clasificar las relaciones semánticas entre entidades en un texto; a diferencia de FSP, usualmente se enfoca en relaciones binarias entre entidades nombradas.

Un estudio sobre el rendimiento de GPT-3 (Brown et al. 2020) en tareas de NER y RE en textos biomédicos se presenta por Jimenez Gutierrez et al. (2022). En ese trabajo se comparan modelos de lenguaje basados en BERT, ajustados a la tarea respectiva, con la aplicación de ICL sobre GPT-3 con una cantidad de ejemplos que varía entre uno y tres, concluyendo que sigue siendo más efectivo el ajuste fino que ICL para ese problema.

Sin embargo, Wan et al. (2023) exploran el uso de ICL también con el modelo GPT-3 para la extracción de relaciones con resultados diversos. Los autores proponen una solución, denominada GPT-RE, que utiliza un LLM para determinar si existe una relación entre un par de entidades. Los ejemplos para las demostraciones son seleccionados mediante un proceso de recuperación basado en similitud con la entrada, variando su cantidad entre 5 y 30 ejemplos. Los experimentos demuestran que GPT-RE puede lograr un rendimiento competitivo en comparación con los modelos ajustados específicamente para la tarea en conjuntos como el de Semeval 2010 y TACRED (Hendrickx et al. 2010; Y. Zhang et al. 2017).

El uso de LLM para la extracción de relaciones sin proveer ejemplos (*zero-shot*) es explorado en los trabajos de G. Li et al. (2023) y K. Zhang et al. (2023). Los autores proponen un método basado en plantillas para generar *prompts* que guían al LLM a extraer relaciones entre entidades nombradas en un texto. Los experimentos se realizan con modelos GPT-3-5, ChatGPT, y FLAN-T5 (Chung et al. 2024), y los resultados muestran que el enfoque puede alcanzar un rendimiento competitivo con modelos ajustados específicamente para la tarea.

En el artículo de Devasier et al. (2025) se presenta un enfoque para la extracción de argumentos con LLM basados en FrameNet, utilizando tanto ICL como ajuste fino. En particular, para el aprendizaje con ICL, se emplea exclusivamente el conjunto de textos con anotaciones completas de FrameNet, y se incorporan hasta cinco ejemplos anotados por caso. Los marcos y sus objetivos ya se asumen identificados y, por tal motivo, los objetivos se señalan en la entrada del modelo. El mejor resultado de F_1 reportado con ICL es de 0.592, con GPT-4, mientras que para ajuste fino es de 0.794 con Qwen2.5-14B. Estos resultados indican que el ajuste fino mantiene una ventaja en el rendimiento sobre ICL, cuyo resultado es considerablemente inferior.

Los trabajos de C. Zheng et al. (2023) y Devasier et al. (2025) incorporan conocimiento de FrameNet y muestran cómo aprovechar sus definiciones y ejemplos anotados para guiar la extracción de argumentos. No obstante, ambos enfoques se centran en la detección de argumentos partiendo de marcos ya identificados y no abordan de forma integral el problema completo de FSP —esto es, la identificación de marcos, objetivos y sus argumentos asociados—.

En síntesis, aunque el aprendizaje en contexto ha sido explorado en diversas tareas de PLN —como resumen automático, análisis de sentimientos, traducción y extracción de relaciones (Agarwal et al. 2025; Dong et al. 2024; Xu et al. 2024)—, su aplicación al FSP permanece todavía poco desarrollada. Esta brecha evidencia una oportunidad clara de investigación: explorar cómo combinar ICL con el conocimiento semántico estructurado de FrameNet para abordar de manera integral la extracción tanto de marcos como de roles semánticos, con el objetivo de reducir la dependencia de ajustes costosos y ampliar la aplicabilidad en dominios específicos, en particular, en esta investigación, el dominio legal.

3.3. FSP en Contexto

El objetivo propuesto es abordar la tarea de análisis de marcos semánticos aprovechando el conocimiento estructurado de FrameNet mediante aprendizaje en contexto, evitando así la necesidad de un entrenamiento adicional.

Se exploran dos estrategias alternativas para resolver el análisis. La primera aborda la tarea en un único paso: dado un texto, se espera que el LLM identifique y etiquete todos los marcos y sus elementos. La segunda estrategia divide la tarea en dos etapas: (a) «identificación de marco y objetivo», donde el modelo identifica los marcos evocados en el texto junto con sus disparadores léxicos; y (b) «identificación de argumentos», en la que los marcos y objetivos detectados se reutilizan como entrada para guiar la extracción de sus argumentos.

Para la resolución de estas dos estrategias se proponen, entonces, tres tipos de tareas a resolver con ICL:

1. Análisis de marcos semánticos (FSP): se detectan, en un texto dado, todos los marcos y sus argumentos en una única inferencia. La entrada es, entonces, un texto y la salida es una lista de marcos detectados junto con sus argumentos. Esto resuelve, de forma completa y en una llamada, la tarea FSP.
2. Identificación de marcos y objetivos (FI): se identifican las unidades léxicas asociadas con los diferentes marcos buscados. La entrada es un texto y la salida esperada es una lista de pares marco–objetivo. Esta tarea combina las subtareas de «detección de objetivos» e «identificación de marcos», abordadas por separado en trabajos previos como los de Roth y Lapata (2015) y Bastianelli et al. (2020).
3. Extracción de argumentos (FSRL): se identifican todos los argumentos asociados a los marcos identificados en un texto. La entrada consiste en un texto y una lista de pares marco–objetivo, previamente detectados en ese texto, y la salida son los argumentos correspondientes a cada uno de los marcos.

En la Figura 3.3 se ilustra la entrada y salida esperada para cada una de estas tres tareas.

Para cada una de las tareas, se propone la generación automática de las instrucciones y ejemplos a ser incluidas como parte de la entrada (*prompts*) de

Análisis de marcos semánticos (FSP)

Texto:

They had killed or captured about a quarter of the
enemy's known leaders.

Salida:

```
[
  {
    "target": "killed",
    "killer": "They",
    "victim": "about a quarter of the enemy's known leaders"
  }
]
```

Identificación de marcos y objetivos (FI)

Texto:

They had killed or captured about a quarter of the
enemy's known leaders.

Salida:

```
[ ("Killing", "killed") ]
```

Extracción de argumentos (FSRL)

Texto:

They had killed or captured about a quarter of the
enemy's known leaders.

Objetivos:

```
[ ("Killing", "killed") ]
```

Salida:

```
[
  {
    "target": "killed",
    "killer": "They",
    "victim": "about a quarter of the enemy's known leaders"
  }
]
```

Figura 3.3: Ejemplos de entrada y salida para las tres tareas FSP, FI y FSRL con el marco «Asesinato».

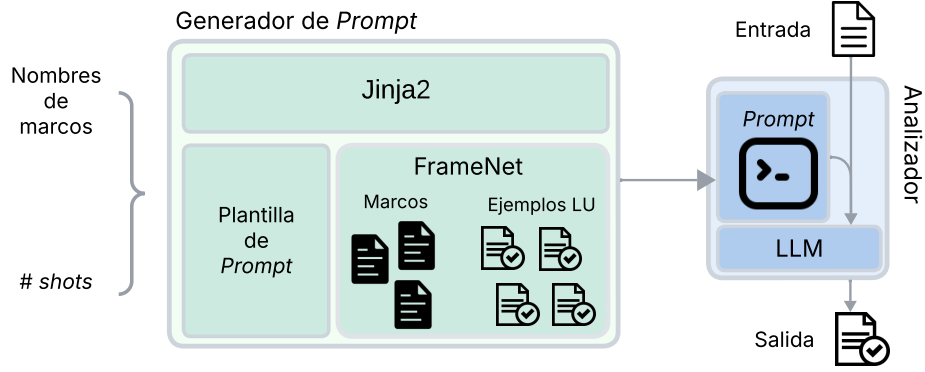


Figura 3.4: FSP en contexto: los marcos de FrameNet, junto con sus ejemplos anotados, se combinan de acuerdo a una plantilla para construir una guía específica para la tarea. Esta guía, junto con un texto, se proporcionan a un LLM como entrada para generar la salida correspondiente.

un LLM. Estas guías, junto con los textos a analizar, se suministran al modelo para producir como salida los análisis correspondientes. La Figura 3.4 muestra un diagrama general de la solución propuesta.

La guía generada incluye la descripción de la tarea, las definiciones de los marcos y los ejemplos anotados, en inglés *shots*⁷. Esto puede verse como una función que, dado un subconjunto de todos los marcos disponibles $\mathcal{F}$ y un subconjunto de los ejemplos E anotados para una tarea T , genera un *prompt* específico para resolver esa tarea:

$$Prompt_T : 2^{\mathcal{F}} \times 2^E \mapsto String$$

De los marcos de FrameNet se extraen tanto sus definiciones como los elementos centrales, incluyendo las descripciones de cada uno de ellos. Para construir los *shots*, se utilizan los ejemplos anotados de las unidades léxicas asociadas a los marcos seleccionados, como el mostrado en la Figura 3.1. Cada ejemplo se procesa para generar un par entrada–salida: la entrada consiste en el texto original, junto con la lista marco–objetivo si corresponde, y la salida contiene la lista de marcos y sus elementos, o bien los pares marco–objetivo,

⁷En ICL se suele referenciar a los ejemplos usados para demostrar la tarea a resolver como *shots*. Así, *n-shots*, *few-shot* o *many-shot* refieren a casos en donde se agregan al *prompts* exactamente *n*, algunos pocos o varios ejemplos de resolución. Un caso especial es *zero-shot*. Como su nombre lo indica, no se proporciona ningún ejemplo al modelo, el cual, tan solo con una especificación de su tarea, debe resolverla con el conocimiento adquirido durante su entrenamiento.

según la tarea a resolver.

La principal limitación de este enfoque radica en las restricciones impuestas por los modelos a la longitud de sus entradas, lo que condiciona la cantidad de marcos y ejemplos que pueden incluirse en un solo *prompt*. Este tipo de restricción es señalado frecuentemente como una desventaja en el uso de ICL (P. Liu et al. 2023; Zhao et al. 2021).

Para limitar el tamaño del *prompt*, se acota la cantidad de marcos a identificar a una selección arbitraria de todos aquellos disponibles en FrameNet. La desventaja evidente de este enfoque es que, para detectar todos los marcos de un texto, el LLM debe ser llamado varias veces, cada una con un nuevo *prompt* que cubra un subconjunto distinto de marcos, hasta cubrirlos todos.

En la primera y segunda tarea, los marcos presentes en el texto son *a priori* desconocidos, por lo que el *prompt* debe incluir ejemplos para todos los marcos que se quieran detectar, independientemente de su presencia efectiva en el texto. Esto limita la cantidad de ejemplos por marco y acota, en consecuencia, el número de marcos a procesar simultáneamente.

En contraste, la tarea de FSRL se beneficia de una etapa previa de FI. Esto permite personalizar el *prompt* específicamente para los marcos detectados previamente, incluyendo definiciones y ejemplos enfocados exclusivamente a ellos. Como desventaja, esta tarea requiere una llamada adicional al modelo para la identificación de los marcos, y su desempeño depende del resultado de esa inferencia previa.

Todas las tareas se evalúan utilizando un número creciente de *shots* para analizar el impacto de la cantidad de ejemplos en el rendimiento de cada uno de los modelos. En particular, para la tercera tarea se realizaron dos evaluaciones: una utilizando la lista de objetivos de referencia —garantizando que la entrada no contenga errores, salvo aquellos cometidos por los propios anotadores de FrameNet—, y otra utilizando la salida de la tarea FI, generada por el mismo modelo, propagando así los errores cometidos en el paso previo.

La cantidad de ejemplos disponibles en FrameNet para cada marco se presenta, en consecuencia, como otra limitante. Dado que los ejemplos se dividen en conjuntos de entrenamiento y prueba, la cantidad de ejemplos disponibles para construir los *shots* puede ser insuficiente para ciertos marcos, especialmente cuando se requieren cientos de ejemplos por marco en el *prompt*.

Finalmente, respecto a los textos a analizar, en la solución propuesta cada elemento de entrada se procesa de forma individual: un conjunto de prueba con

k textos de entrada da lugar exactamente a k inferencias del modelo, sin importar cuántos *shots* o marcos se inyecten en el *prompt*. Si bien agrupar múltiples textos de entrada en un solo *prompt* reduciría la cantidad de inferencias, esta estrategia se evita deliberadamente para prevenir posibles interacciones entre ejemplos que puedan sesgar las salidas del modelo. Además, reducir los textos a procesar a uno por inferencia libera espacio en el *prompt* para incluir una mayor cantidad de *shots* y definiciones de marcos.

A continuación, en la Sección 3.3.1, se presenta la implementación de la generación de *prompts* para las tres tareas propuestas. Los experimentos con un conjunto seleccionado de marcos y distintos números de *shots* se describen en las secciones 3.3.2 y 3.3.3.

3.3.1. Generación de *Prompts*

Para cada una de las tres tareas planteadas —FSP, FI y FSRL— se crea un *prompt* específico a partir de una plantilla que se completa con la información extraída de FrameNet.

En primer lugar, la base de datos de FrameNet —almacenada en múltiples archivos XML— se procesa con el fin de extraer las definiciones de marcos, sus elementos y los ejemplos anotados de todas las unidades léxicas asociadas a cada *frame* a detectar. Los ejemplos se obtienen únicamente de las unidades léxicas que evocan los marcos seleccionados para el experimento. Se evita el uso del conjunto de textos completos de FrameNet⁸, ya que podrían contener múltiples marcos no deseados e incluso algunos duplicados con las anotaciones de las LU.

Los ejemplos anotados se dividen en conjuntos de entrenamiento y de prueba: el primero se utiliza para proporcionar ejemplos en contexto destinados a la construcción del *prompt*, mientras que el segundo sirve como referencia para evaluar la salida del modelo.

La plantilla —que contiene la definición de la tarea y las reglas para renderizar las distintas secciones del *prompt*— junto a los datos y los ejemplos de FrameNet se combinan mediante el motor de plantillas Jinja2⁹. Las planti-

⁸FrameNet incluye una colección de artículos cuyos textos se encuentran completamente anotados (*full-text annotation*), y han sido usados como conjunto de entrenamiento desde su puesta a disposición en SemEval-07 (C. Baker et al. 2007). Fragmentos de estos textos son, además, utilizados como ejemplos de las distintas unidades léxicas.

⁹Jinja2 es un motor de plantillas para Python. Disponible en: <https://palletsprojects.com/p/jinja/>

- (a)
- ```
GOAL:
Your goal is to detect which events, also called frames, are present in a text, by
↪ identifying which is the word that evokes the event or frame, called trigger.

Only the following events are of interest:
- Killing
- Robbery
- Violence
- Rape
- Shoot_projectiles
- Abusing
```
- (b)
- ```
# GOAL:
Your goal is to extract the arguments of the events present in a given text.
You will get a list of words, the triggers, that evoke the events or frame, the
↪ triggers, and for each trigger you must output the appropriate arguments.
```

Figura 3.5: Ejemplos de sección «Objetivo» para las tareas de (a) Identificación de Marcos y (b) Extracción de Argumentos.

llas no varían según el modelo a utilizar, y no fueron previamente ajustadas u optimizadas para ninguno de ellos.

Cada plantilla consta de cuatro secciones principales:

1. **Objetivo** (*Goal*): Define el propósito de la tarea. Esta sección es específica de la tarea pero independiente de los marcos y ejemplos usados.
2. **Eventos** (*Events*): Describe los marcos relevantes para el experimento junto con sus argumentos, y se basa únicamente en las definiciones de los marcos.
3. **Directrices** (*Guidelines*): Proporciona instrucciones generales para el modelo. Al igual que la sección «Objetivo», es independiente de los marcos concretos.
4. **Ejemplos** (*Examples*): Lista de pares entrada-salida derivados de los ejemplos anotados. En particular, esta sección se omite en los experimentos *zero-shot*.

Las secciones «Objetivo» y «Directrices» definen principalmente la tarea y especifican el formato de salida requerido por el modelo, con referencias mínimas a marcos concretos. En contraste, las secciones «Eventos» y «Ejemplos» son específicas de los marcos, construidas a partir de las definiciones y los ejemplos anotados.

- (a)
- ```
EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an
↳ adjective, an adverb.

Below there are the definitions of each frame and a list of triggers for each event
↳ of interest, but synonyms of those words might fit the same.

Rape
- Definition: The words in this frame describe situations in which the Perpetrator
↳ has sexual intercourse with the Victim forcibly (or by threatening the use of
↳ force) and without his or her consent.
- Possible triggers: rape, rapist, raped
- Text example for this frame with trigger "raped": "Does the press make it more
↳ difficult for raped women to obtain justice?"
...
```
- (b)
- ```
# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an
↳ adjective, an adverb; arguments are usually propositions.

Below there are the definitions of each frame and its possible arguments.

## Rape
- Definition: The words in this frame describe situations in which the Perpetrator
↳ has sexual intercourse with the Victim forcibly (or by threatening the use of
↳ force) and without his or her consent.
- Arguments: The entities you might find on the text for Rape events are:
  -- trigger: The verb, noun or adjective that marks the event of violence.
  -- perpetrator: This is the person who forces another person to have sexual
  ↳ intercourse with him or her against the other person's will.
  -- victim: This is the person who is forced to engage in sexual intercourse
  ↳ against his or her will.
- Example for "Rape" frame with trigger "raped":
  Text: "Does the press make it more difficult for raped women to obtain justice?"
  Arguments: {"trigger": "raped", "victim": "women"}
```

Figura 3.6: Ejemplo de la sección «Eventos» para (a) detección de marcos y (b) extracción de argumentos. En el primer caso se detallan posibles valores para los objetivos, mientras que en el segundo se provee la lista de posibles elementos asociados a un marco.

La sección «Objetivo» es la más simple, ya que enuncia de forma directa el propósito general de la tarea. En las tareas FSP y FI, se incluye, además, una lista predefinida de nombres de marcos, con el fin de acotar el alcance de la tarea. La Figura 3.5 presenta ejemplos de la sección «Objetivo» para las tareas FI y FSRL. En particular, como se describe más adelante, los marcos listados en la Figura 3.5(a) son los elegidos para los experimentos de evaluación.

La sección «Eventos» aporta la definición de cada marco, una oración ilustrativa y la lista de los posibles elementos o *targets* del marco, según la tarea. Cada entidad va acompañada, a su vez, de su definición. La Figura 3.6 muestra ejemplos de esta sección para las tareas FI y FSRL.

La sección de «Directrices» contiene instrucciones sobre cómo realizar la

```
# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of detected pairs (frame, trigger):
  -- If N number of events are found, you must answer with one list of N records.
  -- If no event is detected, the list must be empty.
- Do NOT reply with any explanation or any other text but one list of records.
- The output must be of the form [[frame 1, trigger 1],[frame 2, trigger 2]...[frame N,
  ↪ trigger N]] formatted as a JSON string.
- Triggers must be substrings of the input.
```

Figura 3.7: Directrices para la tarea de Identificación de Marco.

```
# EXAMPLES:
## Example 1
Text: A woman staying at a local hotel returned to find her room ransacked and another
↪ motorist had her car broken into .
Triggers: [{ "frame": "Robbery", "trigger": "ransacked" }]
Output: [{ "source": "her room", "trigger": "ransacked" }]

## Example 2
...
```

Figura 3.8: Sección de ejemplos para la tarea FSRL.

tarea y restringe el formato de salida para minimizar errores y facilitar su procesamiento posterior. Por ejemplo, se prohíben explicaciones adicionales, se exige que la salida sea parte de la entrada —intentando evitar alucinaciones¹⁰— y se agregan otras restricciones que aseguren la consistencia de la salida. Aunque las directrices varían según la tarea, son independientes de los marcos específicos considerados en cada experimento. En la Figura 3.7 se presenta un ejemplo de esta sección para la tarea FI.

La sección de «Ejemplos» contiene pares de textos de entrada y sus salidas esperadas, extraídos de FrameNet. El número de *shots* incluidos varía según el experimento; su cantidad es arbitraria, aunque, como se menciona previamente, se encuentra limitada por la longitud máxima de entrada del modelo. La Figura 3.8 presenta un ejemplo para la tarea FSRL.

Como se detalla en la Sección 3.1, cada ejemplo de unidad léxica en FrameNet está anotado con una y solo una instancia de un evento. Los ejemplos de entrada y salida contruidos a partir de estos datos reflejan esta limitación, lo que implica que cada ejemplo contiene exactamente un único evento.

¹⁰En el contexto de los grandes modelos de lenguaje, el término «alucinación» refiere a la generación, por parte de un modelo, de contenido que carece de sentido, es incoherente, repetitivo, incorrecto o no está respaldado por los datos de entrada (Ji et al. 2023).

Esta limitante puede llevar a que el modelo, por imitación, extraiga únicamente un evento, incluso cuando se le instruya devolver todos los que encuentre, o que maneje incorrectamente casos en los que no haya eventos presentes en el texto de entrada¹¹. Para mitigar este problema, por un lado se combinan los últimos dos ejemplos de FrameNet en uno y, por otro, se añade al final un ejemplo «ficticio» sin eventos¹². De esta forma se expone al modelo —aunque mínimamente— a ejemplos sin y con múltiples eventos.

En resumen, la generación de *prompts* permite construir de forma sistemática las entradas necesarias para cada tarea —FSP, FI y FSRL—, incorporando las directrices generales de la tarea junto a las definiciones y ejemplos extraídos de FrameNet. Esta aproximación garantiza que los modelos reciban la información estructurada y contextualizada requerida para realizar la tarea de manera consistente. El Apéndice A contiene las plantillas completas de cada tarea y muestra los resultados de su aplicación para distintas configuraciones de *shots*.

3.3.2. Experimentación

Para la fase de experimentación se plantean dos objetivos: en primer lugar, evaluar el rendimiento y la viabilidad del enfoque propuesto; y en segundo lugar, analizar el impacto que tiene la cantidad *shots* incluidos en el *prompt* en el desempeño del modelo en las distintas tareas planteadas. En consecuencia, se varían tanto los modelos como la cantidad de *shots*, sobre un conjunto seleccionado de marcos. La Tabla 3.2 resume la configuración de los distintos experimentos realizados.

Si bien la solución diseñada puede aplicarse para cualquier conjunto de marcos, los experimentos se centran en un subconjunto relacionado con eventos de violencia, que son, en definitiva, los más relevantes pensando en la Base de Jurisprudencia Nacional. Específicamente, se consideran los marcos «Abuso» (*Abusing*), «Asesinato» (*Killing*), «Violación» (*Rape*), «Robo» (*Robbery*), «Disparo» (*Shoot Projectiles*) y «Violencia» (*Violence*).

En los experimentos solo se tienen en cuenta los atributos centrales de

¹¹Este nunca es el caso de la tarea FSRL, en donde los pares marco-objetivo ya fueron identificados y son parte de la entrada. Un texto sin marcos detectados carece de sentido como entrada para esta tarea.

¹²El texto elegido es «*Ahab and the crew continue their eventful journey and encounter a number of obstacles along the way.*», extraído de la novela «*Moby Dick*» de Herman Melville. Se opta por este texto dado que no contiene eventos relacionados con los marcos seleccionados para la experimentación.

cada marco. Por ejemplo, para «Asesinato», se extraen «Asesino», «Causa», «Víctima», «Instrumento» y «Medio», junto con el «Objetivo», que indica la presencia del marco «Asesinato» en el texto.¹³

Para la primera tarea, se evalúan cuatro modelos distintos: Claude Haiku 3.5, Claude Sonnet 3.5, GPT-4o Mini y GPT-4o (Anthropic, 2024; OpenAI, 2024). Para el resto de las tareas, la evaluación se amplía para incluir DeepSeek Chat y DeepSeek Reasoner (DeepSeek-AI et al. 2025).

Todos estos modelos se basan en la arquitectura *transformers*, la cual permite el procesamiento paralelo de secuencias de *tokens* mediante mecanismos de atención (Vaswani et al. 2017). Esta arquitectura ha demostrado ser altamente efectiva para tareas de generación y comprensión de lenguaje natural. Sin embargo, ni Anthropic ni OpenAI han divulgado la configuración interna exacta de estos modelos, tales como el número de capas, el tamaño de los vectores de palabras o el número total de parámetros. En la Tabla 3.1 se resumen algunas características técnicas disponibles públicamente para estos modelos.

Para la experimentación, todos los parámetros de los LLM se mantienen en sus valores por defecto, excepto la temperatura, que se fija en 0.01. Esta temperatura baja se elige para minimizar la aleatoriedad de las salidas y hacerlas más predecibles. Esta configuración se mantiene constante en todos los experimentos.

El rendimiento se mide mediante las métricas estándar de precisión (P), cobertura (R) y medida F_1 .¹⁴ Se adopta un enfoque de evaluación *micro*, en el cual las métricas se calculan sobre los elementos extraídos, en lugar de un enfoque *macro*, que considera las relaciones como unidades completas. Este último criterio exige que todos los atributos se identifiquen correctamente para considerar la extracción como correcta.

La evaluación es además *estricta*: se requiere que los límites de los elementos reconocidos coincidan exactamente con la salida esperada. Aunque existen métricas de evaluación más laxas —por ejemplo, basadas en dependencias semánticas (C. Baker et al. 2007) o medidas de similitud entre cadenas (Breton et al. 2025)—, no se aplican en estos experimentos. Tampoco se corrigen inconsistencias del conjunto de datos, como variaciones en la inclusión de preposiciones.

¹³En particular, para este marco se puede encontrar una definición más detallada de sus elementos en la Sección 3.1 y en las figuras I.5 y I.6.

¹⁴Se recuerda que P , R y F_1 están definidas en la Figura 2.4.

Modelo	Descripción
GPT-4o	Ventana de contexto: $\sim 128\,000$ <i>tokens</i> . Rendimiento destacado en <i>benchmarks</i> médicos y de razonamiento. Número de parámetros no divulgado.
GPT-4o Mini	Versión reducida de GPT-4o. Ventana de contexto: $\sim 128\,000$ <i>tokens</i> . Menor costo por <i>token</i> . Número de parámetros no divulgado.
Claude Sonnet 3.5	Ventana de contexto: $\sim 200\,000$ <i>tokens</i> . Optimizado para razonamiento y codificación.
Claude Haiku 3.5	Variante más ligera y rápida de Sonnet. Ventana de contexto: $\sim 128\,000$ <i>tokens</i> . Optimizado para baja latencia y tareas de <i>chat</i> y PLN estándar.
DeepSeek Reasoner	Arquitectura «mezcla de expertos» con $\sim 671\,000$ millones de parámetros totales; solo $\sim 37\,000$ millones activos por <i>token</i> . Ventana de contexto: $\sim 128\,000$ <i>tokens</i> . <i>Chain-of-thought</i> explícito.
DeepSeekChat	Variante conversacional de DeepSeek. Optimizada para <i>chat</i> y generación rápida.

Tabla 3.1: Descripciones de los grandes modelos de lenguaje evaluados según la información disponible públicamente.

Exp.	Tarea	# <i>Shots</i>	Modelos	Marcos
1	FSP	0–400	Claude, GPT	Asesinato
2	FI	0–150	Claude, Deepseek, GPT	Abuso, Asesinato, Disparo, Robo, Violación, Violencia
3a	FSRL (referencia)			
3b	FI + FSRL	150		

Tabla 3.2: Configuraciones de los experimentos de las distintas tareas con ICL. Claude, GPT y Deepseek se refieren a ambos modelos evaluados por familia: Haiku 3.5 y Sonnet 3.5, GPT-4o y GPT-4o Mini, y Chat y Reasoner, respectivamente.

En todos los experimentos se construyen dos conjuntos de datos: uno de *shots* para la generación de *prompts*, —que actúa como conjunto de entrenamiento—, y otro para evaluar la salida de los modelos. Para mantener una distribución equilibrada, los ejemplos se seleccionan mediante un muestreo estratificado basado en la distribución de las unidades léxicas de los marcos considerados en el experimento.

Para analizar el impacto de la cantidad de ejemplos sobre el rendimiento de los modelos, se experimenta con un número creciente de *shots*; los ejemplos se van incorporando de forma incremental, preservando su orden de aparición en el *prompt*. En consecuencia, en una misma tarea, los primeros n ejemplos son idénticos y aparecen en el mismo orden en todos los *prompts* que contienen al menos n *shots*, y es, a su vez, el mismo independientemente del modelo particular del escenario considerado.

Sin embargo, por un tema de costos, no se evalúa la misma secuencia de *shots* en todos los modelos: para los modelos más grandes —Claude Sonnet, GPT-4o y DeepSeek Reasoner— se realizan menos experimentos en comparación con los modelos más pequeños —Claude Haiku, GPT-3.5, GPT-4o Mini y DeepSeek Chat—. Comenzando en cero ejemplos —o de 0, 5 y 10, en algunos casos— el número de *shots* se incrementa de a 25 para los modelos pequeños y de a 50 para los modelos grandes.

A diferencia del conjunto de *shots*, para cada tarea se fija un subconjunto independiente de 100 ejemplos para la evaluación, que es el mismo para todos los escenarios de experimentación planteados.

En particular, para la tarea FSP, se experimenta con un escenario simplificado en el que se considera un único marco, pero con un mayor número de ejemplos posibles. El marco seleccionado es «Asesinato», con una cantidad máxima de 400 *shots*.

Por otra parte, para la tarea FSRL, se llevan a cabo dos experimentos: uno utilizando el estándar de referencia, esto es, los pares marco-objetivo anotados manualmente de FrameNet, y otro utilizando los pares generados a partir de la salida de la tarea FI. Ambos utilizan los mismos textos de evaluación, pero difieren en qué pares marco-objetivo se emplean. El segundo escenario permite realizar una evaluación punta a punta, que abarca toda el proceso FSP, desde el texto de entrada hasta el conjunto final de relaciones extraídas. En este último escenario, el número de *shots* se fija en exactamente 150 para las dos subtareas que lo componen.

Modelo	Exp. 1 FSP	Exp. 2 FI	Exp. 3a FSRL	Exp. 3b FI+FSRL
Claude Haiku 3.5	(275) 0.711	(100) 0.909	(75) 0.689	0.535
Claude Sonnet 3.5	(250) 0.829	(150) 0.931	(150) <u>0.763</u>	<u>0.686</u>
DeepSeek Chat	-	(100) 0.943	(150) 0.723	0.686
DeepSeek Reasoner	-	(50) <u>0.935</u>	(150) 0.774	0.749
GPT-4o Mini	(375) 0.687	(125) 0.905	(150) 0.671	0.532
GPT-4o	(400) <u>0.802</u>	(100) 0.929	(100) 0.739	0.669

Tabla 3.3: Mejores valores de F_1 obtenidos en los experimentos, indicando entre paréntesis la cantidad de *shots*. Se muestran resultados para FSP (Exp. 1), FI (Exp. 2), FSRL con pares marco-objetivo (Exp. 3a) y FSRL con resultados de FI como entrada (Exp. 3b). Los valores más altos se indican en negrita y los segundos mejores se encuentran subrayados.

Los experimentos descriptos permiten evaluar de manera sistemática el desempeño de los modelos seleccionados y el efecto de la cantidad de *shots*, permitiendo una comparación de resultados entre tareas, modelos y condiciones de entrada. En particular, esta evaluación se realiza sobre marcos relacionados con eventos de violencia, de interés en el contexto de esta tesis.

3.3.3. Resultados

Como se expone en la sección anterior, se llevan a cabo un total de cuatro experimentos, especificados en la Tabla 3.2. Al examinar los resultados globales, presentados en la Tabla 3.3, se observa que la cantidad de *shots* y el desempeño de los modelos varían según la tarea.

La tarea de FSP, centrada en la extracción completa de relaciones, presenta puntajes de F_1 más bajos que las tareas de FI y FSRL, reflejando la mayor complejidad que implica identificar correctamente todos los elementos de un evento dentro de un texto.

Aunque no es posible realizar una comparación directa con trabajos previos debido a diferencias en los diseños experimentales, los resultados obtenidos son igualmente prometedores. Otros estudios (Ai y Tu, 2024; Bastianelli et al. 2020; Devasier et al. 2025; Roth y Lapata, 2015; C. Zheng et al. 2023) reportan valores de F_1 para la tarea de Extracción de Argumentos que oscilan

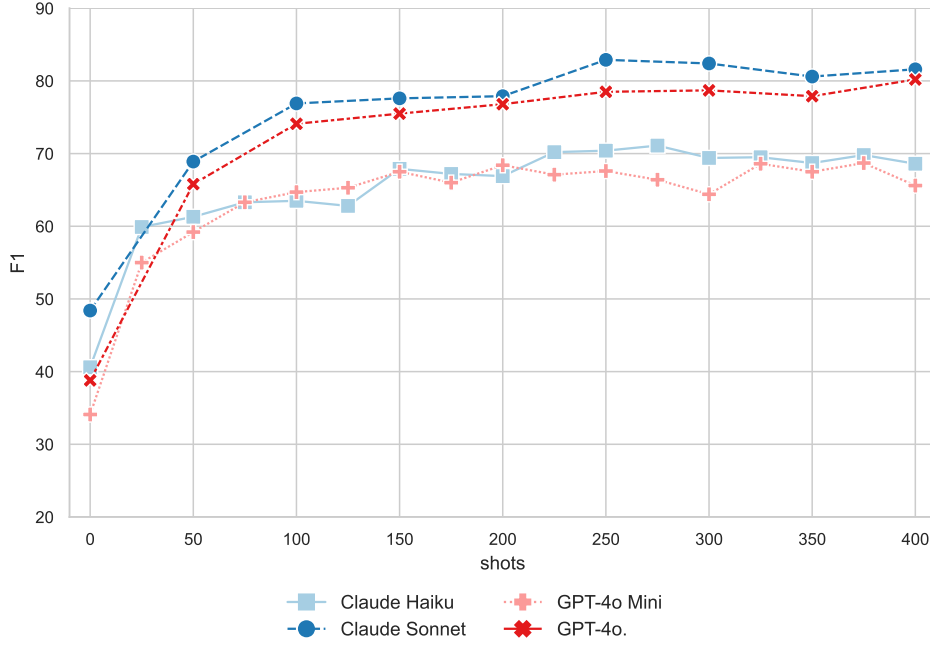


Figura 3.9: Resultados micro F_1 según el número de ejemplos para la tarea FSP (Experimento 1).

entre 0.756 y 0.794 sobre el conjunto FrameNet 1.7, utilizando objetivos y marcos del conjunto de referencia. En contraste, nuestro mejor F_1 alcanza 0.774, logrado por el modelo DeepSeek Reasoner con 150 ejemplos. Este valor resulta particularmente alto en comparación con los obtenidos por Devasier et al. (2025) para su experimento de ICL, con un F_1 de 0.592 con GPT-4.

Bastianelli et al. (2020) también informan un F_1 de 0.768 para la tarea de identificación de objetivos y una precisión de 0.901 para la identificación de marcos, esta última con los objetivos de referencia marcados en la entrada. En este trabajo, para la tarea conjunta de identificación de marcos y objetivos, el modelo DeepSeek Chat alcanza un F_1 máximo de 0.943 con 100 ejemplos; si distinguimos entre marcos y objetivos, los valores son de 0.953 y 0.943, respectivamente.

Para entender mejor estas diferencias, se analiza en detalle el Experimento 1, que evalúa la capacidad de los modelos para extraer todas las relaciones asociadas al marco «Asesinato». Este análisis permite observar cómo cada modelo maneja la identificación de atributos individuales y la evolución del rendimiento a medida que se incorporan más ejemplos.

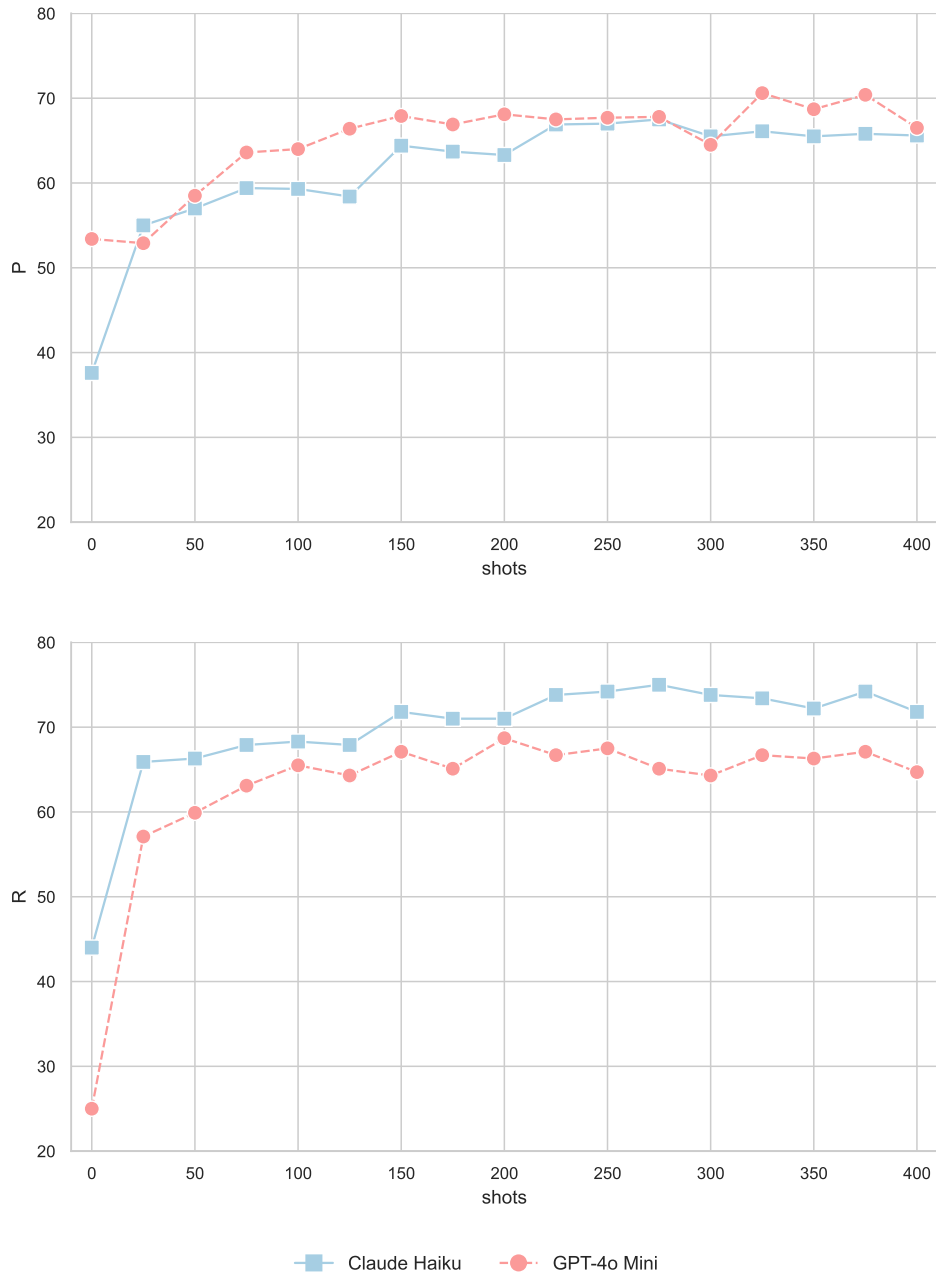


Figura 3.10: P (arriba) y R (abajo) para Claude Haiku 3.5 y GPT-4o Mini según número de ejemplos para la tarea FSP (Experimento 1).

Dado un texto de entrada, se espera que el modelo devuelva todas las instancias de relaciones de «Asesinato» presentes en el texto. En este experimento, el número de *shots* varía entre 0 y 400, y se evalúan cuatro modelos: Claude Haiku 3.5, Claude Sonnet 3.5, GPT-4o Mini y GPT-4o. El número de ejemplos se incrementa de 25 en 25 para los modelos más pequeños y de 50 en 50 para los más grandes.

La Figura 3.9 muestra los valores de F_1 en función del número de ejemplos, considerando tanto los elementos como el objetivo del marco «Asesinato». Claude Sonnet alcanza el mayor rendimiento con 250 ejemplos, con un puntaje F_1 de 0.829, mientras que GPT-4o alcanza su pico de 0.802 con 400 ejemplos. Sonnet supera consistentemente a GPT-4o en todos los conteos de ejemplos y supera el rendimiento máximo de GPT-4o a partir de los 250 ejemplos.

Como era de esperar, los modelos más pequeños tienen un rendimiento inferior comparado con sus contrapartes más grandes. Claude Haiku alcanza su F_1 máximo de 0.711 con 275 ejemplos, mientras que GPT-4o Mini alcanza un pico de 0.687 con 375 ejemplos. Cabe destacar que Haiku 3.5 y GPT-4o Mini muestran tendencias opuestas entre precisión y cobertura: Haiku tiene sistemáticamente mayor cobertura pero menor precisión que GPT-4o Mini para el mismo número de ejemplos (véase Figura 3.10).

Con la adición de ejemplos, los cuatro modelos muestran mejoras significativas en precisión y cobertura en comparación con su versión *zero-shot*, con ganancias de F_1 que van de 0.30 a 0.41 puntos. Esto sugiere que el conocimiento adquirido durante el entrenamiento de los modelos por sí solo es insuficiente para completar la tarea de FSP utilizando solamente las definiciones de FrameNet.

Las figuras 3.11 y 3.12 ilustran los valores individuales F_1 para «Objetivo», «Víctima», «Asesino» y «Causa», según la cantidad de *shots*, mientras que la Tabla 3.4 resume los mejores resultados para cada uno de estos argumentos. En la Figura 3.13 se presentan los mejores valores de F_1 para Sonnet 3.5 y GPT-4o por argumento

Los elementos más exitosamente extraídos son «Objetivo» y «Víctima», con picos de F_1 de 0.970, y 0.833, por GPT-4o y Claude Sonnet, respectivamente. Para «Asesino», GPT-4o registra el mayor F_1 de 0.714 con 400 ejemplos, mientras que «Causa» se extrae mejor con Claude Sonnet, 0.681 de F_1 con 250 ejemplos. Por último, el rendimiento para los argumentos «Medio» e «Instrumento» es bajo: sus valores máximos de F_1 son de 0.500 en ambos casos. En

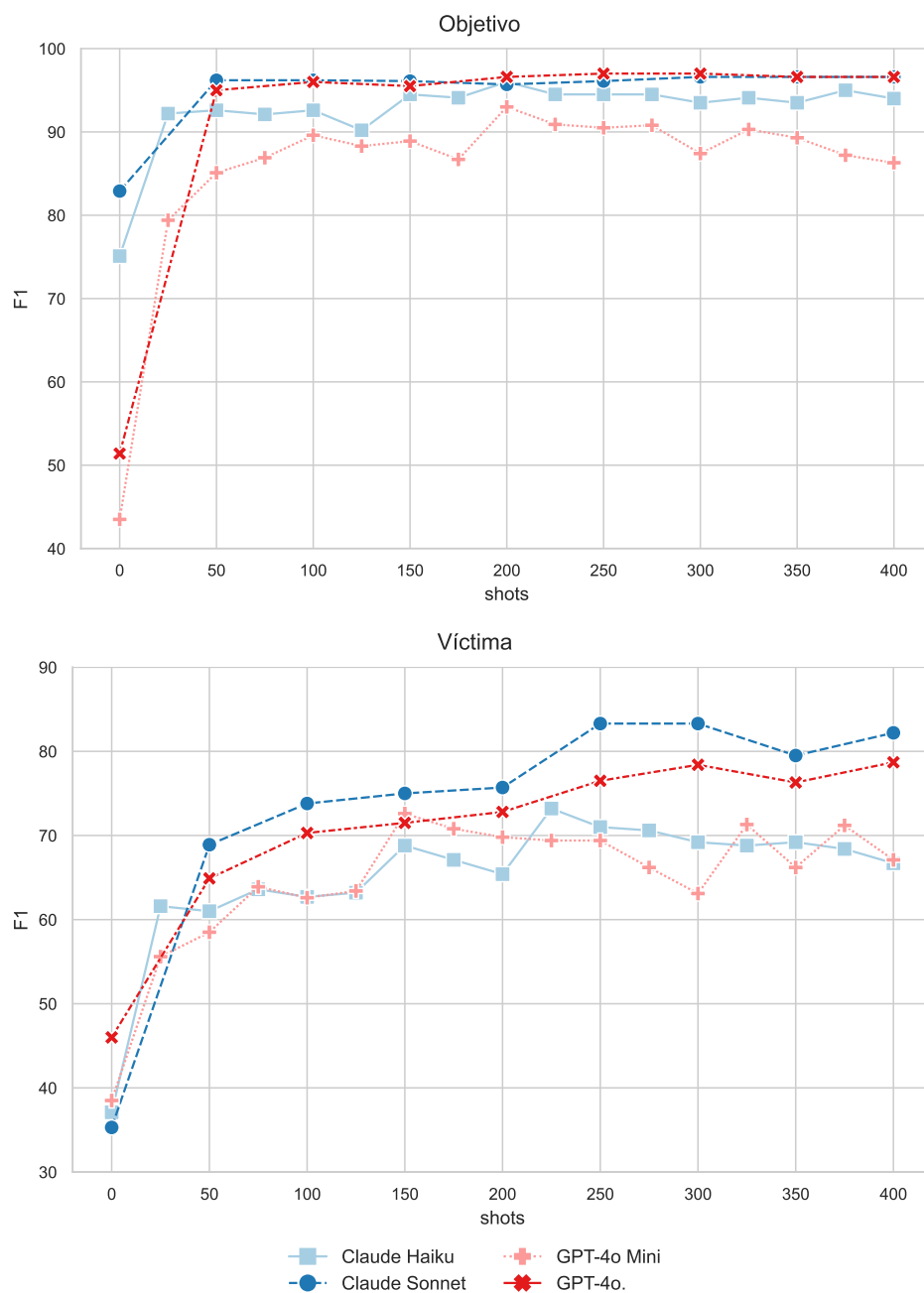


Figura 3.11: F_1 para los argumentos «Objetivo» y «Víctima» según el número de *shots* para el Experimento 1.

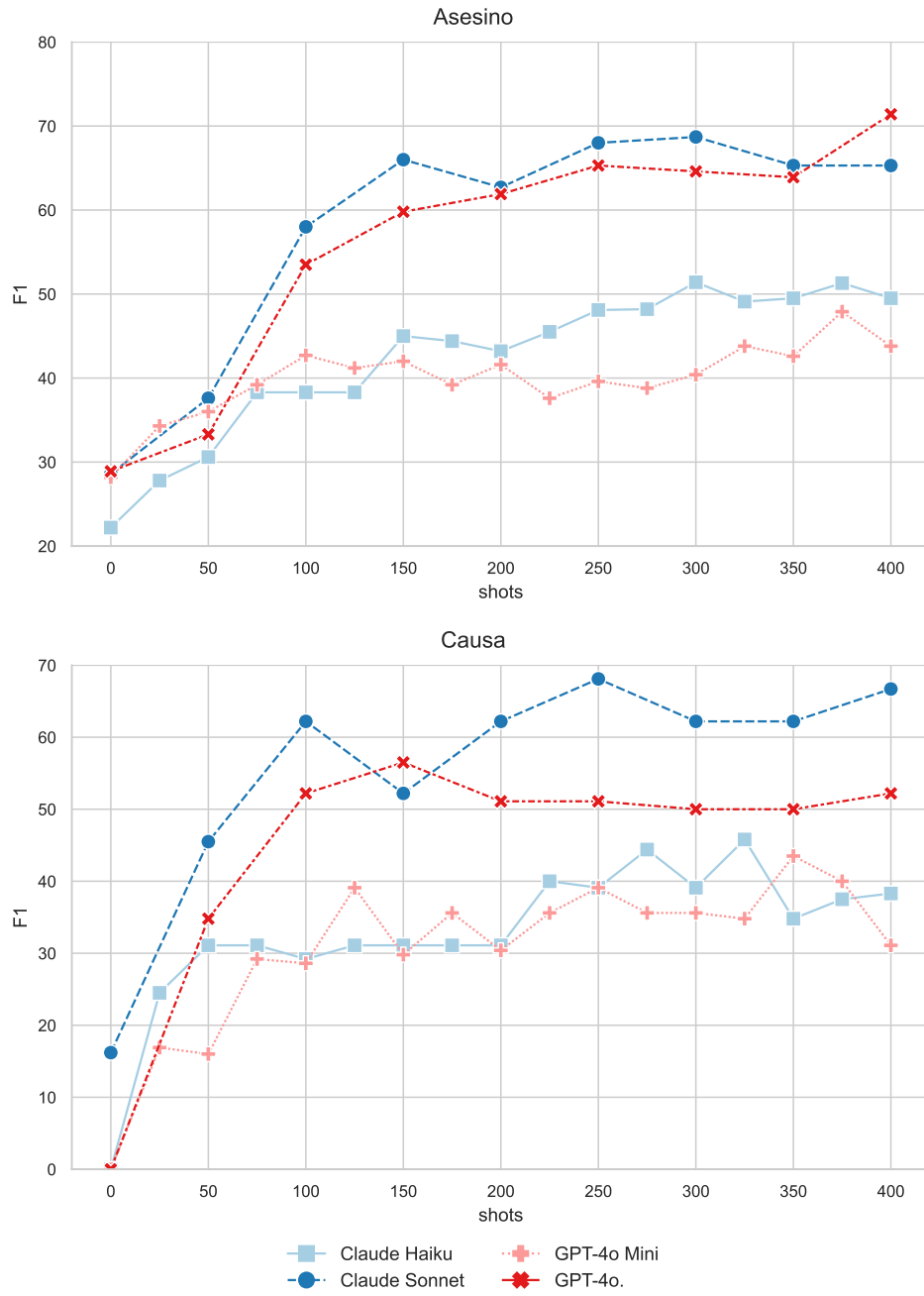


Figura 3.12: F_1 para los argumentos «Asesino» y «Causa» según el número de *shots* para el Experimento 1.

Argumento	Haiku 3.5	Sonnet 3.5	GPT-4o Mini	GPT-4o
Objetivo	(200) 0.960	(300, 350, 400) <u>0.966</u>	(200) 0.930	(250, 300) 0.970
Víctima	(225) 0.732	(250, 300) 0.833	(150) 0.726	(400) <u>0.787</u>
Asesino	(300) 0.514	(300) <u>0.687</u>	(375) 0.479	(400) 0.714
Causa	(325) 0.458	(250) 0.681	(350) 0.435	(150) <u>0.565</u>
Medio	(275) 0.222	(300) 0.500	0.0	(100) 0.500
Instrumento	(275) <u>0.333</u>	(250, 400) 0.500	0.0	(350,400) 0.222

Tabla 3.4: F_1 por argumento para distintos modelos en el Experimento 1. Los números entre paréntesis indican la cantidad de *shots* donde se alcanza el rendimiento máximo.

particular, GPT-4o Mini no logra extraer ninguno de estos dos argumentos en ningún escenario.

Tras caracterizar el desempeño de los modelos en la tarea FSP para un marco particular, el Experimento 2 amplía el análisis hacia la identificación de marcos y sus objetivos. Esta etapa se centra en la detección de pares marco-objetivo en un conjunto de marcos de violencia, permitiendo evaluar la capacidad de los modelos para reconocer correctamente los elementos clave sin requerir la extracción de todos los argumentos asociados.

En este escenario, dado un texto de entrada, se espera que el modelo produzca una lista de objetivos identificados y a qué marco pertenecen, representados como pares de la forma marco-objetivo. La evaluación incluye los seis modelos seleccionados, con el número de ejemplos variando de 0 a 150. Específicamente, se consideran 0, 5 y 10 ejemplos, seguidos de incrementos de 25 para los modelos pequeños y de 50 para los modelos grandes, comenzando desde 25 y 50 ejemplos respectivamente.

La Tabla 3.5 presenta el mejor desempeño alcanzado para los pares marco-objetivo, así como los valores para cada componente por separado. Las Figuras 3.14 y 3.15 ilustran los puntajes F_1 al variar el número de *shots*.

Cabe señalar que todos los modelos alcanzan un F_1 superior al 0.900 para el par marco-objetivo completo, con DeepSeek Chat liderando con un F_1 de 0.943 con 100 ejemplos. Al considerar los componentes por separado, las posiciones se mantienen casi iguales, con DeepSeek Chat manteniendo el primer lugar con 0.953 y 0.943 para Marcos y Objetivos, respectivamente.

Al igual que en el experimento anterior, todos los modelos muestran mejoras

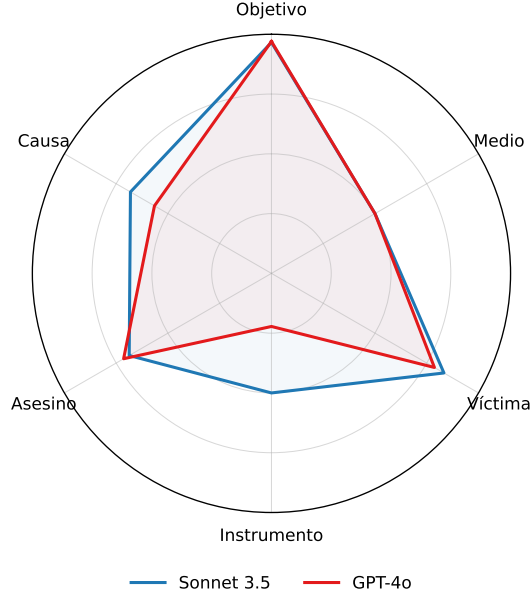


Figura 3.13: Mejores valores de F_1 para los modelos Sonnet 3.5 y GPT-4o por argumento para el Experimento 1.

Modelo	Marco–Obj.		Marco		Objetivo	
Claude Haiku 3.5	(100)	0.909	(100)	0.918	(100)	0.909
Claude Sonnet 3.5	(150)	0.931	(150)	0.940	(150)	0.931
DeepSeek Chat	(100)	0.943	(100)	0.953	(10,100)	0.943
DeepSeek Reasoner	(50)	<u>0.935</u>	(50)	<u>0.944</u>	(50)	<u>0.935</u>
GPT-4o Mini	(125)	0.905	(125)	0.924	(100)	0.910
GPT-4o	(100)	0.929	(100)	0.938	(100)	0.929

Tabla 3.5: F_1 para el Experimento 2. Se reportan los valores para el par marco–objetivo, y para cada componente por separado. Los números entre paréntesis indican la cantidad de ejemplos en el que se alcanza el rendimiento máximo.

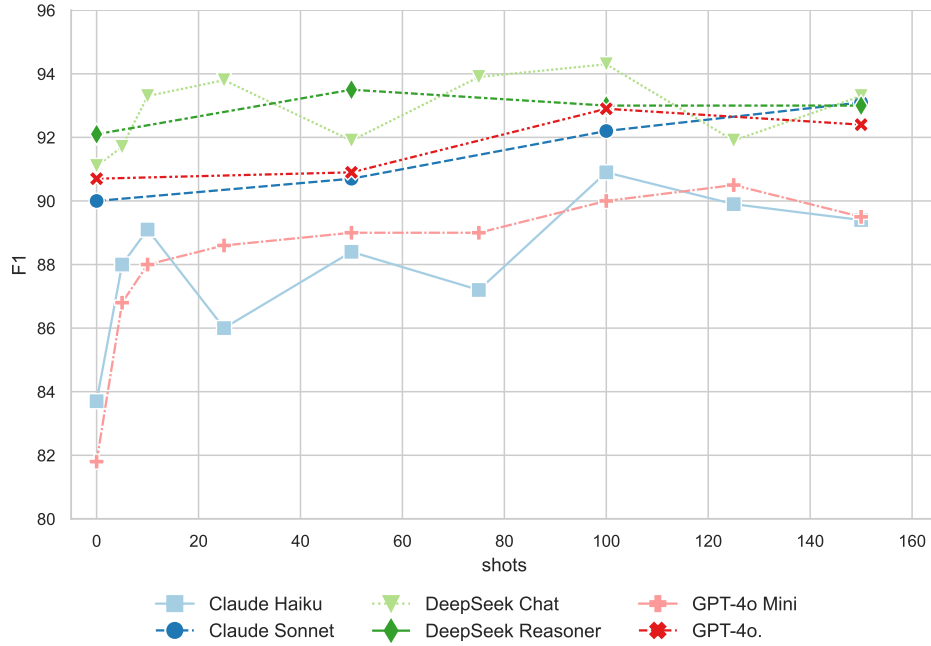


Figura 3.14: F_1 para la detección del par marco–objetivo según número de ejemplos para el Experimento 2.

en F_1 a medida que se agregan ejemplos. Sin embargo, partiendo de valores superiores 0.900 de F_1 en el escenario *zero-shot*, las ganancias son relativamente modestas.

El Experimento 3, en sus dos versiones, evalúa la extracción de argumentos asumiendo que los objetivos y sus marcos ya están disponibles. Se espera que el modelo identifique y agregue todos los argumentos asociados a los pares marco–objetivo que se proporcionan junto al texto de entrada. Como en experimentos previos, el número de ejemplos comienza en 0, 5 y 10, y luego se incrementa de 25 en 25 para los modelos pequeños y de 50 en 50 para los modelos grandes, comenzando desde 25 y 50 respectivamente.

En particular, en el Experimento 3a, los pares marco–objetivo se extraen del conjunto de referencia. Esto permite aislar el desempeño de los modelos en la identificación de argumentos sin la influencia de errores en la detección de marcos y objetivos, proporcionando un análisis más preciso de la capacidad de los modelos para extraer información semántica.

Al igual que en el Experimento 1, los modelos grandes superan a sus contrapartes más pequeñas por un margen de 0.05 a 0.07 puntos. Aunque esta

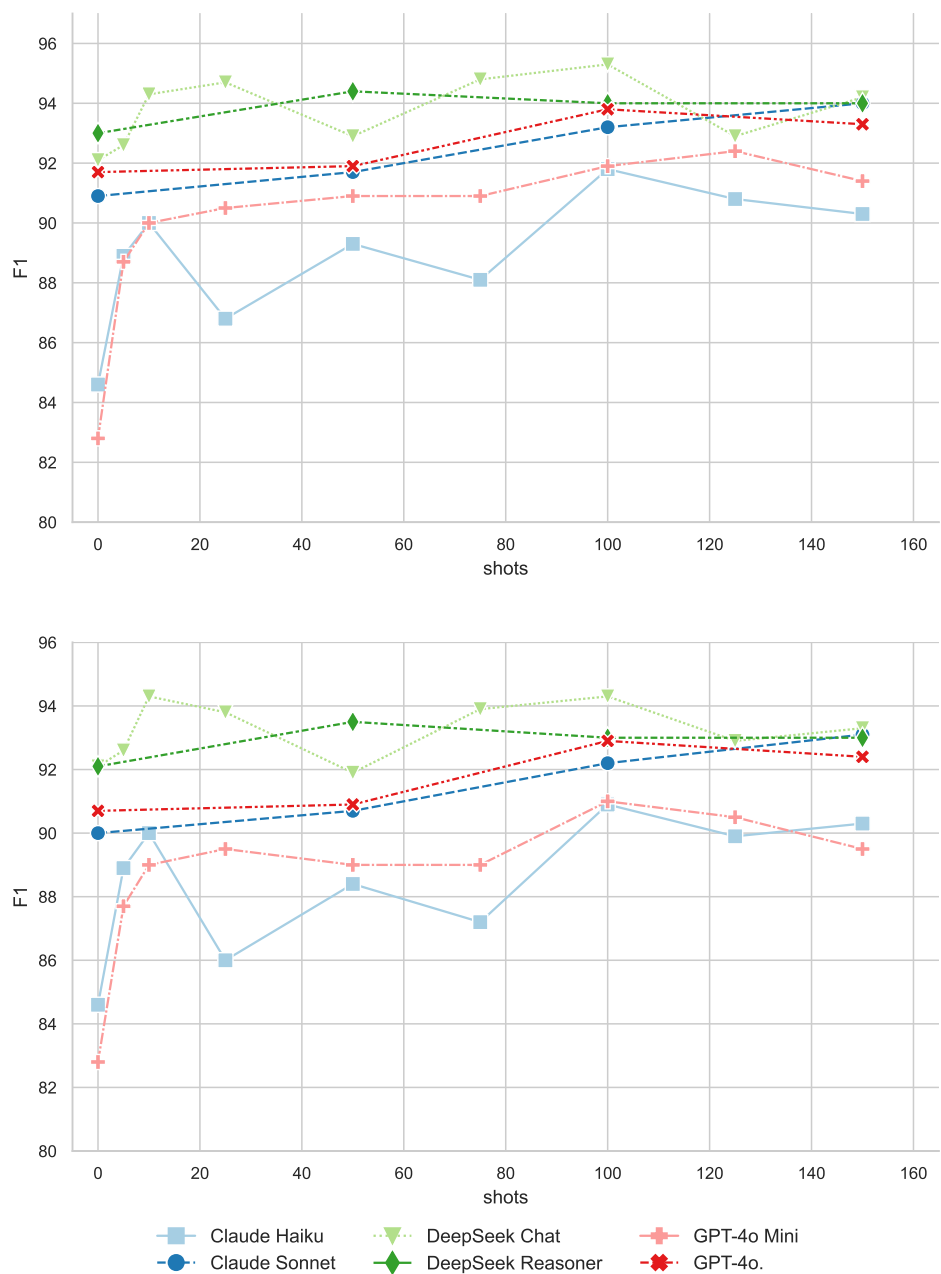


Figura 3.15: F_1 para el Marco (arriba) y Objetivo (abajo) según número de ejemplos para el Experimento 2.

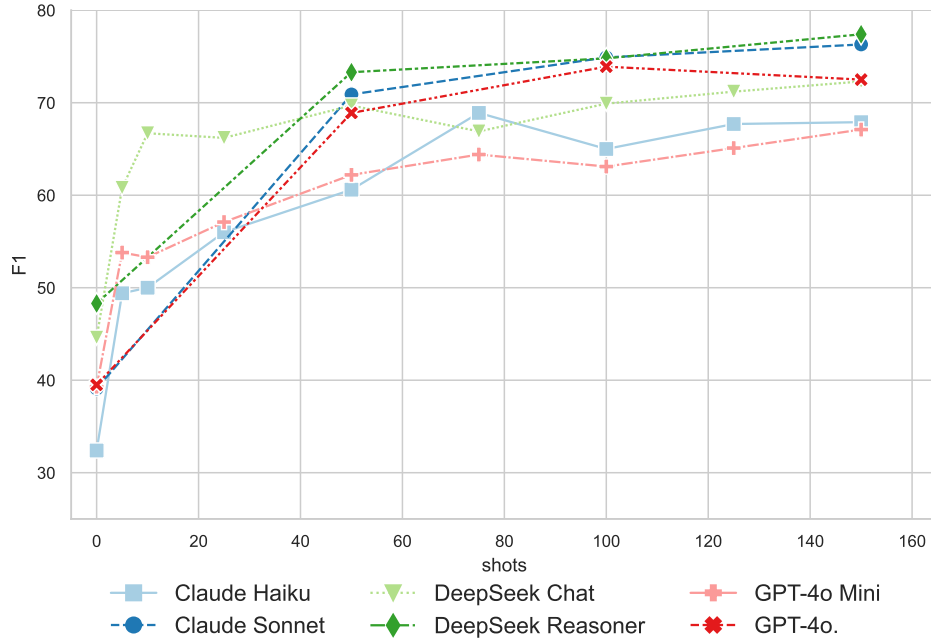


Figura 3.16: F_1 para el Experimento 3a según número de ejemplos con pares marco–objetivo de referencia.

diferencia es considerable, es menos pronunciada que en el primer experimento.

Las figuras 3.16 y 3.17 presentan F_1 , precisión y cobertura al variar la cantidad de ejemplos; persiste el mismo comportamiento de la precisión y cobertura que se observa en el Experimento 1: mientras los modelos GPT tienden a lograr una mayor precisión, los modelos Claude y DeepSeek muestran una mayor cobertura.

Finalmente, el Experimento 3b integra las tareas de identificación de marcos y extracción de argumentos, evaluando la solución en un escenario punta a punta. A diferencia del Experimento 3a, la lista de pares marco–objetivo se deriva de la salida del Experimento 2.

No se realiza evaluación cruzada entre modelos; cada modelo utiliza su propia salida como entrada. Estos resultados reflejan la capacidad de cada modelo para manejar la tarea integrada, donde el resultado de la etapa previa impacta directamente en la calidad del resultado en FSRL. Como antes, los seis modelos son incluidos, evaluándose con 150 ejemplos tanto para FI como para FSRL.

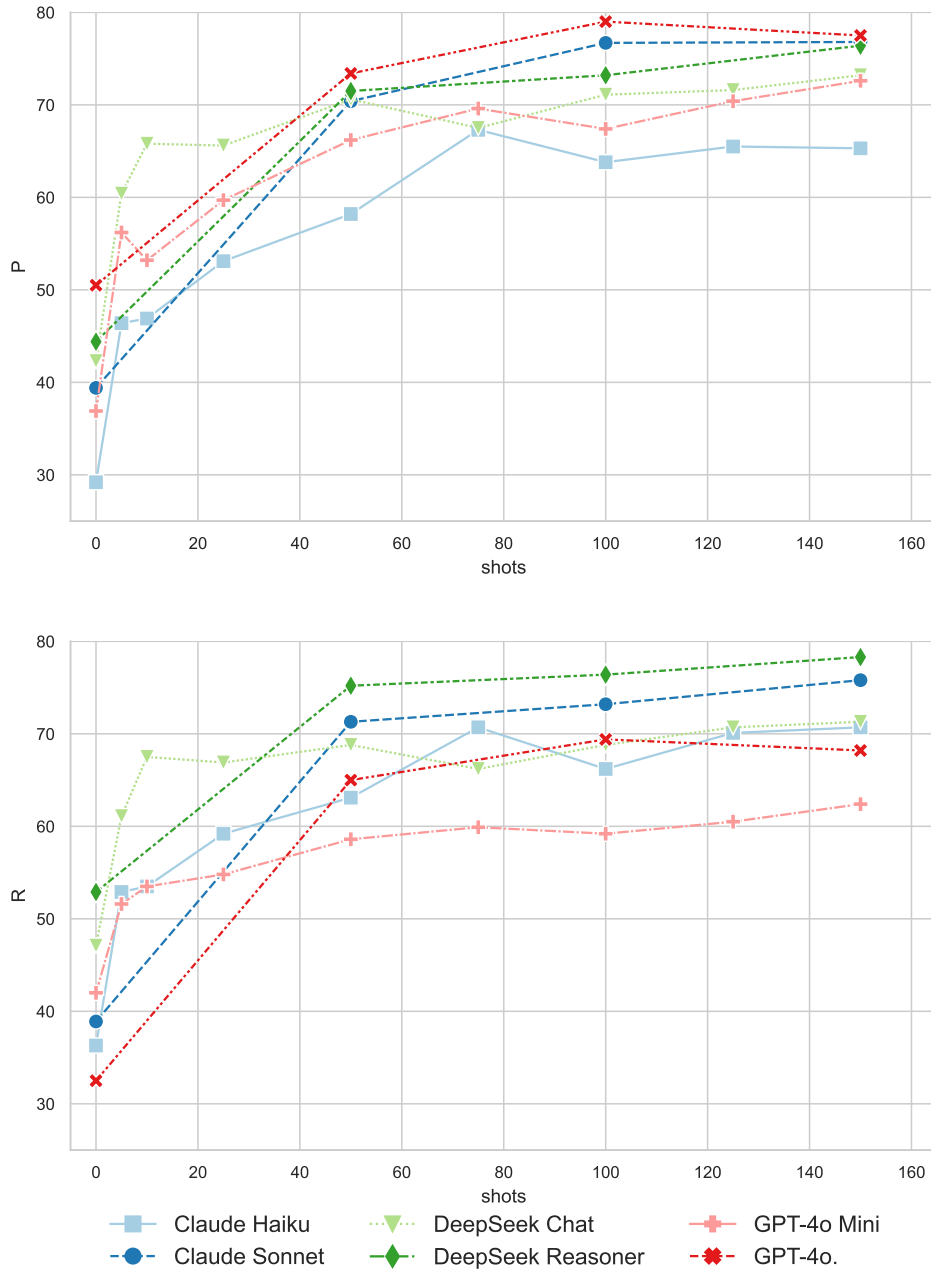


Figura 3.17: P (arriba) y R (abajo) para el Experimento 3a según número de ejemplos con pares marco-objetivo de referencia.

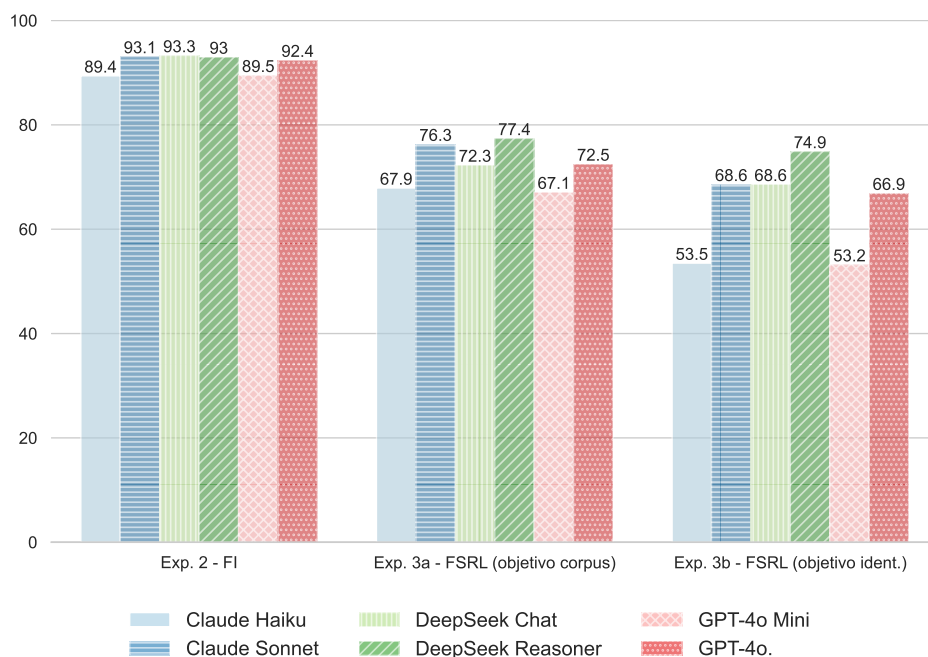


Figura 3.18: F_1 , con 150 ejemplos, para FI (Experimento 2) y FSRL con pares marco–objetivo de referencia (Experimento 3a) y extraídos en la etapa previa (Experimento 3b).

Las figuras 3.18 y 3.19 presentan los puntajes F_1 , P y R para 150 ejemplos en los experimentos 2, 3a y 3b. Como es esperable, todos los modelos presentan un mejor rendimiento en el Experimento 3a, que parte de pares marco–objetivo de referencia, comparado con Experimento 3b, donde estos elementos se identifican automáticamente. La caída en F_1 varía entre 0.025 y 0.144 puntos entre los modelos, como se muestra en la Tabla 3.6.

Los modelos DeepSeek destacan por la estabilidad de sus resultados, con una caída de solo 0.025 puntos para Reasoner y 0.035 para Chat. Para GPT-4o, la caída es de 0.056 puntos, mientras que GPT-4o Mini experimenta una caída de 0.139 puntos. Ambos modelos Claude también muestran disminuciones importantes, con Sonnet cayendo 0.077 puntos y Haiku 0.144 puntos. Comparando precisión y cobertura, GPT-4o Mini y Claude Haiku destacan por su bajo rendimiento, mostrando caídas sustanciales de más de 0.118 puntos en cobertura —comparado con menos de 0.067 puntos para otros modelos— y más de 0.15 puntos en precisión —frente a menos de 0.042 puntos para los tres mejores modelos—.

Comparando los resultados de los experimentos 1 y 3b, se observa que

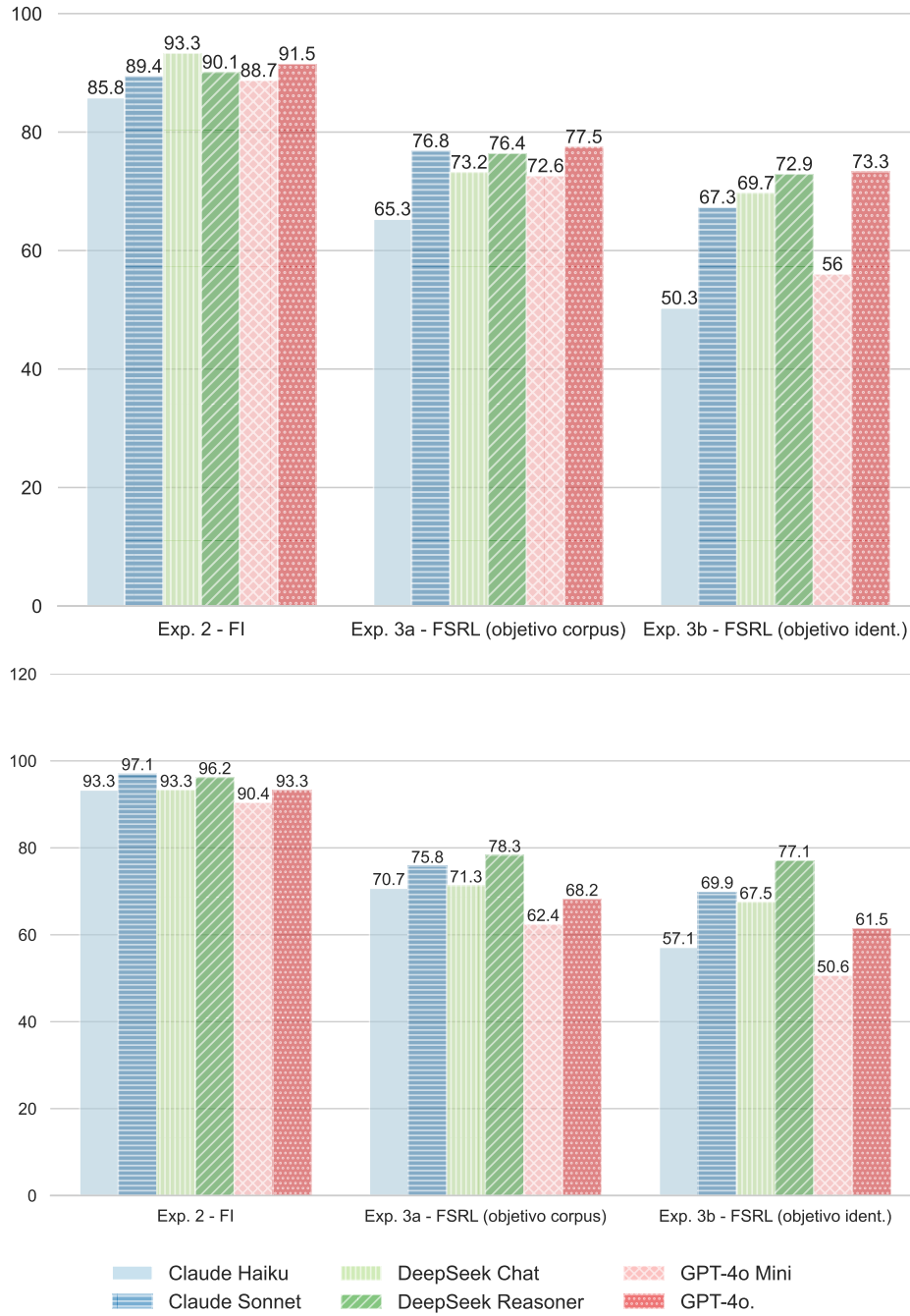


Figura 3.19: P (arriba) y R (abajo), con 150 ejemplos, para FI (Experimento 2) y FSRL con pares marco-objetivo de referencia (Experimento 3a) y extraídos en la etapa previa (Experimento 3b).

Modelo		Exp. 1	Exp. 3b
Claude Haiku 3.5	(275)	0.711	0.676
Claude Sonnet 3.5	(250)	0.829	<u>0.786</u>
DeepSeek Chat	-		0.785
DeepSeek Reasoner	-		0.825
GPT-4o Mini	(375)	0.687	0.668
GPT-4o	(400)	<u>0.802</u>	0.780

Tabla 3.6: Resultados para los experimentos 1 y 3b para la tarea de FSP. Para el Experimento 1 se detalla entre paréntesis la cantidad de ejemplos en que se obtuvo el resultado; para el Experimento 3b esa cantidad es de 150 ejemplos.

los modelos grandes mantienen un desempeño competitivo incluso cuando se enfrentan a múltiples marcos y a la propagación de errores de la detección automática de los marcos y objetivos.

Más allá de la tarea de extracción de argumentos, la solución encadenada que integra FI y FSRL también aborda el mismo objetivo que el Experimento 1, la tarea FSP completa. La Tabla 3.6 muestra los puntajes F_1 considerando la extracción de objetivo y argumentos para los seis modelos. DeepSeek Reasoner alcanza el mayor F_1 con 0.825. Claude Sonnet sigue con 0.786 y DeepSeek Chat llega a 0.785. GPT-4o obtiene 0.780, mientras que Claude Haiku y GPT-4o Mini logran 0.668 y 0.780, respectivamente. Estos resultados demuestran que los modelos grandes tienden a rendir mejor en la tarea integrada, con DeepSeek Reasoner mostrando el manejo más efectivo del proceso completo.

Dado que el conjunto de frames y los datos de evaluación difieren del Experimento 1, los valores no son directamente comparables entre experimentos. Sin embargo, incluso considerando múltiples marcos en lugar de solo «Asesinato», todos los modelos alcanzan resultados competitivos usando significativamente menos ejemplos —particularmente GPT-4o, que requiere tres veces menos ejemplos— con una disminución de rendimiento entre 0.019 y 0.043 puntos de F_1 .

En general, los resultados confirman un desempeño competitivo de los modelos evaluados, con variaciones consistentes según la tarea y el tamaño del modelo. DeepSeek Reasoner y Claude Sonnet 3.5 se destacan en las tareas más complejas, mientras que los modelos más pequeños mantienen un rendimiento adecuado en tareas más sencillas, como la identificación de marcos.

Tarea	Modelo	a. <i>Zero-shot</i> Sin info.	b. <i>Zero-shot</i> Solo def.	c. <i>Zero-shot</i>	d. <i>Best-shot</i>
Exp. 1 FSP	Claude Haiku	0.0	0.0	<u>40.6</u>	(275) 71.1
	Claude Sonnet	0.0	0.0	<u>48.4</u>	(250) 82.9
	GPT-4o Mini	0.0	0.0	<u>34.1</u>	(375) 68.7
	GPT-4o	0.0	0.0	<u>38.8</u>	(400) 80.2
Exp. 2 FI	Claude Haiku	69.9	74.5	<u>83.7</u>	(100) 90.9
	Claude Sonnet	82.2	83.1	<u>90.0</u>	(150) 93.1
	DeepSeek Chat	90.4	<u>91.7</u>	91.1	(100) 94.3
	DeepSeek Reasoner	<u>92.6</u>	92.2	92.1	(50) 93.5
	GPT-4o Mini	67.3	63.5	<u>81.8</u>	(125) 90.5
	GPT-4o	74.1	71.3	<u>90.7</u>	(100) 92.9
Exp. 3a FSRL	Claude Haiku	22.2	<u>33.0</u>	32.4	(75) 68.9
	Claude Sonnet	35.0	<u>39.1</u>	<u>39.1</u>	(150) 76.3
	DeepSeek Chat	28.0	42.4	<u>44.6</u>	(150) 72.3
	DeepSeek Reasoner	41.3	<u>50.9</u>	48.3	(150) 77.4
	GPT-4o Mini	3.9	36.4	<u>39.3</u>	(150) 67.1
	GPT-4o	25.6	32.1	<u>39.5</u>	(100) 73.9

Tabla 3.7: F_1 por modelo y estrategia en los experimentos 1, 2 y 3a con ablación: (a) sin información de los marcos ni ejemplos en contexto; (b) únicamente con la definición del marco; (c) con información completa de los marcos, pero sin ejemplos en contexto; (d) con información completa de los marcos junto con ejemplos en contexto. En el caso (d), el número entre paréntesis indica la cantidad de ejemplos utilizados con el mejor resultado. Los valores más altos se muestran en negrita y los segundos mejores, subrayados.

3.3.4. Ablación

Para evaluar el impacto en las distintas tareas de la información derivada de FrameNet, se consideran dos configuraciones adicionales. En la primera, los *prompts* incluyen únicamente la descripción de la tarea e instrucciones generales, sin contenido específico de FrameNet. En la segunda, los *prompts* se amplían para incluir la definición de los marcos, pero se omiten los detalles o ejemplos de sus argumentos asociados. En el Apéndice A, se exponen ejemplos de los *prompts* utilizados para estas dos variantes.

La Tabla 3.7 presenta los valores de F_1 obtenidos para estas configuraciones, junto con los resultados que incorporan tanto las definiciones de los marcos como las de sus argumentos, con y sin ejemplos en contexto.

Los resultados muestran que la inclusión de las definiciones de los marcos y ejemplos mejora de forma consistente el rendimiento en todas las tareas, especialmente en FSP. Para tareas más simples, como FI, los modelos DeepSeek logran igualar o superar el desempeño en configuración *zero-shot* incluso con información limitada sobre los marcos, aunque con diferencias marginales.

En conjunto, los resultados confirman el impacto positivo de la información de FrameNet en el rendimiento de los modelos, aunque la magnitud de su beneficio depende tanto de la complejidad de la tarea como de las capacidades del modelo.

3.4. Síntesis

Teniendo como objetivo la extracción de eventos en la Base de Jurisprudencia Nacional, y tomando como base la Teoría de Marcos de Fillmore, se investiga la viabilidad de aplicar aprendizaje en contexto a la tarea de análisis de marcos semánticos. Este capítulo explora, mediante una serie de experimentos, el desempeño de grandes modelos de lenguaje en FSP, incluyendo la identificación de los marcos, sus objetivos y sus elementos, utilizando la información disponible en FrameNet.

Para abordar las distintas subtareas, la solución propuesta genera de forma automática *prompts* adaptados a cada tarea y al conjunto de marcos seleccionados, basándose exclusivamente en las definiciones y ejemplos disponibles en FrameNet. Debido a las limitaciones de longitud de entrada inherentes a este enfoque, la evaluación se realiza sobre un conjunto de marcos asociados a eventos de violencia: «Abuso», «Asesinato», «Robo», «Disparo» y «Violencia», utilizando seis modelos distintos.

Todos los modelos presentan mejoras al incluir ejemplos específicos, indicando que el conocimiento adquirido durante su entrenamiento no es suficiente para resolver estas tareas. A su vez, el grado de mejora varía según el modelo y la complejidad de la tarea abordada. Los experimentos muestran que el uso de ICL para tareas de FSP puede alcanzar valores de F_1 competitivos en un conjunto limitado de marcos, lo que sugiere que el enfoque es especialmente adecuado para dominios acotados, como los eventos de violencia en la BJN, ofreciendo una alternativa viable al entrenamiento de modelos específicos.

En particular, los resultados para FSRL muestran que aquellos elementos del marco menos frecuentes son los más difíciles de extraer, sin importar el

número de ejemplos incluidos en el *prompt*. Asimismo, la evaluación estricta —los segmentos deben coincidir exactamente con los esperados— o las posibles inconsistencias u omisiones presentes en las anotaciones de FrameNet contribuyen a que el desempeño sea menor al deseado.

Si bien los resultados obtenidos son prometedores, existen varias áreas que merecen una investigación más profunda. Entre ellas se incluyen la optimización de la selección de ejemplos, buscando aquellos que sean lo más parecidos posibles al ejemplo a tratar, o estudiar el orden de los ejemplos, que puede afectar la calidad de la respuesta (Devasier et al. 2025; P. Liu et al. 2023; Zhao et al. 2021).

En conjunto, los resultados confirman la factibilidad del uso de aprendizaje en contexto para el análisis de marcos semánticos, estableciendo un punto de partida para su aplicación en la extracción de eventos en la Base de Jurisprudencia Nacional.

Los resultados de esta investigación se presentan en la conferencia arbitrada «Foundation and Large Language Models» (Garat et al. 2025). Cabe destacar que, salvo el trabajo paralelo de Devasier et al. (2025) relacionado únicamente a la subtarea FSRL, no se han reportado estudios previos que apliquen explícitamente ICL a tareas de FSP utilizando FrameNet.

Capítulo 4

Eventos de violencia

Una vez concluida la evaluación de análisis de marcos semánticos (FSP) con FrameNet utilizando aprendizaje en contexto (ICL), desarrollada en el Capítulo 3, se trabaja sobre el siguiente objetivo: la extracción de eventos relevantes en las sentencias judiciales contenidas en la Base de Jurisprudencia Nacional de Uruguay (BJN).

En particular, esta etapa del trabajo toma como caso de estudio la detección de eventos de violencia —como homicidios, violaciones o golpes, entre otros— junto con sus participantes y circunstancias. Con la identificación de estos eventos de violencia, se pretende favorecer a la accesibilidad y aprovechamiento de la información contenida en BJN: la información estructurada y relevante puede contribuir a la generación de resúmenes automáticos, a estudios estadísticos sobre fenómenos sociales relacionados con la violencia o a la mejora de la recuperación de la información jurídica contenida en las sentencias judiciales, entre otras aplicaciones (Boella et al. 2019; Habba et al. 2023; Louis et al. 2023; Van Opijnen y Santos, 2017).

En el fragmento del Ejemplo 4.1 se muestra el resultado de identificar un hecho violento —un golpe—, al agente y a su víctima —AA y BB—, junto con el instrumento —la culata del arma—.

- (4.1) ...cuando ingresaron AA^{agente}, de 18 años, junto a CC de 15 años junto a otros adolescentes portando una escopeta y un revólver de juguete; con la culata del arma^{instr.} golpea a BB^{víctima} (fs 18) —ocasionándole las lesiones certificadas por médico forense a fs 20 por las que formuló instancia—...

Sentencia 55/2014,
Tribunal de Apelaciones Penal de 1º Turno

La presente etapa de la investigación se orienta, entonces, en la aplicación de análisis de marcos semánticos sobre documentos pertenecientes a la BJN, basados en el trabajo de la etapa previa con FrameNet (C. F. Baker et al. 1998) y aprendizaje en contexto.

Dado que la BJN se encuentra en español, lo deseable sería utilizar definiciones de marcos para este idioma. Sin embargo, aunque existen trabajos sobre definiciones de FrameNet en español o incluso multilingües, no se encuentran bases de datos con la cantidad de ejemplos anotados como las que ofrece el proyecto de la Universidad de Berkeley para el inglés (C. F. Baker y Lorenzi, 2020; Subirats Rüggeberg y Petruck, 2010).

Ante esta situación, se decide investigar la capacidad multilingüe de los grandes modelos de lenguaje (LLM) mediante el uso de ejemplos tanto en inglés como en español para el aprendizaje en contexto (Lin et al. 2022; Soltan et al. 2022; Xue et al. 2021). Este tipo de estrategias multilingües demuestra ser eficaz en dominios legales de bajos recursos (Moro et al. 2023).

Tanto para construir el contexto de los modelos, así como para la evaluación de los resultados en la BJN, se requiere de un conjunto de ejemplos previamente anotados con las relaciones de violencia a extraer, en un formato análogo al utilizado por FrameNet. Al igual que para el problema de desidentificación, se anota manualmente un conjunto de sentencias con los eventos de violencia, al cual se lo denomina ViolenciaR.

La anotación en el corpus ViolenciaR se basa en una lista de verbos, obtenida a partir de la compilación de verbos del español del proyecto ADESSE de la Universidad de Vigo (García-Miguel et al. 2010). Esta lista es refrendada empíricamente con aquellos verbos que efectivamente se encuentran presentes en 5000 sentencias seleccionadas de la BJN. Los eventos y sus respectivos argumentos son anotados manualmente en un subconjunto de sentencias judiciales, siguiendo pautas similares a las empleadas en FrameNet.

Teniendo en cuenta el carácter sensible de los textos de la BJN, se procede a la desidentificación manual de los ejemplos para preservar la privacidad de las personas involucradas. Este proceso incluye la eliminación o modificación de nombres propios, direcciones, fechas y cualquier otra información que pueda permitir la identificación de individuos específicos.

Al igual que en la etapa previa, se utilizan varios grandes modelos de lenguaje durante la experimentación. En este caso, se recurre a Claude Sonnet 3.7, GPT-4o y DeepSeek Reasoner (Anthropic, 2024; DeepSeek-AI et al. 2025;

OpenAI, 2024).

En cambio, a diferencia de la etapa previa en la cual tanto los textos de entrada como los *shots* provienen de FrameNet y se encuentran redactados en inglés, en este nuevo escenario los textos a analizar se encuentran en español, y los *shots* son una mezcla entre los ejemplos en inglés de FrameNet y ejemplos en español de ViolenciaR. Además, se analiza el desempeño de los modelos según la distribución de los ejemplos utilizados como *shots*, variando la proporción de ejemplos provenientes de cada una de las fuentes. En todos los experimentos se utiliza como conjunto de evaluación a una parte del corpus ViolenciaR apartada para tal fin.

Este capítulo se organiza de la siguiente forma: a continuación, en la Sección 4.1, se presentan los trabajos relacionados con la extracción de información en textos judiciales. La Sección 4.2 desarrolla el proceso de confección de la lista de verbos de violencia, mientras que la Sección 4.3 describe el proceso de anotación en la BJN para la creación del corpus ViolenciaR. En la Sección 4.4 se detallan los experimentos realizados con ICL y sus resultados. Finalmente, en la Sección 4.5 se encuentra una síntesis de este capítulo.

4.1. Trabajos Relacionados

El trabajo del presente capítulo se enmarca en el análisis semántico de textos legales, con un esquema de anotación de eventos de violencia basado en FrameNet, utilizando técnicas de aprendizaje por contexto con grandes modelos de lenguaje para la extracción de información. En el capítulo anterior se presenta una revisión de los trabajos relacionados con el uso de grandes modelos de lenguaje y FrameNet para el análisis semántico. Esta sección complementa esa revisión con algunos trabajos recientes que exploran el uso de grandes modelos de lenguaje en el dominio legal, destacando sus enfoques y resultados.

Savelka (2023) utiliza el modelo GPT-3.5 con *zero-shot* para la categorización de pequeños textos legales. Los autores comparan los resultados de GPT-3.5 con un clasificador basado en bosques aleatorios (*Random Forests*) y otro obtenido por ajuste fino utilizando RoBERTa (Y. Liu et al. 2019). En un corpus consistente de 510 contratos legales, clasificados manualmente por abogados profesionales, obtienen un resultado de 0.860 de medida F_1 para GPT-3.5, mientras que el clasificador basado en bosques aleatorios al-

canza 0.830 y el modelo ajustado con RoBERTa obtiene 0.920. Sin embargo, para textos provenientes de un corpus de disposiciones legales y reglamentarias, GPT-3.5 alcanza un F_1 de 0.540, mientras que el clasificador basado en bosques aleatorios obtiene 0.630 y el modelo ajustado con RoBERTa alcanza 0.720. Aunque en todos los casos el modelo ajustado supera a GPT-3.5, los resultados obtenidos varían según el tipo de documento a clasificar.

En el trabajo de Adhikary et al. (2024) se explora el etiquetado secuencial usando un modelo BiLSTM-CRF para extraer información de sentencias de la Suprema Corte de la India. La investigación se centra en identificar segmentos específicos en las sentencias judiciales, etiquetándolos en una de siete categorías predefinidas, tales como «Testimonio de experto», «Testimonio de testigo», «Asalto» o «Disturbios», entre otras. Debido a la necesidad, para entrenar al modelo supervisado, de ejemplos en las categorías menos frecuentes, los autores proponen un enfoque de aprendizaje débil basado en ejemplos generados utilizando ICL. Dentro de los modelos utilizados se encuentran GPT-3.5, GPT-4 y LLaMA-2 (Touvron et al. 2023). Aunque los mejores resultados globales se obtienen con LLaMA-2, ningún modelo logra ser el primero en todas las categorías, siendo superados en algunas de ellas por aquellos entrenados sin aprendizaje débil.

Italiani et al. (2025) también utilizan un LLM para la generación de datos sintéticos en el dominio legal, con la finalidad de entrenar un modelo supervisado para la respuesta a preguntas legales (*Legal Question Answering*, LQA). En su enfoque, denominado ACE-ATTORNEY, un modelo como LLaMA-2-13B genera ejemplos artificiales de pares pregunta-respuesta, a partir de resúmenes sobre juicios civiles y políticas de privacidad. Estos ejemplos se utilizan para el ajuste fino de un modelo de menor porte, como BART (Lewis et al. 2019) o FLAN-T5 (Chung et al. 2024), que actúa como modelo estudiante, logrando resultados comparables a los obtenidos por modelos de mayor porte.

Bakker et al. (2025) se plantean como objetivo el etiquetado de roles de oraciones provenientes de textos legales, y toman en cuenta únicamente los roles generales: acción, actor, objeto y receptor. En su investigación, realizan un ajuste fino de varios modelos de lenguaje basados en BERT (Devlin et al. 2019), utilizando 1575 oraciones de regulaciones de la Unión Europea. Además, exploran el uso de un sistema de reglas basadas en estructuras sintácticas y de *zero-shot* y *5-shot* con GPT-3.5 y GPT-4. Los modelos ajustados alcanzan un F_1 de 0.822, superando significativamente a los enfoques basados en reglas

o LLM, cuyos mejores resultados se obtienen con GPT-4 para *zero-shot* y *5-shot*, con un F_1 de 0.176 y 0.197, respectivamente.

Por su parte, Ribeiro de Faria et al. (2025) exploran el uso de GPT-4 para la extracción de información de sentencias del Tribunal Laboral del Reino Unido, con el fin de facilitar el análisis de disputas laborales y predecir sus resultados. En particular, su trabajo abarca la extracción de los hechos del caso, los reclamos de cada parte, las referencias a estatutos legales y precedentes, el resultado general del caso, y los motivos de la decisión. La evaluación de la extracción se realiza de forma manual por parte de un experto legal, asignando una puntuación binaria de 0 o 1 a cada uno de los ocho elementos extraídos de 260 documentos, dependiendo de si la información es precisa o no. Los resultados de GPT-4 en todas las categorías superan los 0.91 puntos de F_1 , con algunas categorías, como la de referencias a estatutos legales y a precedentes, alcanzando el máximo valor.

La extracción de términos legales es el objeto del estudio de Breton et al. (2025), quienes evalúan ICL y ajuste fino para esta tarea con grandes modelos de lenguaje. El estudio se realiza sobre un corpus de 1339 fragmentos de texto en francés, extraídos de 200 artículos de la Ley de Tráfico de Luxemburgo (Sleimi et al. 2018), en los cuales expertos de dominio anotaron catorce conceptos legales. De esos catorce conceptos, los autores seleccionan ocho para su trabajo, entre los que se encuentran acción, actor, objeto, y lugar. Se prueban varios enfoques: *zero-shot* y *1-shot* con GPT-4, Miqu-1-70b, basado en LLaMa-2, Mixtral-8x7b y Mistral-7b, así como ajuste fino de Legal-CamemBERT (A. Q. Jiang et al. 2023, 2024). El mejor resultado de F_1 macro¹ de 0.690 se obtiene con *1-shot* con GPT-4, superando al ajuste fino de Legal-CamemBERT, que alcanza un F_1 macro de 0.664. Estos resultados, sin embargo, no logran competir con un enfoque basado en reglas desarrollado por Sleimi et al. (2018), que alcanza un F_1 macro de 0.965.

Breton et al. (2025) plantean la importancia de distinguir entre el acierto perfecto y parcial, donde este último refiere a anotaciones donde se reconoce la entidad pero sus límites no coinciden exactamente con los de la anotación manual, sino que se encuentra comprendida por ella. Los autores afirman que los grandes modelos de lenguaje sufren particularmente de problemas con los límites de sus extracciones, que podrían estar relacionados con la naturale-

¹Promedio de las F_1 de las distintas categorías. A diferencia de su versión micro, la distribución de las instancias en las distintas categorías no afecta al resultado.

za generalista debido al amplio espectro de textos durante su entrenamiento. Por ejemplo, en su trabajo, la solución basada en reglas encuentra 175 instancias parciales, mientras que los LLM encuentran aproximadamente 415. En su análisis introducen la distancia de Levenshtein² normalizada como métrica para evaluar la similitud entre las anotaciones de los modelos y las anotaciones manuales en los casos de acierto parcial. Si se toma un umbral de esa distancia para considerar por bueno un acierto parcial, se puede observar cómo varía el F_1 macro de los modelos según su valor. Considerando un umbral de 0.46, el F_1 macro de GPT-4 en 1-*shot* aumenta de 0.690 a 0.907, aproximándose al rendimiento del sistema basado en reglas.

Con respecto a la extracción de información en la Base de Jurisprudencia Nacional, existe como antecedente el trabajo de Fulloni et al. (2020), desarrollado como tesis de grado, en el cual se utiliza un enfoque basado en reglas y el uso de redes convolucionales para la extracción de personas consideradas referencias doctrinales, junto a leyes y sentencias. Con esas referencias extraídas se genera un grafo de relaciones, permitiendo su análisis y la mejora de las consultas aprovechando estos vínculos.

Finalmente, no se encuentran trabajos previos que utilicen la BJN como fuente de datos para la experimentación con modelos de lenguaje, ni tampoco en tareas de análisis semántico, como la que se presenta en este capítulo.

4.2. Verbos de violencia en la BJN

En este trabajo se consideran a los eventos de violencia como aquellos eventos que involucran actos de agresión física o psicológica entre individuos o grupos, y pueden manifestarse en diversas formas, como homicidios, asaltos o violaciones, entre otros.

El límite entre un acto de violencia y otro de distinto tipo puede ser difuso, dependiendo del contexto y el alcance que se le quiere dar al análisis. Por ejemplo, el verbo «discutir» puede referirse a un intercambio verbal acalorado que no necesariamente implica violencia física, aunque en ciertos contextos puede ser un acto violento o terminar escalando a uno. Un robo puede involucrar violencia física si se utiliza la fuerza o amenazas —por ejemplo, una rapiña—

²La medida de Levenshtein mide la similitud entre dos textos, calculando la cantidad mínima de alteraciones necesarias para transformar un texto en el otro. Es una de las medidas consideradas para la resolución de correferencias en la Sección 2.5.2.

pero también puede no ser considerado como tal si se realiza sin confrontación directa.

Con la finalidad de definir empíricamente cuáles son los eventos de violencia a extraer de las sentencias judiciales, se realiza un trabajo exploratorio sobre la BJN, trabajando, en un principio, con aquellos eventos señalados por verbos. De esta forma, se consideran únicamente eventos expresados mediante grupos verbales, como es el caso de «golpea» en el Ejemplo 4.1 anterior.

La anotación basada en verbos deja de lado otras formas de ocurrencias de eventos en los textos, como, por ejemplo, la nominalización del evento en el fragmento del Ejemplo 4.2. De ese texto se puede extraer, tanto al hecho, «asesinato», como a la víctima, «JJ». Sin embargo, este evento es omitido en este primer análisis por figurar como parte de un grupo nominal, «el asesinato de JJ», y no de un grupo verbal, como sí sería el caso en, por ejemplo, «JJ fue asesinado el 7 de noviembre de 2010».

- (4.2) ...una relación de causalidad entre la fuga del menor GG del hogar “XX” del INAU —acaecida el 18 de agosto de 2010— y **el asesinato**^{asesinato} **de JJ**^{víctima} ocurrido el 7 de noviembre de 2010, como para mantener la condena impuesta a dicho organismo. ...

Sentencia 2.034/2017 Definitiva,
Suprema Corte de Justicia, Recurso de Casación

Se busca, entonces, establecer cuáles son los «verbos de violencia» a extraer, entendiéndose por «verbo de violencia» a aquel que refiere a un hecho violento cuando se utiliza en su acepción correspondiente. Esta lista se define mediante un proceso de exploración manual sobre un conjunto de 5000 documentos de la BJN, enriqueciendo esta búsqueda con la categorización de verbos del proyecto ADESSE (García-Miguel et al. 2010). Además, se determinan los complementos de esos verbos —como puede ser el agente, la víctima, el instrumento, etc.— de forma de contemplar las principales dimensiones de un evento de violencia. Para esto, se consideran a los distintos complementos que define FrameNet para los verbos elegidos según su marco de referencia.

La lista de verbos de violencia se crea a través de un proceso iterativo semiautomático que consta, básicamente, de tres etapas.

En una primera etapa, se genera una lista con todas las ocurrencias de los verbos presentes en 5000 documentos de la BJN. Para esta finalidad, se realiza el etiquetado de categorías gramaticales de los textos con SpaCy (Honnibal y Montani, 2017) para detectar los segmentos correspondientes a los verbos. El

Pos.	Lema	Frec.	Pos.	Lema	Frec.	Pos.	Lema	Frec.
1	tener	22 158	17	recibir	5 061	35	compartir	3 547
2	hacer	13 911	19	solicitar	5 036	36	referir	3 510
3	decir	12 845	20	llevar	5 017	37	reconocer	3 421
4	dar	10 847	21	establecer	4 853	38	proceder	3 419
5	resultar	9 239	22	señalar	4 838	39	mantener	3 400
6	existir	8 505	23	declarar	4 806	40	determinar	3 387
7	corresponder	8 095	24	acordar	4 739	41	ajustar	3 362
8	considerar	7 674	25	realizar	4 620		fallar	
9	encontrar	7 179	26	sostener	4 602		ingresar	
10	disponer	6 927	27	citar	4 440	44	admitir	3 195
11	expresar	6 534	28	imponer	4 342	45	cometer	3 186
12	ver	6 239	29	quedar	4 145	46	tomar	3 107
13	tratar	6 163	30	estar	4 021	47	franquear	3 091
14	surgir	6 076	31	cumplir	3 980	48	manifestar	3 027
15	entender	5 576	32	permitir	3 898	49	interponer	3 010
16	condenar	5 486	33	pasar	3 658	50	concluir	3 005
17	confirmar	5 061	34	revocar	3 598			

Tabla 4.1: Lista de los 50 verbos más frecuentes del total de 8 764 lemas detectados en 5 000 documentos de la BJN según SpaCy, ordenados por cantidad de ocurrencias.

resultado de este paso es una lista de 734 112 ocurrencias de verbos, con la referencia al documento del que es extraído y su respectiva ubicación; la lista contiene 8764 lemas únicos.

En la Tabla 4.1 se listan los cincuenta verbos más frecuentes obtenidos en la primera etapa; ninguno de ellos puede considerarse como candidato para la lista de verbos de violencia. A modo de anticipo de los resultados finales, los cinco primeros verbos violentos ordenados según su frecuencia, —amenazar, violar, pegar, golpear y matar— aparecen recién en los puestos 136, 215, 245, 247 y 253, respectivamente.

En una segunda etapa, se procede a la búsqueda específica de verbos de violencia. Esto se realiza mediante un proceso iterativo: partiendo de una lista primaria de verbos —atacar, golpear, matar, robar y violar—, se detectan aquellos otros que coocurren con los verbos de esta lista en las mismas sentencias. Los nuevos verbos se ordenan por frecuencia y se revisan manualmente, extrayendo aquellos que puedan referir a eventos de violencia. Estos últimos

Pos. Lema	Frec.	Pos. Lema	Frec.	Pos. Lema	Frec.
1 sustraer	1 243	12 atacar	368	23 drogar	24
2 amenazar	1 173	13 violentar	246	24 secuestrar	21
3 violar	813	14 hurtar	245	25 desnudar	18
4 pegar	693	15 delinquir	203	26 asesinar	17
5 golpear	688	16 perpetrar	199	27 degollar	12
6 matar	669	17 atar	189	prostituir	
7 someter	656	18 penetrar	144	29 traficar	10
8 discutir	654	19 rapiñar	128	30 estrangular	8
9 agredir	481	20 forzar	106	31 regentear	2
10 lesionar	413	21 apuñalar	73	32 vilipendiar	1
11 robar	401	22 fotografiar	38	sodomizar	

Tabla 4.2: Lista de los 33 verbos, ordenados por frecuencia, resultantes del proceso iterativo de búsqueda de la segunda etapa.

pasan a engrosar la lista de verbos y se repite nuevamente el procedimiento.

El resultado de la segunda etapa es una lista de 33 verbos que se observan en la Tabla 4.2. Cabe señalar que algunos de ellos no se encuentran ligados necesariamente siempre a un hecho violento, como pueden ser los verbos «violar» o «discutir».

Finalmente, se decide expandir la lista mediante un método alternativo: la búsqueda basada en los verbos compilados y clasificados en el proyecto ADESSE. En concreto, se buscan en ADESSE veinte de los verbos obtenidos en la etapa anterior y se recolectan aquellos verbos que coincidan en una misma clase según esta taxonomía.

Por ejemplo, el verbo «asesinar» se encuentra en la clase «Vida», definida como: «Un ser vivo (prototípicamente un animal, y de forma especial un ser humano) permanece en un estado o experimenta un proceso o un cambio de estado directamente relacionado con su existencia biológica»³. En esta clase se encuentra junto a otros verbos como «degollar», «estrangular» o «aniquilar», que son agregados a la lista de verbos de violencia. En el Apéndice B se presenta una descripción detallada de este proceso.

El resultado de estas tres etapas es la lista de 59 verbos de la Tabla 4.3. Se observa que los verbos de este cuadro se utilizan un total de 8350 veces en las 5000 sentencias judiciales analizadas y representan poco más del 1 % de los

³Disponible en <https://adesse.webs.uvigo.es/data/clases.php?clase=53>

Pos.	Lema	Frec.	Pos.	Lema	Frec.	Pos.	Lema	Frec.
136	amenazar	1 173	994	luchar	56	2117	exterminar	7
215	violar	813	1040	asaltar	49	2392	desfigurar	5
245	pegar	693	1111	embestir	41		incapacitar	
247	golpear	688	1175	atropellar	36		sojuzgar	
253	matar	669	1194	acometer	34	2574	apuñalear	4
263	discutir	654	1224	patear	32		deshonrar	
312	disparar	542	1425	ahorcar	21	3372	acuchillar	2
335	agredir	481		secuestrar			ajusticiar	
362	lesionar	413	1538	asesinar	17		electrocutar	
391	atacar	368		asfixiar			espetar	
490	herir	247	1576	morder	16		raptar	
588	forcejear	171	1686	estafar	13		regentear	
598	pelear	165	1737	degollar	12		reñir	
618	lastimar	155		prostituir		4384	abofetear	1
680	rapiñar	128	1785	aniquilar	11		chantajear	
687	propinar	126	1855	extorsionar	10		envenenar	
724	atentar	115	2013	desangrar	8		vilipendiar	
730	dañar	114		estrangular			sodomizar	
843	manosear	83		fusilar			yugular	
890	apuñalar	73	2117	arrollar	7			

Tabla 4.3: Lista de los 59 verbos resultantes de la tercera etapa, ordenados según su conteo en los 5 000 documentos de la BJN. La columna posición indica el lugar por frecuencia en la lista original de verbos.

734 112 verbos detectados en la primera etapa por SpaCy. Los cinco primeros verbos acumulan un 48 % de las ocurrencias, los 10 primeros un 78 %, mientras que los 20 primeros ya llegan al 94 % de las ocurrencias de los verbos de la lista, como se grafica en la Figura 4.1.

4.3. Corpus ViolenciaR

Para la anotación de los eventos de violencia y sus argumentos en los textos del corpus, se selecciona un conjunto de verbos en función de su uso efectivo en la BJN y de su capacidad para referir a hechos de carácter violento.

De la lista de 59 verbos se seleccionan, entonces, entre los más frecuentes,

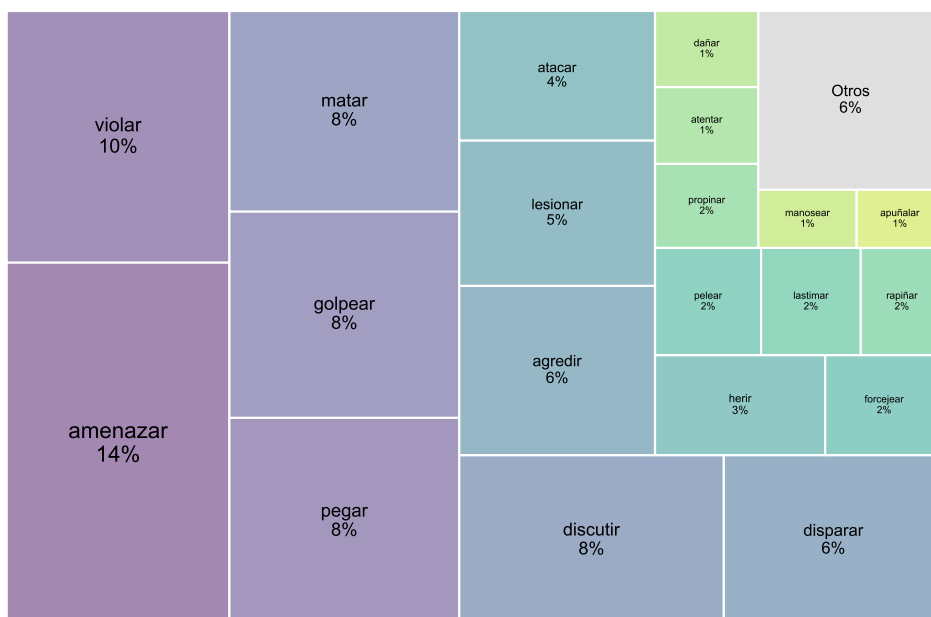


Figura 4.1: Mapa de áreas de los verbos violentos seleccionados en la Base Nacional de Jurisprudencia, donde se detalla el porcentaje de cada verbo sobre el total de ocurrencias de la lista. Los 20 verbos más frecuentes se muestran individualmente; la categoría «Otros» agrupa a los verbos restantes.

los siguientes: amenazar, violar, pegar, golpear, matar y disparar.⁴ Con un total de 4578 ocurrencias, estos verbos representan el 55 % de las 8350 instancias de los verbos de la Tabla 4.3 y un 0.6 % del total de verbos.

En continuidad con lo presentado en el Capítulo 3, el proceso de anotación se realiza con un esquema de anotación basado en los marcos semánticos de FrameNet (C. F. Baker et al. 1998). En correspondencia con esos seis verbos, se seleccionan de FrameNet los siguientes marcos semánticos: «Asesinato» (*Killing*), «Robo» (*Robbery*), «Violación» (*Rape*), «Disparo» (*Shoot Projectiles*), «Impacto» (*Hit Target*), «Daño» (*Cause Harm*), «Arma de fuego» (*Use Firearm*). La descripción de estos marcos, tanto en español como en inglés, se encuentra disponible en el Anexo I.

Un caso aparte resulta el verbo «amenazar», cuyo marco en FrameNet es

⁴Se deja de lado el verbo «discutir», que ocurre frecuentemente en el corpus pero que normalmente no refiere en sí mismo a un hecho violento; salvo excepciones, a lo sumo, es un preludio de tal evento. Algo análogo sucede con el verbo «violar»: la gran mayoría de sus ocurrencias en el corpus no refiere a un hecho violento, sino al quebrantamiento de una norma; sin embargo, es innegable su carácter violento cuando efectivamente ocurre como tal.

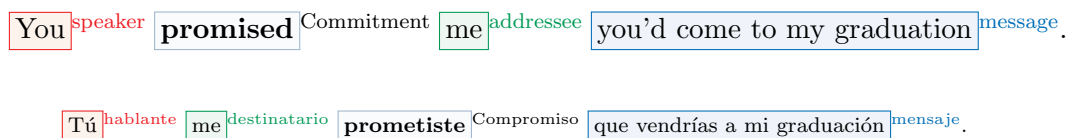


Figura 4.2: Ejemplo extraído de FrameNet para el marco «Compromiso» (*Commitment*), con unidad léxica «prometer» (*promise*). Se marcan los elementos «hablante» (*speaker*), «destinatario» (*addressee*) y «mensaje» (*message*), resaltados en rojo, verde y azul, respectivamente.

«Compromiso» (*Commitment*). Este marco incluye una amplia variedad de situaciones que no se limitan a las amenazas, sino también a juramentos, como los señalados por las unidades léxicas «jurar» (*swear*), «prometer» (*promise*), «consentir» (*consent*) o «comprometerse» (*undertake*), entre otras. En la Figura 4.2 se presenta un ejemplo extraído de FrameNet para la unidad léxica «prometer».

Como esas otras unidades léxicas no refieren a eventos violentos, ni siquiera criminales, se decide definir un nuevo marco específico denominado «Amenaza» (*Threaten*), que toma como base el marco «Compromiso» pero que se especializa en las situaciones de este tipo. A este nuevo marco, se le agrega, además, un nuevo argumento, «Instrumento» (*Instrument*), para relacionar, por ejemplo, las armas que en ocasiones son utilizadas para amedrentar, como en el Ejemplo 4.3. Se observa que, en este ejemplo particular, el mensaje de la amenaza se encuentra implícito.

- (4.3) El día 24/1/2013, próximo a la hora 21.50, los adolescentes ascendieron al ómnibus de Coetc línea 407 y amenazaron al guarda-conductor con arma blanca, circunstancia en la cual se apoderan de la recaudación.

Sentencia 11-000173/2013 Definitiva,
Tribunal Apelaciones Familia 2º Tº, Proceso Penal Especial

Las unidades léxicas consideradas para el marco «Amenaza» son el verbo «amenazar» (*threaten*) y el sustantivo «amenaza» (*threat*), reutilizando los ejemplos de «Compromiso». En el Apéndice C se presenta la definición completa de este nuevo marco.

Para la recolección de ejemplos, se premarcan en el corpus todas las ocurrencias de los verbos seleccionados. Luego, se procede a la revisión manual de las 717 anotaciones en 274 documentos. Durante este proceso de revisión

manual, se identifican y marcan nuevos eventos de violencia que no hayan sido premarcados inicialmente. Por ejemplo, se anotan aquellos eventos en los que el «Objetivo» corresponde a un grupo nominal en lugar de a un verbo. Este proceso se limita a oraciones, al igual que las unidades léxicas de FrameNet.

Asimismo, se etiquetan explícitamente como falsos positivos aquellas ocurrencias que, si bien contienen un verbo vinculado a un marco de violencia, no refieren a un evento violento. Por ejemplo, este sería el caso de los verbos «violar» o «golpear» en expresiones como «violar una ley» o «golpear la puerta», como es el caso en los ejemplos 4.4 y 4.5. Estos casos son incorporados como ejemplos negativos en el corpus, con el objetivo de que el modelo aprenda a distinguirlos de los casos positivos y no extraiga erróneamente eventos.

- (4.4) La Ley No. 18.831 en su artículo segundo viola^{violación} el principio constitucional de irretroactividad de la Ley penal más gravosa. . .

Sentencia 4/2014 Definitiva,
Suprema Corte de Justicia, Proceso de Inconstitucionalidad

- (4.5) Nos acostamos y como a las 22.00 horas nos golpean^{golpear} la puerta, eran las nenas. . .

Sentencia 15/2022,
Tribunal de Apelaciones Penal de 1º Tº

En resumen, en este proceso de análisis, para cada verbo de violencia premarcado automáticamente, se realizan manualmente los siguientes pasos:

1. Se determina si el verbo marcado es efectivamente usado en su acepción violenta y se asigna el atributo «violencia» según corresponda.
2. En caso de que el verbo sea violento, se marcan todas las entidades relacionadas con el evento: persona, instrumento, etc.
3. Se establecen las distintas relaciones entre las entidades marcadas y el evento: agente, víctima, instrumento, etc.
4. Se buscan otras unidades léxicas asociadas a marcos de violencia en la misma oración y para cada uno de ellos se repiten los pasos anteriores.

Durante este proceso se incorpora además la anotación del marco «Robo» (*Robbery*) ya que, aunque no es en todos los casos un evento violento en sí mismo, en muchas ocasiones coocurre en la BJN con otros eventos que sí los son, como es el caso en el Ejemplo 4.3.

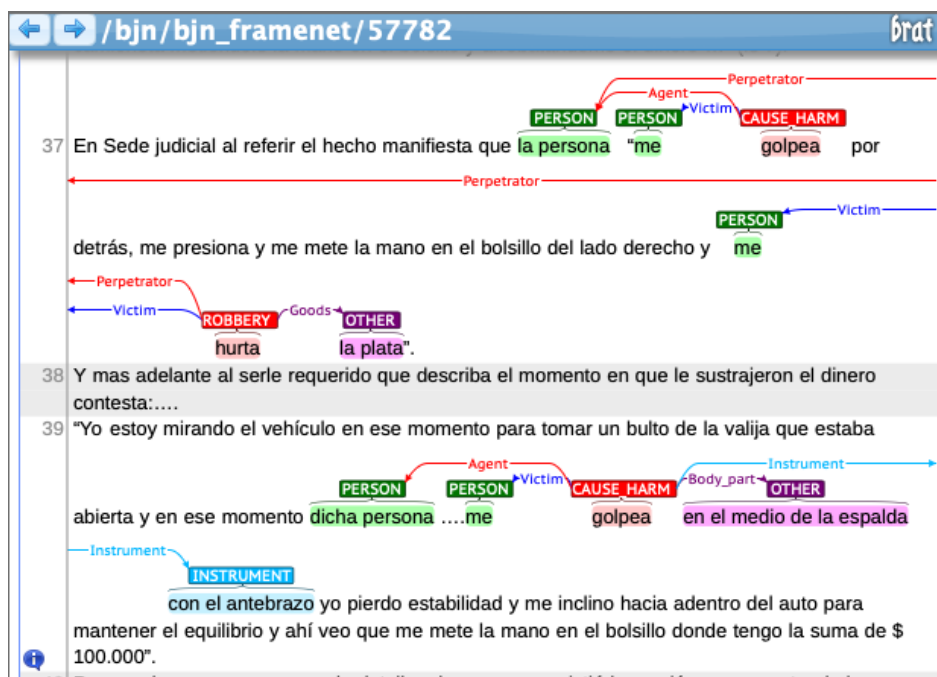


Figura 4.3: Marcado de eventos de violencia con BRAT. Los verbos están marcados automáticamente; los complementos y las relaciones son anotadas de forma manual.

Los ejemplos son revisados manualmente y se completa la anotación de los argumentos semánticos correspondientes con el apoyo de la herramienta BRAT (Stenetorp et al. 2012). En la Figura 4.3 se presenta una captura de pantalla de la herramienta BRAT con el marcado de tres eventos de violencia en una sentencia judicial.

Para la generación de los ejemplos, las sentencias se segmentan en fragmentos más breves, al igual que los ejemplos de las unidades léxicas de FrameNet⁵. Luego de la segmentación, se eliminan aquellos ejemplos repetidos, provenientes de formulaciones que son de estilo en sentencias judiciales, como son los ejemplos 4.4 y 4.6.

- (4.6) Conforme lo dispuesto por el artículo 258 de la Constitución de la República, los comparecientes se encuentran legitimados para promover la declaración de inconstitucionalidad por vía de excepción, por ser la Ley impugnada una norma que, como se verá, viola violación su interés directo, personal y legítimo.

Sentencia 595/2017 Definitiva,
Suprema Corte de Justicia, Recurso de Casación

⁵Para esto se utiliza un algoritmo *ad-hoc* de segmentación adaptado al dominio, dados los muy malos resultados obtenidos con los segmentadores, como se ejemplifica en la Tabla 2.2.

Marco	Violencia		No Violencia		Total	
Amenaza	(24.6 %)	259	(0.5 %)	5	(25.0 %)	264
Daño	(22.0 %)	232	(2.6 %)	27	(24.5 %)	259
Violación	(2.5 %)	26	(10.4 %)	110	(12.9 %)	136
Robo	(11.7 %)	123	(0.2 %)	2	(11.8 %)	125
Asesinato	(8.6 %)	91	(3.3 %)	35	(11.9 %)	126
Impacto	(4.6 %)	49	(0.1 %)	1	(4.7 %)	50
Arma de fuego	(7.1 %)	75	(0.3 %)	3	(7.4 %)	78
Disparo	(1.6 %)	17	(0.0 %)	0	(1.6 %)	17
Total	(82.6 %)	872	(17.4 %)	183	(100 %)	1055

Tabla 4.4: Distribución de los eventos anotados por marco semántico. Se presenta el conteo de casos correspondientes a violencia y a no violencia, así como el total por marco. Entre paréntesis se indica el porcentaje que representa cada valor sobre el total de eventos.

Aunque las sentencias, en su versión desidentificada, son de acceso público y anónimo en el portal de la BJN, se realiza un proceso adicional de desidentificación para garantizar la ausencia de cualquier dato personal en los ejemplos anotados. Esto es crucial considerando la sensibilidad de la información y, en particular, la naturaleza de los modelos a utilizar, al ser servicios provistos por terceros fuera del país. Este proceso busca eliminar cualquier referencia que pueda permitir la identificación de personas, empresas, fechas o ubicaciones específicas en los textos anotados.

Se procede, entonces, a la desidentificación manual de todos los ejemplos, sustituyendo los nombres propios y apellidos por nombres de pila frecuentes —Juan, Ana, José, etc.—, las empresas por nombres ficticios —Acme, Lumen, ...—, las fechas por otras inexistentes —29 de febrero, 31 de junio, ...—, y las calles por letras genéricas —A, B, C, ...—, manteniendo la coherencia interna del texto. En el Apéndice D se presenta en detalle el proceso de desidentificación manual.

El resultado de este proceso es un conjunto de 717 ejemplos extraídos de 274 sentencias. Se compone de un total de 1055 posibles eventos —cada ejemplo puede contener cero, uno o más eventos—, de los cuales un 83 % son efectivamente eventos de violencia. La distribución de los ejemplos anotados, discriminando por marco y por polaridad (positiva o negativa), se presenta en la Tabla 4.4. La Figura 4.4 muestra esta distribución por marco semántico.

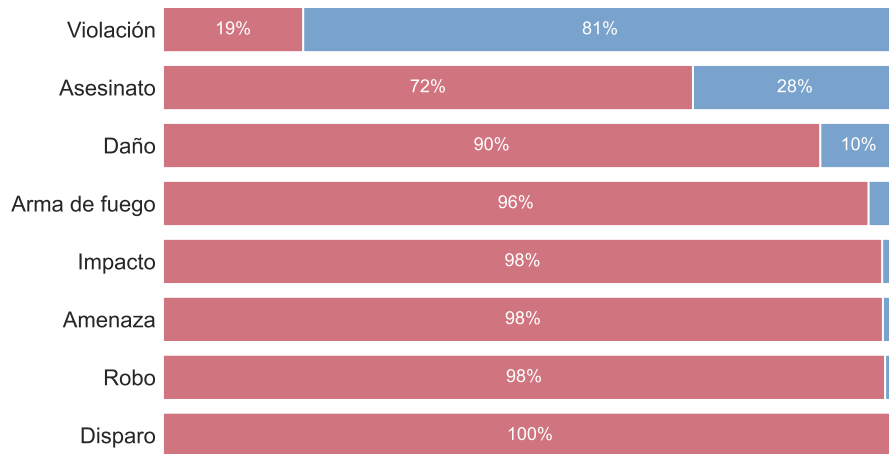


Figura 4.4: Porcentaje de eventos de violencia y no violencia por marco. Cada barra representa un marco y está normalizada al 100 %, mostrando la proporción de casos violentos y no violentos, en rojo y azul respectivamente.

4.4. ICL-FSP en la BJN

Como siguiente etapa, se evalúa el desempeño de los modelos en tareas de análisis semántico sobre textos de la BJN. Estos textos, a diferencia de los utilizados en la investigación del Capítulo 3, se encuentran en español y en jerga judicial, en lugar de basarse en artículos generalistas y en inglés.

En particular, se formulan dos preguntas de investigación: (a) ¿son los ejemplos de FrameNet suficientes para obtener resultados comparables a los obtenidos con textos en inglés cuando se aplican sobre textos en español?, y (b) ¿en qué medida la incorporación de ejemplos en español mejora la capacidad del modelo para identificar y etiquetar correctamente los argumentos de los marcos semánticos? Esto permite cuantificar el impacto de la transferencia interlingüística en tareas de análisis semántico y, a su vez, explorar el valor agregado que aporta el uso de datos anotados en el dominio y lengua objetivo.

La metodología seguida es similar a la presentada en el Capítulo 3, adaptada al nuevo escenario. En este caso, la tarea consiste en la identificación completa de los eventos de violencia, esto es, la identificación del marco y el etiquetado de sus argumentos semánticos asociados, en textos provenientes de la BJN. Esto se corresponde con la tarea de análisis de marcos semánticos (FSP), presentada en el Capítulo 3.

A diferencia del trabajo previo sobre FrameNet, en este caso se combinan ejemplos en inglés y en español: mientras los ejemplos en inglés son tomados de FrameNet, los ejemplos en español pertenecen al corpus ViolenciaR presentado en la Sección 4.3. El desempeño de los modelos se evalúa en un conjunto de evaluación exclusivamente en español, con ejemplos de ViolenciaR.

Para esta experimentación, se seleccionan los modelos Claude Sonnet 3.7, DeepSeek Reasoner y GPT-4o, dejando de lado los modelos de menor capacidad dados los resultados en la experimentación previa.

A continuación, se describen los detalles de la experimentación: en la Sección 4.4.1 se explica la generación de los *prompts* utilizados, y en la Sección 4.4.2 se detallan los aspectos relacionados con la experimentación. Finalmente, en la Sección 4.4.3 se presentan y analizan los resultados obtenidos.

4.4.1. *Prompt* para ICL en la BJN

El *prompt* utilizado en esta experimentación es similar al presentado en la Sección 3.3.1 para la tarea de análisis de marcos semánticos, con las adaptaciones necesarias para el nuevo escenario bilingüe.

De las cuatro secciones planteadas en la Sección 3.3.1 —Objetivo, Eventos, Directrices y Ejemplos—, la sección «Eventos» se mantiene igual, dado que las definiciones de los marcos semánticos siguen proveniente de FrameNet en inglés.

En cuanto a las secciones «Objetivo» y «Directrices», se adaptan al nuevo escenario, con cambios mínimos, enfatizando el hecho de que los textos a analizar están en español y que los ejemplos pueden estar en ambos idiomas. Por ejemplo, en «Objetivo» se agrega «*You will get an input text written in Spanish and your output should be a list of all frames found in the text as JSON records.*»⁶, y en directrices se incluye «*You will receive an input text written in Spanish.*»⁷.

La otra sección del *prompt* que se mantiene casi sin cambios es la de «Ejemplos». Aunque el resultado de aplicar la plantilla en su estructura no difiere del caso anterior —salvo por el idioma de los ejemplos provenientes de ViolenciaR— la selección de los ejemplos es distinta. En este caso se incluyen ejemplos tanto en inglés como en español, para evaluar el efecto de la

⁶Vas a recibir la entrada escrita en español y tu salida debe ser una lista de todos los marcos encontrados en el texto como registros JSON.

⁷Vas a recibir un texto de entrada escrito en español

transferencia interlingüística en la resolución de extracción de eventos en la BJN.

Así, en los experimentos *n-shot*, se incluyen ejemplos en inglés tomados de FrameNet y ejemplos en español extraídos del corpus ViolenciaR.⁸ Estos ejemplos son intercalados en el *prompt*, manteniendo cierta proporción entre ambos idiomas, que puede variar según la configuración del experimento. Así, en una configuración de cinco ejemplos con proporción 80–20, se obtienen bloques de cuatro ejemplos de FrameNet seguidos de uno de ViolenciaR, repitiéndose este patrón hasta completar el número total de *shots*.

Por otra parte, a diferencia de los experimentos anteriores, no se agregan al final ni un ejemplo múltiple —combinando los últimos dos ejemplos— ni el ejemplo sin extracciones, dada la naturaleza de los textos de ViolenciaR, que pueden contener cero, uno o más eventos.

En el Apéndice E se presenta la plantilla del *prompt* utilizado en esta experimentación junto con ejemplos concretos para configuraciones *zero* y *n-shot*.

4.4.2. Experimentación

Para la experimentación se decide utilizar los modelos de mayor capacidad disponibles en ese momento: Claude Sonnet 3.7⁹, DeepSeek Reasoner y OpenAI GPT-4o. Se dejan de lado los modelos de menor capacidad dados los resultados en la experimentación previa, en donde los modelos de mayor capacidad sistemáticamente superan a sus contrapartes de menor capacidad en la tarea FSP.

Los marcos considerados son los ocho anotados en el corpus ViolenciaR que, se recuerda, pueden aparecer mezclados en un mismo texto. Para el proceso de extracción, los marcos no son considerados de forma individual, sino que el modelo debe extraerlos conjuntamente del texto del ejemplo, indicando todos los eventos presentes y sus argumentos.

En cuanto al corpus ViolenciaR, las 717 instancias se distribuyen en 517 para *shots* y las restantes 200 para la evaluación, de manera estratificada, según los eventos, por marco y polaridad. En la Tabla 4.5 se presenta la distribución

⁸En los experimentos *zero-shot* no se incluyen ejemplos en ninguna de las dos lenguas.

⁹No se encuentra disponible información detallada sobre los cambios en esta versión respecto a Sonnet 3.5, utilizada en los experimentos previos. Se decide el cambio de versión según los reportes generales de mejora de desempeño.

Marco	Entrenamiento			Evaluación		
	Positivos	Negativos	Total	Positivos	Negativos	Total
Amenaza	(24.1 %) 189	(0.3 %) 2	(24.4 %) 191	(25.7 %) 70	(1.1 %) 3	(26.8 %) 73
Daño	(20.6 %) 161	(2.7 %) 21	(23.2 %) 182	(26.1 %) 71	(2.2 %) 6	(28.3 %) 77
Robo	(13.3 %) 104	(0.3 %) 2	(13.6 %) 106	(7.0 %) 19	(0.0 %) 0	(7.0 %) 19
Asesinato	(9.2 %) 72	(3.4 %) 27	(12.6 %) 99	(10.3 %) 19	(2.9 %) 8	(13.2 %) 27
Violación	(2.0 %) 16	(10.1 %) 79	(12.1 %) 95	(3.7 %) 10	(11.4 %) 31	(15.1 %) 41
Arma de fuego	(7.5 %) 59	(0.4 %) 3	(7.9 %) 62	(5.9 %) 16	(0.0 %) 0	(5.9 %) 16
Impacto	(4.3 %) 34	(0.0 %) 0	(4.3 %) 34	(5.5 %) 15	(0.2 %) 1	(5.7 %) 16
Disparo	(1.8 %) 14	(0.0 %) 0	(1.8 %) 14	(1.1 %) 3	(0.0 %) 0	(1.1 %) 3
Total	(82.9 %) 649	(17.1 %) 134	(100 %) 783	(82.0 %) 223	(18.0 %) 49	(100 %) 272

Tabla 4.5: Distribución de los eventos positivos y negativos por marco semántico en los 517 y 200 textos de entrenamiento y evaluación, respectivamente. Entre paréntesis, el porcentaje en el total del conjunto respectivo.

de los eventos en los conjuntos de entrenamiento y evaluación.

A diferencia de la experimentación presentada en el Capítulo 3, en donde el estudio mide el desempeño de los distintos modelos según la cantidad de *shots* provistas en el *prompt*, para este caso se fija esa cantidad pero se varía la proporción de su procedencia.

En particular, el escenario planteado considera 500 *shots* en total, organizados en 100 bloques de 5 ejemplos, dentro de los cuales se varía progresivamente la proporción de ejemplos provenientes de ViolenciaR y de FrameNet. La proporción entre una y otra fuente en cada bloque se ajusta desde un 0 % BJN y 100 % FrameNet, hasta 100 % BJN y 0 % FrameNet, en incrementos de 20 %.

La Tabla 4.6 muestra la cantidad de argumentos y objetivos para el conjunto de evaluación y para los conjuntos de 500 *shots* en los extremos: cuando todos los ejemplos provienen únicamente de FrameNet en inglés o cuando todos provienen del corpus ViolenciaR en español.

Además, para cada uno de los tres modelos evaluados, se plantea un escenario *zero-shot*, a fin de comparar su desempeño en ausencia de ejemplos, esto es, la extracción de eventos en la BJN utilizando únicamente las definiciones de los marcos semánticos provenientes de FrameNet.

En la Tabla 4.7 se resumen las configuraciones utilizadas en la experimentación.

Al igual que en los experimentos del Capítulo 3, el rendimiento se mide mediante las métricas estándar de precisión (P), cobertura (*recall*, R) y medida

Marco	Entrenamiento				Evaluación	
		100 % FN		100 % BJN		
Arma	(0.7 %)	10	(0.4 %)	8	(0.7 %)	5
Asesino	(4.2 %)	60	(1.3 %)	24	(1.6 %)	12
Bienes	(0.3 %)	4	(3.6 %)	65	(4.4 %)	32
Blanco	(0.1 %)	2	(1.8 %)	33	(1.8 %)	13
Causa	(3.0 %)	42	(0.2 %)	4	(0.4 %)	3
Destinatario	(0.3 %)	4	(8.6 %)	157	(9.2 %)	67
Fuente	(0.7 %)	10	(0.7 %)	12	(0.4 %)	3
Hablante	(0.7 %)	10	(5.5 %)	100	(6.3 %)	46
Instrumento	(4.4 %)	63	(9.8 %)	179	(9.6 %)	70
Medios	(0.3 %)	4	(0.1 %)	1	(0.0 %)	0
Mensaje	(0.8 %)	11	(2.4 %)	44	(2.5 %)	18
Meta	(0.8 %)	11	(1.4 %)	26	(1.6 %)	12
Objetivo	(35.2 %)	500	(33.6 %)	611	(33.8 %)	247
Parte del cuerpo	(6.1 %)	87	(2.4 %)	44	(1.4 %)	10
Perpetrador	(2.2 %)	31	(3.7 %)	68	(3.4 %)	25
Proyectil	(1.8 %)	25	(0.6 %)	11	(0.8 %)	6
Subregión	(0.0 %)	0	(0.5 %)	9	(0.5 %)	4
Víctima	(24.4 %)	347	(14.4 %)	262	(14.9 %)	109
Total	(100 %)	1422	(100 %)	1820	(100 %)	731

Tabla 4.6: Número de argumentos en los conjuntos de 500 ejemplos para las distribuciones 100 % FrameNet y 100 % BJN, así como en el conjunto de evaluación. Entre paréntesis se indica el porcentaje que representa cada valor sobre el total de argumentos de cada conjunto.

Exp.	Shots	Prop. Ejemplos	Modelos	Marcos
1	500	0, 0.2, 0.4, 0.6, 0.8 y 1; bloques de 5 ejemplos	Sonnet 3.7, DeepSeek R., GPT-4o	Amenaza, Arma de fuego, Asesinato, Daño, Disparo, Impacto, Robo, Violación
2	0	—		

Tabla 4.7: Configuraciones utilizadas para la experimentación de ICL-FSP en la BJN. La evaluación se realiza sobre un conjunto de 200 ejemplos provenientes exclusivamente de ViolenciaR.

Modelo	P	R	F_1
Claude Sonnet 3.7	0.650	0.704	0.676
DeepSeek Reasoner	0.546	0.554	0.551
OpenAI GPT-4o	0.643	0.508	0.561

Tabla 4.8: Mejores resultados de F_1 , junto con sus valores de precisión y cobertura, para los modelos evaluados. En todos los modelos, el mejor F_1 se obtiene utilizando 500 ejemplos provenientes exclusivamente de la BJN, sin incorporar ejemplos de FrameNet. Los valores más altos se destacan en negrita y los segundos mejores se encuentran subrayados.

F_1 , en su versión micro. Además, la evaluación de los argumentos se realiza, nuevamente, de un modo estricto.¹⁰

4.4.3. Resultados

En la Tabla 4.8 se presentan los valores de F_1 , P y R con los 500 ejemplos provenientes exclusivamente de la BJN, que es la distribución en la que se obtiene el mejor desempeño, independientemente del modelo considerado. Claude Sonnet resulta ser el modelo con mejor desempeño, con un F_1 de 0.676, más de 0.11 puntos por encima de GPT-4o, con 0.561, y DeepSeek Reasoner, con 0.551.

Los resultados obtenidos al variar la distribución de ejemplos se presenta en las figuras 4.5 y 4.6; en la primera se presentan los valores de precisión (P) y cobertura (R), mientras que en la segunda se observa el desempeño en términos de F_1 para Claude Sonnet 3.7, GPT-4o y DeepSeek Reasoner.

En la Figura 4.6 se observa que, en todos los modelos, el mejor desempeño según la medida F_1 se obtiene utilizando exclusivamente ejemplos provenientes de la BJN, y el peor, cuando los ejemplos provienen únicamente de FrameNet. La ganancia, sin embargo, varía según el modelo: mientras Claude Sonnet 3.7 experimenta una mejora significativa de 0.261 puntos al incorporar ejemplos de la BJN, los otros dos modelos muestran incrementos más modestos de 0.156 y 0.076 para GPT-4o y DeepSeek Reasoner, respectivamente.

Por otra parte, en la Figura 4.5 se observa que, en todos los modelos la ausencia de ejemplos de la BJN conduce a una baja precisión, sin embargo,

¹⁰Se refiere al lector a las secciones 2.4.2 y 3.3.2 para una descripción detallada de las métricas utilizadas y los criterios de evaluación micro y estricto.

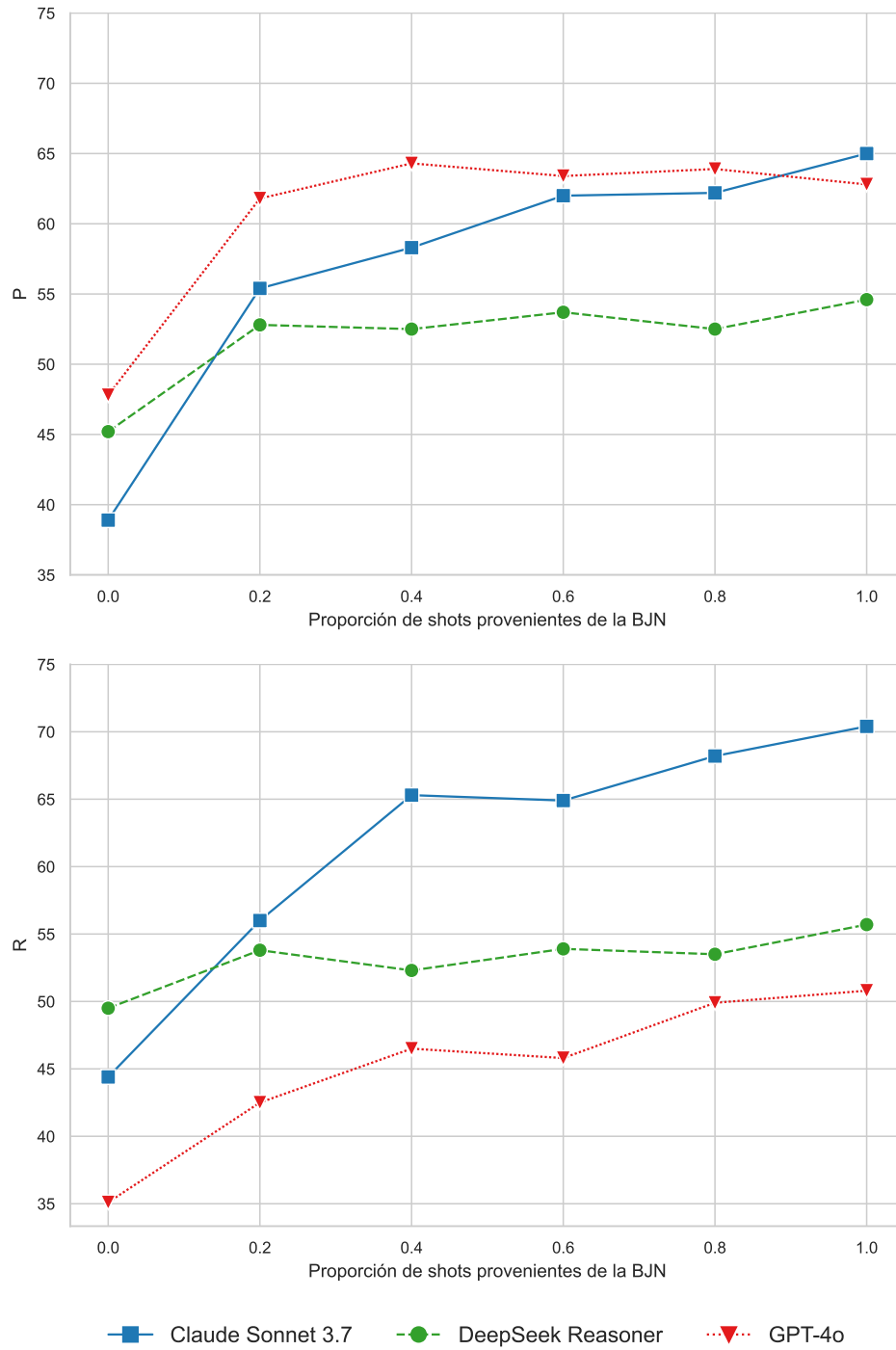


Figura 4.5: P (arriba) y R (abajo) en función de la proporción relativa de ejemplos provenientes de la BJT y FrameNet en el conjunto de entrenamiento para Claude Sonnet 3.7, GPT-4o y DeepSeek Reasoner (0: 100 % FrameNet; 1: 100 % BJT).

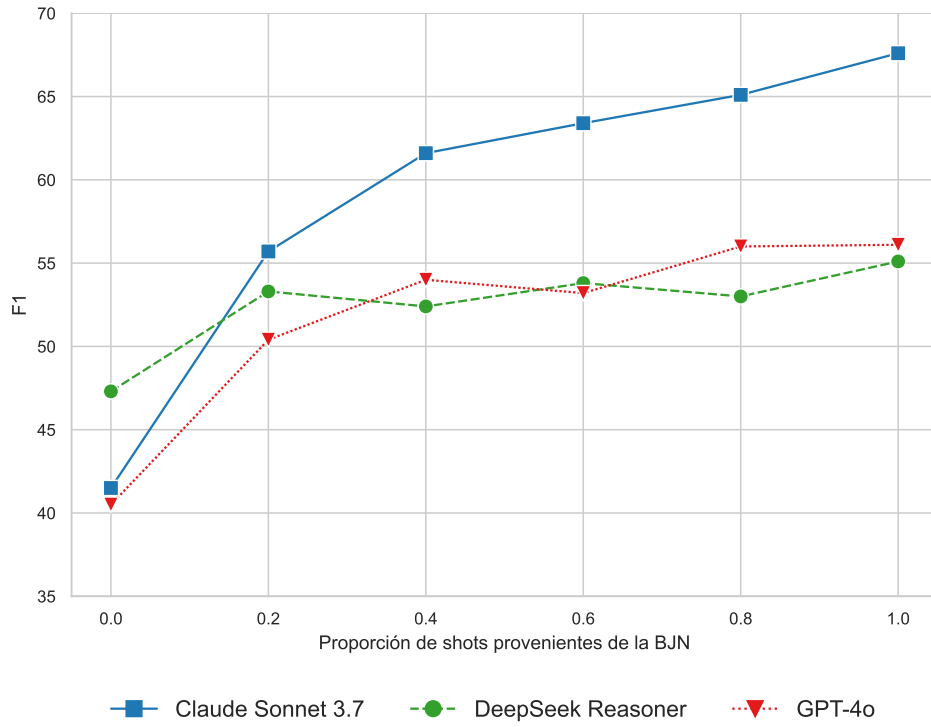


Figura 4.6: F_1 en función de la proporción relativa de ejemplos provenientes de la BJT y FrameNet en el conjunto de entrenamiento para Claude Sonnet 3.7, GPT-4o y DeepSeek Reasoner (0: 100 % FrameNet; 1: 100 % BJT).

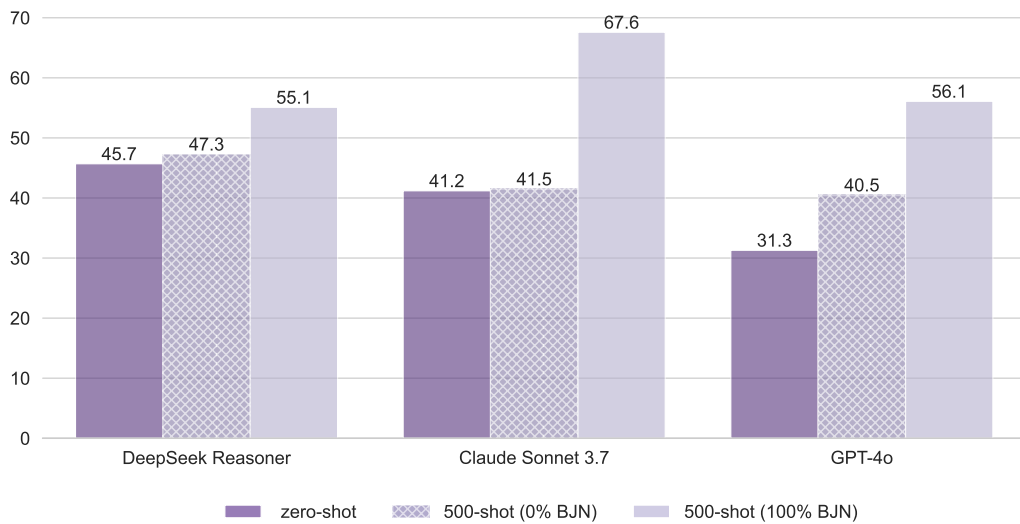


Figura 4.7: F_1 para *zero-shot* y *500-shot* con 0 % y 100 % de ejemplos provenientes de la BJT para cada uno de los tres modelos utilizados.

para DeepSeek y GPT-4o, aumentar su proporción no siempre resulta en una mejora significativa de este valor. En cuanto a la cobertura, es llamativo el comportamiento de DeepSeek, cuyos valores no mejoran sustancialmente al aumentar la proporción de ejemplos de la BJN, en contraste con los otros dos modelos. El caso contrario es Claude Sonnet 3.7, donde la cobertura aumenta de manera constante al incrementar la proporción de ejemplos de la BJN, lo que le permite alcanzar el mejor desempeño en F_1 de los tres modelos.

En comparación con los resultados obtenidos en la experimentación previa sobre FrameNet, disponibles en la Sección 3.3.3, se observa una caída significativa en el desempeño de los modelos al aplicarlos sobre textos en español de la BJN. En particular, el mejor desempeño obtenido en este escenario, 0.676 de F_1 para Claude Sonnet 3.7, se encuentra 0.149 puntos por debajo del 0.825 reportado para el Experimento 3b con textos en inglés de FrameNet. Esta diferencia de 0.149 puntos evidencia la dificultad de transferir el desempeño de modelos entre idiomas y dominios distintos.

Mientras que el primer experimento considera *500-shots* y varía su proporción según la fuente de datos, el segundo experimento evalúa el desempeño de los modelos en un enfoque *zero-shot*, es decir, sin utilizar ejemplos en el *prompt*. En este escenario, los modelos alcanzan valores de F_1 de 0.313, 0.412 y 0.457 para GPT-4o, Claude Sonnet 3.7 y DeepSeek Reasoner, respectivamente.

Bajo este enfoque, todos los modelos obtienen un desempeño inferior al observado con *500-shot*, independientemente de la distribución de los ejemplos según su origen. En la Figura 4.7 se comparan los resultados *zero-shot*, con los de *500-shot* en sus configuraciones de mejor y peor desempeño. Para DeepSeek y Sonnet, la diferencia entre el escenario *zero-shot* y el de *500-shot* con ejemplos exclusivamente de FrameNet es inferior a 0.015 puntos, mientras que para GPT-4o la diferencia es de 0.092 puntos.

A partir de los resultados de ambos experimentos, se observa que la mejora significativa se produce al incorporar ejemplos de la BJN, lo que resalta la importancia de contar con ejemplos en el dominio y lengua objetivo para mejorar el desempeño en tareas de análisis semántico.

En la Tabla 4.9 se presentan los mejores resultados de F_1 obtenidos por cada modelo para cada uno de los argumentos semánticos considerados, junto con la proporción de ejemplos utilizada para alcanzarlos; mientras que en la Tabla 4.10 se presentan los valores de F_1 obtenidos para cada argumento cuando se utilizan exclusivamente ejemplos provenientes de la BJN.

Argumento	# Ej.		Claude Sonnet 3.7		DeepSeek Reasoner		OpenAI GPT-4o
Parte del cuerpo	10	(1.0)	<u>0.800</u>	(0.2)	0.857	(0.4)	0.737
Blanco	13	(1.0)	0.769	(0.8)	0.381	(0.6)	<u>0.500</u>
Objetivo	247	(1.0)	0.766	(0.6)	<u>0.685</u>	(1.0)	0.680
Destinatario	67	(1.0)	0.754	(0.2)	<u>0.746</u>	(0.8)	0.652
Fuente	3	(0.8)	0.750	(0.6)	0.222	(0.4)	<u>0.667</u>
Meta	12	(0.8)	0.737	(1.0)	0.429	(0.6)	<u>0.600</u>
Víctima	109	(0.8)	0.700	(0.8)	0.519	(1.0)	<u>0.567</u>
Bienes	32	(1.0)	0.687	(1.0)	0.455	(0.6, 0.8)	<u>0.625</u>
Instrumento	70	(1.0)	0.662	(0.4)	0.551	(0.4)	<u>0.571</u>
Hablante	46	(1.0)	0.578	(0.4, 1.0)	<u>0.559</u>	(0.8)	0.420
Arma	5	(0.4)	0.222	(0.6)	<u>0.462</u>	(0.2, 0.4)	0.571
Perpetrador	25	(1.0)	0.519	(0.2)	0.400	(1.0)	<u>0.468</u>
Causa	3	(≥ 0.4)	0.500		0.000	(≥ 0.4)	0.500
Asesino	12	(0.6, 0.8)	<u>0.471</u>	(0.8)	0.435	(0.0)	0.500
Subregión	4	(0.6, 1.0)	<u>0.400</u>	(0.6)	<u>0.400</u>	(0.8)	0.500
Mensaje	18	(1.0)	0.486	(0.8, 1.0)	0.381	(0.8)	<u>0.471</u>
Agente	49	(0.8)	0.446	(0.2)	<u>0.438</u>	(0.2)	0.346
Proyectil	6	(0.4, 1.0)	0.400		0.000		0.000
Medios	0		0.000		0.000		0.000

Tabla 4.9: Mejor F_1 por argumento y modelo. El resultado para cada argumento se muestra en negrita y el segundo mejor se subraya. Entre paréntesis, se indican las proporciones de *shots* de la BJN con las que se produce ese resultado.

De forma gráfica, la Figura 4.8 presenta una comparación visual de los resultados de F_1 obtenidos por cada modelo para los ocho argumentos semánticos con mejor resultado bajo dos condiciones: su mejor valor independientemente de la distribución de los *shots* —que se corresponde al de la Tabla 4.9— y la distribución con 100 % de *shots* provenientes de la BJN —que se corresponde al de la Tabla 4.10—.

Se observa que, salvo en algunos casos puntuales, los mejores resultados se obtienen con el modelo Sonnet 3.7, con proporciones elevadas de ejemplos provenientes de la BJN. Una excepción notable es el argumento «Parte del cuerpo», en donde DeepSeek con un F_1 de 0.857 no solo alcanza su mejor desempeño sino el mejor F_1 global, con un quinto de *shots* provenientes de la BJN.

Argumento	# Ej.	Claude Sonnet 3.7	DeepSeek Reasoner	OpenAI GPT-4o
Parte del cuerpo	10	0.800	0.800	<u>0.667</u>
Blanco	13	0.769	0.273	<u>0.400</u>
Objetivo	247	0.766	0.671	<u>0.680</u>
Destinatario	67	0.754	<u>0.741</u>	0.627
Víctima	109	0.692	0.507	<u>0.567</u>
Meta	12	0.696	0.429	<u>0.500</u>
Bienes	32	0.687	0.455	<u>0.623</u>
Instrumento	70	0.662	<u>0.530</u>	0.496
Hablante	46	0.578	<u>0.559</u>	0.390
Perpetrador	25	0.519	0.373	<u>0.468</u>
Causa	3	0.500	0.000	0.500
Asesino	12	0.375	0.333	0.375
Mensaje	18	0.486	0.381	<u>0.457</u>
Arma	5	0.000	0.400	0.400
Subregión	4	0.400	0.200	<u>0.250</u>
Proyectil	6	0.400	0.000	0.000
Agente	49	0.444	<u>0.385</u>	0.333
Fuente	3	<u>0.333</u>	0.000	0.500
Medios	0	0.000	0.000	0.000

Tabla 4.10: Valores de F_1 por argumento y modelo para la configuración con 100 % de ejemplos provenientes de la BJN. El mejor resultado por argumento se indica en negrita y el segundo mejor se subraya.

Tanto DeepSeek como GPT-4o logran sus mejores resultados en varios argumentos con proporciones moderadas de ejemplos de la BJN, pero estos resultados no alcanzan el 70 % de F_1 en la mayoría de los casos —las excepciones son los argumentos «Destinatario» para DeepSeek, y «Parte del cuerpo» para Deepseek y GPT-4o—, lo que contrasta con los resultados de Sonnet 3.7, que logra superar ese umbral en siete de los argumentos evaluados.

En particular, para «Objetivo», que es el que indica la presencia del marco semántico, Sonnet 3.7 alcanza un F_1 de 0.766 con ejemplos exclusivamente de la BJN, mientras que DeepSeek y GPT-4o logran valores de 0.685 y 0.680, respectivamente. Estos valores se encuentran 0.200 puntos por debajo de, por ejemplo, los obtenidos en el Capítulo 3 para el Experimento 1, en donde Sonnet 3.5 alcanza un F_1 de 0.966 y GPT-4o de 0.90, los del Experimento 2, con 0.935,

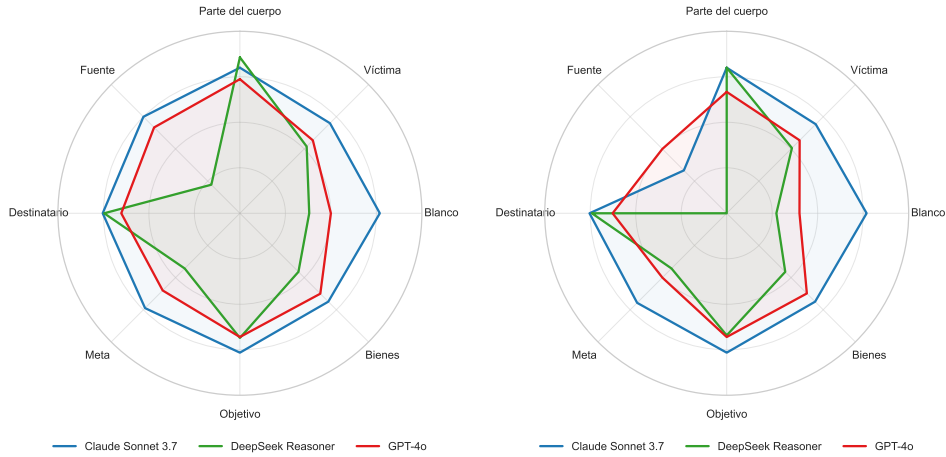


Figura 4.8: Valores de F_1 para los ocho atributos con mejor desempeño: a la izquierda, mejor valor absoluto, independientemente de la proporción de ejemplos; a la derecha, valores correspondientes a la distribución con 100 % de ejemplos provenientes de la BJD.

0.931 y 0.929, para DeepSeek Reasoner, Sonnet 3.5 y GPT-4o, respectivamente. Aunque esos escenarios difieren en la cantidad de marcos o tarea a realizar, nuevamente la diferencia en el desempeño resalta las dificultades adicionales que surgen al trabajar con textos en un idioma diferente y en un dominio específico como el judicial.

Las figuras 4.9 y 4.10 muestran los valores de F_1 obtenidos para «Objetivo» y para tres de los argumentos con mejor desempeño —«Parte del cuerpo», «Destinatario» y «Víctima»—, según la proporción de ejemplos de cada origen. A diferencia del F_1 global presentado en la Figura 4.6, que presenta variaciones moderadas al aumentar los ejemplos de la BJD, estos argumentos muestran oscilaciones más pronunciadas.

Estos resultados sugieren que el desempeño en argumentos específicos puede verse afectado por la distribución de los ejemplos incluidos en el *prompt*. Parte de esta inestabilidad podría explicarse por el reducido tamaño del conjunto de evaluación, que amplifica las oscilaciones observadas, en particular para «Parte del cuerpo», que cuenta con solo 10 ejemplos en el conjunto de evaluación, pero en menor medida para «Objetivo», «Destinatario» y «Víctima», que cuentan con 247, 67 y 109 ejemplos, respectivamente.

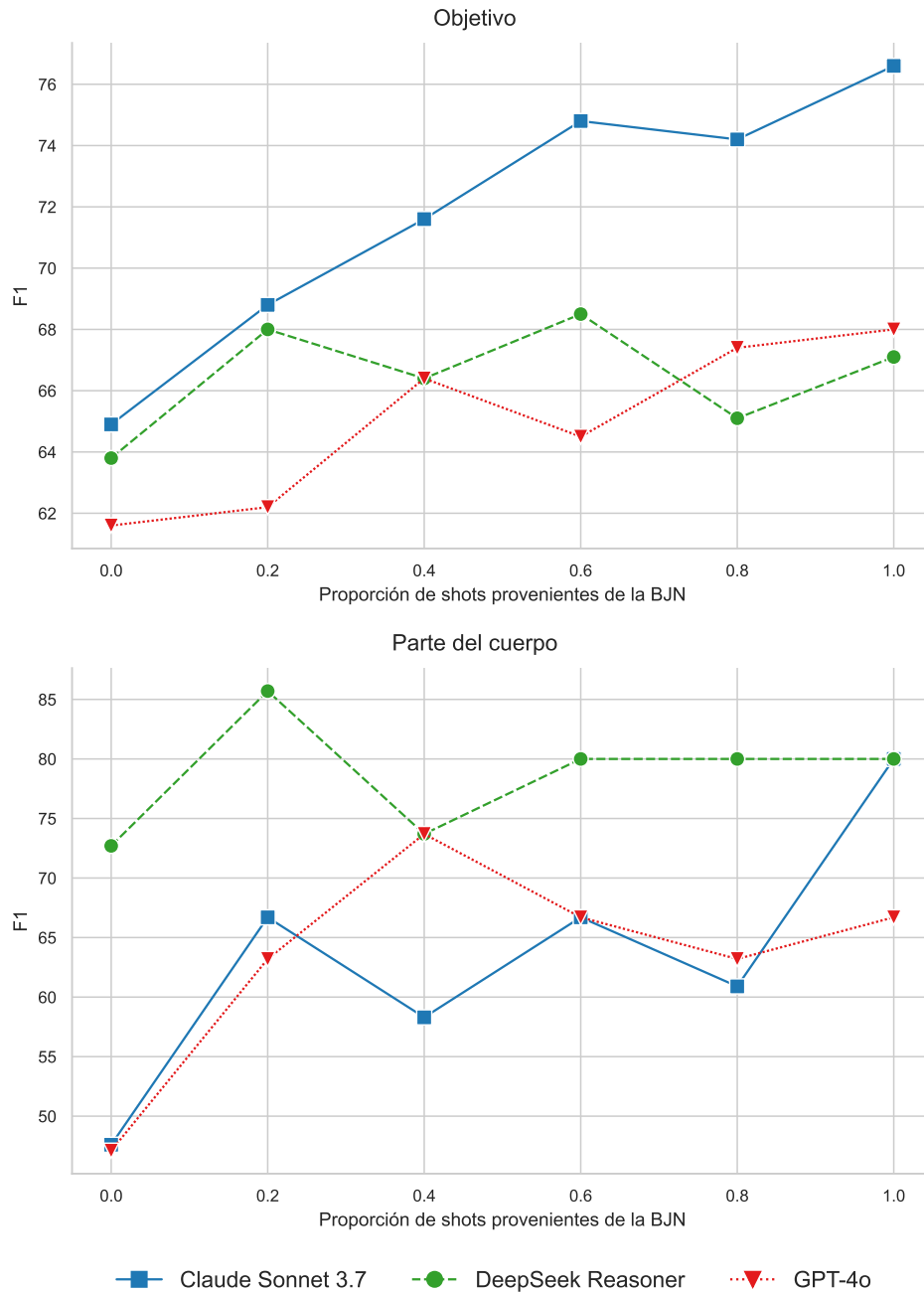


Figura 4.9: F_1 para los argumentos «Objetivo» y «Parte del cuerpo» al variar la proporción de *shots* de la BJT y FrameNet. En el eje de las abscisas se indica la proporción de ejemplos de la BJT, donde 0 corresponde a 100 % FrameNet y 1 a 100 % BJT.

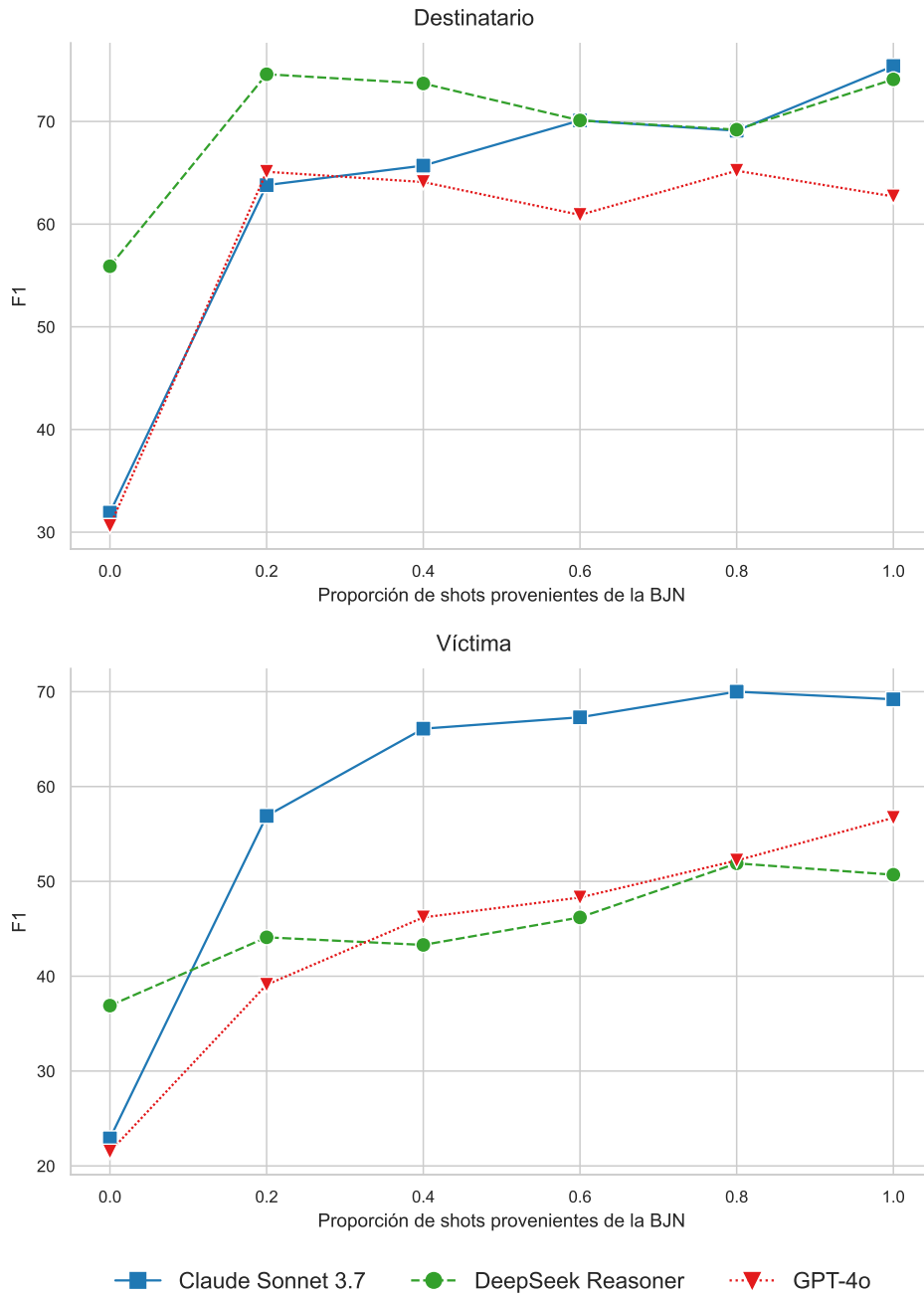


Figura 4.10: F_1 para los argumentos «Destinatario» y «Vítima» al variar la proporción de *shots* de la BJA y FrameNet. En el eje de las abscisas se indica la proporción de ejemplos de la BJA, donde 0 corresponde a 100 % FrameNet y 1 a 100 % BJA.

4.4.4. Ejemplos de extracciones

En la presente sección se muestran algunos ejemplos de extracciones realizadas por el modelo con mejores resultados, Claude Sonnet 3.7, utilizando 500 ejemplos provenientes exclusivamente de la BJN. En casos de múltiples eventos en un mismo texto, se agrega un superíndice que indica a cuál evento se encuentra asociado cada segmento marcado.

Los ejemplos 4.7 y 4.8 ilustran extracciones realizadas por el modelo en las que se identifican adecuadamente tanto los eventos como los argumentos semánticos asociados. Cada ejemplo tiene dos eventos y, en particular, en el segundo ejemplo se observa la capacidad del modelo para distinguir «El otro partícipe» en su doble rol de hablante en el primer evento y de perpetrador en el segundo.

(4.7) **Amenazó**⁽¹⁾ a la comerciante^{dest.(1)} con un arma de fuego^{instr.(1)}, y la^{víctima(2)} **despojó**⁽²⁾ de la suma de \$300^{bienes(2)}.-

(4.8) El otro partícipe^{hablante(1)/perpetr.(2)} se dirigió a la caja y **amenazó**⁽¹⁾ a la empleada^{dest.(1)}, **sustrayendo**⁽²⁾ la suma de \$ 3.000 (tres mil pesos)^{bienes(2)}.-

Los casos a continuación se centran en algunas diferencias con las anotaciones o errores observados en la evaluación. Aunque no representan una lista exhaustiva, ilustran algunos de los límites del enfoque. En cada ejemplo se presentan dos textos anotados: el primero corresponde a la anotación manual, mientras que el segundo muestra la salida según el modelo.

En el Ejemplo 4.9, el modelo encuentra los dos eventos anotados manualmente —marcados por «amenazó» y «apoderó»— pero agrega un tercer evento que no está presente en la anotación. En este caso, el modelo interpreta «llevándose» como un indicador de robo, lo cual podría ser válido.

- (4.9) a. Pedro cerró la puerta y le pidió que se identificar y **el procesado**^{hablante(1)} con la intención de huir y de asegurarse su impunidad, **lo**^{dest.(1)} **amenazó**⁽¹⁾ **con un cuchillo**^{instr.(1)} y huyó por la banderola, llevándose los referidos efectos, los cuales no fueron recuperados, 3) El día 30 de febrero, en horas de la noche, **el encausado**^{perpetr.(2)} ingresó a **la finca sita en la calle A N° 9876, domicilio de Juan**^{origen(2)}, donde se **apoderó**⁽²⁾ **de diversos efectos (plancha para pelo, DVD, anillo de oro y plata, etc)**^{bienes(2)} lo que fue avaluado por el damnificado en la suma de \$24.000 pesos.
- b. Pedro cerró la puerta y le pidió que se identificar y **el procesado**^{hablante(1)/perpetrador(2)} con la intención de huir y de asegurarse su impunidad, **lo**^{dest.(1)} **amenazó**⁽¹⁾ **con un cuchillo**^{instr.(1)} y huyó por la banderola, **llevándose**⁽²⁾ **los referidos efectos**^{bienes(2)}, los cuales no fueron recuperados, 3) El día 30 de febrero, en horas de la noche, **el encausado**^{perpetr.(3)} ingresó a **la finca sita en la calle A N° 9876, domicilio de Juan**^{origen(3)}, donde se **apoderó**⁽³⁾ **de diversos efectos (plancha para pelo, DVD, anillo de oro y plata, etc)**^{bienes(3)} lo que fue avaluado por el damnificado en la suma de \$24.000 pesos.

En Ejemplo 4.10 el modelo detecta correctamente el evento de disparo, pero confunde al agente con el destinatario del disparo; este error es, en cierta forma, inesperado, dado que el pronombre de tercer persona «le» representa al destinatario de la acción. De forma similar, en el Ejemplo 4.11, el modelo confunde al asesino con la víctima.

- (4.10) a. Alcanza con analizar la respuesta para detectar de inmediato que no existe motivo alguno vinculado a la situación que se desarrolló en el día de los hechos, ya que para empezar ni siquiera sabe por qué **le**^{dest.} **disparó** y el resto de su respuesta hace referencia al nacimiento de su hija que en ese momento ya contaba con once años de edad por lo cual mal podría tener relación una cosa con otra.-
- b. Alcanza con analizar la respuesta para detectar de inmediato que no existe motivo alguno vinculado a la situación que se desarrolló en el día de los hechos, ya que para empezar ni siquiera sabe por qué **le**^{agente} **disparó** y el resto de su respuesta hace referencia al nacimiento de su hija que en ese momento ya contaba con once años de edad por lo cual mal podría tener relación una cosa con otra.-

- (4.11) a. IV) La prueba de cargo reunida es plena para condenar por los hechos atribuidos; en tanto no deja el menor atisbo de duda sobre el frustrado accionar **homicida** **del acusado**^{agente}.-
- b. IV) La prueba de cargo reunida es plena para condenar por los hechos atribuidos; en tanto no deja el menor atisbo de duda sobre el frustrado accionar **homicida** **del acusado**^{víctima}.-

En el Ejemplo 4.12, la confusión del modelo ya no es con los roles, sino con los eventos: identifica incorrectamente «golpear la puerta» como un evento de violencia, asignando, además, el rol de víctima al objeto inanimado.

- (4.12) a. El guardia de seguridad del bar, Juan, a fs. 72 y ss dijo que a Pedro lo consideraba fiolo por cómo las controlaba con los clientes (golpeaba la puerta y les decía “tiempo”).-
- b. El guardia de seguridad del bar, Juan, a fs. 72 y ss dijo que a Pedro lo consideraba fiolo por cómo las controlaba con los clientes (**golpeaba** **la puerta**^{víctima} y les decía “tiempo”).-

Al igual que en el trabajo de Breton et al. (2025), se detectan errores en los límites de los argumentos. En el Ejemplo 4.13 el modelo incluye como parte del objeto con el que se golpea a la víctima la expresión «tirones de pelo, etc.». Aunque estos elementos son, efectivamente, parte de una serie, no estarían ligados a «golpear», sino a las acciones adicionales admitidas por «la apelante». En el Ejemplo 4.13, el modelo no logra delimitar correctamente al agente, eligiendo en particular el inciso explicativo sobre «el imputado». En este último caso, sin embargo, desde un punto de vista semántico, la salida del modelo es correcta, ya que «el imputado» y «Juan» se refieren a la misma entidad.

- (4.13) a. **La apelante**^{agente} admitió ante la Magistrada actuante haber **golpeado** **a Juan**^{víctima} **con un cinto**^{instr.}, tirones de pelo, etc.
- b. **La apelante**^{agente} admitió ante la Magistrada actuante haber **golpeado** **a Juan**^{víctima} **con un cinto, tirones de pelo, etc.**^{instr.}
- a. **El imputado, Juan**^{hablante}, declara en lo medular que fue hasta el comercio y **amenazó** **a la dueña**^{dest.} **con un arma de juguete**^{instr.}, y ésta le entregó la suma de \$1200 pesos, que luego los gastó en droga.
- b. El imputado, **Juan**^{hablante}, declara en lo medular que fue hasta el comercio y **amenazó** **a la dueña**^{dest.} **con un arma de juguete**^{instr.}, y ésta le entregó la suma de \$1200 pesos, que luego los gastó en droga.

Los ejemplos 4.14 y 4.15 ilustran dos textos con múltiples eventos.

- (4.14) a. “Yo estoy mirando el vehículo en ese momento para tomar un bulto de la valija que estaba abierta y en ese momento **dicha persona**^{agente} ... **me**^{víctima} **golpea** en el medio de la espalda^{cuerpo} con el antebrazo^{instr.} yo pierdo estabilidad y me inclino hacia adentro del auto para mantener el equilibrio y ahí veo que me mete la mano en el bolsillo donde tengo la suma de \$ 100.000”.
- b. “Yo estoy mirando el vehículo en ese momento para tomar un bulto de la valija que estaba abierta y en ese momento **dicha persona**^{agente(1)/perpetr.(2)} ... **me**^{víctima(1)} **golpea**⁽¹⁾ en el medio de la espalda con el antebrazo^{instr.(1)} yo pierdo estabilidad y me inclino hacia adentro del auto para mantener el equilibrio y ahí veo que **me**^{víctima(2)} **mete la mano**⁽²⁾ en el bolsillo donde tengo la suma de \$ 100.000^{origen(2)}”.

En el primero, el modelo infiere otro evento de violencia, distinto al de «golpear», al interpretar «mete la mano en el bolsillo» como un evento de robo. Sin embargo, admitiendo que «meter la mano» es en el texto sinónimo de «robar»¹¹, los eventos asignados son los correctos, incluyendo a «dicha persona» como agente/perpetrador de ambos eventos.

- (4.15) a. Que la rapiña quedó consumada, desde que **los jóvenes adolescentes**^{hablante(1)/perpetrador(2)}, ingresaron al local de comercio, **amenazando**⁽¹⁾ con un revólver calibre 38^{instr.(1)}, a las personas que allí se encontraban^{dest.(1)}, **sustrayendo**⁽²⁾ el dinero y cartones de cigarrillos^{bienes(2)}, guardando todo en una mochila que portaban.
- b. Que la rapiña quedó consumada, desde que **los jóvenes adolescentes**^{perpetr.(2)}, ingresaron al local de comercio, **amenazando**⁽¹⁾ con un revólver calibre 38^{instr.(1)}, a las personas que allí se encontraban^{dest.(1)}, **sustrayendo**⁽²⁾ el dinero y cartones de cigarrillos^{bienes(2)}, guardando todo en una mochila que portaban.

En el Ejemplo 4.15, el modelo identifica correctamente los dos eventos de violencia y asigna correctamente los elementos correspondientes a cada uno

¹¹«Meter la mano» es, efectivamente, una locución que puede referirse a un robo. Sin embargo, de la lectura del texto no se desprende que efectivamente se produzca el hurto aunque esa pareciera ser la intención.

de ellos, salvo por «los jóvenes adolescentes» que son hablantes en el primer evento y perpetradores en el segundo, asignándoles únicamente el segundo de los roles.

En definitiva, los ejemplos presentados ilustran algunos de los errores cometidos por el modelo, que incluyen errores en la delimitación de los argumentos, confusiones entre los roles y dificultades en la detección o asignación de los elementos. Estos errores reflejan los desafíos inherentes a la tarea de análisis semántico en textos legales complejos con grandes modelos de lenguaje y muestran áreas potenciales de mejora para futuros trabajos.

4.5. Síntesis

En este capítulo se presenta la experimentación realizada para evaluar el desempeño de tres grandes modelos de lenguaje —Claude Sonnet 3.7, DeepSeek Reasoner y OpenAI GPT-4o— en la tarea de análisis semántico aplicada a textos legales en español. Continuando el trabajo presentado en el Capítulo 3, se adopta un enfoque de aprendizaje en contexto mediante ejemplos para guiar a los modelos en la tarea de identificación de eventos basados en los marcos semánticos de FrameNet.

Para delimitar los tipos de eventos a detectar, se realiza un estudio previo en donde se exploran los verbos asociados a violencia, teniendo en cuenta tanto la clasificación de verbos de ADESSE como la comprobación empírica de su uso en la BJN. Además, se introduce un nuevo marco semántico, «Amenaza», diseñado específicamente para capturar los eventos y roles referidos a los actos de intimidación. Aunque este tipo de eventos se encuentra abarcado por el marco «Compromiso» de FrameNet, el uso de aprendizaje en contexto requiere una definición más precisa para no generar ambigüedades en la tarea.

A partir de la lista de verbos, se anotan manualmente un corpus de 717 ejemplos en español, denominado ViolenciaR, extraídos de 274 sentencias de la Base de Jurisprudencia Nacional. Debido a la posible multiplicidad de eventos por ejemplo, estos textos contienen un total 1055 eventos: un 83 % son efectivamente eventos de violencia, mientras que el resto son usos en otras acepciones o sentidos no relacionados con violencia, como «violar una norma» o «golpear una puerta». Todos los ejemplos de ViolenciaR se encuentran desidentificados manualmente para proteger la privacidad de las personas involucradas.

Para la experimentación, el corpus ViolenciaR se divide en un conjunto de

517 ejemplos para inyectar como *shots*, dejando 200 ejemplos para la evaluación. Para evaluar el impacto de la procedencia de los ejemplos, se plantean experimentos variando la proporción de ejemplos provenientes de FrameNet y de la BJN en los *shots*, desde 0 % —todos los ejemplos de FrameNet— hasta 100 % —todos los ejemplos de la BJN—.

Los resultados obtenidos muestran una disminución significativa en el desempeño de los modelos al compararlos con experimentos previos realizados en textos en inglés de FrameNet. En particular, el mejor desempeño alcanzado en este escenario, un F_1 de 0.676 para Claude Sonnet 3.7, se encuentra 0.149 puntos por debajo del obtenido en los experimentos del Capítulo 3 con textos en inglés. Además, los mejores resultados globales, independientemente del modelo, se dan cuando todos los ejemplos provienen de la BJN, aunque esto no siempre es así cuando se consideran los argumentos individualmente. Desde un análisis cualitativo, se identifican errores en la delimitación de los argumentos, confusiones entre los roles y dificultades en la detección o asignación de los elementos.

Estos resultados, entonces, evidencian desafíos en el análisis semántico de textos legales complejos con grandes modelos de lenguaje y sugieren futuras investigaciones, como la incorporación de técnicas de ajuste fino o la exploración de arquitecturas híbridas que combinen el poder de los grandes modelos de lenguaje con enfoques más tradicionales de procesamiento del lenguaje natural.

Capítulo 5

Conclusiones

Desde hace décadas, el área de la inteligencia artificial aplicada al derecho constituye un campo de investigación activo, orientado al desarrollo de métodos computacionales para la representación, el análisis y la automatización de tareas jurídicas complejas. En particular, el procesamiento del lenguaje natural (PLN) en el dominio jurídico se ha consolidado como un área de investigación clave, dado que gran parte de la información jurídica se encuentra codificada en textos legales, como leyes, reglamentos, sentencias judiciales y contratos. Estos textos presentan desafíos únicos debido a su complejidad lingüística, semántica y discursiva.

La presente tesis aborda el desarrollo y la evaluación de métodos de PLN sobre textos jurídicos en español provenientes de la Base de Jurisprudencia Nacional (BJN) del Poder Judicial de Uruguay. La BJN es una colección en constante crecimiento de más de 100 000 sentencias judiciales en formato digital y constituye, por su volumen y relevancia institucional, un valioso recurso para la investigación en PLN aplicada al derecho.

En particular, esta tesis se plantea dos objetivos principales: (a) el estudio de la desidentificación automática de las sentencias judiciales de la BJN, con el fin de publicar las sentencias sin dañar la privacidad de las personas involucradas en los casos; (b) la extracción de eventos relevantes en sentencias judiciales, con el fin de facilitar la búsqueda, el análisis y la explotación de la información contenida en la BJN.

Para el primer objetivo, el trabajo de esta tesis se centra en los nombres de las personas como la categoría de información más crítica a desidentificar en los textos legales. En este trabajo se estudia la sustitución de los nombres de

las personas por identificadores ficticios, de forma que a una misma persona se le asigne siempre un mismo identificador. Esto permite la desidentificación del texto sin perder su coherencia narrativa.

La solución propuesta se organiza en dos subtareas: (a) el reconocimiento de entidades nombradas para identificar las menciones de nombres en el texto, y (b) la resolución de correferencias para agrupar todas las menciones a una misma persona, que luego se sustituyen de forma conjunta por su valor de reemplazo. Con el fin de entrenar y evaluar soluciones que resuelvan estas tareas, se construye el corpus B*N. Este corpus contiene 10 102 menciones a nombres propios y su correspondiente nuevo identificador, anotadas manualmente en 997 sentencias judiciales.

Para la primera subtarea, se experimenta reentrenando al módulo NER de la biblioteca SpaCy; se obtiene así un modelo con un F_1 de 0.902 que supera al valor reportado para ese mismo modelo en su corpus de entrenamiento original. Para resolver las correferencias, se experimenta con un algoritmo de agrupamiento aglomerativo y diversas distancias entre textos, utilizando al corpus para fijar su umbral de fusión. La mejor solución en este caso obtiene una medida *ARI* de 0.96 y una precisión por documento de 0.81.

Se construye, entonces, una solución para la desidentificación de textos basada en herramientas de código abierto y de bajo costo computacional, lo que permite su integración con sistemas existentes. Sin embargo, la precisión obtenida en ambas subtareas implica que el sistema comete errores que pueden afectar la privacidad de los datos o la coherencia del texto desidentificado. Por ejemplo, un error en la detección de una mención a un nombre puede dejar expuesta información sensible, mientras que errores en la agrupación de menciones pueden llevar a asignar identificadores incorrectos, afectando así la comprensión del texto. Esto descarta su uso completamente automatizado, pero sí lo habilita como apoyo en un esquema semiautomatizado, orientado a facilitar la labor de un operador humano que revise y corrija las salidas del sistema.

Para el segundo objetivo, la extracción de eventos en la BJN, se investiga la viabilidad del uso de aprendizaje en contexto (ICL) con grandes modelos de lenguaje para la tarea de análisis de marcos semánticos (FSP), tomando como base la Teoría de Marcos Semánticos de Fillmore. En particular, se elige centrar el estudio en los eventos de violencia, dada su relevancia social y jurídica.

Como primer acercamiento a la tarea, se realiza una serie de experimentos

preliminares utilizando textos en inglés de FrameNet, con el fin de establecer una línea base y comprender las capacidades de los modelos en un escenario controlado. Se explora la resolución de subtareas clave de FSP, incluyendo la identificación de los marcos, sus unidades léxicas objetivo y sus elementos.

La solución propuesta genera de forma automática *prompts* adaptados a cada tarea y al conjunto de marcos seleccionados. Estos *prompts* incluyen las definiciones de los marcos a extraer y ejemplos obtenidos de FrameNet.

La evaluación se realiza utilizando seis grandes modelos de lenguaje: GPT-4, GPT-4 Mini, Claude Sonnet 3.5, Claude Haiku 3.5, DeepSeek Reasoner y DeepSeek Chat. Se experimenta, además, con la cantidad de *shots* en el *prompt*, variando desde un enfoque sin ejemplos (*zero-shot*) hasta la inclusión de 150 o 400 *shots*, dependiendo del problema.

Los resultados obtenidos alcanzan valores de F_1 competitivos en varias subtareas de FSP, y muestran que el uso de ICL con ejemplos específicos mejora significativamente el desempeño de los modelos en comparación con un enfoque sin ejemplos.

A partir de estos resultados preliminares, se avanza hacia la aplicación de la metodología en textos legales de la BJN. Para ello, se construye el corpus ViolenciaR, compuesto de 717 textos anotados manualmente, extraídos de 274 sentencias judiciales, que contienen 1055 eventos de violencia. Los ejemplos de ViolenciaR se desidentifican manualmente para evitar el envío de información sensible a los modelos utilizados.

La selección de los marcos a extraer se realiza a partir de un estudio de los verbos asociados a violencia, utilizando la clasificación de verbos de ADESS-SE y la comprobación empírica de su uso en la BJN. Asimismo, se define un nuevo marco semántico, «Amenaza», para capturar eventos específicos de intimidación, especificando el marco «Compromiso» de FrameNet para evitar ambigüedades al incorporar otros verbos sin relación con eventos de violencia.

Para la evaluación se consideran tres modelos: GPT-4, Claude Sonnet 3.7 y DeepSeek Reasoner. A diferencia del caso anterior, se fija la cantidad de *shots* en 500, pero se varía la proporción de ejemplos provenientes de FrameNet y de la BJN, con el objetivo de evaluar el impacto de la procedencia de los ejemplos en el rendimiento de los modelos.

De estos experimentos se destacan dos resultados principales: (a) los mejores desempeños se obtienen cuando todos los ejemplos provienen de la BJN, independientemente del modelo; y (b) el mejor desempeño alcanzado en este

escenario, un F_1 de 0.676 para Claude Sonnet 3.7, se encuentra 0.149 puntos por debajo del obtenido en los experimentos preliminares. Estos resultados evidencian los desafíos asociados al uso de grandes modelos de lenguaje en el análisis semántico en el dominio jurídico debido a la complejidad de los textos legales.

Ambas tareas se realizan sobre textos provenientes de la BJN, que presentan particularidades lingüísticas y estructurales propias: lenguaje técnico específico, transcripciones de declaraciones y testimonios con distintos registros, que abarcan desde lo formal hasta lo coloquial, a menudo combinados en un mismo texto. Estas particularidades afectan la aplicación directa de herramientas entrenadas en dominios generales —desde la segmentación de oraciones hasta el reconocimiento de entidades nombradas y los grandes modelos de lenguaje— y constituyen un desafío adicional para el desarrollo de las técnicas propuestas.

Mediante el desarrollo de los dos objetivos planteados, este trabajo aporta avances en el área de PLN aplicada al dominio jurídico, con especial énfasis en la Base de Jurisprudencia Nacional.

Primeramente, se crean dos recursos lingüísticos anotados manualmente: los corpus B*N y ViolenciaR. Estos recursos constituyen aportes significativos al procesamiento automático del lenguaje jurídico en español y, específicamente, a la BJN. El trabajo de anotación para B*N es aceptado como ponencia en «Workshop on Legal Annotation, JURIX» (Garat y Wonsever, 2023).

El corpus B*N permite desarrollar y evaluar modelos automáticos para la desidentificación de sentencias judiciales, entrenando modelos para reconocimiento de entidades nombradas y métodos de agrupamiento aglomerativo para la resolución de correferencias. La solución es integrada como un prototipo en el proyecto del Fondo Sectorial de Investigación a partir de Datos de la ANII-MIEM. Asimismo, se publica este trabajo en la revista arbitrada «Information» (Garat y Wonsever, 2022).

Por su parte, el corpus ViolenciaR es el primer recurso anotado manualmente para la extracción de eventos de violencia en sentencias judiciales en español, basado en la Teoría de Marcos Semánticos. A esto se agrega una nueva definición de marco semántico, «Amenaza», adaptado a las características de los textos legales.

Este recurso habilita el uso de grandes modelos de lenguaje en tareas de análisis semántico en el dominio jurídico mediante ICL. Además, permite evaluar la influencia del origen y lengua de los ejemplos en el desempeño de los

modelos. Este estudio se basa en el análisis semántico con ICL y FrameNet en inglés, para el cual, salvo el trabajo paralelo de Devasier et al. (2025) para la extracción de argumentos, no se registran antecedentes. Los resultados de esta investigación se presentan en la conferencia arbitrada «Foundation and Large Language Models» (Garat et al. 2025).

En conclusión, este trabajo aporta nuevos recursos lingüísticos y metodologías innovadoras que abren nuevas líneas de investigación en tareas complejas como la desidentificación automática y la extracción de eventos. Es una contribución tanto al desarrollo tecnológico como al estudio académico del PLN aplicado al derecho en español y, en particular, a la Base de Jurisprudencia Nacional.

Estos resultados también abren oportunidades para investigaciones futuras que amplíen y profundicen en el análisis semántico y la desidentificación en textos jurídicos; se identifican a continuación posibles líneas de trabajo futuro.

En primer lugar, la desidentificación automática puede beneficiarse de la incorporación de nuevas categorías de información sensible más allá de los nombres de personas, completando la lista de categorías establecidas en la Acordada 7564 (Suprema Corte de Justicia, República Oriental del Uruguay, 2006). Además, siguiendo el trabajo de Cruces y Ferraro (2023), se pueden explorar otras técnicas de reconocimiento de entidades nombradas y resolución de correferencias, incluyendo enfoques basados en modelos de lenguaje entrenados a partir de textos de la BJN y ajustados a esta tarea específica.

La selección de ejemplos para la experimentación con ICL, aunque refleja la distribución de los marcos en el corpus, se realiza al azar y no toma en cuenta la entrada a procesar. Mejoras a investigar en este proceso de selección incluyen, por ejemplo, su optimización a partir de ejemplos relacionados con la entrada, o estudiando cómo afecta el orden de presentación de los ejemplos en la calidad de la respuesta (Devasier et al. 2025; Jimenez Gutierrez et al. 2022; P. Liu et al. 2023; Zhao et al. 2021). Más allá de la selección de ejemplos, la optimización de la construcción en sí misma de los *prompts* en las distintas tareas puede ser objeto de estudio, dado que pequeñas variaciones en su redacción pueden impactar significativamente en el desempeño de los modelos (Sclar et al. 2024; Tam et al. 2024; Zhao et al. 2021).

La evaluación realizada en este trabajo se basa en métricas automáticas que comparan las salidas de los modelos con las anotaciones manuales de referencia. El carácter estricto de la evaluación —en la que los segmentos deben coincidir

exactamente con los esperados— junto con posibles inconsistencias u omisiones presentes en las anotaciones contribuyen a subestimar el desempeño de los modelos. Aun considerando otras métricas automáticas más flexibles —basadas en dependencias semánticas (C. Baker et al. 2007) o que incluyen medidas de similitud entre cadenas (Breton et al. 2025)—, la comparación continúa apoyándose fuertemente en criterios predominantemente sintácticos. Por ello, una línea futura de investigación se orienta a la incorporación de medidas de carácter semántico, así como de evaluaciones humanas que consideren aspectos cualitativos, como la coherencia, relevancia y utilidad de las extracciones.

El corpus ViolenciaR, aunque constituye un recurso valioso para la investigación en PLN aplicada al derecho, presenta limitaciones en cuanto a su tamaño y al hecho de haber sido anotado por un único anotador que no es experto en el dominio. La ampliación del corpus mediante la incorporación de un mayor número de sentencias judiciales, así como su anotación y validación por parte de expertos, permitiría una evaluación y un entrenamiento más robustos y generalizables.

Los grandes modelos de lenguaje provistos por terceros facilitan la experimentación con modelos cuyas cantidades de parámetros se sitúan en el orden de los miles de millones. Sin embargo, presentan problemas legales y éticos a tener en cuenta, más si se tiene en consideración la naturaleza sensible de los textos legales. En este sentido, una línea futura de investigación consiste en explorar modelos de lenguaje abiertos que puedan ser entrenados localmente, evitando así el envío de datos sensibles a terceros. Técnicas como el ajuste fino permitirían, por ejemplo, adaptar estos modelos a las particularidades del dominio jurídico y del idioma español, mejorando así su desempeño en tareas específicas como la extracción de eventos (Chalkidis et al. 2020; Huang et al. 2023; Xiao et al. 2021). Asimismo, queda abierta la exploración de posibles soluciones que combinen estos modelos con enfoques más tradicionales de procesamiento del lenguaje natural.

Finalmente, la Base de Jurisprudencia Nacional se consolida como un recurso de gran valor para la investigación en PLN aplicada al derecho en español. Su continua expansión, junto con la creación de nuevos recursos anotados y el desarrollo de metodologías innovadoras, permitirá profundizar en el estudio y la automatización de tareas complejas en el ámbito jurídico, contribuyendo de manera sostenida al avance del campo de la inteligencia artificial aplicada al derecho.

Referencias bibliográficas

- Adhikary, S., Sen, P., Roy, D., y Ghosh, K. (2024). A case study for automated attribute extraction from legal documents using large language models. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-024-09425-7>
- Agarwal, R., Singh, A., Zhang, L., Bohnet, B., Rosias, L., Chan, S., Zhang, B., Anand, A., Abbas, Z., Nova, A., Co-Reyes, J. D., Chu, E., Behbahani, F., Faust, A., y Larochelle, H. (2025). Many-shot in-context learning. *Proceedings of the 38th International Conference on Neural Information Processing Systems*.
- Ai, C., y Tu, K. (2024). Frame Semantic Role Labeling Using Arbitrary-Order Conditional Random Fields. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17638-17646. <https://doi.org/10.1609/aaai.v38i16.29715>
- Aleven, V., y Ashley, K. D. (1997). Evaluating a Learning Environment for Case-Based Argumentation Skills. *International Journal of Artificial Intelligence in Education*, 8, 87-113.
- Anthropic. (2024). Introducing the next generation of Claude: The Claude 3 family [Model card and evaluations for Claude 3 family]. <https://www.anthropic.com/news/claude-3-family>
- Aramaki, E., Imai, T., Miyo, K., y Ohe, K. (2006). Automatic deidentification by using sentence features and label consistency. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, 10-11.
- Ariai, F., Mackenzie, J., y Demartini, G. (2025). Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges [arXiv:2410.21306]. <https://arxiv.org/abs/2410.21306>
- Ashley, K. D. (1990). *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*. MIT Press.

- Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., y Auli, M. (2019). Cloze-driven Pretraining of Self-attention Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5360-5369. <https://doi.org/10.18653/v1/D19-1539>
- Bagga, A., y Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. *Proceedings of the Linguistics Coreference Workshop at the First International Conference on Language Resources and Evaluation (LREC)*, 563-566.
- Baker, C., Ellsworth, M., y Erk, K. (2007, junio). SemEval-2007 Task 19: Frame Semantic Structure Extraction. En E. Agirre, L. Màrquez y R. Wicentowski (Eds.), *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 99-104). Association for Computational Linguistics.
- Baker, C. F., Fillmore, C. J., y Lowe, J. B. (1998). The Berkeley FrameNet Project. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 86-90. <https://doi.org/10.3115/980845.980860>
- Baker, C. F., y Lorenzi, A. (2020, mayo). Exploring Crosslinguistic Frame Alignment. En T. T. Torrent, C. F. Baker, O. Czulo, K. Ohara y M. R. L. Petruck (Eds.), *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet* (pp. 77-84). European Language Resources Association.
- Bakker, R. M., Schoevers, A. J., van Drie, R. A. N., Schraagen, M. P., y de Boer, M. H. T. (2025). Semantic role extraction in law texts: a comparative analysis of language models for legal information extraction. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-025-09437-x>
- Bastianelli, E., Vanzo, A., y Lemon, O. (2020). Encoding Syntactic Constituency Paths for Frame-Semantic Parsing with Graph Convolutional Networks [arXiv:2011.13210]. <https://arxiv.org/abs/2011.13210>
- Bayardo, R. J., y Agrawal, R. (2005). Data privacy through optimal k-anonymization. *Proceedings - International Conference on Data Engineering*, (Icde), 217-228. <https://doi.org/10.1109/ICDE.2005.42>

- Bex, F. J. (2024). AI, Law and beyond. A transdisciplinary ecosystem for the future of AI & Law. *Artif. Intell. Law*, 33(1), 253-270. <https://doi.org/10.1007/s10506-024-09404-y>
- Bird, S., Loper, E., y Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
- Blair-Stanek, A., Holzenberger, N., y Van Durme, B. (2023). Can GPT-3 Perform Statutory Reasoning? *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 22-31. <https://doi.org/10.1145/3594536.3595163>
- Boella, G., Caro, L., y Leone, V. (2019). Semi-automatic knowledge population in a legal document management system. *Artif. Intell. Law*, 27(2), 227-251. <https://doi.org/10.1007/s10506-018-9239-8>
- Bohnet, B., Alberti, C., y Collins, M. (2023). Coreference Resolution through a seq2seq Transition-Based System. *Transactions of the Association for Computational Linguistics*, 11, 212-226. https://doi.org/10.1162/tacl_a-00543
- Breton, J., Billami, M. M., Chevalier, M., Nguyen, H. T., Satoh, K., Trojahn, C., y Zin, M. M. (2025). Leveraging LLMs for legal terms extraction with limited annotated data. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-025-09448-8>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*.
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., y Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. *Workshop on Practical Machine Learning for Developing Countries (PML4DC) at ICLR 2020*.
- Cardie, C., y Wagstaff, K. (1999). Noun Phrase Coreference as Clustering. *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. <https://www.aclweb.org/anthology/W99-0611>
- Carreras, X., y Màrquez, L. (2005, junio). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. En I. Dagan y D. Gildea (Eds.),

- Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)* (pp. 152-164). Association for Computational Linguistics.
- Carvalho Dias, F. M. (2016). *Multilingual Automated Text Anonymization* [Tesis doctoral, Instituto Superior Técnico, Universidade de Lisboa].
- Centers for Medicare & Medicaid Services, United States of America. (1996). The Health Insurance Portability and Accountability Act (HIPAA).
- Chalapathy, R., Zare Borzeshi, E., y Piccardi, M. (2016). Bidirectional LSTM-CRF for Clinical Concept Extraction. *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, 7-12.
- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., y Androutsopoulos, I. (2020, noviembre). LEGAL-BERT: The Muppets straight out of Law School. En T. Cohn, Y. He y Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898-2904). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.261>
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tai, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., ... Wei, J. (2024). Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).
- Cruces, M., y Ferraro, N. (2023). *Aprendizaje profundo para la anonimización de textos legales* [Proyecto de grado]. Facultad de Ingeniería, Universidad de la República. <https://hdl.handle.net/20.500.12008/41686>
- Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., y Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. *Symmetry*, 13(8). <https://doi.org/10.3390/sym13081490>
- D'Amato, A. (1977). Can/Should Computers Replace Judges? *Georgia Law Review*, 11(4), 1277-1301.
- Das, D., Chen, D., Martins, A. F. T., Schneider, N., y Smith, N. A. (2014). Frame-Semantic Parsing. *Computational Linguistics*, 40(1), 9-56. https://doi.org/10.1162/COLI_a_00163
- DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., ... Pan, Z. (2025). DeepSeek-V3 Technical Report [arXiv:2412.19437]. <https://arxiv.org/abs/2412.19437>

- Dernoncourt, F., Lee, J. Y., Uzuner, Ö., y Szolovits, P. (2016). De-identification of Patient Notes with Recurrent Neural Networks. *CoRR*, *abs/1606.03475*. <http://arxiv.org/abs/1606.03475>
- Devasier, J., Mediratta, R., y Li, C. (2025, noviembre). Can LLMs Extract Frame-Semantic Arguments? En C. Christodoulopoulos, T. Chakraborty, C. Rose y V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 30597-30610). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.emnlp-main.1557>
- Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
- Domingo-Ferrer, J., Sánchez, D., y Rufian-Torrell, G. (2013). Anonymization of nominal data based on semantic marginality. *Information Sciences*, *242*, 35-48. <https://doi.org/10.1016/j.ins.2013.04.021>
- Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., y Sui, Z. (2024, noviembre). A Survey on In-context Learning. En Y. Al-Onaizan, M. Bansal y Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 1107-1128). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.64>
- Everitt, B., Landau, S., Leese, M., y Stahl, D. (2011). *Cluster Analysis*. Wiley. <https://doi.org/10.1002/9780470977811>
- Ferraro, G., y Suominen, H. (2020, diciembre). Transformer Semantic Parsing. En M. Kim, D. Beck y M. Mistica (Eds.), *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association* (pp. 121-126). Australasian Language Technology Association. <https://aclanthology.org/2020.alta-1.16/>
- Fillmore, C. J. (1976). Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences*, *280*(1), 20-32. <https://doi.org/10.1111/j.1749-6632.1976.tb25467.x>
- FitzGerald, N., Täckström, O., Ganchev, K., y Das, D. (2015, septiembre). Semantic Role Labeling with Neural Network Factors. En L. Màrquez, C. Callison-Burch y J. Su (Eds.), *Proceedings of the 2015 Conference*

- on *Empirical Methods in Natural Language Processing* (pp. 960-970). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D15-1112>
- Friedrich, M., Köhn, A., Wiedemann, G., y Biemann, C. (2019). Adversarial Learning of Privacy-Preserving Text Representations for De-Identification of Medical Records. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5829-5839. <https://doi.org/10.18653/v1/P19-1584>
- Fulloni, A., Langone, D., y Wonsever, D. (2020). A Citations Network for Legal Decisions. *Fourteenth International Workshop on Juris-Informatics (JURISIN 2020)*.
- Garat, D., Moncecchi, G., y Wonsever, D. (2025). In-Context Learning for Frame-Semantic Parsing. *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, 629-635. <https://doi.org/10.1109/FLLM67465.2025.11390933>
- Garat, D., y Wonsever, D. (2019). Towards De-identification of Legal Texts [arXiv:1910.03739]. <https://arxiv.org/abs/1910.03739>
- Garat, D., y Wonsever, D. (2022). Automatic Curation of Court Documents: Anonymizing Personal Data. *Information*, 13(1). <https://doi.org/10.3390/info13010027>
- Garat, D., y Wonsever, D. (2023). Annotating legal texts in interaction with software tools [Workshop on Legal Annotation, JURIX, Maastricht].
- García-Miguel, J. M., Vaamonde, G., y Domínguez, F. G. (2010, mayo). ADESSÉ, a Database with Syntactic and Semantic Annotation of a Corpus of Spanish. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner y D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). <https://aclanthology.org/L10-1593/>
- Gardner, J., Xiong, L., y Science, C. (2008). HIDE: An Integrated System for Health Information DE-identification. *Framework*, (100), 2005-2008. <https://doi.org/10.1109/CBMS.2008.129>
- Gildea, D., y Jurafsky, D. (2002). Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3), 245-288. <https://doi.org/10.1162/089120102760275983>

- Glaser, I., Schamberger, T., y Matthes, F. (2021). Anonymization of German Legal Court Rulings. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 205-209. <https://doi.org/10.1145/3462757.3466087>
- Governatori, G., Bench-Capon, T., Verheij, B., Araszkiewicz, M., Francesconi, E., y Grabmair, M. (2022). Thirty years of Artificial Intelligence and Law: the first decade. *Artif. Intell. Law*, 30(4), 481-519. <https://doi.org/10.1007/s10506-022-09329-4>
- Greenstein, S. (2022). Preserving the rule of law in the era of artificial intelligence (AI). *Artif. Intell. Law*, 30(3), 291-323. <https://doi.org/10.1007/s10506-021-09294-4>
- Grishman, R., y Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. <https://aclanthology.org/C96-1079/>
- Gupta, D., Saul, M., y Gilbertson, J. (2004). Evaluation of a Deidentification (De-Id) Software Engine to Share Pathology Reports and Clinical Documents for Research. *American Journal of Clinical Pathology*, 121(2), 176-186. <https://doi.org/10.1309/E6K33GBPE5C27FYU>
- Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., Armentano-Oller, C., Rodríguez-Penagos, C., Gonzalez-Agirre, A., y Villegas, M. (2022). MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 39-60. <https://doi.org/10.26342/2022-68-3>
- Habba, E., Keydar, R., Bareket, D., y Stanovsky, G. (2023). The Perfect Victim: Computational Analysis of Judicial Attitudes towards Victims of Sexual Violence. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 111-120. <https://doi.org/10.1145/3594536.3595168>
- Heatherly, R., Rasmussen, L. V., Peissig, P. L., Pacheco, J. A., Harris, P., Denny, J. C., y Malin, B. A. (2016). A multi-institution evaluation of clinical profile anonymization. *Journal of the American Medical Informatics Association*. <https://doi.org/10.1093/jamia/ocv154>
- Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., y Szpakowicz, S. (2010, julio). SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between

- Pairs of Nominals. En K. Erk y C. Strapparava (Eds.), *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 33-38). Association for Computational Linguistics. <https://aclanthology.org/S10-1006/>
- Honnibal, M., y Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. <https://spacy.io>
- Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., y Feng, Y. (2023). Lawyer LLaMA Technical Report [arXiv:2305.15062]. <https://arxiv.org/abs/2305.15062>
- Hubert, L., y Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, 2(1), 193-218. <https://doi.org/10.1007/BF01908075>
- Italiani, P., Moro, G., y Ragazzi, L. (2025). Enhancing legal question answering with data generation and knowledge distillation from large language models. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-025-09463-9>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., y Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12). <https://doi.org/10.1145/3571730>
- Jia, C., Liang, X., y Zhang, Y. (2019). Cross-Domain NER using Cross-Domain Language Modeling. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2464-2474. <https://doi.org/10.18653/v1/P19-1236>
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., y Sayed, W. E. (2023). Mistral 7B [arXiv:2310.06825]. <https://arxiv.org/abs/2310.06825>
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., ... Sayed, W. E. (2024). Mixtral of Experts [arXiv:2401.04088]. <https://arxiv.org/abs/2401.04088>
- Jiang, Y., Li, W., He, C., Ding, Y., y Tu, K. (2021). Automated Concatenation of Embeddings for Structured Prediction. *Proceedings of the 59th An-*

nual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2597-2608.

- Jimenez Gutierrez, B., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., y Su, Y. (2022, diciembre). Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. En Y. Goldberg, Z. Kozareva y Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 4497-4512). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.findings-emnlp.329>
- Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., y Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans (M. Johnson, B. Roark y A. Nenkova, Eds.). *Transactions of the Association for Computational Linguistics*, 8, 64-77. https://doi.org/10.1162/tacl_a_00300
- Jurafsky, D., y Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd) [Online manuscript released August 24, 2025]. <https://web.stanford.edu/~jurafsky/slp3/>
- Karle, T., y Vora, D. (2017). PRIVACY preservation in big data using anonymization techniques. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*. <https://doi.org/10.1109/ICDMAI.2017.8073538>
- Karttunen, L. (1969). Discourse Referents. *International Conference on Computational Linguistics COLING 1969: Preprint No. 70*. <https://aclanthology.org/C69-7001/>
- Katz, D. M., Bommarito, M. J., Gao, S., y Arcuri, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2265), 20230254. <https://doi.org/10.1098/rsta.2023.0254>
- Kim, Y. (2014, octubre). Convolutional Neural Networks for Sentence Classification. En A. Moschitti, B. Pang y W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746-1751). Association for Computational Linguistics. <https://doi.org/10.3115/v1/D14-1181>
- Krishnan, V., y Manning, C. D. (2006). An Effective Two-stage Model for Exploiting Non-local Dependencies in Named Entity Recognition. *Proceedings of the 21st International Conference on Computational Linguis-*

- tics and the 44th Annual Meeting of the Association for Computational Linguistics*, 1121-1128. <https://doi.org/10.3115/1220175.1220316>
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., y Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 260-270. <https://doi.org/10.18653/v1/N16-1030>
- Lecun, Y., Bottou, L., Bengio, Y., y Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. <https://doi.org/10.1109/5.726791>
- Lee, J. (2025). Legal text classification in Korean sexual offense cases: from traditional machine learning to large language models with XAI insights. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-025-09454-w>
- Lee, J. Y., Deroncourt, F., Uzuner, Ö., y Szolovits, P. (2016). Feature-Augmented Neural Networks for Patient Note De-identification. *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, 17-22.
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., y Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [arXiv:1910.13461]. <https://arxiv.org/abs/1910.13461>
- Li, G., Wang, P., y Ke, W. (2023, diciembre). Revisiting Large Language Models as Zero-shot Relation Extractors. En H. Bouamor, J. Pino y K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 6877-6892). Association for Computational Linguistics. <https://aclanthology.org/2023.findings-emnlp.459/>
- Li, J., Chen, X., Hovy, E., y Jurafsky, D. (2016). Visualizing and Understanding Neural Models in NLP. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 681-691. <https://doi.org/10.18653/v1/N16-1082>
- Li, Y., Baldwin, T., y Cohn, T. (2018). Towards Robust and Privacy-preserving Text Representations. *Proceedings of the 56th Annual Meeting of the*

- Association for Computational Linguistics (Volume 2: Short Papers)*, 25-30. <https://doi.org/10.18653/v1/P18-2005>
- Lin, X. V., Mihaylov, T., Artetxe, M., Wang, T., Chen, S., Simig, D., Ott, M., Goyal, N., Bhosale, S., Du, J., Pasunuru, R., Shleifer, S., Koura, P. S., Chaudhary, V., O'Horo, B., Wang, J., Zettlemoyer, L., Kozareva, Z., Diab, M., ... Li, X. (2022, diciembre). Few-shot Learning with Multilingual Generative Language Models. En Y. Goldberg, Z. Kozareva y Y. Zhang (Eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9019-9052). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.616>
- Litkowski, K. (2004). Senseval-3 task: Automatic labeling of semantic roles. *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 9-12.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., y Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9). <https://doi.org/10.1145/3560815>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., y Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach [arXiv:1907.11692]. <https://arxiv.org/abs/1907.11692>
- Liu, Z., Chen, Y., Tang, B., Wang, X., Chen, Q., Li, H., Wang, J., Deng, Q., y Zhu, S. (2015). Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of Biomedical Informatics*. <https://doi.org/10.1016/j.jbi.2015.06.009>
- Liu, Z., Tang, B., Wang, X., y Chen, Q. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, 75, S34-S42. <https://doi.org/10.1016/j.jbi.2017.05.023>
- Louis, A., van Dijck, G., y Spanakis, G. (2023, mayo). Finding the Law: Enhancing Statutory Article Retrieval via Graph Neural Networks. En A. Vlachos y I. Augenstein (Eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2761-2776). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.eacl-main.203>

- Luo, X. (2005). On Coreference Resolution Performance Metrics. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 25-32. <https://doi.org/10.3115/1220575.1220579>
- Luo, X., Pradhan, S., Recasens, M., y Hovy, E. (2014). An Extension of BLANC to System Mentions. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 24-29. <https://doi.org/10.3115/v1/P14-2005>
- Ma, X., Liu, Z., y Hovy, E. (2016). Unsupervised Ranking Model for Entity Coreference Resolution. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1012-1018. <https://doi.org/10.18653/v1/N16-1116>
- Mamede, N., Baptista, J., y Dias, F. (2016). Automated anonymization of text documents. *2016 IEEE Congress on Evolutionary Computation, CEC 2016*, 1287-1294. <https://doi.org/10.1109/CEC.2016.7743936>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., y McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Association for Computational Linguistics (ACL) System Demonstrations*, 55-60. <http://www.aclweb.org/anthology/P/P14/P14-5010>
- Màrquez, L., Recasens, M., y Sapena, E. (2013). Coreference Resolution: An Empirical Study Based on SemEval-2010 Shared Task 1. *Language Resources and Evaluation*, 47(3), 661-694. <https://doi.org/10.1007/s10579-012-9194-z>
- Martinelli, G., Barba, E., y Navigli, R. (2024, agosto). Maverick: Efficient and Accurate Coreference Resolution Defying Recent Trends. En L.-W. Ku, A. Martins y V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13380-13394). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.722>
- Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. *Artif. Intell. Law*, 32(3), 447-462.
- Mitra, R., Jeuniaux, P., Angheluta, R., y Moens, M.-F. (2003). Progressive fuzzy clustering for noun phrase coreference resolution. <https://lirias.kuleuven.be/1710555>

- Moosavi, N. S., y Strube, M. (2016). Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 632-642. <https://doi.org/10.18653/v1/P16-1060>
- Moro, G., Piscaglia, N., Ragazzi, L., y Italiani, P. (2023). Multi-language transfer learning for low-resource legal case summarization. *Artif. Intell. Law*, 32(4), 1111-1139. <https://doi.org/10.1007/s10506-023-09373-8>
- Newhauser, W., Jones, T., Swerdloff, S., Newhauser, W., Cilia, M., Carver, R., Halloran, A., y Zhang, R. (2014). Anonymization of DICOM electronic medical records for radiation therapy. *Computers in Biology and Medicine*. <https://doi.org/10.1016/j.compbiomed.2014.07.010>
- Oksanen, A., Tamper, M., Tuominen, J., Hietanen, A., y Hyvönen, E. (2019). Anoppi: A pseudonymization service for Finnish court documents. En *Legal Knowledge and Information Systems* (pp. 251-254). IOS Press.
- OpenAI. (2024). GPT-4 Technical Report [arXiv:2303.08774]. <https://arxiv.org/abs/2303.08774>
- Padró, L., y Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*.
- Palmer, M., Gildea, D., y Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 71-106. <https://doi.org/10.1162/0891201053630264>
- Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., y Ji, H. (2017, julio). Cross-lingual Name Tagging and Linking for 282 Languages. En R. Barzilay y M.-Y. Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1946-1958). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P17-1178>
- Parlamento Europeo y Consejo de la Unión Europea. (2016). Reglamento (UE) 2016/679. <http://eur-lex.europa.eu/legal-content/ES/LSU/?uri=CELEX:32016R0679>
- Patil, D., Mohapatra, R. K., y Babu, K. S. (2017). Evaluation of generalization based K-anonymization algorithms. *2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)*, 171-175. <https://doi.org/10.1109/SSPS.2017.8071586>

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., y Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Phuong, N. D., y Chau, V. T. N. (2016). Automatic de-identification of medical records with a multilevel hybrid semi-supervised learning approach. *2016 IEEE RIVF International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for the Future, RIVF 2016 - Proceedings*. <https://doi.org/10.1109/RIVF.2016.7800267>
- Poder Legislativo, República Oriental del Uruguay. (2008). Ley 18331. Protección de Datos Personales y Acción de "Habeas Data". <https://parlamento.gub.uy/documentosleyes/leyes/ley/18331>
- Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., y Zhang, Y. (2012, julio). CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. En S. Pradhan, A. Moschitti y N. Xue (Eds.), *Joint Conference on EMNLP and CoNLL - Shared Task* (pp. 1-40). Association for Computational Linguistics. <https://aclanthology.org/W12-4501/>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., y Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Ragazzi, L., Moro, G., Guidi, S., y Frisoni, G. (2024). LAWSUIT: a Large expert-Written SUMmarization dataset of ITalian constitutional court verdicts. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-024-09414-w>
- Ribeiro de Faria, J., Xie, H., y Steffek, F. (2025). Information extraction from employment tribunal judgments using a large language model. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-025-09443-z>
- Rosenberg, A., y Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 410-420. <https://www.aclweb.org/anthology/D07-1043>

- Roth, M., y Lapata, M. (2015). Context-aware Frame-Semantic Role Labeling (M. Collins y L. Lee, Eds.). *Transactions of the Association for Computational Linguistics*, 3, 449-460. https://doi.org/10.1162/tacl_a_00150
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., y Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice* (inf. téc.) (Revised November 1, 2016. Distributed with the FrameNet data). International Computer Science Institute. Berkeley, California. <https://framenet.icsi.berkeley.edu>
- Sachoulidou, A. (2023). Going beyond the “common suspects”: to be presumed innocent in the era of algorithms, big data and artificial intelligence. *Artificial Intelligence and Law*, 1-54. <https://doi.org/10.1007/s10506-023-09347-w>
- Sartor, G., Araszkievicz, M., Atkinson, K., Bex, F., van Engers, T., Francesconi, E., Prakken, H., Sileno, G., Schilder, F., Wyner, A., y Bench-Capon, T. (2022). Thirty years of Artificial Intelligence and Law: the second decade. *Artificial Intelligence and Law*, 30(4), 521-557. <https://doi.org/10.1007/s10506-022-09326-7>
- Savelka, J. (2023). Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts. *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 447-451. <https://doi.org/10.1145/3594536.3595161>
- Schweitzer, S., y Conrads, M. (2024). The digital transformation of jurisprudence: an evaluation of ChatGPT-4’s applicability to solve cases in business law. *Artificial Intelligence and Law*, 32(4), 555-591. <https://doi.org/10.1007/s10506-024-09406-w>
- Sclar, M., Choi, Y., Tsvetkov, Y., y Suhr, A. (2024). Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. <https://openreview.net/forum?id=RLu5lyNXjT>
- Seow, W. L., Chaturvedi, I., Hogarth, A., Mao, R., y Cambria, E. (2025). A review of named entity recognition: from learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, 58(10), 315. <https://doi.org/10.1007/s10462-025-11321-8>
- Sharma, S., Gamoura, S., Prasad, D., y Aneja, A. (2021). Emerging Legal Informatics towards Legal Innovation: Current status and future challenges

- and opportunities. *Legal Information Management*, 21(3–4), 218–235. <https://doi.org/10.1017/S1472669621000384>
- Siino, M., Falco, M., Croce, D., y Rosso, P. (2025). Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access*, 13, 18253–18276. <https://doi.org/10.1109/ACCESS.2025.3533217>
- Sleimi, A., Sannier, N., Sabetzadeh, M., Briand, L., y Dann, J. (2018). Automated Extraction of Semantic Legal Metadata using Natural Language Processing. *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 124–135. <https://doi.org/10.1109/RE.2018.00022>
- Soltan, S., Ananthakrishnan, S., FitzGerald, J., Gupta, R., Hamza, W., Khan, H., Peris, C., Rawls, S., Rosenbaum, A., Rumshisky, A., Prakash, C. S., Sridhar, M., Triefenbach, F., Verma, A., Tur, G., y Natarajan, P. (2022). AlexaTM 20B: Few-Shot Learning Using a Large-Scale Multilingual Seq2Seq Model [arXiv:2208.01448]. <https://arxiv.org/abs/2208.01448>
- Song, Y., Jiang, J., Zhao, W. X., Li, S., y Wang, H. (2012). Joint Learning for Coreference Resolution with Markov Logic. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1245–1254. <https://www.aclweb.org/anthology/D12-1114>
- Soon, W. M., Ng, H. T., y Lim, D. C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4), 521–544. <https://doi.org/10.1162/089120101753342653>
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., y Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>
- Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., y Varga, D. (2006, mayo). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. En N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk y D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA). <https://aclanthology.org/L06-1196/>
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., y Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. *Pro-*

- ceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 102-107.
- Subirats Rüggeberg, C., y Petruck, M. (2010). Surprise: Spanish FrameNet [Proceedings of the International Congress of Linguists, Prague]. *Estudios de Lingüística del Español*, 31, 1-12.
- Suprema Corte de Justicia, República Oriental del Uruguay. (2006). Acordada 7564. Tratamiento de datos en el ámbito del Poder Judicial. <https://www.poderjudicial.gub.uy/documentos/46-acordadas-y-circulares/4537-008-2006-acordada-7564-tratamiento-de-datos-en-el-ambito-del-pj.html>
- Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. *AMIA Annu Symp Proc*, 333-337.
- Tagarelli, A., y Simeri, A. (2022). Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artif. Intell. Law*, 30(3), 417-473. <https://doi.org/10.1007/s10506-021-09301-8>
- Tam, Z. R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H.-y., y Chen, Y.-N. (2024, noviembre). Let Me Speak Freely? A Study On The Impact Of Format Restrictions On Large Language Model Performance. En F. Dernoncourt, D. Preoțiuc-Pietro y A. Shimorina (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 1218-1236). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-industry.91>
- Taulé, M., Martí, M. A., y Recasens, M. (2008, mayo). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odiijk, S. Piperidis y D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA). <https://aclanthology.org/L08-1222/>
- Touvron, H., Martin, L., Stone, K., ... y Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models [arXiv:2307.09288]. <https://arxiv.org/abs/2307.09288>
- Unanue, I. J., Borzeshi, E. Z., y Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of Biomedical Informatics*, 76(November), 102-109. <https://doi.org/10.1016/j.jbi.2017.11.007>

- Uzuner, Ö., Sibanda, T. C., Luo, Y., y Szolovits, P. (2008). A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine*, 42(1), 13-35. <https://doi.org/10.1016/j.artmed.2007.10.001>
- Van Opijnen, M., y Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artif. Intell. Law*, 25(1), 65-87. <https://doi.org/10.1007/s10506-017-9195-8>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., y Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 5998-6008. <https://arxiv.org/abs/1706.03762>
- Vilain, M., Burger, J., Aberdeen, J., Connolly, D., y Hirschman, L. (1995). A Model-theoretic Coreference Scoring Scheme. *Proceedings of the 6th Conference on Message Understanding*, 45-52. <https://doi.org/10.3115/1072399.1072405>
- Villata, S., Araszkievicz, M., Ashley, K., Bench-Capon, T., Branting, L. K., Conrad, J. G., y Wyner, A. (2022). Thirty years of artificial intelligence and law: the third decade. *Artif. Intell. Law*, 30(4), 561-591. <https://doi.org/10.1007/s10506-022-09327-6>
- Waltl, B., Bonczek, G., Scepankova, E., y Matthes, F. (2019). Semantic types of legal norms in German laws: classification and analysis using local linear explanations. *Artif. Intell. Law*, 27(1), 43-71. <https://doi.org/10.1007/s10506-018-9228-y>
- Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., y Kurohashi, S. (2023, diciembre). GPT-RE: In-context Learning for Relation Extraction using Large Language Models. En H. Bouamor, J. Pino y K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 3534-3547). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.214>
- Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., y Han, W. (2024). ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT [arXiv:2302.10205]. <https://arxiv.org/abs/2302.10205>
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1), 36-45. <https://doi.org/10.1145/365153.365168>

- Welles, O. (1970). *The Trial* [Screenplay based on the novel by Franz Kafka]. Simon; Schuster. <https://cinephiliabeyond.org/wp-content/uploads/2023/09/The-Trial.pdf>
- Wu, W., Wang, F., Yuan, A., Wu, F., y Li, J. (2020, julio). CorefQA: Coreference Resolution as Query-based Span Prediction. En D. Jurafsky, J. Chai, N. Schluter y J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6953-6963). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.622>
- Xiao, C., Hu, X., Liu, Z., Tu, C., y Sun, M. (2021). Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents [arXiv:2105.03887]. <https://arxiv.org/abs/2105.03887>
- Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., y Chen, E. (2024). Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6), 186357. <https://doi.org/10.1007/s11704-024-40555-y>
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., y Raffel, C. (2021, junio). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. En K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty y Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 483-498). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.41>
- Yadav, S., Ekbal, A., Saha, S., y Bhattacharyya, P. (2016). Deep Learning Architecture for Patient Data De-identification in Clinical Records. *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, 32-41.
- Yamada, H., Teufel, S., y Tokunaga, T. (2019). Building a corpus of legal argumentation in Japanese judgement documents: towards structure-based summarisation. *Artif. Intell. Law*, 27(2), 141-170. <https://doi.org/10.1007/s10506-019-09242-3>
- Yang, Z., Ma, J., Chen, H., Zhang, J., y Chang, Y. (2024). Context-Aware Attentive Multilevel Feature Fusion for Named Entity Recognition.

- IEEE Transactions on Neural Networks and Learning Systems*, 35(1), 973-984. <https://doi.org/10.1109/TNNLS.2022.3178522>
- Zhang, K., Jimenez Gutierrez, B., y Su, Y. (2023, julio). Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors. En A. Rogers, J. Boyd-Graber y N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 794-812). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.50>
- Zhang, X., Yang, L. T., Liu, C., y Chen, J. (2014). A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. *IEEE Transactions on Parallel and Distributed Systems*, 25(2), 363-373. <https://doi.org/10.1109/TPDS.2013.48>
- Zhang, Y., Zhong, V., Chen, D., Angeli, G., y Manning, C. D. (2017). Position-aware Attention and Supervised Data Improve Slot Filling. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, 35-45. <https://nlp.stanford.edu/pubs/zhang2017tacred.pdf>
- Zhao, T. Z., Wallace, E., Feng, S., Klein, D., y Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 12697-12706. <https://proceedings.mlr.press/v139/zhao21c/zhao21c.pdf>
- Zheng, C., Wang, Y., y Chang, B. (2023). Query your model with definitions in FrameNet: an effective method for frame semantic role labeling. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, 14029-14037. <https://doi.org/10.1609/aaai.v37i11.26642>
- Zheng, J., Chapman, W. W., Crowley, R. S., y Savova, G. K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of Biomedical Informatics*, 44(6), 1113-1122. <https://doi.org/10.1016/j.jbi.2011.08.006>

Glosario

ADESSE Recurso léxico-semántico del español que describe el significado verbal mediante esquemas de predicación, roles semánticos y patrones de realización sintáctica, a partir de la clasificación y anotación de verbos en corpus.

agrupamiento aglomerativo Técnica de agrupamiento jerárquico que comienza con cada punto de datos como un clúster individual y luego fusiona iterativamente los clústeres más cercanos hasta que se alcanza un criterio de detención predefinido.

alternancia diatética En el caso de los verbos, propiedad que habilita la expresión de un mismo evento mediante distintas configuraciones sintácticas, con cambios sistemáticos en la realización de sus argumentos, sin modificar sustancialmente el significado del predicado.

anotación Proceso de añadir información estructurada a un texto, como etiquetas sintácticas, semánticas o discursivas, para su análisis sistemático o automático.

análisis de marcos semánticos (*Frame Semantic Parsing*, FSP) Técnica de procesamiento de lenguaje natural que utiliza marcos semánticos para interpretar el significado de las oraciones, identificando las relaciones entre los elementos y su función dentro del contexto.

aprendizaje débil Método de aprendizaje automático que utiliza datos con etiquetas incompletas, imprecisas o ruidosas para entrenar modelos, aprovechando técnicas como el aprendizaje semi-supervisado, el autoaprendizaje y la generación automática de etiquetas.

aprendizaje en contexto (*In-Context Learning*, ICL) Técnica que consiste en incluir una descripción de la tarea y ejemplos representativos en el

contexto de entrada de un modelo de lenguaje, con el fin de guiar su comportamiento, sin requerir un proceso de entrenamiento adicional.

aprendizaje no supervisado Método de aprendizaje automático en el que el modelo aprende patrones y estructuras a partir de datos no etiquetados, sin recibir retroalimentación explícita sobre las respuestas correctas.

bosque aleatorio (*Random Forest*) Técnica de aprendizaje automático que utiliza múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste, combinando sus resultados para la clasificación o regresión.

corpus Conjunto de textos recopilados para servir como base empírica del análisis lingüístico y del entrenamiento o evaluación de modelos.

desidentificación Proceso de eliminar u ofuscar la información personal o sensible en un conjunto de datos para proteger la privacidad de las personas.

dropout Método utilizado en el entrenamiento de redes neuronales que consiste en desactivar aleatoriamente un porcentaje de neuronas durante cada iteración, con el objetivo de reducir el sobreajuste del modelo.

elemento del marco (*Frame Element*, FE) Rol participante definido dentro de un marco semántico que especifica una función particular en la situación o evento representado, siguiendo la definición de FrameNet.

etiquetado de categorías gramaticales (*POS Tagging*) Proceso de asignar categorías gramaticales a cada palabra en un texto, como sustantivo, verbo, adjetivo, etc., basado en su función dentro de la oración.

etiquetado de roles semánticos (*Frame Semantic Role Labeling*, FSRL) Técnica de procesamiento de lenguaje natural que asigna roles semánticos a los argumentos de un marco semántico, identificando las funciones que desempeñan en relación con el evento o situación descrita.

evento de violencia Evento que implica el uso explícito o implícito de fuerza física o coerción contra una entidad, identificado a partir de predicados, participantes y contexto discursivo.

extracción de eventos Proceso de identificar y estructurar eventos específicos en un texto, incluyendo predicados, participantes y relaciones tem-

porales o causales, con el fin de analizarlos o clasificarlos en tareas de procesamiento de lenguaje natural.

extracción de información (*Information Extraction*, IE) Proceso de identificar y extraer información estructurada a partir de datos no estructurados, como textos, para facilitar su análisis y uso.

FrameNet Recurso léxico-semántico basado en la Teoría de Marcos Semánticos que modela el significado de unidades léxicas mediante marcos semánticos, roles participantes (elementos del marco) y patrones de valencia, a partir de anotaciones sobre corpus.

gran modelo de lenguaje (*Large Language Model*, LLM) Arquitectura computacional basada en redes neuronales profundas —generalmente del tipo transformador— entrenada sobre grandes corpus textuales para modelar distribuciones lingüísticas. Estos modelos se caracterizan por poseer desde cientos de millones hasta cientos de miles de millones de parámetros, lo que les permite capturar patrones complejos y desempeñar tareas avanzadas de procesamiento del lenguaje.

grupo nominal Unidad sintáctica que tiene como núcleo un sustantivo o pronombre y puede incluir determinantes, adjetivos y otros modificadores relacionados con el sustantivo.

grupo verbal Unidad sintáctica que tiene como núcleo un verbo y puede incluir complementos y modificadores relacionados con la acción o estado expresado por el verbo.

identificación de marco (*Frame Identification*, FI) Proceso de asignar un marco semántico específico a una palabra o frase en un texto, basándose en su contexto y significado.

lema Forma base o canónica de una palabra, utilizada como referencia en diccionarios y análisis lingüísticos. Por ejemplo, el lema de las palabras «corriendo», «corrí» y «correrá» es «correr».

marco semántico Estructura cognitiva que representa una situación, evento o concepto específico, incluyendo sus participantes y las relaciones entre ellos. Los marcos semánticos son utilizados para comprender el significado de las palabras y frases en un contexto dado.

modelo estudiante Modelo de aprendizaje automático que se entrena para imitar el comportamiento de un modelo más grande y complejo (modelo maestro), con el objetivo de lograr un rendimiento similar pero con menor costo computacional.

nominalización Proceso lingüístico mediante el cual un verbo o adjetivo se convierte en un sustantivo, permitiendo que la acción o cualidad sea tratada como una entidad.

procesamiento del lenguaje natural (PLN) Rama de la inteligencia artificial que se enfoca en la interacción entre las computadoras y el lenguaje humano, permitiendo a las máquinas comprender, interpretar y generar texto, voz y otros modos de comunicación natural.

prompt Entrada diseñada para guiar la respuesta de un modelo de lenguaje. Un *prompt* puede incluir instrucciones, contexto y ejemplos para ayudar al modelo a generar la salida deseada.

PropBank Recurso léxico-semántico que describe la estructura argumental de predicados verbales mediante roles semánticos numerados y esquemas de valencia, a partir de anotaciones sobre corpus.

realización sintáctica En el caso de los verbos, las distintas configuraciones sintácticas mediante las cuales un mismo evento o predicado verbal se expresa en la oración, en función de la disposición y las funciones gramaticales de sus argumentos. Estas son las formas concretas en que se materializa la alternancia diatética; por ejemplo, el verbo «romper» puede realizarse en una construcción transitiva, «el niño rompió el vaso», o intransitiva, «el vaso se rompió».

reconocimiento de entidades nombradas (*Named Entity Recognition*, NER) Técnica de procesamiento de lenguaje natural que identifica y clasifica entidades mencionadas en un texto, como nombres de personas, organizaciones, ubicaciones, fechas, etc.

resolución de correferencias Tarea de procesamiento de lenguaje natural que consiste en identificar y agrupar todas las expresiones en un texto que se refieren a la misma entidad o concepto.

rol semántico Función que desempeña una entidad en un evento o situación

expresados por un verbo o un marco semántico, como agente, paciente, instrumento, entre otros.

ruido Información irrelevante o errónea que puede afectar negativamente el entrenamiento o el rendimiento de los modelos de aprendizaje automático.

shot Término utilizado en el aprendizaje en contexto para referirse a una demostración individual incluida en el *prompt*, típicamente compuesta por un par de entrada y salida. Expresiones como *zero-shot* o *n-shot* indican la ausencia o el número de estas demostraciones.

tokenización Proceso de dividir un texto en unidades más pequeñas llamadas tokens, que usualmente son palabras, subpalabras o caracteres.

transferencia interlingüística (*Cross-Lingual Transfer*) Técnica en procesamiento de lenguaje natural que permite aplicar conocimientos o modelos entrenados en un idioma a otro idioma, facilitando tareas como la traducción automática, el análisis de sentimientos y el reconocimiento de entidades nombradas en múltiples lenguas.

unidad léxica (*Lexical Unit*, LU) Asociación entre una forma léxica y un sentido específico, definida en relación con un marco semántico o un esquema de predicación.

valencia En el caso de un verbo, propiedad que indica el número y tipo de argumentos que requiere o puede admitir para formar una oración gramaticalmente completa. Los argumentos incluyen sujetos, objetos directos e indirectos, y, en algunos análisis, complementos circunstanciales obligatorios.

APÉNDICES

Apéndice A

ICL para Análisis de Marcos Semánticos

En este apéndice se presentan los *prompts* generados para las distintas tareas de análisis de marcos semánticos con aprendizaje en contexto.

Como se detalla en la Sección 3.3, se realizan tres tipos de tareas:

1. Análisis de marcos semánticos (FSP): en una única llamada al modelo, se extraen los marcos presentes junto a sus objetivos y sus argumentos.
2. Identificación de marco y objetivo (FI): se detectan las unidades léxicas y los marcos a los que evocan, esto es, las subtareas de «detección de objetivo» e «identificación de marco» consideradas conjuntamente.
3. Etiquetado de roles semánticos (FSRL): dada una lista de pares marco–objetivo previamente detectados en un texto, se identifican todos los argumentos correspondientes a cada uno de esos marcos.

A su vez, y considerando los experimentos generales y de ablación de la Sección 3.3, para cada tarea se consideran distintos escenarios incrementales, en los que se varían la información de FrameNet proporcionada respecto a los marcos semánticos a extraer y la cantidad de ejemplos dados como contexto:

1. Sin información: no se proporciona información de los marcos ni ejemplos en contexto; solo se establece la definición de la tarea a resolver y los marcos a extraer.
2. Solo definición: se agrega únicamente la definición de los marcos a extraer, pero no se incluyen ejemplos en contexto.

Escenario	FSP	FI	FSRL
Sin información	A.2	A.7	A.12
Solo definición	A.3	A.8	A.13
<i>Zero-shot</i>	A.4	A.9	A.14
<i>n-shot</i>	A.5	A.10	A.15

Tabla A.1: Escenarios experimentales y figuras asociadas a los *prompts* generados para cada tarea

3. *Zero-shot*: la información anterior se complementa con la descripción de los argumentos, pero aún sin ejemplos en contexto.
4. *n-shot*: se proporciona toda la información disponible sobre los marcos, junto con n ejemplos en contexto.

A continuación se presentan ejemplos de *prompts* utilizados para cada una de las tareas y escenarios mencionados. En particular, para la tarea FSP se ejemplifica con el marco semántico «Asesinato», mientras que para las tareas FI y FSRL se agregan además los marcos «Robo» y «Violación». En el caso de los escenarios *n-shot*, se detallan los resultados de la generación para tres ejemplos.

En las figuras [A.1](#), [A.6](#) y [A.11](#) se presentan las plantillas Jinja utilizadas para generar los *prompts* para las tareas FSP, FI y FSRL, respectivamente. La Tabla [A.1](#) resume las figuras con ejemplos de *prompts* generados para los distintos escenarios experimentales y tareas.

```

{% set p_frame = prompt_frames|first %}
{% set frame = p_frame.frame %}
## GOAL:
Your goal is to extract all {{ frame.name }} events present in an english text as JSON
↪ records.
{% if render_definitions %}

## {{ frame.name }} EVENT:
A {{ frame.name }} event is defined as:

    {{ frame.definition }}
{% if render_definition_example %}

For example, in the following input:

    {{ frame.definition_example.text }}

this {{ frame.name }} event is found:

    {{ p_frame.definition_example | tojson }}

{% endif %}
{% endif %}
{% if render_definition_arguments %}
### ENTITIES:
The entities you might find on the text for {{ frame.name }} events are:

{% for entity in p_frame.prompt_entities %}
    - {{ entity.name }}: {{ entity.definition }}
{% endfor %}
{% endif %}

## GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

# You can only use the information provided in the input to generate your answer.
# Your answer must be one list of JSON records of all extracted {{ frame.name }} events:
    - If N number of events are found, you must answer with one list of N records.
    - If no event is detected, the list must be empty.
    - All keys in the output records must be one of the defined entities, all values in the
      ↪ record must be actual segments of the input text.
# Do NOT reply with any explanation or any other text but one list of records.

{% if prompt_examples %}
## EXAMPLES:

{% for example in prompt_examples %}
    ### Example {{ loop.index }}
    Input: {{ example.text | tojson }}
    Output: {{ example.output | tojson }}

{% endfor %}
    ## Example {{ prompt_examples|length +1 }}
    Input: Ahab and the crew continue their eventful journey and encounter a number of
    ↪ obstacles along the way.
    Output: [ ]

{% endif %}

```

Figura A.1: Plantilla Jinja para el *prompt* de la tarea FSP.

```

## GOAL:
Your goal is to extract all Killing events present in an english text as JSON records.

## GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

# You can only use the information provided in the input to generate your answer.
# Your answer must be one list of JSON records of all extracted Killing events:
  - If N number of events are found, you must answer with one list of N records.
  - If no event is detected, the list must be empty.
  - All keys in the output records must be one of the defined entities, all values in the
    ↪ record must be actual segments of the input text.
# Do NOT reply with any explanation or any other text but one list of records.

```

Figura A.2: *Prompt* para la tarea FSP, escenario sin información, ejemplificado con el marco «Asesinato».

```

## GOAL:
Your goal is to extract all Killing events present in an english text as JSON records.

## Killing EVENT:
A Killing event is defined as:

    A Killer or Cause causes the death of the Victim.

## GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

# You can only use the information provided in the input to generate your answer.
# Your answer must be one list of JSON records of all extracted Killing events:
  - If N number of events are found, you must answer with one list of N records.
  - If no event is detected, the list must be empty.
  - All keys in the output records must be one of the defined entities, all values in the
    ↪ record must be actual segments of the input text.
# Do NOT reply with any explanation or any other text but one list of records.

```

Figura A.3: *Prompt* para la tarea FSP, escenario solo definición, ejemplificado con el marco «Asesinato».

```

## GOAL:
Your goal is to extract all Killing events present in an english text as JSON records.

## Killing EVENT:
A Killing event is defined as:

    A Killer or Cause causes the death of the Victim.

For example, in the following input:

    John drowned Martha.

this Killing event is found:

    {"killer": "John", "trigger": "drowned", "victim": "Martha"}

### ENTITIES:
The entities you might find on the text for Killing events are:

- trigger: The verb, noun or adjective that marks the event of violence.
- killer: The person or sentient entity that causes the death of the Victim.
- victim: The living entity that dies as a result of the killing.
- cause: An inanimate entity or process that causes the death of the Victim.
- means: The method or action that the Killer or Cause performs resulting in the death
    ↪ of the Victim.
- instrument: The device used by the Killer to bring about the death of the Victim.

## GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

# You can only use the information provided in the input to generate your answer.
# Your answer must be one list of JSON records of all extracted Killing events:
- If N number of events are found, you must answer with one list of N records.
- If no event is detected, the list must be empty.
- All keys in the output records must be one of the defined entities, all values in the
    ↪ record must be actual segments of the input text.
# Do NOT reply with any explanation or any other text but one list of records.

```

Figura A.4: *Prompt* para la tarea FSP, escenario *zero-shot*, ejemplificado con el marco «Asesinato».

```

## GOAL:
Your goal is to extract all Killing events present in an english text as JSON records.

## Killing EVENT:
A Killing event is defined as:

    A Killer or Cause causes the death of the Victim.

For example, in the following input:

    John drowned Martha.

this Killing event is found:

    {"killer": "John", "trigger": "drowned", "victim": "Martha"}

### ENTITIES:
The entities you might find on the text for Killing events are:

    - trigger: The verb, noun or adjective that marks the event of violence.
    - killer: The person or sentient entity that causes the death of the Victim.
    - victim: The living entity that dies as a result of the killing.
    - cause: An inanimate entity or process that causes the death of the Victim.
    - means: The method or action that the Killer or Cause performs resulting in the death
    ↪ of the Victim.
    - instrument: The device used by the Killer to bring about the death of the Victim.

## GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

# You can only use the information provided in the input to generate your answer.
# Your answer must be one list of JSON records of all extracted Killing events:
    - If N number of events are found, you must answer with one list of N records.
    - If no event is detected, the list must be empty.
    - All keys in the output records must be one of the defined entities, all values in the
    ↪ record must be actual segments of the input text.
# Do NOT reply with any explanation or any other text but one list of records.

## EXAMPLES:

### Example 1
Input: "Asked if the death penalty should be brought back for terrorist murderers , a
↪ majority supported its return . "
Output: [{"killer": "terrorist", "trigger": "murderers"}]

### Example 2
Input: "Here we have a possible explanation of why the Cro-Magnons might have
↪ exterminated the Neanderthals : the Neanderthals were very like them , and therefore
↪ constituted a threat . "
Output: [{"killer": "the Cro-Magnons", "trigger": "exterminated", "victim": "the
↪ Neanderthals"}]

### Example 3
Input: Ahab and the crew continue their eventful journey and encounter a number of
↪ obstacles along the way.

```

Figura A.5: *Prompt* para la tarea FSP, escenario 3–*shot*, ejemplificado con el marco «Asesinato».

```

# GOAL:
Your goal is to detect which events, also called frames, are present in a text, by
↪ identifying which is the word that evokes the event or frame, called trigger.

Only the following events are of interest:
{% for prompt_frame in prompt_frames %}
    - {{ prompt_frame.frame.name }}
{% endfor %}

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an
↪ adverb.

{% if render_definitions %}
Below there are the definitions of each frame {% if render_definition_triggers -%} and a
↪ list of triggers for each event of interest, but synonyms of those words might fit the
↪ same{% - endif %}.

{% for prompt_frame in prompt_frames %}
{% if not render_definition_example and not render_definition_triggers %}
## {{ prompt_frame.frame.name }}: {{ prompt_frame.frame.definition | replace("\n", " ")}}

{% else %}
## {{ prompt_frame.frame.name }}
- Definition: {{ prompt_frame.frame.definition | replace("\n", " ")}}
{% if render_definition_triggers %}
- Possible triggers: {% for trigger in prompt_frame.trigger_examples -%}{{ trigger }}{% if
↪ not loop.last %}, {% endif %}{%- endfor %}

{% endif %}
{% if render_definition_example %}
- Text example for this frame with trigger
↪ "{{prompt_frame.frame.definition_example.relations[0].arguments.trigger.text}}": {{
↪ prompt_frame.frame.definition_example.text | tojson }}
{% endif %}

{% endif %}
{% endfor %}
{% endif %}

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of detected pairs (frame, trigger):
    -- If N number of events are found, you must answer with one list of N records.
    -- If no event is detected, the list must be empty.
- Do NOT reply with any explanation or any other text but one list of records.
- The output must be of the form [[frame 1, trigger 1],[frame 2, trigger 2]...[frame N, trigger
↪ N]] formatted as a JSON string.
- Triggers must be substrings of the input.

{% if prompt_examples %}
# EXAMPLES:

{% for example in prompt_examples %}
    ## Example {{ loop.index }}
    Input: {{ example.text | tojson }}
    Output: {{ example.output | tojson }}

{% endfor %}
    ## Example {{ prompt_examples| length +1}}
    Input: Ahab and the crew continue their eventful journey and encounter a number of
    ↪ obstacles along the way.
    Output: [ ]

{% endif %}

```

Figura A.6: Plantilla Jinja para el *prompt* de la tarea FI.

```

# GOAL:
Your goal is to detect which events, also called frames, are present in a text, by
↪ identifying which is the word that evokes the event or frame, called trigger.

Only the following events are of interest:
- Killing
- Robbery
- Rape

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an
↪ adverb.

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of detected pairs (frame, trigger):
  -- If N number of events are found, you must answer with one list of N records.
  -- If no event is detected, the list must be empty.
- Do NOT reply with any explanation or any other text but one list of records.
- The output must of the form [[frame 1, trigger 1],[frame 2, trigger 2]...[frame N, trigger
↪ N]] formatted as a JSON string.
- Triggers must be substrings of the input.

```

Figura A.7: *Prompt* para la tarea FI, escenario sin información, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

```

# GOAL:
Your goal is to detect which events, also called frames, are present in a text, by
↳ identifying which is the word that evokes the event or frame, called trigger.

Only the following events are of interest:
- Killing
- Robbery
- Rape

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an
↳ adverb.

Below there are the definitions of each frame.

## Killing: A Killer or Cause causes the death of the Victim.

## Robbery: These are words describing situations in which a Perpetrator wrongs a Victim by
↳ taking something (Goods) from them. In some cases, the Source location from which the
↳ Goods were taken is used to metonymically refer to the Victim. The robbery can be done
↳ in a particular Manner (for example, forcefully) and via a specific Means (for example,
↳ by threatening the Victim). Victim and on the Source location from which the Goods have
↳ been removed, rather than the Goods themselves.

## Rape: The words in this frame describe situations in which the Perpetrator has sexual
↳ intercourse with the Victim forcibly (or by threatening the use of force) and without
↳ his or her consent.

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of detected pairs (frame, trigger):
  -- If N number of events are found, you must answer with one list of N records.
  -- If no event is detected, the list must be empty.
- Do NOT reply with any explanation or any other text but one list of records.
- The output must be of the form [[frame 1, trigger 1],[frame 2, trigger 2]...[frame N, trigger
↳ N]] formatted as a JSON string.
- Triggers must be substrings of the input.

```

Figura A.8: *Prompt* para la tarea FI, escenario solo definición, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

```

# GOAL:
Your goal is to detect which events, also called frames, are present in a text, by
↳ identifying which is the word that evokes the event or frame, called trigger.

Only the following events are of interest:
- Killing
- Robbery
- Rape

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an
↳ adverb.

Below there are the definitions of each frame and a list of triggers for each event of
↳ interest, but synonyms of those words might fit the same.

## Killing
- Definition: A Killer or Cause causes the death of the Victim.
- Possible triggers: holocaust, carnage, silence, murder, annihilation, take (someone's)
↳ life, suicide, fratricide, slay, butcher, take out, asphyxiate, kill, blood-bath,
↳ garrotte, fatality, smother, massacre, assassin, slayer, infanticide, drown, decapitate,
↳ deadly, shooting, liquidate, destroy, decapitation, homicide, crucify, butchery,
↳ slaughterer, assassinate, assassination, pogrom, behead, lethal, crucifixion, regicide,
↳ starve, extermination, euthanasia, fatal, matricide, dispatch, immolation, murderer,
↳ lynch, slaughter, killing, slaying, annihilate, exterminate, genocide, bloodshed,
↳ killer, beheading
- Text example for this frame with trigger "drowned": "John drowned Martha."

## Robbery
- Definition: These are words describing situations in which a Perpetrator wrongs a Victim
↳ by taking something (Goods) from them. In some cases, the Source location from which
↳ the Goods were taken is used to metonymically refer to the Victim. The robbery can be
↳ done in a particular Manner (for example, forcefully) and via a specific Means (for
↳ example, by threatening the Victim). Victim and on the Source location from which the
↳ Goods have been removed, rather than the Goods themselves.
- Possible triggers: mugger, ransack, robber, mugging, rob, stick-up, hold-up, robbery, mug
- Text example for this frame with trigger "held": "Half of the team held up the tellers
↳ while the others were stationed to report the arrival of police."

## Rape
- Definition: The words in this frame describe situations in which the Perpetrator has
↳ sexual intercourse with the Victim forcibly (or by threatening the use of force) and
↳ without his or her consent.
- Possible triggers: rape, rapist, raped
- Text example for this frame with trigger "raped": "Does the press make it more difficult
↳ for raped women to obtain justice?"

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of detected pairs (frame, trigger):
  -- If N number of events are found, you must answer with one list of N records.
  -- If no event is detected, the list must be empty.
- Do NOT reply with any explanation or any other text but one list of records.
- The output must be of the form [[frame 1, trigger 1],[frame 2, trigger 2]...[frame N, trigger
↳ N]] formatted as a JSON string.
- Triggers must be substrings of the input.

```

Figura A.9: *Prompt* para la tarea FI, escenario *zero-shot*, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

```

# GOAL:
Your goal is to detect which events, also called frames, are present in a text, by identifying which is the word that
↳ evokes the event or frame, called trigger.

Only the following events are of interest:
- Killing
- Robbery
- Rape

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an adverb.

Below there are the definitions of each frame and a list of triggers for each event of interest, but synonyms of those
↳ words might fit the same.

## Killing
- Definition: A Killer or Cause causes the death of the Victim.
- Possible triggers: holocaust, carnage, silence, murder, annihilation, take (someone's) life, suicide, fratricide, slay,
↳ butcher, take out, asphyxiate, kill, blood-bath, garrote, fatality, smother, massacre, assassin, slayer, infanticide,
↳ drown, decapitate, deadly, shooting, liquidate, destroy, decapitation, homicide, crucify, butchery, slaughterer,
↳ assassinate, assassination, pogrom, behead, lethal, crucifixion, regicide, starve, extermination, euthanasia, fatal,
↳ matricide, dispatch, immolation, murderer, lynch, slaughter, killing, slaying, annihilate, exterminate, genocide,
↳ bloodshed, killer, beheading
- Text example for this frame with trigger "drowned": "John drowned Martha."

## Robbery
- Definition: These are words describing situations in which a Perpetrator wrongs a Victim by taking something (Goods)
↳ from them. In some cases, the Source location from which the Goods were taken is used to metonymically refer to the
↳ Victim. The robbery can be done in a particular Manner (for example, forcefully) and via a specific Means (for
↳ example, by threatening the Victim). Victim and on the Source location from which the Goods have been removed, rather
↳ than the Goods themselves.
- Possible triggers: mugger, ransack, robber, mugging, rob, stick-up, hold-up, robbery, mug
- Text example for this frame with trigger "held": "Half of the team held up the tellers while the others were stationed
↳ to report the arrival of police."

## Rape
- Definition: The words in this frame describe situations in which the Perpetrator has sexual intercourse with the Victim
↳ forcibly (or by threatening the use of force) and without his or her consent.
- Possible triggers: rape, rapist, raped
- Text example for this frame with trigger "raped": "Does the press make it more difficult for raped women to obtain
↳ justice?"

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of detected pairs (frame, trigger):
  -- If N number of events are found, you must answer with one list of N records.
  -- If no event is detected, the list must be empty.
- Do NOT reply with any explanation or any other text but one list of records.
- The output must of the form [[frame 1, trigger 1],[frame 2, trigger 2]...[frame N, trigger N]] formatted as a JSON
↳ string.
- Triggers must be substrings of the input.

# EXAMPLES:

## Example 1
Input: "During my researches in the library of the Royal Geographical Society I learnt that in 1875 an Egyptian army
↳ commanded by Munzinger , a Swiss mercenary who had served under Gordon in Equatoria , had set out to invade
↳ Abyssinia from Tajura , and been exterminated by the Danakil before they reached Aussa . "
Output: [["Killing", "exterminated"]]

## Example 2
Input: "A woman staying at a local hotel returned to find her room ransacked and another motorist had her car broken
↳ into . Trial of police chief -- Legal decision on status of foetus -- Massacre by Sydney gunman"
Output: [["Robbery", "ransacked"], ["Killing", "Massacre"]]

## Example 3
Input: Ahab and the crew continue their eventful journey and encounter a number of obstacles along the way.
Output: [ ]

```

Figura A.10: *Prompt* para la tarea FI, escenario 3-shot, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

```

# GOAL:
Your goal is to extract the arguments of the events present in a given text.
You will get a list of words, the triggers, that evoke the events or frame, and for each
↪ trigger you must output the appropriate arguments.

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an
↪ adverb; arguments are usually propositions.

{% if render_definitions %}
Below there are the definitions of each frame {% if render_definition_triggers -%} and its
↪ possible arguments{% - endif %}.

{% for prompt_frame in prompt_frames %}
{% if not render_definition_example and not render_definition_triggers %}
## {{ prompt_frame.frame.name }}: {{ prompt_frame.frame.definition | replace("\n", " ") }}

{% else %}
## {{ prompt_frame.frame.name }}
- Definition: {{ prompt_frame.frame.definition | replace("\n", " ") }}
{% if render_definition_arguments %}
- Arguments: The entities you might find on the text for {{ prompt_frame.frame.name }}
↪ events are:
{% for entity in prompt_frame.prompt_entities %}
-- {{ entity.name }}: {{ entity.definition }}
{% endfor %}
{% endif %}
{% if render_definition_example %}
- Example for "{{ prompt_frame.frame.name }}" frame with trigger
↪ "{{ prompt_frame.frame.definition_example.relations[0].arguments.trigger.text }}":
Text: {{ prompt_frame.frame.definition_example.text | tojson }}
Arguments: {{ prompt_frame.frame.definition_example | tojson }}
{% endif %}

{% endif %}
{% endfor %}
{% endif %}
# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You will receive an input text and a list of detected pairs of frames and triggers.
- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of extracted arguments:
-- If N is the number of events, you must answer with one list of N records.
-- Each record is of the form {name_1: value_1,...}, where "name_N" is the name of the
↪ argument (trigger, instrument, ...) and "value_N", its value. The trigger should
↪ always be present.
-- If no event is detected, the output list must be empty.
{% if render_definition_arguments %}
-- All keys in the output records must be defined in its frame's definition.
{% endif %}
-- All values in the record must be actual segments of the input text.
- Do NOT add new events to the output.
- Do NOT add new arguments, not explicitly defined in the frame's definition.
- Do NOT reply with any explanation or any other text but one list of records.

{% if prompt_examples %}
# EXAMPLES:
{% for example in prompt_examples %}
## Example {{ loop.index }}
Text: {{ example.text }}
Triggers: {{ example.triggers | tojson }}
Output: {{ example.output | tojson }}

{% endfor %}
{% endif %}

```

Figura A.11: Plantilla Jinja para el *prompt* de la tarea FSRL.

```

# GOAL:
Your goal is to extract the arguments of the events present in a given text.
You will get a list of words, the triggers, that evoke the events or frame, and for each
↪ trigger you must output the appropriate arguments.

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an
↪ adverb; arguments are usually propositions.

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You will receive an input text and a list of detected pairs of frames and triggers.
- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of extracted arguments:
  -- If N is the number of events, you must answer with one list of N records.
  -- Each record is of the form {name_1: value_1,...}, where "name_N" is the name of the
  ↪ argument (trigger, instrument, ...) and "value_N", its value. The trigger should
  ↪ always be present.
  -- If no event is detected, the output list must be empty.
  -- All values in the record must be actual segments of the input text.
- Do NOT add new events to the output.
- Do NOT add new arguments, not explicitly defined in the frame's definition.
- Do NOT reply with any explanation or any other text but one list of records.

```

Figura A.12: *Prompt* para la tarea FSRL, escenario sin información, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

```

# GOAL:
Your goal is to extract the arguments of the events present in a given text.
You will get a list of words, the triggers, that evoke the events or frame, and for each
↳ trigger you must output the appropriate arguments.

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an
↳ adverb; arguments are usually propositions.

Below there are the definitions of each frame .

## Killing: A Killer or Cause causes the death of the Victim.

## Robbery: These are words describing situations in which a Perpetrator wrongs a Victim by
↳ taking something (Goods) from them. In some cases, the Source location from which the
↳ Goods were taken is used to metonymically refer to the Victim. The robbery can be done
↳ in a particular Manner (for example, forcefully) and via a specific Means (for example,
↳ by threatening the Victim). Victim and on the Source location from which the Goods have
↳ been removed, rather than the Goods themselves.

## Rape: The words in this frame describe situations in which the Perpetrator has sexual
↳ intercourse with the Victim forcibly (or by threatening the use of force) and without
↳ his or her consent.

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You will receive an input text and a list of detected pairs of frames and triggers.
- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of extracted arguments:
  -- If N is the number of events, you must answer with one list of N records.
  -- Each record is of the form {name_1: value_1,...}, where "name_N" is the name of the
  ↳ argument (trigger, instrument, ...) and "value_N", its value. The trigger should
  ↳ always be present.
  -- If no event is detected, the output list must be empty.
  -- All values in the record must be actual segments of the input text.
- Do NOT add new events to the output.
- Do NOT add new arguments, not explicitly defined in the frame's definition.
- Do NOT reply with any explanation or any other text but one list of records.

```

Figura A.13: *Prompt* para la tarea FSRL, escenario solo definición, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

```

# GOAL:
Your goal is to extract the arguments of the events present in a given text.
You will get a list of words, the triggers, that evoke the events or frame, and for each trigger you must output the
↳ appropriate arguments.

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an adverb; arguments are usually
↳ propositions.

Below there are the definitions of each frame and its possible arguments.

## Killing
- Definition: A Killer or Cause causes the death of the Victim.
- Arguments: The entities you might find on the text for Killing events are:
  ↳ -- trigger: The verb, noun or adjective that marks the event of violence.
  ↳ -- killer: The person or sentient entity that causes the death of the Victim.
  ↳ -- victim: The living entity that dies as a result of the killing.
  ↳ -- cause: An inanimate entity or process that causes the death of the Victim.
  ↳ -- means: The method or action that the Killer or Cause performs resulting in the death of the Victim.
  ↳ -- instrument: The device used by the Killer to bring about the death of the Victim.
- Example for "Killing" frame with trigger "drowned":
  Text: "John drowned Martha."
  Arguments: {"killer": "John", "trigger": "drowned", "victim": "Martha"}

## Robbery
- Definition: These are words describing situations in which a Perpetrator wrongs a Victim by taking something (Goods)
↳ from them. In some cases, the Source location from which the Goods were taken is used to metonymically refer to the
↳ Victim. The robbery can be done in a particular Manner (for example, forcefully) and via a specific Means (for
↳ example, by threatening the Victim). Victim and on the Source location from which the Goods have been removed, rather
↳ than the Goods themselves.
- Arguments: The entities you might find on the text for Robbery events are:
  ↳ -- trigger: The verb, noun or adjective that marks the event of violence.
  ↳ -- perpetrator: The Perpetrator is the person (or other agent) that takes the Goods away by force or threats.
  ↳ -- victim: The Victim is the person (or other sentient being or group) that owns the Goods before they are taken away
  ↳ by the Perpetrator.
  ↳ -- source: Source is the initial location of the goods, before they change location.
- Example for "Robbery" frame with trigger "held":
  Text: "Half of the team held up the tellers while the others were stationed to report the arrival of police."
  Arguments: {"perpetrator": "Half of the team", "trigger": "held", "victim": "the tellers"}

## Rape
- Definition: The words in this frame describe situations in which the Perpetrator has sexual intercourse with the Victim
↳ forcibly (or by threatening the use of force) and without his or her consent.
- Arguments: The entities you might find on the text for Rape events are:
  ↳ -- trigger: The verb, noun or adjective that marks the event of violence.
  ↳ -- perpetrator: This is the person who forces another person to have sexual intercourse with him or her against the
  ↳ other person's will.
  ↳ -- victim: This is the person who is forced to engage in sexual intercourse against his or her will.
- Example for "Rape" frame with trigger "raped":
  Text: "Does the press make it more difficult for raped women to obtain justice?"
  Arguments: {"trigger": "raped", "victim": "women"}

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You will receive an input text and a list of detected pairs of frames and triggers.
- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of extracted arguments:
  ↳ -- If N is the number of events, you must answer with one list of N records.
  ↳ -- Each record is of the form {name_1: value_1,...}, where "name_N" is the name of the argument (trigger, instrument,
  ↳ ...) and "value_N", its value. The trigger should always be present.
  ↳ -- If no event is detected, the output list must be empty.
  ↳ -- All keys in the output records must be defined in its frame's definition.
  ↳ -- All values in the record must be actual segments of the input text.
- Do NOT add new events to the output.
- Do NOT add new arguments, not explicitly defined in the frame's definition.
- Do NOT reply with any explanation or any other text but one list of records.

```

Figura A.14: *Prompt* para la tarea FSRL, escenario *zero-shot*, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

```

# GOAL:
Your goal is to extract the arguments of the events present in a given text.
You will get a list of words, the triggers, that evoke the events or frame, and for each trigger you must output the
↳ appropriate arguments.

# EVENTS:
That word that evokes an event could be of any syntax type: a noun, a verb, an adjective, an adverb; arguments are usually
↳ propositions.

Below there are the definitions of each frame and its possible arguments.

## Killing
- Definition: A Killer or Cause causes the death of the Victim.
- Arguments: The entities you might find on the text for Killing events are:
  ↳ -- trigger: The verb, noun or adjective that marks the event of violence.
  ↳ -- killer: The person or sentient entity that causes the death of the Victim.
  ↳ -- victim: The living entity that dies as a result of the killing.
  ↳ -- cause: An inanimate entity or process that causes the death of the Victim.
  ↳ -- means: The method or action that the Killer or Cause performs resulting in the death of the Victim.
  ↳ -- instrument: The device used by the Killer to bring about the death of the Victim.
- Example for "Killing" frame with trigger "drowned":
  Text: "John drowned Martha."
  Arguments: {"killer": "John", "trigger": "drowned", "victim": "Martha"}

## Robbery
- Definition: These are words describing situations in which a Perpetrator wrongs a Victim by taking something (Goods)
↳ from them. In some cases, the Source location from which the Goods were taken is used to metonymically refer to the
↳ Victim. The robbery can be done in a particular Manner (for example, forcefully) and via a specific Means (for
↳ example, by threatening the Victim). Victim and on the Source location from which the Goods have been removed, rather
↳ than the Goods themselves.
- Arguments: The entities you might find on the text for Robbery events are:
  ↳ -- trigger: The verb, noun or adjective that marks the event of violence.
  ↳ -- perpetrator: The Perpetrator is the person (or other agent) that takes the Goods away by force or threats.
  ↳ -- victim: The Victim is the person (or other sentient being or group) that owns the Goods before they are taken away
  ↳ by the Perpetrator.
  ↳ -- source: Source is the initial location of the goods, before they change location.
- Example for "Robbery" frame with trigger "held":
  Text: "Half of the team held up the tellers while the others were stationed to report the arrival of police."
  Arguments: {"perpetrator": "Half of the team", "trigger": "held", "victim": "the tellers"}

## Rape
- Definition: The words in this frame describe situations in which the Perpetrator has sexual intercourse with the Victim
↳ forcibly (or by threatening the use of force) and without his or her consent.
- Arguments: The entities you might find on the text for Rape events are:
  ↳ -- trigger: The verb, noun or adjective that marks the event of violence.
  ↳ -- perpetrator: This is the person who forces another person to have sexual intercourse with him or her against the
  ↳ other person's will.
  ↳ -- victim: This is the person who is forced to engage in sexual intercourse against his or her will.
- Example for "Rape" frame with trigger "raped":
  Text: "Does the press make it more difficult for raped women to obtain justice?"
  Arguments: {"trigger": "raped", "victim": "women"}

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You will receive an input text and a list of detected pairs of frames and triggers.
- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of extracted arguments:
  ↳ -- If N is the number of events, you must answer with one list of N records.
  ↳ -- Each record is of the form {name_1: value_1,...}, where "name_N" is the name of the argument (trigger, instrument,
  ↳ ...) and "value_N", its value. The trigger should always be present.
  ↳ -- If no event is detected, the output list must be empty.
  ↳ -- All keys in the output records must be defined in its frame's definition.
  ↳ -- All values in the record must be actual segments of the input text.
- Do NOT add new events to the output.
- Do NOT add new arguments, not explicitly defined in the frame's definition.
- Do NOT reply with any explanation or any other text but one list of records.

# EXAMPLES:

## Example 1
Text: During my researches in the library of the Royal Geographical Society I learnt that in 1875 an Egyptian army
↳ commanded by Munzinger , a Swiss mercenary who had served under Gordon in Equatoria , had set out to invade
↳ Abyssinia from Tajura , and been exterminated by the Danakil before they reached Aussa .
Triggers: [{"frame": "Killing", "trigger": "exterminated"}]
Output: [{"killer": "by the Danakil", "trigger": "exterminated", "victim": "an Egyptian army commanded by Munzinger ,
↳ a Swiss mercenary who had served under Gordon in Equatoria ,"}]

## Example 2
Text: Rebels entering the capital have found the bodies of dozens of apparently recent victims of violence at
↳ hospitals and military bases .
Triggers: [{"frame": "Violence", "trigger": "violence"}]
Output: [{"trigger": "violence"}]

## Example 3
Text: A woman staying at a local hotel returned to find her room ransacked and another motorist had her car broken
↳ into .
Triggers: [{"frame": "Robbery", "trigger": "ransacked"}]
Output: [{"source": "her room", "trigger": "ransacked"}]

```

Figura A.15: *Prompt* para la tarea FSRL, escenario 3-shot, ejemplificado para los marcos «Asesinato», «Robo» y «Violación».

Apéndice B

Verbos de violencia

Durante el proceso de descubrimiento de los verbos relacionados con actos de violencia que ocurren en la BJN, se decide utilizar la recopilación de verbos del proyecto «Base de datos de verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español»¹ (ADESSE, García-Miguel et al. 2010).

ADESSE contiene 3450 lemas verbales diferentes, basados en un corpus anotado manualmente de 159 000 oraciones. Los verbos tienen asignada una serie de características sintácticas y semánticas, incluyendo la información sobre los roles semánticos de sus argumentos, sus distintas alternancias diatéticas y las realizaciones sintácticas que pueden adoptar. La valencia de un verbo se describe en función de su voz, sus funciones sintácticas y sus roles semánticos.

Al igual que FrameNet, en ADESSE se parte de la premisa de que «*each individual verb evokes a conceptual frame, that is, a complex conceptual representation which includes some basic participants in a scene*»², y se crea una jerarquía cuyo objetivo principal es «*to represent generalizations over these types of conceptual frames evoked by individual verbs*»³ (García-Miguel et al. 2010).

La clasificación de ADESSE es jerárquica, con un primer nivel que consiste en las clases —«Mental», «Relacional», «Material», «Verbal», «Existencial» y «Modulación»—, que se van refinando hasta llegar, en algunas categorías, hasta un cuarto nivel. Por ejemplo, la clase «Mental» se divide en las subclases «Sentimiento», «Percepción» y «Cognición», y esta última en un tercer

¹Disponible en <https://adesse.webs.uvigo.es/>.

²cada verbo individual evoca un marco conceptual, es decir, una representación conceptual compleja que incluye algunos participantes básicos en una escena

³representar generalizaciones sobre estos tipos de marcos conceptuales evocados por verbos individuales

#	Verbo	Categoría	#	Verbo	Categoría
1	agredir	333 – Contacto	11	lesionar	322 – Modificación
2	amenazar	41 – Comunicación	12	matar	53 – Vida
3	apuñalar	322 – Modificación	13	pegar	333 – Contacto
4	asesinar	53 – Vida	14	prostituir	343 – Rel. Sociales
5	atacar	338 – Competición	15	rapiñar	
6	degollar	53 – Vida	16	regentear	343 – Rel. Sociales
7	discutir	41 – Comunicación	17	secuestrar	343 – Rel. Sociales
8	disparar	311 – Desplazamiento	18	sodomizar	
9	estrangular	53 – Vida	19	vilipendiar	
10	golpear	333 – Contacto	20	violar	342 – Fisiología

Tabla B.1: Lista de los 20 verbos de violencia de la lista inicial con su correspondiente categoría en ADESSE.

nivel con las subclases «Conocimiento» y «Creencia», dentro de las cuales se encuentran los verbos «saber» y «creer», respectivamente.

Con el propósito de expandir la lista inicial de verbos de violencia de la BJN, se procede, entonces, a la búsqueda de verbos semánticamente equivalentes, utilizando para esto la clasificación de verbos de ADESSE. Para ello, se identifican en primer lugar las clases semánticas a las que pertenecen una lista inicial de veinte verbos para luego extraer de esas clases a todos aquellos relacionados con eventos de violencia.

Los verbos semilla iniciales se basan en los verbos de violencia más frecuentes en la BJN⁴, a los que se les suma el verbo «disparar», por su relevancia en el contexto de violencia. Para esta etapa se descartan algunos verbos de esa tabla como «desnudar», «fotografiar» o «hurtar» al no considerarlos necesariamente delitos violentos aunque estén ligados a ellos y podrían introducir errores en la expansión de la lista.

En la Tabla B.1 se listan estos verbos junto a su categoría de pertenencia. De esa lista inicial, los verbos «rapiñar», «sodomizar» y «vilipendiar» no se encuentran clasificados en ADESSE. En la Tabla B.2 se presentan las definiciones de las distintas categorías según ADESSE, junto con los verbos extraídos de esas categorías; en negrita se indican los verbos que ya formaban parte de la lista inicial.

⁴La lista de verbos junto a su frecuencia de aparición en la BJN puede encontrarse en la Tabla 4.2

-
- 311 – Desplazamiento:** una entidad (A1) se desplaza desde una localización inicial (A2) hacia una localización final (A3) recorriendo un trayecto (A4).
Verbos: **disparar**.
- 322 – Modificación:** una entidad (A1) experimenta algún tipo de alteración de sus propiedades físicas. Con frecuencia existe otra entidad (A0) que actúa sobre A1 y es la responsable de esta transformación.
Verbos: acuchillar, **apuñalar**, apuñalear, atentar, dañar, desfigurar, despedazar, despellejar, destripar, desvirgar, envenenar, herir, incapacitar, intoxicar, lastimar, **lesionar**, pervertir, yugular, lapidar, linchar.
- 333 – Contacto:** una entidad (A1) entra en contacto físico con otra entidad (A2), sin que ello implique necesariamente la modificación de esta última.
Verbos: abofetear, **agredir**, aporrear, arrollar, atizar, atropellar, azotar, embestir, **golpear**, manosear, morder, patear, **pegar**, pisotear, propinar, trompear, zurrar.
- 338 – Competición:** una entidad (A1) (normalmente un ser humano o un grupo de seres humanos) participa en una actividad cuyo objetivo suele ser imponerse a otro(s) (A2) demostrando mayor fuerza o habilidad.
Verbos: acometer, arremeter, asaltar, **atacar**, contraatacar, contratacar, **discutir**, embestir, forcejear, luchar, pelear.
- 342 – Fisiología:** un animal realiza o experimenta un proceso directamente relacionado con su constitución (psico-)fisiológica.
Verbos: desangrar, deshorrar, **violar**.
- 343 – Relaciones Sociales:** una entidad humana (A1) establece o mantiene de forma activa algún tipo de relación con otra u otras entidades humanas (A2), frecuentemente regulada o valorada socialmente.
Verbos: chantajear, esquilmar, estafar, extorsionar, **prostituir**, raptar, regentar, **regentear**, **secuestrar**, sojuzgar, **violar**.
- 41 – Comunicación:** una entidad (A1) dotada de capacidad comunicativa transfiere información (A2) por medio de cualquier sistema semiótico a otra entidad (A3).
Verbos: **amenazar**, **discutir**, espetar, reñir.
- 53 – Vida:** un ser vivo (prototípicamente un animal y, de forma especial, un ser humano) (A1) permanece en un estado o experimenta un proceso o un cambio de estado directamente relacionado con su existencia biológica.
Verbos: ahorcar, ajusticiar, aniquilar, **asesinar**, asfixiar, **degollar**, electrocutar, **estrangular**, exterminar, fusilar, lapidar, linchar, **matar**.
-

Tabla B.2: Definición de las categorías en ADESSE de los verbos semilla. Se listan los verbos de violencia asociados a cada categoría. Los verbos en negrita pertenecen a la lista inicial; mientras que aquellos subrayados no figuran en el corpus.

#	Lema	Frec.	#	Lema	Frec.	#	Lema	Frec.
1	amenazar	1 173	21	luchar	56	40	exterminar	7
2	violar	813	22	asaltar	49	42	desfigurar	5
3	pegar	693	23	embestir	41		incapacitar	
4	golpear	688	24	atropellar	36		sojuzgar	
5	matar	669	25	acometer	34	45	apuñalar	4
6	discutir	654	26	patear	32		deshonrar	
7	disparar	542	27	ahorcar	21	47	acuchillar	2
8	agredir	481		secuestrar			ajusticiar	
9	lesionar	413	29	asesinar	17		electrocutar	
10	atacar	368		asfixiar			espetar	
11	herir	247	31	morder	16		raptar	
12	forcejear	171	32	estafar	13		regentear	
13	pelear	165	33	degollar	12		reñir	
14	lastimar	155		prostituir		54	abofetear	1
15	rapiñar	128	35	aniquilar	11		chantajear	
16	propinar	126	36	extorsionar	10		envenenar	
17	atentar	115	37	desangrar	8		vilipendiar	
18	dañar	114		estrangular			sodomizar	
19	manosear	83		fusilar			yugular	
20	apuñalar	73	40	arrollar	7			

Tabla B.3: Lista de los 59 verbos de violencia ordenados según su conteo en la BJN.

La expansión de la lista inicial de verbos de violencia a partir de las categorías de ADESSE da como resultado un total de 78 lemas únicos —contando los tres que no se encuentran categorizados—. A su vez, para conservar únicamente aquellos verbos que efectivamente figuran en corpus de trabajo, esta lista se valida contra 5000 documentos de la BJN, eliminando 19 verbos, los cuales se subrayan en la Tabla B.2.

El resultado es una lista depurada de 59 verbos de violencia, presentada en la Tabla B.3.

Apéndice C

Marco de Amenaza

En este apéndice se presenta el marco «Amenaza» (*Threaten*), definido para la anotación de eventos de violencia relacionados con amenazas. Este tipo de eventos resulta de interés en el análisis de la BJN debido a su frecuencia de aparición, como se observa en los resultados de la Tabla 4.5.

En FrameNet, los eventos de amenaza están comprendidos en el marco «Compromiso»¹ (*Commitment*), el cual abarca no solo eventos de amenazas, sino también aquellos relacionados con promesas, juramentos y votos. Por ejemplo, entre las unidades léxicas asociadas a este marco se encuentran verbos como «prometer» (*promise*), «jurar» (*swear*) y «comprometerse» (*undertake*), o los sustantivos como «promesa» (*pledge*), «voto» (*vow*) y «juramento» (*oath*).

Para la aplicación de aprendizaje en contexto, esto último representa un problema, debido a que tanto la descripción como los ejemplos en la definición del marco, así como los posibles *shots* extraídos de las unidades léxicas asociadas, introducen ruido en la información de contexto proporcionada a los modelos.

Además, en la BJN los eventos de amenaza presentan frecuentemente, como elemento complementario, un arma —un cuchillo, una pistola, un corte carcelario², etc.—; sin embargo, el marco «Compromiso» no contempla la presencia de este tipo de elementos.

(C.1) Amenazaba^{amenaza} a María^{dest.} con “molerle la cabeza a
palos”^{mensaje} .-

211/2013 Definitiva,
Tribunal de Apelaciones Penal 1º Tº

¹Disponible en <https://framenet.icsi.berkeley.edu/frames/Commitment>.

²Arma blanca de fabricación artesanal con materiales obtenidos en el centro de reclusión.

(C.2) Con un arma de juguete^{instr.} amenazaron^{amenaza} a los
empleados^{dest.}, se llevaron \$ 12.000 y fugaron en moto.

248/2013 Definitiva,
Tribunal de Apelaciones Penal 1º Tº

(C.3) Durante los hechos el compañero del procesado^{hablante}
amenazaba^{amenaza} a la víctima^{dest.} con el revólver^{instr.}
diciéndole que la iba a matar^{mensaje}, sintiendo que engatillaba el
arma.

102/2013 Definitiva,
Tribunal de Apelaciones Penal 4º Tº

Se observa que tanto el instrumento como el mensaje son opcionales o pueden ocurrir simultáneamente, como sucede en los ejemplos C.1, C.2 y C.3.

En consecuencia, se decide crear, a partir de «Compromiso», un nuevo marco «Amenaza» (*Threaten*) que se adapte a las necesidades del análisis en la BJN. Al marco de referencia se le agrega, entonces, un nuevo elemento denominado «Instrumento» (*Instrument*), el cual permite capturar la presencia de armas u otros dispositivos utilizados para llevar a cabo o reforzar una amenaza, y se eliminan todas las unidades léxicas que no se encuentren relacionadas con eventos de amenaza, dejando únicamente el sustantivo «amenaza» (*threat.n*) y el verbo «amenazar» (*threaten.v*).

El resultado se encuentra en la Figura C.1, donde, además de la definición, se muestran los elementos del marco que fueron efectivamente utilizados en el etiquetado sobre la BJN. Esta definición, al igual que la de los otros marcos, se encuentra en inglés, y se utiliza para la creación de *prompts* durante la experimentación. La Figura C.2 presenta su traducción al español.

Threaten

A **Speaker** makes a threat to an **Addressee** to carry out some future action. The action is not desirable to the **Addressee**.

You **threatened** **me** **to kill my dog** **with a knife**.

Frame Elements

- **Speaker** The Speaker is the person who threatens to do something. Speaker is usually the External Argument of predicative uses of the target or the Genitive modifier of a noun. **Jo** made a **threat** to me. **Jo** **threatens** me she would tell my wife. **Jo's** **threat** to kill my dog came to nothing.
- **Addressee** The Speaker's threat can be made to an Addressee. With those words which allow this frame element to be expressed, Addressee usually occurs as a PP Complement headed by to or the (Indirect) Object of verbs: Kim made a **threat** **to me**. Kim **threatened** **me** with expulsion.
- **Message** An expression of the threat made by the Speaker expresses the frame element Message. Message is expressed as a finite or non-finite clausal Complement or an NP Object. I **threatened** **that I would kill all of them**. They **threatened** **revenge**.
- **Instrument** A device used to enforce a threat. They **threatened** all the present journalists **with a gun**.

Frame-frame Relations

- Inherits from: Commitment

Lexical Units

- threat.n: a statement of an intention to inflict injury, damage, or other hostile action as retribution.
- threaten.v: make or express a threat to (someone) or to do (something).

Figura C.1: Fragmento de la definición del marco «Amenaza» en inglés, utilizada como insumo para la creación de *prompts*.

Amenaza

Un **Hablante** realiza una amenaza a un **Destinatario** de llevar a cabo una acción futura. La acción no es deseable para el **Destinatario**.

Tú **me** **amenazaste** **con matar a mi perro** **con un cuchillo**.

Elementos del Marco

- **Hablante** El Hablante es la persona que realiza la amenaza. Suele corresponder al argumento externo en usos predicativos del núcleo léxico o al modificador genitivo de un sustantivo. **Jo** me hizo una **amenaza**. **Jo** **amenazó** con decirle a mi esposa. La **amenaza** de **Jo** de matar a mi perro no tuvo consecuencias.
- **Destinatario** La amenaza puede dirigirse a un Destinatario. Cuando este elemento puede expresarse, suele aparecer como complemento preposicional encabezado por *a* o como objeto indirecto del verbo. Kim hizo una **amenaza** **contra mí**. Kim **me** **amenazó** con la expulsión.
- **Mensaje** La expresión de la amenaza realizada por el Hablante corresponde al elemento Mensaje. El Mensaje se puede expresar mediante un complemento oracional finito o no finito, o como un sintagma nominal objeto. Yo **amenacé** **con matarlos a todos ellos**. Ellos **amenazaron** **con venganza**.
- **Instrumento** Un dispositivo utilizado para llevar a cabo la amenaza. Ellos **amenazaron** a todos los periodistas presentes **con un arma**.

Relaciones marco-marco

- Hereda de: Compromiso

Unidades Léxicas

- **amenaza.n**: enunciado que expresa la intención de infligir daño, perjuicio u otra acción hostil como represalia.
- **amenazar.v**: realizar o expresar una amenaza hacia alguien o de realizar alguna acción.

Figura C.2: Fragmento en español del marco «Amenaza», basado en su definición original en inglés.

Apéndice D

Desidentificación de ViolenciaR

Para el uso de grandes modelos de lenguaje en el análisis de textos a través de servicios en Internet, es crucial garantizar la privacidad y confidencialidad de la información sensible contenida en los textos, especialmente cuando se trata de datos protegidos, como lo son los textos de la Base de Jurisprudencia Nacional.

En consecuencia, a los textos del corpus ViolenciaR se les aplica un proceso de desidentificación para anonimizar los datos y proteger la identidad de los individuos involucrados. En lugar de seguir los lineamientos de la Acordada 7564 de la Suprema Corte de Justicia (2006), se opta por un enfoque de sustitución de nombres de personas y organizaciones, fechas, ubicaciones y otros identificadores únicos con etiquetas de fantasía.¹ Este método permite mantener la estructura y el contexto del texto original, facilitando el análisis posterior sin comprometer la privacidad de los participantes.

El proceso de desidentificación se lleva a cabo de forma manual, revisando uno por uno los ejemplos del corpus y reemplazando los elementos identificatorios. Estos textos se encuentran divididos por categorías —nombres, direcciones, fechas, etc.— y su sustitución se define mediante un mapa entre los textos originales en una sentencia y los textos de fantasía correspondientes. En la Tabla D.1 se presenta la lista de sustitutos utilizados para la desidentificación en las distintas categorías.

Así, en los ejemplos D.1 y D.2 se muestra el antes y el después de aplicar el proceso de desidentificación a una sentencia del corpus ViolenciaR; los datos cambiados se encuentran resaltados en ambos textos. Se observa que el nivel de

¹Por el detalle del proceso de desidentificación utilizado en la BJN se remite al lector a la Sección 2.1.

Nombres masculinos: Juan, Pedro, José, Pablo, Santiago, Nicolás, Luis, Jorge, Carlos, Daniel, Matías, Facundo, Federico, Lucas, Martín.

Nombres femeninos: María, Lucía, Ana, Sofía, Camila, Florencia, Natalia, Valentina.

Nombres de familia: Rodríguez, Fernández, Hernández.

Alias: X, Y, Z.

Organizaciones: ACME, Lumon, Soylent, LexCorp.

Lugares: Macondo, Comala, Santa María.

Calles: A, B, C, D, E, F, G, H, J, K, L.

Números de puerta: 9876, 9999, 8888, 7777.

Fechas: 30 de febrero, 31 de febrero, 31 de abril, 31 de junio, 31 de septiembre, 31 de noviembre.

Años: 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939.

Números telefónicos: 0303456, 0353456, 5303456, 5353456.

Placas de vehículos: ZZZ 0000.

Números: 999, 888, 777, 666.

Tabla D.1: Lista de sustitutos de fantasía utilizados para la desidentificación según su categoría.

desidentificación es más alto que el aplicado en la publicación de esta sentencia en la BJA, en donde solo se modifican los nombres de las personas involucradas. Por otro lado, las siglas dobles, que reemplazan indistintamente nombres de pila, apellidos o nombres completos, en ViolenciaR se sustituyen siempre por nombres de pila. Las fechas, por su parte, se reemplazan por fechas imposibles, sin referencia al año.

- (D.1) El día 16 de agosto de 2012, en horas de la tarde, AA junto a tres individuos más, portando armas de fuego sustrajeron y se apoderaron de la cartera que llevaba consigo la denunciante BB quien describió a la persona que la amenazó con el arma como AA con quien, en arriesgada acción, forcejeó; ...
- (D.2) El día 30 de febrero, en horas de la tarde, Juan junto a tres individuos más, portando armas de fuego sustrajeron y se apoderaron de la cartera que llevaba consigo la denunciante María quien describió a la persona que la amenazó con el arma como Juan con quien, en arriesgada acción, forcejeó; ...

Sentencia 339/2013 Definitiva,
Tribunal de Apelaciones Penal de 2º Tº

En los ejemplos D.3 y D.4 se presenta otro caso de desidentificación, en donde se observa la sustitución de fechas, nombres, compañías, líneas de transporte, calles y lugares.

- (D.3) El 27/5/13 alrededor de la 21 y 35 hs., en Padre Juan Bonmesadri entre Andrés y Camino Colman, AA, se subió al ómnibus de CUTCSA, línea 147 con destino Aduana y amenazó al conductor con una navaja que portaba, exigiéndole todo el dinero, apoderándose de la suma de \$ 4.405 y se dio a la fuga.
- (D.4) El 31 de junio alrededor de la 21 y 35 hs., en D entre C y Camino E, Juan, se subió al ómnibus de Lumon, línea 888 con destino Santa María y amenazó al conductor con una navaja que portaba, exigiéndole todo el dinero, apoderándose de la suma de \$ 4.405 y se dio a la fuga.

SEF 0010-000032/2014 Definitiva,
Tribunal de Apelaciones de Familia de 1º Tº

Finalmente, en los ejemplos D.5 y D.6 se observa la sustitución de un alias y, en particular, el uso de un apellido genérico en lugar de un nombre de pila.

Esto último se debe a que el testigo utiliza el apellido en un sentido tipificador, funcionando como un estereotipo familiar que permite inferir determinadas conductas (¿violentas?) a partir de la pertenencia al grupo. Para su publicación en la BJN, en cambio, se opta por el uso de las siglas dobles como un nombre cualquiera, lo que fija esta referencia al individuo «BB» particular (cuyo alias es «Bebe»).

- (D.5) d) LES TIENE CIERTO TEMOR POR SER CONOCIDOS EN ESE MEDIO COMO PENDENCIEROS y sospechó que el “Bebe” fuera a pelear por lo que se le pregunta “¿Por qué sospechó que iba a pegar?” y respondió tajantemente “Y ES BB COMO QUIEN DICE...” lo que explica ese lógico temor a represalias y desacredita toda una versión construida un año después a fuerza del ánimo de defensa.
- (D.6) d) LES TIENE CIERTO TEMOR POR SER CONOCIDOS EN ESE MEDIO COMO PENDENCIEROS y sospechó que el “X” fuera a pelear por lo que se le pregunta “¿Por qué sospechó que iba a pegar?” y respondió tajantemente “Y ES RODRÍGUEZ COMO QUIEN DICE...” lo que explica ese lógico temor a represalias y desacredita toda una versión construida un año después a fuerza del ánimo de defensa.

261/2013 Definitiva,
Tribunal de Apelaciones Penal de 2º Tº

Apéndice E

Prompt para FSP en ViolenciaR

A continuación se presentan los *prompts* utilizados para la experimentación en el corpus ViolenciaR, mediante aprendizaje en contexto aplicado al análisis de marcos semánticos, como se detalla en el Capítulo 4.

La tarea que se aborda busca, en una única llamada al modelo, extraer los marcos de violencia presentes, junto con sus objetivos y sus argumentos, en textos provenientes del corpus ViolenciaR. Esta tarea se corresponde con la presentada en el Capítulo 3 como análisis de marcos semánticos (FSP).

Sin embargo, los ejemplos a utilizar como *shots* provienen tanto de ejemplos anotados en FrameNet como de ejemplos anotados en ViolenciaR. Durante la experimentación, se mantiene fija la cantidad total de ejemplos —500 *shots*—, pero se varía la proporción de ejemplos provenientes de cada fuente para analizar su impacto en el desempeño del modelo. La proporción se aplica cada cinco ejemplos, de forma que las dos fuentes se encuentren distribuidas uniformemente a lo largo de la secuencia de ejemplos.

En la Figura E.1 se presenta la plantilla Jinja utilizada para generar los *prompts* para esta tarea. Las figuras E.2 y E.3 muestran los *prompts* generados para el caso *zero-shot* y la sección de ejemplos para el caso *10-shot*, respectivamente. En el caso del *zero-shot*, por una cuestión de espacio, se incluyen únicamente los marcos «Robo», «Violación» y «Amenaza». Se hace notar que, para el caso *10-shot* del Ejemplo E.3, la proporción utilizada es de 20–80 % de ejemplos BJN–FrameNet, lo que resulta en que los *shots* 5 y 10 provengan de ViolenciaR, y el resto de FrameNet.

```

# GOAL:
Your goal is to extract the frames (or events) present in a given text according to the Frame Semantic Theory following
↳ FrameNet syntax.
You will get an input text written in Spanish and your output should be a list of all frames found in the text as JSON
↳ records.
You should only consider these frames and no others:
{% for prompt_frame in prompt_frames -%}
- {{ prompt_frame.frame.name }}
{% endfor %}

# FRAMES:
That word that evokes a frame could be of any syntax type: a noun, a verb, an adjective, an adverb; arguments are usually
↳ propositions.

{% if render_definitions %}
Below there are the definitions of each frame {% if render_definition_triggers -%} and its possible arguments{%- endif%}.

{% for prompt_frame in prompt_frames %}
{% if not render_definition_example and not render_definition_triggers %}
## {{ prompt_frame.frame.name }}: {{ prompt_frame.frame.definition | replace("\n", " ")}}

{% else %}
## {{ prompt_frame.frame.name }}
- Definition: {{ prompt_frame.frame.definition | replace("\n", " ")}}
{% if render_definition_arguments %}
- Arguments: The entities you might find on the text for {{ prompt_frame.frame.name }} events are:
{% for entity in prompt_frame.prompt_entities %}
-- {{ entity.name }}: {{ entity.definition }}
{% endfor %}
{% endif %}
{% if render_definition_example and prompt_frame.frame.definition_example %}
- Example for "{{ prompt_frame.frame.name }}" frame with target
↳ "{{prompt_frame.frame.definition_example.relations[0].arguments.trigger.text}}":
Text: {{ prompt_frame.frame.definition_example.text | safe }}
Arguments: {{ prompt_frame.frame.definition_example | safe }}
{% endif %}

{% endif %}
{% endfor %}
{% endif %}

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You will receive an input text written in Spanish.
- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of extracted arguments:
  -- If N is the number of events, you must answer with one list of N records.
  -- Each record is of the form {name_1: value_1,...}, where "name_N" is the name of the argument (target, instrument,
  ↳ ...) and "value_N", its value. The target should always be present.
  -- If no event is detected, the output list must be empty.
  {% if render_definition_arguments %}
  -- All keys in the output records must be defined in its frame's definition.
  {% endif %}
  -- All values in the record must be actual segments of the input text.
  -- In Spanish, a clitic at the end of the verb is an unstressed pronoun that is placed after and attached to the verb,
  ↳ forming a single word and being called enclitic. You might consider the verb and the enclitic as two separate
  ↳ words.
- Do NOT add new arguments, not explicitly defined in the frame's definition.
- Do NOT reply with any explanation or any other text but one list of records.

{% if prompt_examples %}
# EXAMPLES:
Although your input is in Spanish, the following are mixed examples from Spanish and English texts:

{% for example in prompt_examples %}
## Example {{ loop.index }}
Text: {{ example.text|safe }}
Output: {{example.output|safe }}

{% endfor %}
{% endif %}

```

Figura E.1: Plantilla Jinja para el *prompt* de la tarea FSP sobre el corpus ViolenciaR.

```

# GOAL:
Your goal is to extract the frames (or events) present in a given text according to the Frame Semantic Theory following
↳ FrameNet syntax.
You will get an input text written in Spanish and your output should be a list of all frames found in the text as JSON
↳ records.
You should only consider these frames and no others:
- Robbery
- Rape
- Threaten

# FRAMES:
That word that evokes a frame could be of any syntax type: a noun, a verb, an adjective, an adverb; arguments are usually
↳ propositions.

Below there are the definitions of each frame and its possible arguments.

## Robbery
- Definition: These are words describing situations in which a Perpetrator wrongs a Victim by taking something (Goods)
↳ from them. In some cases, the Source location from which the Goods were taken is used to metonymically refer to the
↳ Victim. The robbery can be done in a particular Manner (for example, forcefully) and via a specific Means (for
↳ example, by threatening the Victim). Victim and on the Source location from which the Goods have been removed, rather
↳ than the Goods themselves.
- Arguments: The entities you might find on the text for Robbery events are:
  -- trigger: The verb, noun or adjective that marks the event of violence.
  -- perpetrator: The Perpetrator is the person (or other agent) that takes the Goods away by force or threats.
  -- victim: The Victim is the person (or other sentient being or group) that owns the Goods before they are taken away
  ↳ by the Perpetrator.
  -- goods: The Goods are anything (including labor, time, or legal rights) that can be taken away.
  -- source: Source is the initial location of the goods, before they change location.
- Example for "Robbery" frame with target "held":
  Text: Half of the team held up the tellers while the others were stationed to report the arrival of police.
  Arguments: {"trigger": "held", "perpetrator": "Half of the team", "victim": "the tellers"}

## Rape
- Definition: The words in this frame describe situations in which the Perpetrator has sexual intercourse with the Victim
↳ forcibly (or by threatening the use of force) and without his or her consent.
- Arguments: The entities you might find on the text for Rape events are:
  -- trigger: The verb, noun or adjective that marks the event of violence.
  -- perpetrator: This is the person who forces another person to have sexual intercourse with him or her against the
  ↳ other person's will.
  -- victim: This is the person who is forced to engage in sexual intercourse against his or her will.
- Example for "Rape" frame with target "raped":
  Text: Does the press make it more difficult for raped women to obtain justice?
  Arguments: {"trigger": "raped", "victim": "women"}

## Threaten
- Definition: A Speaker makes a threat to an Addressee to carry out some future action. The action is not desirable to the
↳ Addressee.
- Arguments: The entities you might find on the text for Threaten events are:
  -- trigger: The verb, noun or adjective that marks the event of violence.
  -- speaker: The Speaker is the person who threatens to do something. Speaker is usually the External Argument of
  ↳ predicative uses of the target or the Genitive modifier of a noun: Jo made a threat to me. Jo threatens me she
  ↳ would tell my wife. Jo's threat to kill my dog came to nothing.
  -- addressee: The Speaker's threat can be made to an Addressee. With those words which allow this frame element to be
  ↳ expressed, Addressee usually occurs as a PP Complement headed by to or the (Indirect) Object of verbs: Kim made a
  ↳ threat to me. Kim threatened me with expulsion.
  -- message: An expression of the threat made by the Speaker expresses the frame element Message. Message is expressed
  ↳ as a finite or non-finite clausal Complement or an NP Object: I threatened that I would kill all of them. They
  ↳ threatened revenge.
  -- instrument: A device used to enforce a threat. They threatened all the present journalists with a gun.
- Example for "Threaten" frame with target "threatened":
  Text: You threatened me to kill my dog with a knife.
  Arguments: {"trigger": "threatened", "speaker": "You", "addressee": "me", "message": "to kill my dog", "instrument":
  ↳ "with a knife"}

# GUIDELINES:
These are the guidelines for your task. These rules are strict and must be complied:

- You will receive an input text written in Spanish.
- You can only use the information provided in the input to generate your answer.
- Your answer must be one list of extracted arguments:
  -- If N is the number of events, you must answer with one list of N records.
  -- Each record is of the form {name_1: value_1,...}, where "name_N" is the name of the argument (target, instrument,
  ↳ ...) and "value_N", its value. The target should always be present.
  -- If no event is detected, the output list must be empty.
  -- All keys in the output records must be defined in its frame's definition.
  -- All values in the record must be actual segments of the input text.
  -- In Spanish, a clitic at the end of the verb is an unstressed pronoun that is placed after and attached to the verb,
  ↳ forming a single word and being called enclitic. You might consider the verb and the enclitic as two separate
  ↳ words.
- Do NOT add new arguments, not explicitly defined in the frame's definition.
- Do NOT reply with any explanation or any other text but one list of records.

```

Figura E.2: *Prompt* para la tarea FSP sobre el corpus ViolenciaR, escenario *zero-shot*, ejemplificado para los marcos «Robo», «Violación» y «Amenaza».

```

# EXAMPLES:
Although your input is in Spanish, the following are mixed examples from Spanish and English
↪ texts:

## Example 1
Text: "The blacks had just seen a film , ` Mississippi Burning \" , in which Klansmen
↪ beat a black child to death . "
Output: [{"agent": "Klansmen", "victim": "a black child", "trigger": "beat"}]

## Example 2
Text: "Not only had he robbed that particular S&L just three days earlier , but the
↪ teller at the window was the same person who had handed him the cash at gunpoint . "
Output: [{"perpetrator": "he", "source": "that particular S&L", "trigger": "robbed"}]

## Example 3
Text: "Vernage then stabbed him again and Sergeant King was dead on arrival at Hospital
↪ . "
Output: [{"agent": "Vernage", "victim": "him", "trigger": "stabbed"}]

## Example 4
Text: "A post mortem showed 80year-old Mr Goult , known as ` Old Jack \" , had been
↪ stabbed many times in the chest at his home in Woolfall Heath Avenue , Huyton ,
↪ Liverpool . "
Output: [{"victim": "80year-old Mr Goult , known as ` Old Jack", "body_part": "in the
↪ chest", "trigger": "stabbed"}]

## Example 5
Text: "Estima, que si se permite una lectura de los distintos escritos presentados por
↪ la Defensora, con matices, hay una idéntica redacción en escritos presentados por
↪ actores y defensora, y no puede pasar por desapercibida por el Tribunal por violar
↪ los artículos 194, 195 y 196 del Código Penal."
Output: []

## Example 6
Text: "The prosecution say that following an armed robbery at a building society near
↪ Gloucester in 1988 White had been due to give evidence against Gardiner and
↪ implicating him in the crime . "
Output: [{"trigger": "robbery"}]

## Example 7
Text: "When the clumsy plotting of the Queen of Scots against Elizabeth finally
↪ exasperated the English beyond endurance , Mary was beheaded at Fotheringhay . "
Output: [{"victim": "Mary", "trigger": "beheaded"}]

## Example 8
Text: "A while later he gets up , tenderly feeling the bump on his head and shrugging
↪ off the dust , and continues onto his friend 's house , the whole time wondering
↪ what the devil that big darn thing was that nearly killed him ."
Output: [{"killer": "that", "victim": "him", "trigger": "killed"}]

## Example 9
Text: "Sheriff Higgins said yesterday when Arthur appeared for sentence that witnesses
↪ said the accused had threatened earlier that he was going to hurt his victim . "
Output: [{"agent": "he", "victim": "his victim", "trigger": "hurt"}]

## Example 10
Text: "Tras amenazar a las empleadas, detonar un disparo y exigirles que se tiraran al
↪ suelo, se opoderaron de cinco equipos deportivos y cuatro remeras para hombre."
Output: [{"addressee": "a las empleadas", "trigger": "amenazar"}, {"projectile": "un
↪ disparo", "trigger": "detonar"}, {"goods": "de cinco equipos deportivos y cuatro
↪ remeras para hombre", "perpetrator": "se", "trigger": "opoderaron"}]

```

Figura E.3: Sección «Ejemplos» para la tarea FSP sobre el corpus ViolenciaR, escenario *n-shot*, para $n = 10$, con un distribución 20–80 % de ejemplos BJN–FrameNet.

ANEXOS

Anexo I

Marcos de FrameNet

Este anexo reproduce definiciones de FrameNet con fines descriptivos y de referencia, utilizadas en el desarrollo de la tesis, tanto en los experimentos realizados exclusivamente con FrameNet, descritos en el Capítulo 3, como en aquellos aplicados con textos de la BJN en el corpus ViolenciaR, presentados en el Capítulo 4. Se incluye, además, el marco «Compromiso», que, aunque finalmente no se utiliza en la extracción de eventos, constituye el punto de partida para la definición del marco «Amenaza», creado específicamente para el análisis de eventos de amenaza.

La información presentada en este anexo es parcial: se incluyen únicamente los marcos que son relevantes para la tesis y, dentro de ellos, solo los elementos extraídos en los experimentos. Los colores para los elementos de cada marco respetan, en la medida de lo posible, los colores originales de FrameNet. Las versiones completas de los marcos pueden consultarse en el sitio web de FrameNet¹.

Se presentan a continuación los marcos en inglés, junto con su traducción al español:

- Abuso (*Abusing*): figuras I.1 y I.2
- Arma de fuego (*Use Firearm*): figuras I.3 y I.4
- Asesinato (*Killing*): figuras I.5 y I.6
- Compromiso (*Commitment*): figuras I.7 y I.8
- Daño (*Cause Harm*): figuras I.9 y I.10

¹Los marcos se encuentran disponibles en: <https://framenet.icsi.berkeley.edu/frames/>

- Disparo (*Shoot Projectiles*): figuras [I.11](#) y [I.12](#)
- Impacto (*Hit Target*): figuras [I.13](#) y [I.14](#)
- Robo (*Robbery*): figuras [I.15](#) y [I.16](#)
- Violación (*Rape*): figuras [I.17](#) y [I.18](#)
- Violencia (*Violence*): figuras [I.19](#) y [I.20](#)

Cabe señalar que en todos los experimentos estas definiciones se incorporan a los *prompts* exclusivamente en inglés. La traducción al español se agrega con fines de completitud y referencia.

Abusing

In this frame an **Abuser** repeatedly treats the **Victim** in a cruel and violent way, including physically harming or forcing the **Victim** to engage in sexual activity against their will. The **Victim** usually lacks the power to resist or fight back. A Type of abuse may be indicated. There may also be a Degree.

Her father verbally **abused** **her mother**.

Frame Elements

- **Abuser**: The Abuser subjects the Victim to repeated verbal, physical, emotional or sexual attacks. **Fred** **abused** Martha by locking her into a closet.
- **Victim**: The Victim is the recipient of the Abuser's actions. **Fred** **abused** **Martha** by locking her into a closet.

Frame-frame Relations

- Child of: Committing Crime
- Using: Cause Harm

Lexical Units

- abuse.n: cruel and violent treatment
- abuse.v: to cruelly mistreat, esp. physically or sexually
- abusive.a: characterized by cruel and violent treatment
- batter.v: subject (one's spouse, partner, or child) to repeated violence and assault
- domestic_violence.n: violence that occurs inside a family
- maltreat.v: treat cruelly or with violence
- maltreatment.n: cruel, improper actions taken upon someone

Figura I.1: Fragmento de la definición del marco «Abuso» (*Abusing*) en inglés.

Abuso

En este marco, un **Agresor** somete de forma reiterada a la **Víctima** a un trato cruel y violento, que incluye el daño físico o la coacción para que la **Víctima** participe en actividades sexuales contra su voluntad. La **Víctima** suele carecer del poder necesario para resistirse o defenderse. Puede indicarse un Tipo de abuso, así como también un Grado.

Su padre abusaba verbalmente de **su madre**.

Elementos del Marco

- **Agresor**: El Agresor somete a la Víctima a ataques verbales, físicos, emocionales o sexuales de manera reiterada. **Fred** abusó de Martha encerrándola en un armario.
- **Víctima**: La Víctima es quien recibe las acciones del Agresor. Fred abusó de **Martha** encerrándola en un armario.

Relaciones marco-marco

- Hijo de: Comisión de delito
- Usa: Causar daño

Unidades Léxicas

- abuso.n: trato cruel y violento
- abusar.v: tratar de manera cruel, especialmente de forma física o sexual
- abusivo.a: caracterizado por un trato cruel y violento
- golpear.v: someter (a un cónyuge, pareja o hijo) a violencia y agresiones reiteradas
- violencia_doméstica.n: violencia que ocurre dentro del ámbito familiar
- maltratar.v: tratar con crueldad o violencia
- maltrato.n: acciones crueles o indebidas ejercidas sobre una persona

Figura I.2: Fragmento de la definición del marco «Abuso» (*Abusing*) en español.

Use Firearm

An **Agent** causes a **Firearm** to discharge, usually directing the projectile from the barrel of the **Firearm** (the Source), along a Path, and to a **Goal**.

The Germans wouldn't stop **shooting** **cannons** at them.

Oswald **fired** twice through the window.

Frame Elements

- **Agent**: The self-directed entity that uses and activates the Firearm.
- **Firearm**: This FE identifies the Firearm that an Agent activates so that it rapidly propels a projectile.
- **Goal**: Goal is the location the Theme ends up in, or is intended to end up in. Who **fired** a pistol **at my client**?

Lexical Units

- discharge.v: fire (a gun or missile)
- fire.v: discharge (a gun or other weapon)
- shoot.v: cause (a gun) to fire
- shot.n: the discharge of a firearm

Figura I.3: Fragmento de la definición del marco «Arma de fuego» (*Use Firearm*) en inglés.

Arma de fuego

Un **Agente** provoca la descarga de un **Arma**, generalmente dirigiendo el proyectil desde el cañón del **Arma** (el origen), a lo largo de una Trayectoria y hacia un **Meta**.

Los alemanes no dejaban de **disparar** **cañones** contra ellos.

Oswald **disparó** dos veces a través de la ventana.

Elementos del Marco

- **Agente**: La entidad dotada de voluntad que utiliza y activa el arma de fuego.
- **Arma**: Este elemento identifica el arma de fuego que el Agente activa para propulsar rápidamente un proyectil.
- **Meta**: La Meta es la ubicación en la que el proyectil termina o se pretende que termine. ¿Quién **disparó** una pistola **contra mi cliente**?

Unidades Léxicas

- disparar.v: efectuar el disparo de un arma de fuego
- accionar.v: activar un arma de fuego
- tirar.v: hacer que un arma dispare
- disparo.n: descarga de un arma de fuego

Figura I.4: Fragmento de la definición del marco «Arma de fuego» (*Use Firearm*) en español.

Killing

A Killer or Cause causes the death of the Victim.

John drowned Martha.

Frame Elements

- Cause: An inanimate entity or process that causes the death of the Victim. The rockslide killed nearly half of the climbers.
- Instrument: The device used by the Killer to bring about the death of the Victim. It's difficult to suicide with only a pocketknife.
- Killer: The person or sentient entity that causes the death of the Victim.
- Means: The method or action that the Killer or Cause performs resulting in the death of the Victim. The flood exterminated the rats by cutting off access to food.
- Victim: The living entity that dies as a result of the killing.

Lexical Units

- annihilate.v: to kill, esp. a large proportion of a group
- annihilation.n: killing, esp. of people
- :
- take (someone's) life.idio: to kill someone
- take_out.v: kill
- terminate.v: to murder (used as a euphemism)

Figura I.5: Fragmento de la definición del marco «Asesinato» (*Killing*) en inglés.

Asesinato

Un **Asesino** o una **Causa** provocan la muerte de la **Víctima**.

John ahogó a **Martha**.

Elementos del Marco

- **Causa**: Una entidad inanimada o un proceso que provoca la muerte de la Víctima. El deslizamiento de rocas **mató** a casi la mitad de los escaladores.
- **Instrumento**: El dispositivo utilizado por el Asesino para provocar la muerte de la Víctima. Es difícil **suicidarse** solo con una navaja.
- **Asesino**: La persona o entidad dotada de agencia que provoca la muerte de la Víctima.
- **Medio**: El método o la acción que el Asesino o la Causa realizan y que resulta en la muerte de la Víctima. La inundación **exterminó** a las ratas al cortar el acceso a los alimentos.
- **Víctima**: La entidad viva que muere como resultado del acto.

Unidades Léxicas

- aniquilar.v: matar, especialmente a una gran proporción de un grupo
- aniquilación.n: matanza, especialmente de personas
- :
- quitar la vida (a alguien).idio: matar a alguien
- eliminar.v: matar
- terminar.v: asesinar (usado como eufemismo)

Figura I.6: Fragmento de la definición del marco «Asesinato» (*Killing*) en español.

Commitment

A **Speaker** makes a commitment to an **Addressee** to carry out some future action. This may be an action desirable (as with promise) or not desirable (as with threaten) to the **Addressee**. Some of the words in this frame allow an **Addressee** to be expressed.

You **promised** **me** **you'd come to my graduation**

Frame Elements

- **Addressee**: The Speaker's commitment can be made to an Addressee. With those words which allow this frame element to be expressed, Addressee usually occurs as a PP Complement headed by to or the (Indirect) Object of verbs: Kim made a **promise** **to me**. Kim **promised** **me** that everything would be ok.
- **Message**: An expression of the commitment made by the Speaker expresses the frame element Message. Message is expressed as a finite or non-finite clausal Complement or an NP Object. I **swore** **hat I would never make the same mistake again**. The owner finally **consented** **to sell the business**. They **threatened** **revenge**.
- **Speaker**: The Speaker is the person who commits him/herself to do something. Speaker is usually the External Argument of predicative uses of the target or the Genitive modifier of a noun: **Jo** made a **promise** to me. **Jo** **promised** me she would cook dinner. Jo's **promise** to cook dinner came to nothing.

Lexical Units

- **commit.v**: to obligate or pledge oneself
- **commitment.n**: a pledge or undertaking
- **:**
- **volunteer.v**: freely offer to do something
- **vow.n**: a solemn promise
- **vow.v**: solemnly promise to do something

Figura I.7: Fragmento de la definición del marco «Compromiso» (*Commitment*) en inglés.

Compromiso

Un **Hablante** asume un compromiso con un **Destinatario** de llevar a cabo una acción futura. Dicha acción puede ser deseable (como en el caso de una promesa) o no deseable (como en el caso de una amenaza) para el **Destinatario**. Algunas de las unidades léxicas de este marco permiten que el **Destinatario** se exprese explícitamente.

Tú **me** **prometiste** **que vendrías a mi graduación**

Elementos del Marco

- **Destinatario**: El compromiso del Hablante puede realizarse hacia un Destinatario. En aquellas unidades léxicas que permiten expresar este elemento, el Destinatario suele aparecer como un complemento preposicional encabezado por *a* o como objeto (indirecto) del verbo: Kim **me** hizo una **promesa**. Kim **me** **prometió** que todo estaría bien.
- **Mensaje**: La expresión del compromiso asumido por el Hablante corresponde al elemento Mensaje. El Mensaje se expresa mediante un complemento oracional finito o no finito, o como un sintagma nominal objeto. Yo **juré** **que nunca volvería a cometer el mismo error**. El propietario finalmente **consintió** **vender el negocio**. Ellos **amenazaron** **con venganza**.
- **Hablante**: El Hablante es la persona que se compromete a realizar una acción. Suele corresponder al argumento externo en usos predicativos del núcleo léxico o al modificador genitivo de un sustantivo: **Jo** me hizo una **promesa**. **Jo** me **prometió** que cocinaría la cena. La **promesa** de **Jo** de cocinar la cena no se cumplió.

Unidades Léxicas

- comprometerse.v: obligarse o prometerse a uno mismo realizar algo
- compromiso.n: promesa u obligación asumida
- :
- ofrecerse.v: ofrecerse voluntariamente a hacer algo
- voto.n: promesa solemne
- jurar.v: prometer solemnemente hacer algo

Figura I.8: Fragmento de la definición del marco «Compromiso» (*Commitment*) en español.

Cause Harm

The words in this frame describe situations in which an **Agent** or a **Cause** injures a **Victim**. The **Body_part** of the **Victim** which is most directly affected may also be mentioned in the place of the **Victim**.

Rod saw his father **beating** **his mother**.

Frame Elements

- **Agent**: Agent is the person causing the Victim's injury. **Jolosa, who broke** a rival player's jaw, was told to model his play on the England striker.
- **Body_part**: The **Body_part** identifies the location on the body where the bodily injury takes place. Someone **bashed** him **on the back of the head** with a heavy smooth object.
- **Cause**: The **Cause** marks expressions that indicate some non-intentional, typically non-human, force that inflicts harm on the Victim. **A falling rock** **crushed** my ankle.
- **Instrument**: This FE is any object used to cause the injury. Someone **bashed** him on the back of the head **with a heavy smooth object**.
- **Victim**: The Victim is the being or entity that is injured. Someone **bashed** **him** on the back of the head with a heavy smooth object.

Lexical Units

- bash.v: strike hard and violently
- batter.v: strike repeatedly with hard blows
- :
- welt.v: strike hard and heavily
- whip.v: beat with a whip
- wound.v: inflict a wound on

Figura I.9: Fragmento de la definición del marco «Daño» (*Cause Harm*) en inglés.

Daño

Las palabras de este marco describen situaciones en las que un **Agente** o una **Causa** lesionan a una **Víctima**. También puede mencionarse la **Parte_del_cuerpo** de la **Víctima** que resulta más directamente afectada, en lugar de la propia **Víctima**.

Rod vio a su padre **golpeando** a **su madre**.

Elementos del Marco

- **Agente**: El Agente es la persona que causa las lesiones de la Víctima. **Jolosa, quien** le **rompió** la mandíbula a un jugador rival, fue instruido para modelar su juego según el delantero inglés.
- **Parte_del_cuerpo**: La Parte_del_cuerpo identifica la localización del cuerpo donde se produce la lesión. Alguien lo **golpeó** en la parte posterior de la cabeza con un objeto pesado y liso.
- **Causa**: La Causa señala expresiones que indican una fuerza no intencional, típicamente no humana, que inflige daño a la Víctima. **Una roca que caía** **aplastó** mi tobillo.
- **Instrumento**: Este elemento corresponde a cualquier objeto utilizado para causar la lesión. Alguien lo **golpeó** en la parte posterior de la cabeza **con un objeto pesado y liso**.
- **Víctima**: La Víctima es el ser o la entidad que resulta lesionada. Alguien **lo** **golpeó** en la parte posterior de la cabeza con un objeto pesado y liso.

Unidades Léxicas

- golpear.v: pegar fuerte y violentamente
- apalea.v: golpear reiteradamente con golpes fuertes
- :
- azotar.v: golpear con fuerza
- flagelar.v: golpear con un látigo
- herir.v: causar una herida

Figura I.10: Fragmento de la definición del marco «Daño» (*Cause Harm*) en español.

Shoot Projectiles

An **Agent** causes a **Firearm** to discharge which induces the rapid motion of the **Projectile**.

The bank robber shot at the clerks with an AK-47.

Shooting a cannonball isn't as easy as it looks.

Note that usages with the **Firearm** as object are out of frame, belonging to "Use Firearm".

Frame Elements

- **Agent**: The Agent is the person performing the intentional act causing the motion of the Projectile.
- **Firearm**: The device used by the Agent to cause the motion of the Projectile.
- **Goal**: This FE is the point at which the Theme ends up as a result of the motion. Kim threw the cat into the garden.
- **Projectile**: The entity whose rapid, dangerous motion is caused by the Agent and/or Firearm.

Lexical Units

- fire.off.v: discharge, shoot, launch (a gun, rocket, etc)
- fire.n: projectiles fired as or from a weapon
- fire.v: propel (a bullet or projectile) from a gun or other weapon
- launch.n: an act or instance of launching something
- launch.v: to shoot projectiles from a gun or other weapon
- shoot.v: cause an object to be projected forcefully from a device (esp. a weapon, e.g., gun, bow). "He shot the bullet into the air."

Figura I.11: Fragmento de la definición del marco «Disparo» (*Shoot Projectiles*) en inglés.

Disparo

Un **Agente** provoca la descarga de un **Arma**, lo que induce el movimiento rápido del **Proyectil**.

El ladrón del banco **disparó** **contra los empleados**
con un AK-47.

Disparar **una bala de cañón** no es tan fácil como parece.

Nótese que los usos en los que el **Arma** aparece como objeto quedan fuera de este marco y corresponden a “Arma de fuego”.

Elementos del Marco

- **Agente**: El Agente es la persona que realiza el acto intencional que provoca el movimiento del Proyectil.
- **Arma**: El dispositivo utilizado por el Agente para causar el movimiento del Proyectil.
- **Meta**: Este elemento corresponde al punto en el que el Tema termina como resultado del movimiento. Kim **arrojó** al gato **al jardín**.
- **Proyectil**: La entidad cuyo movimiento rápido y peligroso es provocado por el Agente y/o el Arma.

Unidades Léxicas

- **hacer_fuego.v**: efectuar la descarga o lanzamiento de un proyectil desde un arma
- **fuego.n**: proyectiles disparados como o desde un arma
- **lanzar.v**: disparar o impulsar proyectiles desde un arma u otro dispositivo
- **lanzamiento.n**: acto o instancia de lanzar algo
- **disparar.v**: provocar que un objeto sea proyectado con fuerza desde un dispositivo (especialmente un arma)

Figura I.12: Fragmento de la definición del marco «Disparo» (*Shoot Projectiles*) en español.

Hit Target

An **Agent** causes the ballistic motion of a projectile and its subsequent impact on a **Target**.

It looks like **he** **hit** **the bull's-eye** **with his first arrow**!

If **he** **shot** **her**, he is fully culpable whether he meant to or not.

Frame Elements

- **Agent**: The individual that directs the Projectile such that it impacts the Target.
- **Instrument**: The entity, most often ammunition or a firearm, which the Agent manipulates to hit the Target.
- **Subregion**: The part of the Target hit by the Agent. He **shot** me **in the arm**!
- **Target**: The entity which is hit by a Projectile directed by the Agent.

Lexical Units

- hit.v: strike (a target)
- pick_off.v: shoot
- shoot.v: to fire a gun such that another object is impacted by a bullet

Figura I.13: Fragmento de la definición del marco «Impacto» (*Hit Target*) en inglés.

Impacto

Un **Agente** provoca el movimiento balístico de un proyectil y su impacto posterior sobre una **Meta**.

Parece que **él** **dio** **en el blanco** **con su primera flecha**.

Si **él** **le** **disparó**, es plenamente culpable, independientemente de si fue intencional o no.

Elementos del Marco

- **Agente**: El individuo que dirige el Proyectil de modo tal que impacta en la Meta.
- **Instrumento**: La entidad —con mayor frecuencia munición o un arma de fuego— que el Agente manipula para impactar la Meta.
- **Subregión**: La parte de la Meta que es alcanzada por el Agente. Me **disparó** **en el brazo**.
- **Meta**: La entidad que es impactada por un Proyectil dirigido por el Agente.

Unidades Léxicas

- dar.v: alcanzar o impactar una meta
- abatir.v: disparar y alcanzar una meta
- disparar.v: efectuar un disparo de modo que un objeto resulte impactado por un proyectil

Figura I.14: Fragmento de la definición del marco «Impacto» (*Hit Target*) en español.

Robbery

These are words describing situations in which a **Perpetrator** wrongs a **Victim** by taking something (**Goods**) from them. In some cases, the **Source** location from which the **Goods** were taken is used to metonymically refer to the **Victim**. The robbery can be done in a particular Manner (for example, forcefully) and via a specific Means (for example, by threatening the **Victim**).

Half of the team held up the tellers while the others were stationed to report the arrival of police.

Frame Elements

- **Goods**: The Goods are anything (including labor, time, or legal rights) that can be taken away. Leslie robbed John of his watch.
- **Perpetrator**: The Perpetrator is the person (or other agent) that takes the Goods away by force or threats. Leslie robbed Kim of her watch.
- **Source**: Source is the initial location of the goods, before they change location. Leslie and Kim robbed a bank.
- **Victim**: The Victim is the person (or other sentient being or group) that owns the Goods before they are taken away by the Perpetrator. Leslie robbed Kim of her watch.

Lexical Units

- hold-up.v: rob someone using the threat of violence
- hold-up.n: the act or an instance of forcibly stopping and robbing
- ⋮
- robbery.n: An act of robbing
- stick-up.v: rob someone or something at gunpoint
- stick-up.n: robbery (slang)

Figura I.15: Fragmento de la definición del marco «Robbery» (*Robo*) en inglés.

Robo

Las palabras de este marco describen situaciones en las que un **Perpetrador** perjudica a una **Víctima** al apoderarse de algo (**Bienes**) que le pertenece. En algunos casos, la ubicación **Fuente** desde la cual se sustraen los **Bienes** se utiliza de manera metonímica para referirse a la **Víctima**. El robo puede realizarse de una determinada Manera (por ejemplo, con violencia) y mediante un Medio específico (por ejemplo, amenazando a la **Víctima**).

La mitad del equipo **asaltó** a **los cajeros** mientras los demás se apostaban para avisar de la llegada de la policía.

Elementos del Marco

- **Bienes**: Los Bienes son cualquier cosa (incluido trabajo, tiempo o derechos legales) que puede ser sustraída. Leslie **robó** a John **su reloj**.
- **Perpetrador**: El Perpetrador es la persona (u otro agente) que se apodera de los Bienes mediante el uso de la fuerza o amenazas. **Leslie** **robó** a Kim su reloj.
- **Fuente**: La Fuente es la ubicación inicial de los bienes, antes de que cambien de localización. Leslie y Kim **robaron** **un banco**.
- **Víctima**: La Víctima es la persona (u otro ser sintiente o grupo) que posee los Bienes antes de que el Perpetrador se los quite. Leslie **robó** a **Kim** su reloj.

Unidades Léxicas

- asaltar.v: robar a alguien mediante amenaza o el uso de la violencia
- asalto.n: acto de detener y robar a alguien por la fuerza
- :
- robo.n: acto de robar
- atracar.v: robar a una persona o establecimiento, especialmente a mano armada
- atraco.n: robo (coloquial)

Figura I.16: Fragmento de la definición del marco «Robbery» (*Robo*) en español.

Rape

The words in this frame describe situations in which the **Perpetrator** has sexual intercourse with the **Victim** forcibly (or by threatening the use of force) and without his or her consent.

Does the press make it more difficult for **raped** **women** to obtain justice?

On their third date **he** took her back to his surgery and **raped** **her**.

Frame Elements

- **Perpetrator**: This is the person who forces another person to have sexual intercourse with him or her against the other person's will.
- **Victim**: This is the person who is forced to engage in sexual intercourse against his or her will.

Lexical Units

- rape.n: an act or the crime of raping someone
- rape.v: (of a man) force (another person) to have sexual intercourse with him against their will
- raped.a: having been forced (by a man) to have sexual intercourse with (him)
- rapist.n: a man who commits rape
- sexually_assault.v: force (another person) to have sexual intercourse with against their will

Figura I.17: Fragmento de la definición del marco «Violación» (*Rape*) en inglés.

Violación

Las palabras de este marco describen situaciones en las que el **Perpetrador** mantiene relaciones sexuales con la **Víctima** de manera forzada (o mediante la amenaza del uso de la fuerza) y sin su consentimiento.

¿La prensa dificulta que las **mujeres** **violadas** obtengan justicia?

En su tercera cita, **él** la llevó de vuelta a su consultorio y **la** **violó**.

Elementos del Marco

- **Perpetrador**: Es la persona que obliga a otra persona a mantener relaciones sexuales con él o ella contra la voluntad de esta última.
- **Víctima**: Es la persona que es forzada a mantener relaciones sexuales contra su voluntad.

Unidades Léxicas

- violación.n: acto o delito de violar a una persona
- violar.v: forzar a otra persona a mantener relaciones sexuales contra su voluntad
- violado.a: que ha sido forzado a mantener relaciones sexuales
- violador.n: persona que comete una violación
- agredir sexualmente.v: forzar a otra persona a mantener relaciones sexuales contra su voluntad

Figura I.18: Fragmento de la definición del marco «Violación» (*Rape*) en español.

Violence

The words in this frame describe acts (or situations characterized by acts) that cause injury or harm. The acts may involve an **Aggressor** or **Cause** injuring a **Victim**, or **Aggressors** causing each other harm.

Research on whether concealed carry laws increase gun violence among gang members has been inconclusive.

They committed brutality against innocent civilians.

The storm's violence last week came without warning and caused unprecedented damage to coastal regions.

Frame Elements

- **Aggressor**: The Aggressor is the person causing the Victim's injury. Your brutality on your enemies seems to know no boundaries.
- **Aggressors**: The Aggressors commit acts of violence upon each other. Hope for a peaceful resolution is gone as violence between the two sides reignites.
- **Cause**: The Cause marks expressions that indicate some non-intentional, typically non-human, force that inflicts harm on the Victim. The storm's violence left many lives in shambles.
- **Victim**: The Victim is the being or entity that is injured. Increasing violence against supporters of government reform have attracted international attention.

Lexical Units

- brutality.n: cruel act of harm
- savagery.n: an act of cruelty
- violence.n: an act that causes harm or injury upon another

Figura I.19: Fragmento de la definición del marco «Violencia» (*Violence*) en inglés.

Violencia

Las palabras de este marco describen actos (o situaciones caracterizadas por actos) que causan daño o lesiones. Los actos pueden involucrar a un Agresor o a una Causa que lesionan a una Víctima, o bien a Agresores que se causan daño entre sí.

Las investigaciones sobre si las leyes de porte oculto de armas aumentan la violencia armada entre miembros de pandillas no han sido concluyentes.

Ellos cometieron brutalidades contra civiles inocentes.

La violencia de la tormenta de la semana pasada llegó sin previo aviso y causó daños sin precedentes en las regiones costeras.

Elementos del Marco

- **Agresor**: El Agresor es la persona que causa las lesiones de la Víctima. Tu brutalidad hacia tus enemigos parece no conocer límites.
- **Agresores**: Los Agresores cometen actos de violencia entre sí. Toda esperanza de una resolución pacífica se desvanece cuando la violencia entre ambas partes se reactiva.
- **Causa**: La Causa señala expresiones que indican una fuerza no intencional, típicamente no humana, que inflige daño a la Víctima. La violencia de la tormenta dejó numerosas vidas devastadas.
- **Víctima**: La Víctima es el ser o la entidad que resulta lesionada. El aumento de la violencia contra partidarios de la reforma gubernamental ha atraído la atención internacional.

Unidades Léxicas

- brutalidad.n: acto cruel que causa daño
- salvajismo.n: acto de crueldad
- violencia.n: acto que causa daño o lesiones a otra persona o entidad

Figura I.20: Fragmento de la definición del marco «Violencia» (*Violence*) en español.