UNIVERSIDAD DE LA REPÚBLICA URUGUAY

FACULTAD DE INGENIERÍA

# Experimentos sobre Transcripción y Traducción Español - Lengua de Señas Uruguaya

Informe de Proyecto de Grado presentado por

Gastón Paiva

en cumplimiento parcial de los requerimientos para la graduación de la carrera de Ingeniería en Computación de Facultad de Ingeniería de la Universidad de la República

Supervisor

Luis Chiruzzo

Montevideo, 11 de septiembre de 2025

# Agradecimientos

Quisiera agradecer a todas las personas que me acompañaron durante toda la carrera. A mis padres, Luis y Nancy, que me apoyaron siempre para lograr mis sueños y ya no están conmigo para ver el fruto de su esfuerzo. A mi hermano Luis , mi cuñada, Gabriela, y mis sobrinos Francisco e Isabella, que siempre me apoyaron y me ayudaron en lo que podían. A mi novia Camila, por estar para ayudarme en el último tramo de la carrera que, sin su apoyo, no hubiera podido terminar este proyecto .

# Resumen

En este proyecto se investiga la viabilidad de construir un sistema de traducción automática que convierta una señal acústica en español uruguayo en una secuencia de glosas correspondientes a la Lengua de Señas Uruguaya (LSU). Las lenguas de señas, al ser lenguajes naturales en modalidad visual-espacial, han recibido históricamente menos atención en el campo del procesamiento de lenguaje natural. En Uruguay, este vacío es aún más pronunciado debido a la escasez de recursos lingüísticos en LSU, lo que dificulta tanto el desarrollo de tecnologías accesibles como la investigación en esta área.

El objetivo principal del trabajo es avanzar en la traducción automática entre el español uruguayo hablado y la LSU. Para ello, se diseñó un sistema dividido en dos componentes fundamentales: primero, un módulo de reconocimiento automático de habla , encargado de transcribir el audio en español uruguayo a texto; y segundo, un módulo de traducción automática que transforma ese texto en una secuencia de glosas, que representan una forma escrita intermedia para expresar LSU.

El proyecto comienza con la recopilación y procesamiento de un corpus de audio local proveniente de emisiones televisivas, que fue transcrito y corregido manualmente para evaluar el rendimiento de distintos sistemas de ASR. Se compararon los resultados de dos herramientas: Google Speech Recognition y Whisper, concluyendo que Whisper ofrece un desempeño significativamente superior para el dominio específico del español uruguayo.

En la etapa de traducción, se utilizaron cuatro conjuntos de datos: dos en Lengua de Señas Española y dos en LSU. Para entrenar los modelos se empleó la herramienta OpenNMT, que implementa técnicas de aprendizaje automático para traducción de secuencia a secuencia. Los resultados se evaluaron utilizando métricas estándar como BLEU (Bilingual Evaluation Understudy) y CHRF (Character F-score).

Como resultado tangible del trabajo, se desarrollaron prototipos funcionales de las dos etapas principales del pipeline: la transcripción automática de audio y la traducción a glosas LSU. Si bien no se implementó un sistema integrado que encadene ambas etapas de forma continua, los componentes desarrollados permiten validar cada parte del proceso de forma independiente y sientan las bases para una integración futura. La generación visual de señas a través de un avatar queda fuera del alcance de este proyecto.

**Palabras clave:** Traducción automática, Lengua de Señas Uruguaya, Reconocimiento automático de habla, Glosas, Aprendizaje automático

# Índice general

# Capítulo 1

# Introducción

La comunicación accesible es un elemento esencial para garantizar la inclusión de todas las personas en la sociedad. En Uruguay, las personas sordas que utilizan la Lengua de Señas Uruguaya (LSU) enfrentan importantes barreras para acceder a contenidos transmitidos en español hablado, particularmente en medios de comunicación y servicios públicos. Esta situación limita sus posibilidades de participación plena en distintos ámbitos sociales, educativos y laborales. Según datos del Censo 2011, en Uruguay hay aproximadamente 120.000 personas con discapacidades auditivas, de las cuales alrededor de 30.000 presentan sordera severa o total (Fondo de Población de las Naciones Unidas (UNFPA), 2011). Sin embargo, actualmente hay menos de 100 intérpretes profesionales registrados en el país, lo que genera una brecha significativa en el acceso a la información y la comunicación inclusiva.

Uno de los principales desafíos para revertir esta situación es la falta de recursos tecnológicos adaptados al contexto uruguayo. Actualmente, las soluciones de traducción entre el español y la LSU son escasas y, en muchos casos, dependen exclusivamente de intérpretes humanos. Aunque existen profesionales capacitados en la interpretación de lengua de señas, su número no alcanza para cubrir la creciente demanda. Esto hace necesario explorar alternativas tecnológicas que permitan complementar, y no reemplazar, el trabajo de los intérpretes, facilitando el acceso automático a la información.

Una de las herramientas que se puede utilizar para revertir esta situación son las glosas. Las glosas son representaciones escritas que describen las señas de un lenguaje de señas mediante palabras o símbolos que representan cada signo o concepto. Funcionan como una forma intermedia entre el lenguaje hablado y el lenguaje de señas, facilitando el trabajo de traducción y análisis, ya que permiten transcribir la información en un formato textual que mantiene la estructura y significado de la LSU.

Un pipeline de punta a punta para la traducción de texto hablado a lengua de señas típicamente comprende varias etapas secuenciales, desde la entrada del audio hasta la generación del video final en lengua de señas. Primero, se realiza la transcripción del audio a texto en el idioma de origen. A continuación,

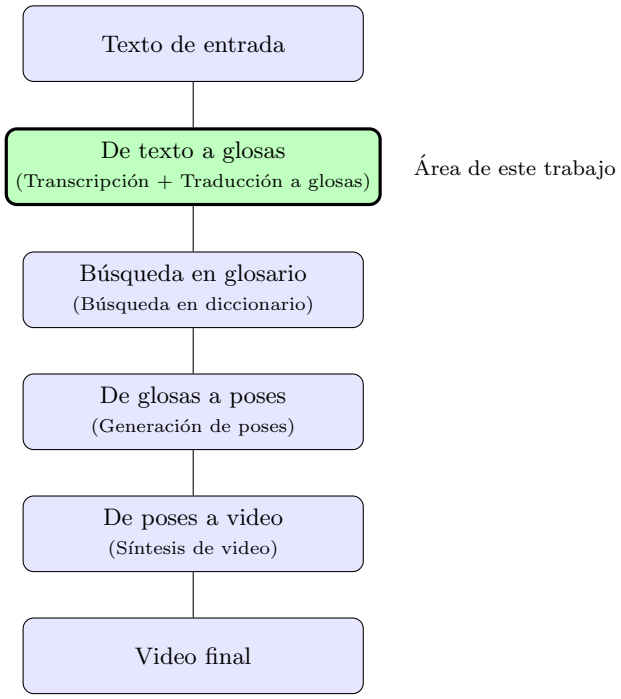


**Figura 1.1:** Pipeline de Moryossef et al.

el texto se convierte en una secuencia de glosas que representan de manera intermedia los signos de la lengua de señas. Estas glosas se transforman en poses corporales mediante un modelo de glosas a poses, que aprende a mapear cada glosa a la posición y movimiento de brazos, manos, cabeza y cuerpo a lo largo del tiempo, generando la secuencia de posturas necesarias para expresar cada signo. Finalmente, las poses se sintetizan en un video continuo que muestra la interpretación en lengua de señas.

En la Figura 1.1 se muestra un flujo de procesamiento típico para la traducción de texto hablado a lengua de señas, tal como lo presenta (Moryossef y cols., 2023). El flujo comienza con el texto de entrada, que puede ser una transcripción de audio, y pasa a la etapa de texto a glosas, donde se convierte la oración en una secuencia de glosas. Posteriormente, la etapa búsqueda en glosario localiza cada glosa en un diccionario de señas, y la etapa de glosas a poses genera una secuencia de posturas corporales correspondientes. Finalmente, la etapa de poses a video sintetiza el video en lengua de señas.

En este trabajo nos focalizamos específicamente en la etapa de texto a glosas, que está resaltada en la figura, desarrollando y evaluando un sistema de transcripción y traducción a glosas a partir de texto en español.

En este contexto, el presente proyecto propone investigar y desarrollar un sistema de traducción automática que procese señales acústicas en español uru-

guayo y las convierta en secuencias de glosas correspondientes en LSU. Para ello, se abordan dos componentes principales de manera independiente: por un lado, la transcripción automática del audio mediante tecnologías de reconocimiento automático de habla (ASR), y por otro, la traducción automática del texto transcrito a glosas utilizando técnicas de procesamiento de lenguaje natural y aprendizaje automático.

## 1.1. Objetivos

El objetivo general es avanzar en la traducción automática entre el español uruguayo y la LSU. Los objetivos específicos incluyen: (1) estudiar el estado del arte en reconocimiento automático de habla (ASR) y traducción automática (MT) aplicados a lenguas de señas, (2) analizar la posibilidad de adaptar o mejorar estos sistemas utilizando datos recopilados en Uruguay, y (3) construir un prototipo funcional que realice la transcripción del habla y la traducción a glosas LSU.

Entre los resultados esperados, se busca realizar un relevamiento técnico actualizado de modelos y herramientas existentes, evaluar su desempeño sobre datos uruguayos y llevar a cabo experimentos que permitan demostrar el funcionamiento de las distintas etapas del pipeline: desde el reconocimiento de audio en español uruguayo hasta su traducción en forma de glosas. Estos experimentos servirán como base para desarrollos más completos en el futuro, incluyendo eventualmente la generación de señas a través de animación visual, que queda fuera del alcance de este proyecto.

Desde una perspectiva científica y técnica, este trabajo representa una contribución pionera al desarrollo de tecnologías lingüísticas accesibles para el contexto uruguayo, al adaptar herramientas existentes al dominio local y explorar nuevos recursos para el tratamiento automático de la LSU.

## 1.2. Estructura del documento

Este documento se organiza de la siguiente manera: en el capítulo 2 se presenta una revisión de antecedentes relevantes; el capítulo 3 describe el proceso de transcripción automática del habla; el capítulo 4 detalla los experimentos y resultados de la traducción a glosas; finalmente, en el capítulo 5 se presentan las conclusiones y líneas de trabajo futuras.

# Capítulo 2

# Revisión de antecedentes

En esta sección se presentan antecedentes relevantes vinculados con el tema de este proyecto. Además, se introducen brevemente los conceptos necesarios para comprender el enfoque adoptado.

## 2.1. Lenguas de Señas

Las lenguas de señas son sistemas de comunicación visual-gestual utilizados principalmente por personas sordas. Estas lenguas poseen gramática, vocabulario y reglas propias, diferentes a las de la lengua oral correspondiente a su región (Johnston, 2010). Al tratarse de lenguas naturales, presentan variación dialectal, estructuras sintácticas complejas y mecanismos lingüísticos específicos del canal visual-espacial. En contextos computacionales, estas particularidades implican desafíos adicionales respecto al procesamiento de lenguas orales, como la falta de sistemas de escritura estandarizados y la escasez de recursos anotados.

### 2.1.1. Lengua de Señas Uruguaya

Este proyecto se centra específicamente en la Lengua de Señas Uruguaya (LSU), la lengua natural de la comunidad sorda en Uruguay. La LSU ha sido poco explorada desde el punto de vista computacional, y los datos disponibles son aún limitados. Entre los recursos existentes, se destaca el dataset LSU-DS (Stassi y cols., 2022), que constituye un primer esfuerzo público para el reconocimiento automático de LSU. Además, se encuentran en desarrollo nuevos corpus y publicaciones que contribuirán a ampliar los recursos disponibles para esta lengua, como el trabajo presentado por (Fojo y cols., 2024), que introduce un corpus de LSU con metadatos organizados para facilitar futuras investigaciones lingüísticas y computacionales .

Por otra parte, en la Figura 2.1 se muestra un ejemplo extraído del corpus Sobreciencia. Este corpus constituye un recurso en construcción que recopila transmisiones televisivas del programa de divulgación científica del mismo nom-

bre, y ha sido anotado manualmente por personas bilingües en español y LSU. En él se transcriben las emisiones orales en español y se alinean con glosas en LSU, lo que permite generar pares paralelos útiles para tareas de traducción automática. Se trata de un proyecto en curso que continúa ampliándose y refinándose, con el fin de proveer datos de alta calidad para la investigación en tecnologías de lengua de señas, y que será utilizado en el presente trabajo como fuente de datos.

**Glosa (LSU)**:
`DOS AÑO ANTES HACER ENCUESTA TEMA ALCOHOL`
`QUINCE AÑO HABER CAMBIAR IMPORTANTE`

**Traducción al español**:
Hicimos una encuesta sobre el consumo de alcohol hace dos años y vimos un punto de quiebre a los 15 años.

**Figura 2.1:** Ejemplo de oración en LSU y su correspondiente traducción al español,extraída del corpus *Sobreciencia*.

Por otra parte, la LSU ha sido objeto de estudios sociolingüísticos y sintácticos que han analizado la variación estructural y los procesos de estandarización, proporcionando antecedentes fundamentales para la comprensión de sus particularidades lingüísticas y culturales (Fojo, González, y Tancredi, 2013; Fojo y Tancredi, 2015).

### 2.1.2. Lengua de Señas Española

Además de recursos en LSU, este trabajo emplea corpora anotados en Lengua de Señas Española (LSE), dado que es una de las lenguas de señas mejor documentadas y con mayor cantidad de datos disponibles en formatos compatibles con tareas de traducción automática. Cabe señalar que la LSU y la LSE no están relacionadas entre sí, a pesar de que ambas conviven con el mismo idioma oral, el español. Por ello, pueden presentar diferencias tanto en el repertorio de señas como en la organización sintáctica de las oraciones. Por ejemplo, un mismo concepto puede representarse con señas distintas en cada lengua, y la estructura de las glosas puede variar en el orden en que se presentan los constituyentes de la oración.

En particular, se utilizó el corpus *id/dl* (San-Segundo y cols., 2008), que contiene diálogos en contextos administrativos como renovaciones de cédula o licencias de conducir, y presenta alineación entre texto y glosas en LSE. También se trabajó con *iSignos* (Cabeza y García-Miguel, 2019), un corpus más extenso que *id/dl*, que incluye segmentos anotados en glosas y ha sido utilizado en investigaciones previas para tareas de traducción.

### 2.1.3. Representaciones Escritas de Lenguas de Señas

Para su representación escrita, una de las principales posibilidades son las glosas, que consisten en transcripciones simplificadas donde se asocia cada seña a una palabra o etiqueta del idioma oral hablado habitualmente en la misma región que la lengua de señas (Stokoe, 2005).

Las glosas no constituyen un sistema de escritura formal ni estandarizado, y su uso presenta limitaciones importantes: se trata de una representación con pérdida, ya que omite información crucial sobre la realización espacial, facial y corporal de las señas, así como aspectos prosódicos y transiciones entre signos. A pesar de ello, su simplicidad y relativa disponibilidad las han convertido en una herramienta ampliamente utilizada en contextos lingüísticos y computacionales, facilitando el análisis textual y el entrenamiento de modelos de traducción automática.

Existen otros sistemas de escritura más expresivos, diseñados para capturar con mayor fidelidad la estructura visual y articulatoria de las lenguas de señas. Entre ellos se destacan HamNoSys (Hamburg Notation System) (Prillwitz, Leven, Zienert, Hanke, y Henning, 1989), un sistema fonético que representa manualidades y rasgos no manuales de las señas, y SignWriting (Sutton, 1995), que permite escribir signos de forma más visual y legible, incluso por usuarios no expertos. Sin embargo, actualmente no existen corpus disponibles en LSU que utilicen estos sistemas de escritura, por lo que quedan fuera del alcance de este trabajo.

En la figura 2.2 se presenta un ejemplo real tomado del corpus *iSignos*, que ilustra las diferencias entre la representación en glosa y su traducción al español. La glosa `YO NACER OYENTE / DOCE AÑO OÍDO-APAGADO` condensa la información mediante etiquetas que omiten artículos, conectores y flexión verbal, reflejando únicamente los elementos clave de la lengua de señas. Por el contrario, la traducción al español "Yo nací oyente, pero a los doce años quedé sordo." incorpora la estructura gramatical completa, incluyendo morfología, conectores y puntuación, adaptándose a las normas del idioma oral.

**Glosa (LSE)**:
`YO NACER OYENTE`
`DOCE AÑO OÍDO-APAGADO`

**Traducción al español**:
Yo nací oyente, pero a los doce años quedé sordo.

**Figura 2.2:** Ejemplo de oración en LSE y su correspondiente traducción al español.

Esta representación intermedia cumple un rol clave en diversos enfoques de traducción automática hacia lenguas de señas, actuando como puente entre la forma textual y la representación visual-gestual.

## 2.2. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural (PLN) es el campo de la inteligencia artificial que estudia la capacidad de las computadoras para procesar y comprender el lenguaje humano, tanto en forma escrita como hablada (Jurafsky y Martin, 2009). Entre las tareas tradicionales del PLN se encuentran el análisis de sentimiento, la traducción automática, el reconocimiento de habla, la transcripción automática y el modelado de lenguaje. Estas tareas constituyen la base de numerosas aplicaciones ampliamente utilizadas en la vida cotidiana, como asistentes virtuales, correctores gramaticales, motores de búsqueda, sistemas de recomendación y plataformas de atención automatizada. También tienen usos relevantes en contextos especializados, como la minería de textos científicos, la automatización del análisis legal o el procesamiento de documentos médicos.

En los últimos años, la irrupción de los grandes modelos de lenguaje (Large Language Models, LLMs) ha transformado profundamente el campo del PLN, ampliando significativamente el rango de tareas que pueden abordarse. A diferencia de los enfoques anteriores, estos modelos son capaces de generalizar conocimientos lingüísticos a gran escala y realizar tareas complejas con un mínimo de ajuste o supervisión específica. Modelos como ChatGPT (Brown y cols., 2020) y LLaMA (Touvron, Martin, Stone, y cols., 2023) han revolucionado el área al demostrar capacidades avanzadas en generación de texto coherente, respuesta a preguntas abiertas, interacción conversacional, razonamiento lógico, resumen automático y traducción multilingüe de alta calidad.

Esta expansión en las capacidades del PLN ha sido posible gracias al uso de redes neuronales profundas, en particular la arquitectura Transformer (Vaswani y cols., 2017), que introdujo mecanismos de atención capaces de modelar relaciones contextuales entre palabras de manera más eficiente que los enfoques recurrentes anteriores. Los Transformers han demostrado ser altamente escalables y efectivos para capturar patrones complejos en secuencias de texto, lo que los convierte en la base de prácticamente todos los LLMs actuales. Su impacto ha sido tal que han redefinido los límites del PLN, impulsando investigaciones interdisciplinarias y aplicaciones novedosas en dominios que hasta hace poco se consideraban fuera del alcance de las técnicas tradicionales.

En el caso de las lenguas de señas, el PLN enfrenta desafíos específicos, ya que se trata de lenguajes visuales y multimodales cuya estructura difiere notablemente de las lenguas orales. Esto requiere adaptar o repensar las técnicas tradicionales del PLN para poder procesar adecuadamente este tipo de lenguajes.

### 2.2.1. Redes Neuronales

Las redes neuronales son una familia de modelos del aprendizaje automático inspirados en la estructura del cerebro humano. Estas redes consisten en funciones parametrizadas que se organizan en capas sucesivas, y que pueden aproximar relaciones complejas entre entradas y salidas a partir de datos de entrenamiento (Goodfellow, Bengio, y Courville, 2016). Durante el proceso de entrenamiento,

los parámetros de la red se ajustan automáticamente para minimizar el error en las predicciones mediante técnicas de optimización, como el descenso por el gradiente.

Existen múltiples arquitecturas de redes neuronales, cada una con características particulares según el tipo de dato o tarea que se desea abordar. Entre las más utilizadas en procesamiento de lenguaje natural se encuentran las redes neuronales recurrentes (RNN), que permiten modelar secuencias de datos, como texto o audio, al incorporar información contextual mediante conexiones entre pasos temporales. Una variante ampliamente empleada de las RNN son las redes LSTM (Long Short-Term Memory) (Hochreiter y Schmidhuber, 1997), que fueron diseñadas para mitigar problemas como el desvanecimiento del gradiente y conservar información relevante a largo plazo. Este tipo de arquitectura será retomada más adelante en el contexto del sistema de traducción OpenNMT.

### 2.2.2. Arquitectura Transformer

Otra arquitectura que domina actualmente muchos de los métodos en procesamiento de lenguaje natural es el modelo Transformer (Vaswani y cols., 2017). A diferencia de las redes recurrentes, el Transformer no procesa las secuencias de forma secuencial, sino que opera en paralelo sobre todas las posiciones de entrada, lo que mejora significativamente la eficiencia computacional y facilita el entrenamiento en grandes volúmenes de datos.

El Transformer se basa en tres componentes fundamentales: representaciones vectoriales, mecanismos de autoatención y una arquitectura general de tipo encoder-decoder. Primero, las palabras o tokens de entrada se transforman en vectores de dimensión fija mediante embeddings. Luego, el mecanismo de autoatención permite que cada posición de la secuencia considere simultáneamente el contexto completo, ponderando dinámicamente la relevancia de otras posiciones al generar una representación contextualizada. Finalmente, la arquitectura se compone de un codificador (encoder), que procesa la secuencia de entrada, y un decodificador (decoder), que genera la salida de forma autoregresiva, prestando atención tanto a los estados del encoder como a los tokens previamente generados.

Esta arquitectura ha demostrado un rendimiento sobresaliente en una amplia variedad de tareas, y es la base de la mayoría de los modelos de lenguaje actuales, incluidos los LLMs como GPT, BERT, LLaMA y T5.

### 2.2.3. Transcripción Automática de Audio

Este proyecto aborda la transcripción automática de audio extraído de videos en español uruguayo y la traducción automática a secuencias de glosas en Lengua de Señas Uruguaya (LSU) como se mencionó en la sección 1.1. En este caso, la entrada es audio extraído de videos y la salida es su representación en glosas, que facilita su posterior interpretación o traducción al formato visual-gestual de la LSU. Este enfoque contempla las particularidades lingüísticas y

dialectales del español uruguayo, lo que presenta desafíos específicos en tareas de reconocimiento automático de voz.

La transcripción automática es el proceso mediante el cual se convierte lenguaje hablado en texto escrito, y constituye una tarea fundamental dentro del procesamiento automático del lenguaje. Sus aplicaciones incluyen desde el subtitulado automático de videos y la generación de contenido accesible, hasta la indexación de archivos audiovisuales o la transcripción de entrevistas y conferencias.

Entre los desarrollos más destacados en esta área se encuentra Whisper, un modelo de transcripción desarrollado por OpenAI que ha demostrado alta precisión en la conversión de audio a texto en múltiples idiomas (Radford y cols., 2023). Whisper está basado en la arquitectura Transformer, la misma que ha revolucionado múltiples tareas de procesamiento de lenguaje natural, y fue entrenado con grandes volúmenes de datos multilingües, lo que lo hace robusto frente a variaciones de acento, ruido de fondo y otros desafíos frecuentes en señales de audio reales.

Además, en este trabajo también se empleó la API de reconocimiento de voz de Google, accesible mediante la librería `SpeechRecognition` de Python (Anthony Zhang et al., 2025), para la transcripción de audio en español uruguayo. La API de Google forma parte del servicio Cloud Speech-to-Text, que permite convertir audio en texto en múltiples idiomas y variantes. Aunque los detalles completos de su arquitectura no han sido divulgados públicamente, se sabe que utiliza modelos de aprendizaje automático entrenados sobre grandes volúmenes de datos de audio y texto, con el objetivo de lograr una alta precisión en tareas de reconocimiento de voz. Estas herramientas permiten obtener transcripciones automáticas sin necesidad de entrenar modelos desde cero. Además de las palabras transcriptas, una tarea complementaria en el procesamiento de audio es la estimación de los turnos de los hablantes, conocida como diarización. Este procedimiento consiste en segmentar la señal en función de los distintos interlocutores, permitiendo identificar sus turnos de habla y conservar la estructura conversacional. Cabe destacar que Whisper no incorpora esta funcionalidad, por lo que en este trabajo la diarización se realizó completamente de forma manual.

En cuanto a la transcripción automática del habla en español, se ha trabajado históricamente en la adaptación de modelos a diferentes variedades dialectales. Por ejemplo, en (Caballero, Mariño, y Moreno, 2002) presentaron un enfoque de modelado multidialectal para el reconocimiento automático del español, incluyendo variantes regionales como la argentina. Estos antecedentes resultan relevantes dado que las diferencias dialectales impactan directamente en la precisión de los modelos de reconocimiento y traducción, y la LSU también presenta particularidades regionales que deben ser consideradas. Más allá de este trabajo para la variedad argentina, que es muy similar a la uruguaya, no se encontraron antecedentes particularmente para la transcripción de la variedad de español uruguaya.

En este trabajo, la comparación de los resultados de la transcripción automática se realiza mediante las siguientes métricas (Morris, Maier, y Green,

2004):

- **WER (Word Error Rate)**: El WER es una de las métricas más utilizadas para medir el rendimiento de un sistema de reconocimiento de voz o traducción automática. Mide el porcentaje de palabras que deben insertarse, eliminarse o reemplazarse para que la predicción coincida con la referencia. Cuanto más bajo sea el valor de WER, mejor será el rendimiento del sistema, siendo 0 el valor óptimo que indica una coincidencia perfecta. A continuación, se presenta la fórmula que define formalmente el WER:

$$\text{WER} = \frac{S + D + I}{N} \tag{2.1}$$

  donde:

  - $S$ = número de sustituciones
  - $D$ = número de eliminaciones
  - $I$ = número de inserciones
  - $N$ = número total de palabras en la referencia

- **MER (Match Error Rate)**: El MER es una métrica que se centra en cuántas palabras coinciden correctamente en una traducción en comparación con una referencia. A diferencia del WER, que se enfoca en los errores en relación con el total de palabras de la referencia, el MER pone el foco en las coincidencias correctas. Es decir, mide la proporción de tokens que realmente coinciden entre la predicción y la referencia, considerando tanto los errores como los aciertos. Esta perspectiva permite evaluar de manera más directa qué tan bien el sistema captura correctamente los elementos esperados, y puede ser especialmente útil cuando se quiere destacar la precisión global del modelo sin penalizar excesivamente por pequeños desajustes en la alineación de tokens. En otras palabras, mientras que WER refleja la "cantidad de errores", MER refleja la "calidad de aciertos" de la predicción. Se calcula teniendo en cuenta las sustituciones, eliminaciones e inserciones de palabras. Al igual que el WER, cuanto más bajo sea el valor del MER, mejor es el rendimiento del sistema, siendo 0 el valor ideal.

  A continuación, se presenta la fórmula que define formalmente el MER

$$\text{MER} = \frac{S + D + I}{S + D + I + C} \tag{2.2}$$

  donde:

  - $S$ = número de sustituciones
  - $D$ = número de eliminaciones
  - $I$ = número de inserciones
  - $C$ = número de tokens correctos

- **WIL (Word Information Lost)**: El WIL se calcula a partir del número de palabras correctas, la cantidad total de palabras en la referencia y las palabras presentes en la predicción. Esta métrica proporciona una estimación de la proporción de información perdida en la transcripción. Un valor de WIL más bajo indica un mejor rendimiento, siendo 0 el valor óptimo que representa una transcripción perfecta, al igual que en las métricas WER y MER.

### 2.2.4. Traducción Automática

La traducción automática (TA) es el proceso mediante el cual se convierte un texto de un idioma a otro de forma automática utilizando modelos computacionales. En este proyecto, el enfoque principal se basa en arquitecturas neuronales recurrentes, en particular las redes LSTM (Long Short-Term Memory), que son adecuadas para modelar secuencias de datos y manejar dependencias temporales en el procesamiento de lenguaje.

Las redes LSTM permiten capturar relaciones contextuales en secuencias de texto y mitigan problemas comunes en redes recurrentes tradicionales, como el desvanecimiento del gradiente. Para entrenar los modelos de traducción automática se utilizó OpenNMT (Klein, Kim, Deng, Senellart, y Rush, 2017), una herramienta de código abierto para la traducción automática desarrollada por el grupo de procesamiento de lenguaje natural de la Universidad de Harvard.

Aunque actualmente los modelos basados en la arquitectura Transformer dominan el estado del arte en traducción automática debido a su capacidad para procesar secuencias en paralelo y mejorar la eficiencia, en este trabajo se optó por LSTM dada su robustez y adaptabilidad para los corpus disponibles y los recursos computacionales con los que se contó.

OpenNMT ofrece flexibilidad para experimentar con diferentes arquitecturas, pero en esta etapa se priorizó la implementación y evaluación de modelos LSTM para la traducción entre texto y glosas, con el fin de obtener una línea base sólida y comprensible para futuros desarrollos.

Para evaluar y comparar el rendimiento de los modelos de traducción automática, se emplean dos métricas habituales en esta tarea: BLEU y CHRF , que son precisamente las utilizadas en este trabajo.

- **BLEU** : es una medida de evaluación de traducción automática que compara la similitud entre la traducción generada y una o más traducciones de referencia. Calcula la proporción de n-gramas en la traducción generada que coinciden con los n-gramas en las traducciones de referencia. Cuanto más alto sea el BLEU score, más similar será la traducción generada a las traducciones de referencia. En particular, en estos experimentos se utilizaron las variantes BLEU-2, BLEU-3 y BLEU-4, que consideran la coincidencia de n-gramas de tamaño 2, 3 y 4, respectivamente, entre las secuencias generadas por el modelo y las referencias esperadas. (Papineni, Roukos, Ward, y Zhu, 2002).

- **CHRF (Character n-gram F-score)**: mide la similitud utilizando n-gramas de caracteres, lo que la hace especialmente útil en contextos donde las traducciones pueden presentar pequeñas variaciones léxicas o morfológicas (Popović, 2015).

Ambas métricas se calcularon utilizando funciones disponibles en la biblioteca de Python `nltk` (Bird, Klein, y Loper, 2009), herramienta ampliamente utilizada en el procesamiento del lenguaje natural.

Estas métricas permiten cuantificar la calidad de las transcripciones y traducciones automáticas producidas, ofreciendo un criterio objetivo para comparar el rendimiento de los distintos modelos entrenados en este trabajo.

# Capítulo 3

# Transcripción Automática

En primera instancia se transcribieron transmisiones de canales de televisión uruguayos con el objetivo de construir un dataset en español uruguayo. Los mismos fueron proporcionados por la Dirección Nacional de Telecomunicaciones (DINATEL) que forma parte del Ministerio de Industria, Energía y Minería.

## 3.1. Pruebas

Para llevar a cabo la transcripción de las transmisiones, se utilizó el lenguaje de programación Python. Python fue elegido debido a su versatilidad y a la amplia disponibilidad de bibliotecas especializadas para el procesamiento de audio y video, así como también de procesamiento de lenguaje natural y se eligieron los siguientes dos módulos que se mencionaron en la sección 2.2.3 .

- Google Speech Recognition (GSR) de SpeechRecognition : Este módulo utiliza el servicio de reconocimiento de voz de Google para transcribir el audio de los vídeos.
- Whisper: es un modelo avanzado desarrollado por OpenAI para el reconocimiento automático de voz, entrenado a partir de grandes cantidades de datos con supervisión débil.

Ambos módulos recibieron como entrada un vídeo. En el caso de Whisper, se retornó la transcripción del audio de cada vídeo, así como los subtítulos del vídeo en un archivo con formato .srt. Este formato, conocido como SubRip Subtitle, es uno de los más utilizados para almacenar subtítulos sincronizados con un recurso audiovisual. Un archivo .srt contiene bloques numerados que incluyen tanto la transcripción textual como las marcas de tiempo de inicio y fin que indican el momento exacto en que cada línea debe aparecer y desaparecer en pantalla. De esta manera, Whisper no solo produce el texto plano de la transcripción, sino también una representación lista para ser empleada como subtítulos en reproductores de video o plataformas de difusión. En cuanto a Google Speech Recognition, solo se obtuvo la transcripción del audio del vídeo.

En concreto, se decidió transcribir las transmisiones del informativo de Canal 5 correspondientes a los días 30 de julio de 2021 y 3 de agosto de 2021. Posteriormente, se amplió la cantidad de datos transcribiendo un bloque del informativo de TV Ciudad del 26 de septiembre de 2022. La inclusión de esta transmisión adicional fue motivada por la necesidad de contar con una mayor cantidad de datos, lo que permitiría obtener un conjunto más robusto.

Los archivos de las transmisiones originales están en los formatos .mp4 y .mxf. Al revisar estos archivos, se identificaron momentos en los cuales no había ninguna persona hablando, los cuales correspondían a las pausas comerciales. Estos segmentos de silencio eran innecesarios para los fines del proyecto y podían interferir en la calidad del dataset. Por esta razón, se procedió a editar estos fragmentos, separando cada transmisión en tres partes distintas, de manera que cada una de ellas quedara sin los momentos de silencio no deseados. Por lo tanto, se transcribieron un total de 238 minutos y 48 segundos.

Las transcripciones obtenidas con Whisper fueron revisadas y corregidas manualmente para generar un conjunto de referencia confiable. Este proceso consistió en verificar la exactitud de cada palabra y ajustar posibles errores de reconocimiento, asegurando que la referencia reflejara fielmente el contenido de las transmisiones. La elección de Whisper como base se debió a que sus resultados iniciales presentaban una mayor calidad, lo que permitió reducir significativamente el esfuerzo de corrección manual y asegurar un gold standard robusto para el proyecto.

Para compararlos, se evaluó el rendimiento de acuerdo a las siguientes métricas : WER (Word Error Rate), MER (Match Error Rate) y WIL (Word Information Lost).

En la Tabla 3.1 se presenta un resumen del tiempo transcripto y del número de hablantes identificados en las distintas transmisiones televisivas consideradas en el proyecto. Cabe destacar que si bien la transcripción inicial se obtuvo utilizando el modelo Whisper, todo el material fue posteriormente corregido manualmente para asegurar su calidad. La diarización se realizó íntegramente de forma manual, dado que el modelo no ofrece esta funcionalidad. El número total de hablantes asciende a 88, dado que algunos aparecen en más de una transmisión y el conteo se realizó considerando únicamente individuos únicos.

**Cuadro 3.1:** Tiempo transcripto, cantidad de hablantes y tokens en cada una de las transmisiones televisivas utilizadas en el proyecto.

| **Origen** | **Tiempo transcripto** | **Hablantes** | **Tokens** |
|---|---|---|---|
| Canal 5 - 30/07/2021 | 85 Min y 30 Seg | 43 | 26,177 |
| Canal 5 - 3/08/2021 | 84 Min y 31 Seg | 38 | 26,866 |
| TV Ciudad B1 - 26/09/2022 | 68 Min y 37 Seg | 21 | 10,752 |
| Total | 238 Min y 48 Seg | 88 | 63,795 |

Los resultados obtenidos para cada una de las métricas correspondientes a los segmentos transcriptos de las transmisiones del 30 de julio y 3 de agosto de

2021 se presentan en la Tabla 3.2.

**Cuadro 3.2:** Métricas WER, MER y WIL para Canal 5 - 30 de Julio y 3 de Agosto 2021.

| **Fecha** | **Segmento** | **WER** | **MER** | **WIL** |
|---|---|---|---|---|
| Canal 5 - 30/07/2021 | Parte 1 - Whisper | 0.05 | 0.05 | 0.07 |
| | Parte 1 - GSR | 0.27 | 0.26 | 0.33 |
| | Parte 2 - Whisper | 0.17 | 0.17 | 0.26 |
| | Parte 2 - GSR | 0.28 | 0.27 | 0.36 |
| | Parte 3 - Whisper | 0.13 | 0.12 | 0.19 |
| | Parte 3 - GSR | 0.24 | 0.24 | 0.31 |
| Canal 5 - 3/08/2021 | Parte 1 - Whisper | 0.10 | 0.10 | 0.15 |
| | Parte 1 - GSR | 0.23 | 0.23 | 0.30 |
| | Parte 2 - Whisper | 0.15 | 0.15 | 0.22 |
| | Parte 2 - GSR | 0.23 | 0.23 | 0.31 |
| | Parte 3 - Whisper | 0.19 | 0.18 | 0.27 |
| | Parte 3 - GSR | 0.25 | 0.25 | 0.33 |
| **Totales(Promedio)** | Whisper | 0.13 | 0.13 | 0.19 |
| | GSR | 0.25 | 0.25 | 0.32 |

## 3.2. Conclusiones

Considerando los resultados obtenidos para cada métrica , se observó consistentemente que Whisper arrojaba mejores resultados en comparación con GSR. No obstante, se detectaron ciertas limitaciones en el desempeño de Whisper, especialmente en el reconocimiento de nombres propios de instituciones, personas o lugares uruguayos, posiblemente debido a su menor presencia en los datos de entrenamiento del modelo. En términos generales, estos resultados refuerzan la decisión de continuar el trabajo exclusivamente con Whisper, ya que ofrece transcripciones más precisas, mejor alineadas y con menor pérdida de información, a pesar de estas limitaciones particulares.

Además, como parte de este experimento, se construyó un recurso de 68 minutos y 37 segundos de audio transcripto, diarizado y corregido manualmente. Este corpus podrá emplearse en futuras investigaciones para continuar evaluando la performance de distintos métodos de transcripción automática sobre audio en español uruguayo, facilitando así la comparación estandarizada entre modelos y fomentando el desarrollo de soluciones mejor adaptadas al contexto local.

# Capítulo 4

# Traducción automática a glosas

En esta fase del proyecto, el objetivo principal fue desarrollar un sistema capaz de traducir texto en español oral , específicamente en español de Uruguay, a sus glosas correspondientes en LSU. Para lograr este propósito, se utilizaron cuatro corpus compuestos por textos en español oral y sus respectivas glosas, los cuales fueron utilizados para entrenar modelos de traducción automática.

Dado que la cantidad de datos anotados específicamente en LSU aún es limitada, en algunos experimentos se utilizarán corpus paralelos entre español oral y LSE como complemento. Si bien es claro que la LSE y la LSU son lenguas distintas, consideramos relevante explorar si es posible aprovechar estos datos para mejorar los modelos de traducción hacia LSU. Esta decisión se justifica en dos aspectos principales:

- la lengua oral es compartida entre ambos contextos;
- las glosas utilizadas en ambos corpus están anotadas con palabras similares, generalmente tomadas del español oral.

A pesar de que estas glosas pueden parecer similares, las lenguas de señas involucradas presentan diferencias gramaticales y estructurales. Por este motivo, resulta de particular interés evaluar si alguno de estos corpus efectivamente contribuye a mejorar el rendimiento del sistema en el escenario de traducción hacia LSU.

## 4.1. Corpus

Para realizar los experimentos se utilizaron los siguientes cuatro corpus:

- *id/dl* (San-Segundo y cols., 2008): corpus que se centra en renovaciones de la cédula o de la licencia de conducir en España. Utiliza la Lengua de Señas Española (LSE).

- *iSignos* (Cabeza y García-Miguel, 2019): corpus más numeroso que *id/dl* y que también emplea la Lengua de Señas Española (LSE).
- *Datos Uruguay*: corpus que trabaja con la Lengua de Señas Uruguaya (LSU).
- *Sobreciencia*: corpus que contiene glosas en LSU correspondientes al programa Sobreciencia de TV Ciudad, orientado a la divulgación científica.

Cada uno de los corpus fue dividido en tres conjuntos: entrenamiento (70 %), prueba (15 %) y validación (15 %). La Tabla 4.2 presenta la cantidad de ejemplos asignados a cada conjunto para cada corpus. Además, en la Tabla 4.1 se muestra un resumen con la cantidad total de oraciones, palabras en lengua oral y glosas en lengua de señas presentes en cada corpus.

**Cuadro 4.1:** Cantidad total de oraciones, palabras (tokens de lengua oral) y glosas (tokens de lengua de señas) por corpus.

| Corpus | Oraciones | Palabras (tokens) | Glosas (tokens) |
|---|---|---|---|
| *Sobreciencia* | 255 | 5241 | 2411 |
| *id/dl* | 416 | 4929 | 4640 |
| *iSignos* | 2798 | 15331 | 7502 |
| *Datos Uruguayos* | 87 | 834 | 511 |

**Cuadro 4.2:** Cantidad de oraciones por conjunto (train, validación y test).

| Corpus | Train | Val | Test | Total |
|---|---|---|---|---|
| *Sobreciencia* | 178 | 38 | 39 | 255 |
| *id/dl* | 266 | 75 | 75 | 416 |
| *iSignos* | 1869 | 462 | 462 | 2793 |
| *Datos Uruguayos* | 60 | 13 | 14 | 87 |

## 4.2. Experimentos

Para el entrenamiento de los modelos se empleó OpenNMT, (ver sección 2.2.4). En los experimentos realizados, se utilizaron los parámetros predeterminados de la herramienta. No se realizaron ajustes específicos de arquitectura ni hiperparámetros, más allá de los definidos por la herramienta de forma predeterminada.Es decir, se utilizó la configuración básica ofrecida por OpenNMT, la cual funciona con una arquitectura encoder-decoder basada en redes LSTM y un mecanismo de atención (modelo atencional).

En el contexto del entrenamiento de modelos de traducción automática, se denomina paso (o step) a una iteración individual en la que el modelo procesa

un lote de ejemplos, ajustando sus pesos en función del error observado. Por otro lado, una época (epoch) corresponde a un recorrido completo por todos los datos del conjunto de entrenamiento. Dado que los experimentos se organizaron en función del número de pasos, este será el término utilizado a lo largo de esta sección.

Durante el proceso de entrenamiento, en todos los experimentos se llevaron a cabo un total de 10,000 pasos de entrenamiento. Se implementó un mecanismo de guardado de checkpoints, donde cada 500 pasos de entrenamiento, una versión del modelo en ese punto se guardaba. Esto permitió tener acceso a múltiples versiones del modelo en diferentes etapas de su entrenamiento, con el objetivo de analizar su progreso y desempeño a lo largo del tiempo. De esa manera, poder determinar si el modelo seguía mejorando su rendimiento con más iteraciones o si el modelo alcanzaba una meseta en algún punto. En todos los casos, el conjunto de validación se utilizó durante el entrenamiento para monitorear el desempeño del modelo en datos no vistos. Esto permitió a OpenNMT ajustar los parámetros y seleccionar los mejores checkpoints, evitando sobreajuste y garantizando que el modelo generalice adecuadamente a nuevos ejemplos.

Se realizaron los siguientes experimentos:

1. Por un lado, se entrenó y evaluó sobre las particiones del mismo corpus, ya sea *id/dl*, *iSignos*, *Datos Uruguay* o *Sobreciencia*. Siempre utilizando la partición de train para el entrenamiento y la partición de test para la evaluación

2. Por otro lado, se decidió realizar experimentos cruzados para evaluar el comportamiento de los modelos al enfrentarse a ejemplos que no forman parte de su corpus. Se eligió el corpus iSignos debido a que es el más numeroso.

    - Entrenar con el conjunto de entrenamiento de *iSignos* y evaluar con datos de *id/dl*.
    - Entrenar con el conjunto de entrenamiento de *iSignos* y evaluar con *Datos Uruguay*.
    - Combinar los corpus de *iSignos* e *id/dl* y evaluar los resultados.

3. Como *Sobreciencia* constituye el corpus más extenso de glosas en LSU utilizado en este estudio se definieron a partir de él los siguientes experimentos, centrados en distintas combinaciones de los corpus para el entrenamiento.:

    - Entrenar con *Sobreciencia* junto con id/dl, y evaluar en el conjunto de prueba de *Sobreciencia*.
    - Entrenar con *Sobreciencia* junto con *Datos Uruguay*, y evaluar en el conjunto de prueba de *Sobreciencia*.
    - Entrenar con la combinación de *Sobreciencia*, *id/dl*, *Datos Uruguay* e *iSignos*, y evaluar en el conjunto de prueba de *Sobreciencia*.

**Cuadro 4.3:** Resumen de los experimentos realizados

| Corpus de entrenamiento | Corpus de test |
|---|---|
| *id/dl* | *id/dl* |
| *iSignos* | *iSignos* |
| *Datos Uruguay* | *Datos Uruguay* |
| *Sobreciencia* | *Sobreciencia* |
| *iSignos* | *id/dl* |
| *iSignos* | *Datos Uruguay* |
| *iSignos + id/dl* | *iSignos + id/dl* |
| *Sobreciencia + id/dl* | *Sobreciencia* |
| *Sobreciencia + Datos Uruguay* | *Sobreciencia* |
| *Sobreciencia + id/dl + Datos Uruguay + iSignos* | *Sobreciencia* |

Para evaluar el rendimiento del modelo, se emplearon dos métricas ampliamente utilizadas en el campo del procesamiento del lenguaje natural: el CHRF score y el BLEU score, (ver sección 2.2.4). Ambas métricas se calcularon utilizando funciones de la biblioteca de Python `nltk`.

Para el cálculo de BLEU se aplicó el método de suavizado method4 provisto por SmoothingFunction también provisto por `nltk`. Este método consiste en agregar una pequeña constante a las cuentas de los n-gramas para evitar que la puntuación global se reduzca a cero cuando ciertos n-gramas de orden superior no aparecen en las referencias, lo cual es común en tareas con secuencias cortas como la generación de glosas. De esta forma, el suavizado permite una evaluación más robusta del rendimiento del modelo.

El objetivo de utilizar tanto la métrica CHRF como BLEU fue obtener una visión más completa del rendimiento del modelo, ya que BLEU evalúa la precisión en la coincidencia de n-gramas de palabras enteras, mientras que CHRF mide la similitud a nivel de n-gramas de caracteres. Esta combinación permite evaluar tanto la coincidencia de palabras como las pequeñas diferencias en la forma de las palabras o errores menores en las traducciones.

## 4.3. Resultados experimentos con el mismo corpus

En esta serie de experimentos, el modelo fue entrenado y evaluado utilizando exclusivamente el mismo corpus en ambos casos. El objetivo principal es observar el comportamiento del modelo cuando se expone de forma consistente a un único conjunto de datos, sin influencias externas. Esto permite analizar su capacidad de aprendizaje, generalización interna y adaptación a las características específicas del corpus en cuestión.

Los experimentos en los que el modelo fue entrenado exclusivamente con el conjunto de entrenamiento del corpus *Sobreciencia* y evaluado sobre su conjunto

de prueba se presentan en la sección 4.5.1 , junto con el resto de experimentos que evalúan el desempeño en este corpus.

### 4.3.1. Corpus *isignos*

**Cuadro 4.4:** Resultados de métricas BLEU y CHRF para los distintos checkpoints del experimento con el corpus *isignos*

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.0007 | 0.0003 | 0.0001 | 0.0726 |
| 1000 | 0.0026 | 0.0006 | 0.0003 | 0.0979 |
| 1500 | 0.0639 | 0.0204 | 0.0053 | 0.1976 |
| 2000 | 0.1327 | 0.0650 | 0.0285 | 0.2445 |
| 2500 | 0.1466 | 0.0763 | 0.0354 | 0.2549 |
| 3000 | 0.1476 | 0.0791 | 0.0400 | 0.2570 |
| 3500 | 0.1405 | 0.0760 | 0.0385 | 0.2564 |
| 4000 | 0.1443 | 0.0746 | 0.0355 | 0.2583 |
| 4500 | 0.1499 | 0.0799 | 0.0403 | 0.2577 |
| 5000 | 0.1460 | 0.0768 | 0.0363 | 0.2598 |
| 5500 | 0.1502 | 0.0799 | 0.0374 | 0.2584 |
| 6000 | 0.1506 | 0.0800 | 0.0375 | 0.2605 |
| 6500 | 0.1485 | 0.0777 | 0.0366 | 0.2597 |
| 7000 | 0.1481 | 0.0793 | 0.0373 | 0.2605 |
| 7500 | 0.1507 | 0.0799 | 0.0374 | 0.2586 |
| 8000 | 0.1494 | 0.0797 | 0.0375 | 0.2616 |
| 8500 | 0.1467 | 0.0773 | 0.0366 | 0.2614 |
| 9000 | 0.1471 | 0.0771 | 0.0364 | 0.2592 |
| 9500 | 0.1502 | 0.0799 | 0.0376 | 0.2621 |
| 10000 | **0.1509** | **0.0803** | **0.0376** | **0.2622** |

En este experimento, se entrenó y evaluó el modelo utilizando únicamente el corpus *isignos*.

La Tabla 4.4 presenta los valores obtenidos para las métricas BLEU-2, BLEU-3, BLEU-4 y CHRF en los distintos checkpoints del entrenamiento con el corpus *iSignos*. Complementariamente, la Figura 4.1 ilustra gráficamente su evolución a lo largo de los pasos de entrenamiento, permitiendo visualizar el comportamiento progresivo y la estabilización de cada métrica.

Los valores de CHRF comienzan en 0.0726 y aumentan de forma sostenida hasta aproximadamente el paso 3000. A partir de ese punto, se estabilizan en una meseta, con ligeras fluctuaciones entre 0.256 y 0.262, manteniéndose consistentes hasta el final del entrenamiento, logrando un máximo de 0.2622 en el paso 10000.

En cuanto a BLEU-2, los valores inician en 0.0007 y presentan un incremento progresivo, alcanzando una estabilización a partir del paso 4000 con cifras que oscilan entre 0.14 y 0.15, alcanzando un valor máximo de 0.1509 en el paso

10000. BLEU-3 comienza en 0.0003, muestra una subida en los primeros pasos y se mantiene alrededor de 0.08 desde aproximadamente el paso 4000, con una leve variación en los valores posteriores. Por su parte, BLEU-4 inicia en 0.0001 y presenta una mejora hasta alcanzar un máximo de 0.0403 en el paso 4500. A partir de allí, los valores se estabilizan en torno a 0.037, con pequeñas oscilaciones hasta el final del entrenamiento.

En general, todas las métricas muestran un crecimiento inicial marcado seguido de una estabilización, y se observa una disminución progresiva en los valores absolutos a medida que aumenta el tamaño del n-grama considerado en BLEU.

La métrica CHRF reporta valores significativamente más altos que las distintas variantes de BLEU; particularmente, los valores de BLEU-3 y BLEU-4 son muy cercanos a cero durante todo el entrenamiento. Esto sugiere que, si bien las traducciones pueden no coincidir exactamente a nivel de n-gramas, sí presentan similitudes relevantes en estructura y contenido, captadas con mayor eficacia por CHRF.

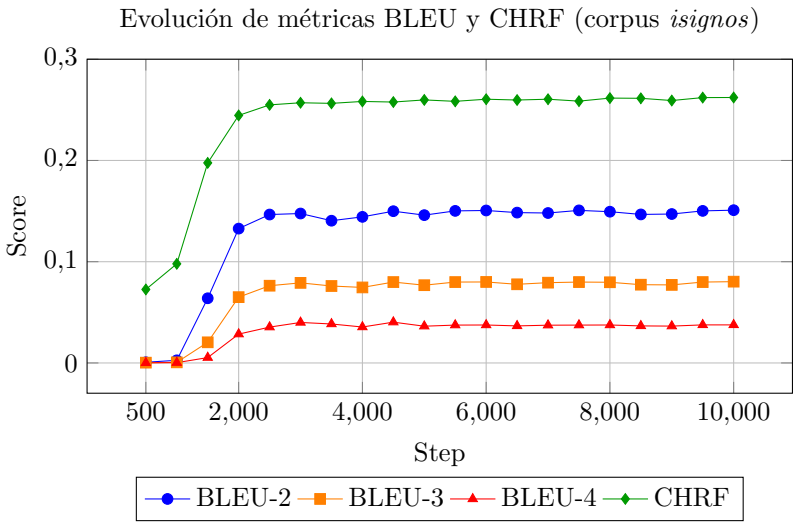


**Figura 4.1:** Comparación de métricas BLEU-2, BLEU-3, BLEU-4 y CHRF en función del número de steps durante el entrenamiento con el corpus *iSignos*.

### 4.3.2. Corpus *id/dl*

En este experimento, se entrenó y evaluó el modelo utilizando únicamente el corpus *id/dl*.

Los resultados obtenidos para cada métrica se presentan en la Tabla 4.5, donde se detallan los valores de las métricas BLEU-2, BLEU-3, BLEU-4 y CHRF en diferentes etapas del entrenamiento. Además, la Figura 4.2 ilustra gráficamente

**Cuadro 4.5:** Resultados de métricas BLEU y CHRF para los distintos checkpoints del experimento con el corpus *id/dl*

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.3631 | 0.2955 | 0.2403 | 0.4448 |
| 1000 | 0.3667 | 0.3107 | 0.2655 | 0.4550 |
| 1500 | **0.4218** | 0.3641 | 0.3143 | **0.5060** |
| 2000 | 0.4114 | 0.3545 | 0.3063 | 0.4949 |
| 2500 | 0.4133 | 0.3577 | 0.3104 | 0.4929 |
| 3000 | 0.4156 | 0.3609 | 0.3148 | 0.4894 |
| 3500 | 0.4070 | 0.3494 | 0.2987 | 0.4882 |
| 4000 | 0.4200 | **0.3660** | **0.3195** | 0.4954 |
| 4500 | 0.4132 | 0.3566 | 0.3085 | 0.4844 |
| 5000 | 0.4138 | 0.3550 | 0.3055 | 0.4900 |
| 5500 | 0.4104 | 0.3525 | 0.3036 | 0.4868 |
| 6000 | 0.4128 | 0.3544 | 0.3051 | 0.4896 |
| 6500 | 0.4049 | 0.3475 | 0.2996 | 0.4777 |
| 7000 | 0.4070 | 0.3486 | 0.3002 | 0.4798 |
| 7500 | 0.4111 | 0.3534 | 0.3050 | 0.4851 |
| 8000 | 0.4122 | 0.3553 | 0.3070 | 0.4845 |
| 8500 | 0.4014 | 0.3468 | 0.3009 | 0.4719 |
| 9000 | 0.3976 | 0.3413 | 0.2942 | 0.4712 |
| 9500 | 0.4058 | 0.3519 | 0.3063 | 0.4760 |
| 10000 | 0.4058 | 0.3519 | 0.3063 | 0.4760 |

esta evolución, permitiendo observar con mayor claridad las tendencias de cada métrica a lo largo del tiempo.

Los valores de CHRF comienzan en 0.4448 y alcanzan su punto máximo de 0.5060 en el paso 1500. A partir de ese momento, se observa una ligera caída seguida de una estabilización parcial. Desde el paso 5000 hasta el 10000, los valores oscilan entre 0.4712 y 0.4950, indicando que el modelo mantiene un rendimiento relativamente constante en esta métrica, aunque sin superar el valor pico inicial.

En cuanto a las métricas BLEU, los valores para las tres variantes comienzan en 0.3631, 0.2955 y 0.2403 respectivamente en el paso 500. Todas ellas aumentan rápidamente en las etapas iniciales, alcanzando su máximo alrededor del paso 4000, con valores de 0.4200 para BLEU-2, 0.3660 para BLEU-3 y 0.3195 para BLEU-4. Luego, a partir de ese punto, los valores entran en una meseta que se extiende aproximadamente hasta el final del entrenamiento, fluctuando en rangos moderadamente altos: BLEU-2 entre 0.3976 y 0.4156, BLEU-3 entre 0.3413 y 0.3609, y BLEU-4 entre 0.2942 y 0.3148. Aunque no se mantiene el valor máximo, el rendimiento se estabiliza sin caídas significativas.

Estas métricas reflejan que el modelo aprende rápidamente en las primeras etapas del entrenamiento, pero no logra mejoras significativas después de esos

puntos. Considerando todas las métricas, el mejor desempeño general se alcanza entre los pasos 1500 y 4000.

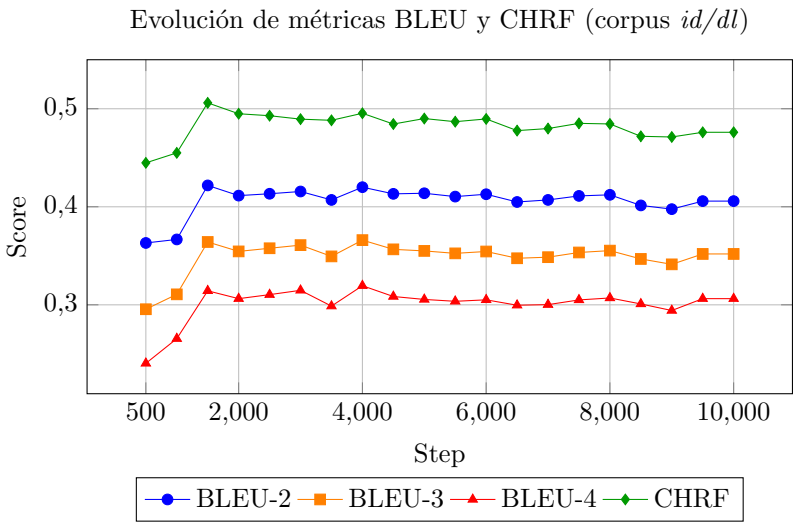


**Figura 4.2:** Comparación de métricas BLEU-2, BLEU-3, BLEU-4 y CHRF en función del número de steps durante el entrenamiento con el corpus *id/dl*.

### 4.3.3. Corpus *Datos Uruguayos*

En este experimento, se entrenó y evaluó el modelo utilizando únicamente el corpus *Datos Uruguayos*.

Los resultados obtenidos para cada métrica se presentan en la Tabla 4.6, donde se detallan los valores de las métricas BLEU-2, BLEU-3, BLEU-4 y CHRF en diferentes etapas del entrenamiento. Además, las Figuras 4.3 y 4.4 ilustran gráficamente la evolución de estas métricas por separado.

El valor de CHRF comienza en 0.1147 y desciende ligeramente en los pasos iniciales, alcanzando un mínimo de 0.1042. Luego se mantiene estable, con muy poca oscilación hasta el final del entrenamiento. Este comportamiento indica que el modelo mostró un desempeño constante en esta métrica, con escasa variación a lo largo del entrenamiento.

En cuanto a las métricas BLEU, se observa que los valores absolutos obtenidos para BLEU-2, BLEU-3 y BLEU-4 son muy bajos en todos los casos, lo que sugiere un rendimiento limitado del modelo. BLEU-2 inicia con un valor de 0.0225, pero desciende rápidamente y luego se estabiliza en torno a 0.0093, con mínimas fluctuaciones. BLEU-3 sigue un comportamiento similar: empieza en 0.0182 y, tras una caída inicial, se mantiene prácticamente constante alrededor de 0.0060. BLEU-4 muestra la misma tendencia, comenzando en 0.0133 y estabilizándose cerca de 0.0040.

**Cuadro 4.6:** Resultados de métricas BLEU y CHRF para los distintos checkpoints del experimento con el corpus *Datos Uruguayos*

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | **0.0225** | **0.0182** | **0.0133** | **0.1147** |
| 1000 | 0.0101 | 0.0063 | 0.0041 | 0.1041 |
| 1500 | 0.0090 | 0.0058 | 0.0039 | 0.1047 |
| 2000 | 0.0094 | 0.0061 | 0.0040 | 0.1042 |
| 2500 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 3000 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 3500 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 4000 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 4500 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 5000 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 5500 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 6000 | 0.0094 | 0.0061 | 0.0040 | 0.1048 |
| 6500 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |
| 7000 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |
| 7500 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |
| 8000 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |
| 8500 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |
| 9000 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |
| 9500 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |
| 10000 | 0.0092 | 0.0060 | 0.0040 | 0.1105 |

En conjunto, estos valores reflejan una baja calidad en las coincidencias exactas de n-gramas y sin mejoras significativas a lo largo del entrenamiento. Esta falta de variación sugiere que el modelo no logró avances sustanciales en este aspecto.

El resultado máximo se alcanza apenas empezado el entrenamiento. Esto es típico cuando se dispone de pocos datos; el modelo aprende a generalizar un poco al principio, pero luego tiende a sobreajustarse rápidamente a los datos de entrenamiento, lo que provoca una caída estrepitosa en la performance sobre el conjunto de prueba.

### 4.3.4. Conclusiones

El rendimiento varió significativamente en función del corpus utilizado durante el entrenamiento. En el caso de *iSignos*, que es el de mayor tamaño, se observó un aprendizaje progresivo en las etapas iniciales, seguido de una estabilización. BLEU-2 alcanzó un máximo de 0.1509, lo que sugiere una adecuada captación de patrones en secuencias cortas. Sin embargo, los valores más bajos en BLEU-3 y BLEU-4 indican dificultades para generar secuencias más extensas de manera coherente. CHRF, por su parte, se mantuvo estable en torno

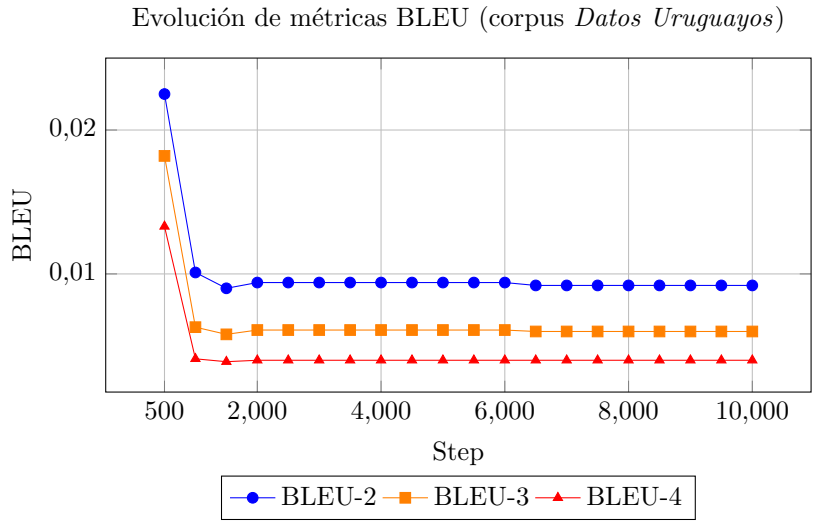


**Figura 4.3:** Evolución de las métricas BLEU-2, BLEU-3 y BLEU-4 en función del número de steps durante el entrenamiento con el corpus *Datos Uruguayos*.

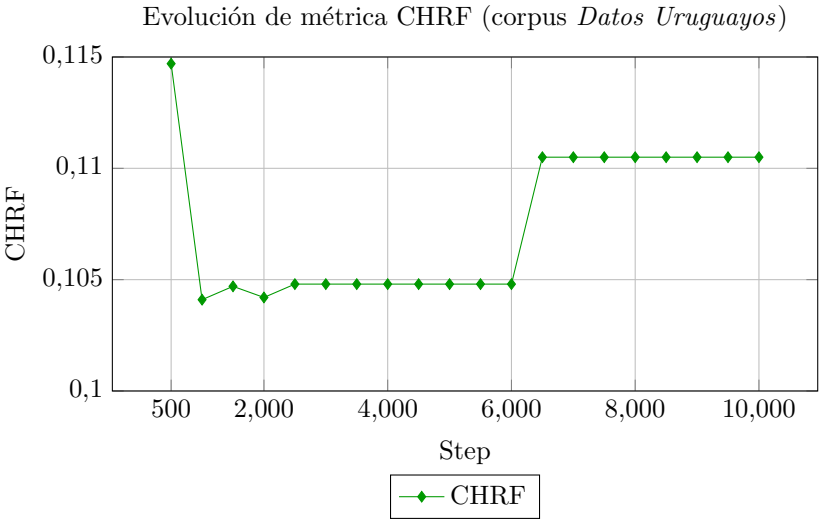


**Figura 4.4:** Evolución de la métrica CHRF en función del número de steps durante el entrenamiento con el corpus *Datos Uruguayos*.

a 0.2622, lo que refleja cierta similitud léxica entre las salidas del sistema y las traducciones de referencia.Los valores máximos de las métricas se alcanzan hacia las últimas etapas del entrenamiento, como se observa en la Tabla 4.4, lo que podría sugerir que continuar el proceso permitiría seguir mejorando el rendimiento del modelo. Sin embargo, al analizar las curvas de aprendizaje, se

advierte que los resultados comienzan a oscilar, lo que indica una posible saturación. En otros experimentos, en cambio, se observa una estabilización más temprana, lo cual sugiere que prolongar el entrenamiento no habría tenido un impacto significativo en la calidad de las predicciones.

En contraste, el corpus *id/dl* presentó el mejor desempeño global. Se obtuvieron los valores más altos en todas las métricas: BLEU-2 (0.4218), BLEU-3 (0.3660), BLEU-4 (0.3195) y CHRF (0.5060). Estos resultados reflejan una mayor capacidad para generar traducciones coherentes y precisas, incluso en secuencias más largas.

Probablemente, el buen desempeño observado en el corpus *id/dl* se deba al dominio acotado de los datos, enfocado principalmente en la renovación de documentos, lo que genera frases bastante repetitivas y fáciles de memorizar para el modelo. En contraste, el corpus *iSignos* presenta una mayor variedad en el contenido y la estructura de las oraciones, por lo que se requieren significativamente más datos para que el modelo pueda aprender y generalizar adecuadamente.

El corpus *Datos Uruguayos*, el más pequeño de los datasets, mostró el rendimiento más bajo. Las métricas BLEU se mantuvieron cercanas a cero durante todo el entrenamiento, con un máximo de apenas 0.0225 en BLEU-2, lo que evidencia una escasa capacidad para reproducir coincidencias exactas con las referencias. En cuanto a CHRF, se registró una leve mejora inicial, pero luego se estabilizó en torno a 0.1147, un valor notablemente inferior al obtenido con los otros dos corpus. En conjunto, estos resultados reflejan un aprendizaje limitado y una baja similitud con las traducciones de referencia.

Al analizar el comportamiento de las métricas en general, se observan patrones consistentes entre los distintos corpus. Dentro de las métricas BLEU, los valores más altos se registraron sistemáticamente en BLEU-2. Esto indica que las coincidencias de bigramas fueron más frecuentes durante el entrenamiento y sugiere que los sistemas tienden a aprender con mayor facilidad patrones cortos y locales. A medida que aumenta la longitud del n-grama evaluado, los valores decrecen considerablemente, reflejando una mayor dificultad para capturar dependencias a largo plazo y generar secuencias más extensas de forma coherente. Este comportamiento es esperable, ya que la generación de n-gramas más largos requiere una mayor capacidad de modelado contextual, así como una cantidad de datos suficiente para cubrir las combinaciones posibles. Esta limitación se manifiesta con mayor claridad en el corpus *Datos Uruguayos*, donde la escasez de datos restringe el aprendizaje de estructuras complejas.

Por otro lado, la métrica CHRF arrojó en todos los casos valores superiores a los obtenidos con BLEU. Al estar basada en coincidencias a nivel de caracteres, CHRF es más sensible a similitudes léxicas parciales, incluso cuando los n-gramas exactos no coinciden. Esto permite capturar aspectos de la calidad de traducción que BLEU podría pasar por alto, como la cercanía morfológica o la estructura interna de las palabras. Además, sus valores más estables a lo largo del entrenamiento indican que refleja mejor las mejoras graduales en fluidez y naturalidad. Por lo tanto, CHRF se consolida como una métrica complementaria fundamental para evaluar traducciones en contextos donde las coincidencias exactas no son el único criterio relevante.

**Cuadro 4.7:** Mejores resultados obtenidos para cada métrica y corpus

| Corpus | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| *iSignos* | 0.1509 | 0.0803 | 0.0403 | 0.2622 |
| *id/dl* | 0.4218 | 0.3660 | 0.3195 | 0.5060 |
| *Datos Uruguayos* | 0.0225 | 0.0182 | 0.0133 | 0.1147 |

La Tabla 4.7 resume los mejores valores obtenidos para cada métrica y corpus, y permite visualizar de forma comparativa el rendimiento alcanzado en cada escenario. De este análisis general se concluye que el corpus *id/dl* proporciona las mejores condiciones para el aprendizaje, permitiendo generar traducciones más precisas y coherentes tanto en secuencias cortas como largas. *iSignos* mostró un progreso más limitado, con un mejor desempeño en la reproducción de bigramas, pero con mayores dificultades en estructuras complejas. Finalmente, *Datos Uruguayos* presentó un rendimiento modesto, evidenciando las restricciones que impone trabajar con conjuntos de datos pequeños y poco diversos.

Estos resultados subrayan la importancia de utilizar corpora amplios y representativos para lograr modelos de traducción automática más efectivos. Asimismo, resaltan la necesidad de explorar estrategias que favorezcan el aprendizaje de secuencias largas y complejas, dado que las métricas basadas en n-gramas mayores siguen siendo uno de los principales desafíos en esta tarea.

## 4.4. Resultados experimentos cruzados

Dado que *iSignos* es el corpus más numeroso, se lo utilizó como base para predecir los otros dos conjuntos de datos. Esta estrategia parte de la premisa de que un modelo entrenado con una mayor cantidad de ejemplos posee una mayor capacidad de generalización, incluso cuando se enfrenta a tareas de traducción en contextos con datos escasos. En este caso, se busca evaluar su desempeño al aplicarlo sobre un conjunto diferente y no visto durante el entrenamiento, lo cual permite analizar su capacidad para transferir el conocimiento adquirido a nuevos dominios. El objetivo es determinar si el modelo puede adaptarse a nuevas variedades lingüísticas o temáticas sin requerir entrenamiento específico en cada corpus reducido.

Cabe señalar que, en algunos de estos experimentos cruzados, los idiomas de destino en la traducción son distintos: mientras que *iSignos* está orientado a la Lengua de Signos Española (LSE), los corpus *id/dl* y *Datos Uruguay* están orientados a la Lengua de Señas Uruguaya (LSU). Si bien en algunos casos comparten glosas con formas superficiales similares, se trata de lenguas diferentes. Por ello, resulta especialmente relevante evaluar si un modelo entrenado en una lengua puede mejorar el rendimiento de traducción al aplicarse sobre otra, analizando así su capacidad de generalización en un escenario multilingüe con recursos limitados.

Adicionalmente, se entrenó un modelo utilizando una combinación de los cor-

pus *id/dl* e *iSignos*, dado que ambos corresponden a traducciones entre español y Lengua de Signos Española (LSE). Esta configuración busca explorar si el entrenamiento conjunto en un espacio lingüístico compartido puede mejorar los resultados con respecto al entrenamiento exclusivo sobre *id/dl*, particularmente en términos de aprovechamiento del volumen y la diversidad de datos.

### 4.4.1. Corpus id/dl con isignos

Este experimento consiste en entrenar el modelo utilizando exclusivamente el corpus *iSignos*, el más grande disponible, y evaluarlo sobre el conjunto de datos *id/dl*. El objetivo es analizar la capacidad de transferencia del conocimiento aprendido en un dominio más amplio hacia otro relacionado, pero con diferentes características y tamaño.

**Cuadro 4.8:** Resultados de métricas BLEU y CHRF para distintos checkpoints del experimento Train Isignos con datos id/dl

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.0000 | 0.0000 | 0.0000 | 0.02 |
| 1000 | 0.0000 | 0.0000 | 0.0000 | 0.03 |
| 1500 | 0.0001 | 0.0001 | 0.0001 | 0.06 |
| 2000 | 0.0003 | 0.0002 | 0.0002 | 0.07 |
| 2500 | 0.0007 | 0.0005 | 0.0003 | **0.08** |
| 3000 | 0.0005 | 0.0004 | 0.0003 | **0.08** |
| 3500 | 0.0006 | 0.0004 | 0.0003 | 0.07 |
| 4000 | 0.0007 | 0.0005 | 0.0003 | **0.08** |
| 4500 | **0.0008** | **0.0006** | **0.0004** | **0.08** |
| 5000 | **0.0008** | 0.0005 | 0.0003 | **0.08** |
| 5500 | 0.0005 | 0.0004 | 0.0003 | 0.07 |
| 6000 | 0.0005 | 0.0004 | 0.0003 | **0.08** |
| 6500 | 0.0006 | 0.0004 | 0.0003 | 0.07 |
| 7000 | 0.0006 | 0.0005 | 0.0003 | **0.08** |
| 7500 | 0.0004 | 0.0003 | 0.0002 | 0.07 |
| 8000 | 0.0006 | 0.0004 | 0.0003 | 0.07 |
| 8500 | 0.0007 | 0.0005 | 0.0003 | **0.08** |
| 9000 | 0.0007 | 0.0005 | 0.0003 | 0.07 |
| 9500 | 0.0007 | 0.0005 | 0.0003 | 0.07 |
| 10000 | 0.0005 | 0.0004 | 0.0003 | 0.07 |

En este experimento, los valores obtenidos para las métricas CHRF y BLEU se mantuvieron muy cercanos a cero a lo largo de todo el entrenamiento, lo que indica un desempeño muy limitado del modelo en términos de calidad de traducción. La métrica CHRF alcanzó un valor máximo de 0.0831, con fluctuaciones dentro de un rango bajo y sin mostrar una mejora sostenida, como se muestra en la Figura 4.6. En el caso de BLEU, las tres variantes evaluadas

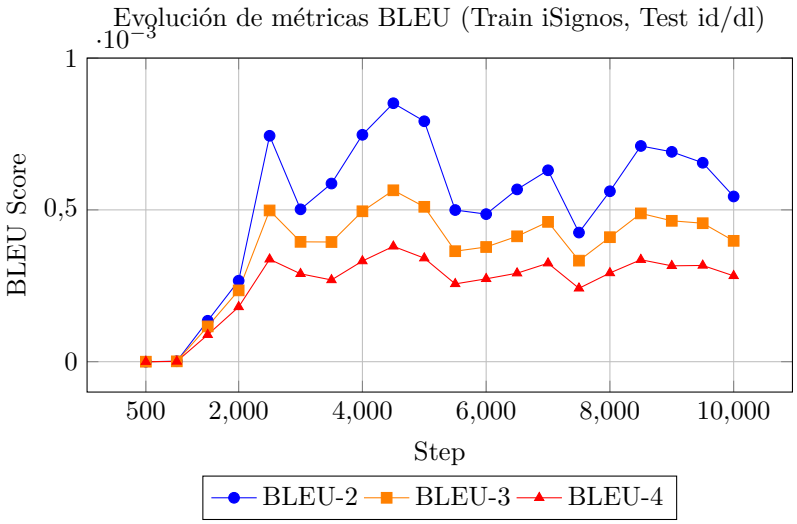


**Figura 4.5:** Evolución de las métricas BLEU-2, BLEU-3 y BLEU-4 durante el entrenamiento con el corpus *iSignos* y evaluación sobre *id/dl*.

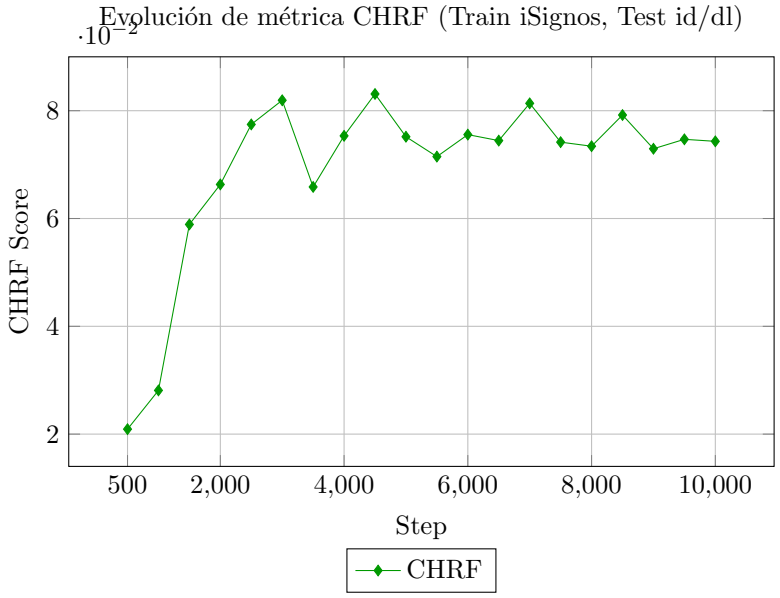


**Figura 4.6:** Evolución de la métrica CHRF durante el entrenamiento con el corpus *iSignos* y evaluación sobre *id/dl*.

(BLEU-2, BLEU-3 y BLEU-4) también reflejaron un rendimiento muy bajo, con valores que no superaron los 0.00085, 0.00056 y 0.00038 respectivamente,

ubicándose varios órdenes de magnitud por debajo de 1, como puede observarse en la Figura 4.5. Estos resultados refuerzan la idea de que el modelo no logra generalizar adecuadamente sobre este conjunto de datos, posiblemente debido a una señal de entrenamiento insuficiente o a diferencias sustanciales entre el corpus de origen y el de destino. La Tabla 4.8 presenta los valores registrados para cada métrica en los distintos checkpoints.

### 4.4.2. Corpus Datos Uruguayos con isignos

Aquí se entrena el modelo con el corpus *iSignos* y se evalúa en el conjunto *Datos Uruguayos*, que es considerablemente más pequeño y diverso. Este experimento busca explorar el desempeño del modelo al aplicarlo a un dominio con escasez de datos, evaluando su generalización y adaptación a contextos más limitados.

**Cuadro 4.9:** Resultados de métricas BLEU y CHRF para distintos checkpoints del experimento: Train iSignos evaluado con Datos Uruguay

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.0000 | 0.0000 | 0.0000 | 0.03 |
| 1000 | 0.00003 | 0.00002 | 0.00001 | 0.04 |
| 1500 | 0.0051 | 0.0035 | 0.0023 | 0.06 |
| 2000 | 0.0050 | 0.0034 | 0.0024 | 0.10 |
| 2500 | **0.0057** | **0.0039** | **0.0026** | **0.11** |
| 3000 | 0.0054 | 0.0036 | 0.0025 | **0.11** |
| 3500 | 0.0027 | 0.0020 | 0.0014 | 0.09 |
| 4000 | 0.0047 | 0.0032 | 0.0022 | 0.10 |
| 4500 | 0.0052 | 0.0035 | 0.0024 | 0.10 |
| 5000 | 0.0048 | 0.0033 | 0.0023 | **0.11** |
| 5500 | 0.0052 | 0.0035 | 0.0025 | 0.10 |
| 6000 | 0.0048 | 0.0032 | 0.0022 | 0.10 |
| 6500 | 0.0050 | 0.0034 | 0.0023 | 0.10 |
| 7000 | 0.0052 | 0.0036 | 0.0024 | 0.10 |
| 7500 | 0.0052 | 0.0035 | 0.0024 | 0.10 |
| 8000 | 0.0050 | 0.0034 | 0.0024 | 0.10 |
| 8500 | 0.0048 | 0.0033 | 0.0022 | **0.11** |
| 9000 | 0.0048 | 0.0033 | 0.0022 | **0.11** |
| 9500 | 0.0048 | 0.0033 | 0.0022 | **0.11** |
| 10000 | 0.0050 | 0.0034 | 0.0024 | **0.11** |

Al igual que en el experimento anterior, las métricas obtenidas en este caso se mantuvieron muy cercanas a cero, lo que indica nuevamente una baja calidad de las traducciones generadas por el modelo. La métrica CHRF alcanzó su valor máximo de 0.1103, siendo este el mejor desempeño observado en toda la serie, como se muestra en la Figura 4.8. A pesar de algunas oscilaciones, los valores

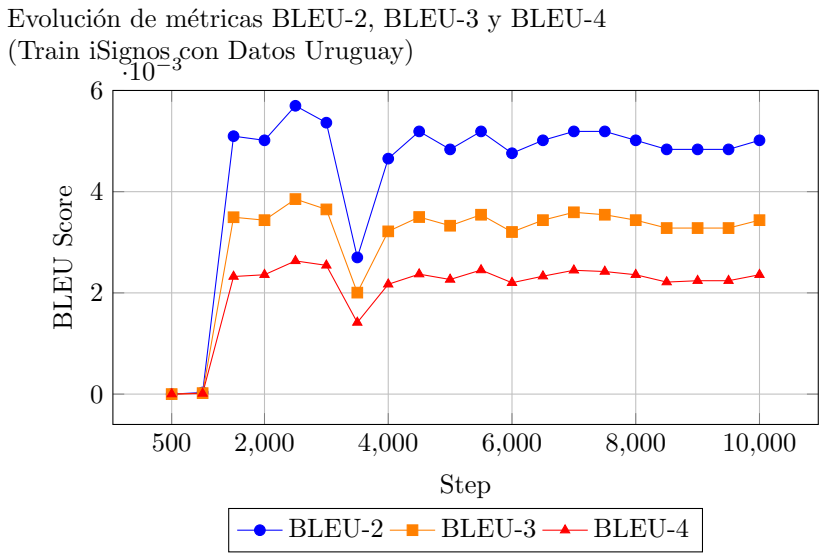


**Figura 4.7:** Evolución de las métricas BLEU-2, BLEU-3 y BLEU-4 durante el entrenamiento con el corpus *iSignos* evaluado sobre *Datos Uruguay*.

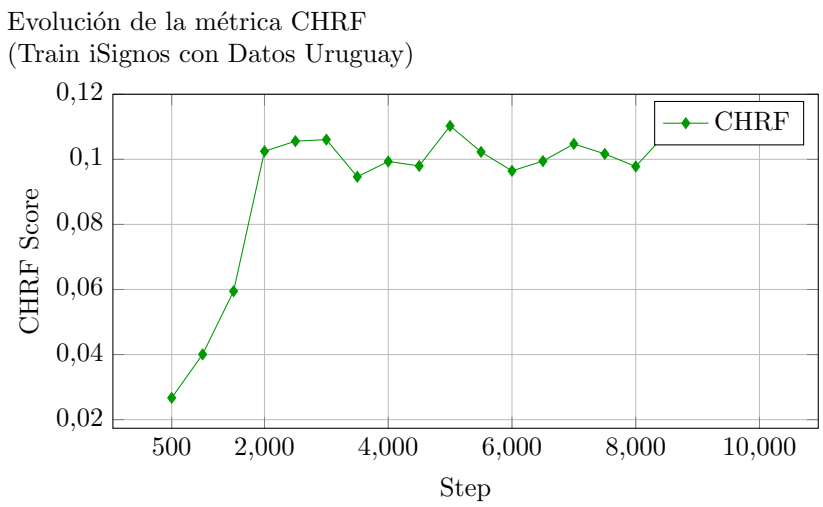


**Figura 4.8:** Evolución de la métrica CHRF durante el entrenamiento con el corpus *iSignos* evaluado sobre *Datos Uruguay*.

permanecieron en un rango bajo y sin una mejora progresiva clara. Por otro lado, las métricas BLEU-2, BLEU-3 y BLEU-4 también mostraron resultados muy bajos, con máximos de 0.0057, 0.0039 y 0.0026 respectivamente, según se observa en la Figura 4.7. Estos valores, además de encontrarse varios órdenes

de magnitud por debajo de 1, evidencian el escaso poder predictivo del modelo en este escenario. La Tabla 4.9 resume los valores registrados para cada métrica a lo largo de los distintos checkpoints del entrenamiento.

### 4.4.3. Corpus id/dl sumado a corpus isignos

En este caso, se entrena un modelo con la combinación de los corpus *iSignos* e *id/dl*, ambos relacionados con la Lengua de Señas Española (LSE). El propósito es investigar si la integración de ambos conjuntos de datos puede mejorar la calidad de la traducción en comparación con el entrenamiento individual, aprovechando la complementariedad y el mayor volumen de información disponible.

**Cuadro 4.10:** Resultados de métricas BLEU y CHRF para distintos checkpoints al combinar los corpus de iSignos e id/dl

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.0026 | 0.0009 | 0.0004 | 0.0630 |
| 1000 | 0.0511 | 0.0263 | 0.0160 | 0.1247 |
| 1500 | 0.0379 | 0.0243 | 0.0159 | 0.2156 |
| 2000 | 0.2001 | 0.1521 | 0.1175 | 0.2544 |
| 2500 | 0.2214 | 0.1589 | 0.1156 | 0.2693 |
| 3000 | 0.2351 | 0.1782 | 0.1391 | 0.2711 |
| 3500 | 0.2242 | 0.1714 | 0.1359 | 0.2693 |
| 4000 | 0.2319 | 0.1783 | 0.1398 | 0.2735 |
| 4500 | 0.2294 | 0.1730 | 0.1346 | 0.2733 |
| 5000 | 0.2297 | 0.1697 | 0.1298 | 0.2778 |
| 5500 | 0.2259 | 0.1709 | 0.1346 | 0.2727 |
| 6000 | 0.2280 | 0.1734 | 0.1356 | 0.2706 |
| 6500 | 0.2141 | 0.1562 | 0.1180 | 0.2667 |
| 7000 | 0.2336 | 0.1758 | 0.1366 | **0.2816** |
| 7500 | 0.2329 | 0.1772 | 0.1411 | 0.2766 |
| 8000 | 0.2305 | 0.1736 | 0.1355 | 0.2730 |
| 8500 | 0.2302 | 0.1741 | 0.1345 | 0.2761 |
| 9000 | 0.2292 | 0.1792 | 0.1443 | 0.2740 |
| 9500 | 0.2192 | 0.1651 | 0.1289 | 0.2722 |
| 10000 | **0.2386** | **0.1850** | **0.1483** | 0.2718 |

Si bien los resultados obtenidos en este experimento muestran una mejora respecto a los demás experimentos cruzados, siguen estando por debajo de los que fueron alcanzados con el corpus id/dl de manera individual. En ese caso, la métrica CHRF alcanzó un valor máximo de 0.5059 en el paso 1500 , mientras que en esta combinación apenas se llegó a 0.2816 en el paso 7000, como se muestra en la Figura 4.10. De manera similar, la métrica BLEU-4 en id/dl tuvo un pico de 0.3143 en el paso 3, mientras que aquí alcanzó solo 0.1483 al final del entrenamiento, según se observa en la Tabla 4.10 y la Figura 4.9.

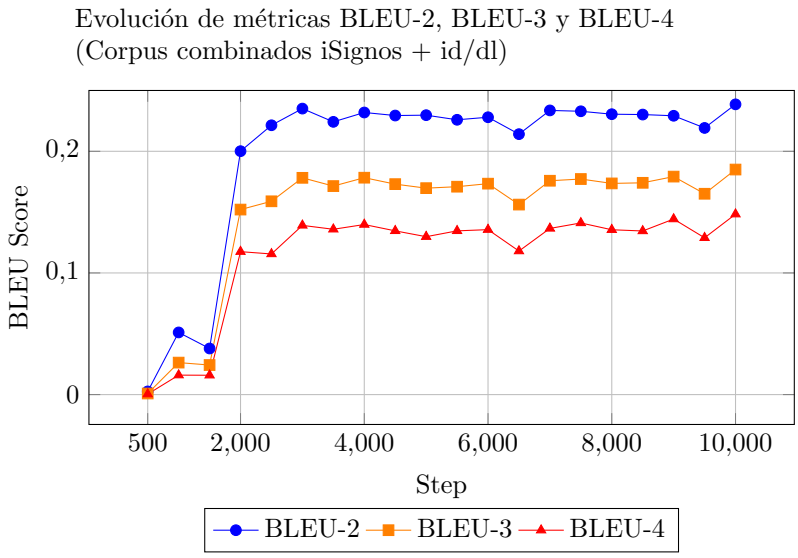


**Figura 4.9:** Evolución de métricas BLEU-2, BLEU-3 y BLEU-4 en Corpus combinados iSignos + id/dl).

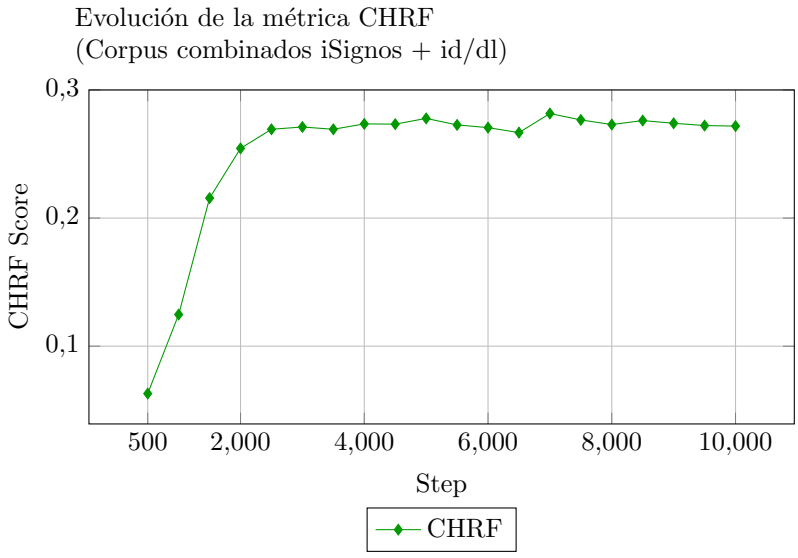


**Figura 4.10:** Evolución de CHRF (Corpus combinados iSignos + id/dl).

Esta disminución en el rendimiento puede deberse a que, al incorporar datos de un corpus más heterogéneo, el modelo enfrenta una mayor variabilidad en los patrones lingüísticos, lo cual dificulta la convergencia y reduce la precisión en la generación de traducciones, especialmente si el corpus adicional es de menor

calidad o más ruidoso.

### 4.4.4. Conclusiones experimentos cruzados

Los resultados del experimento en el que se utilizó *iSignos* como base para predecir sobre los corpus *id/dl* y *Datos Uruguay* muestran que las métricas CHRF y BLEU se mantuvieron muy cercanas a cero en ambos casos. Esto indica que, si bien el modelo fue entrenado con una gran cantidad de datos, no logró generalizar eficazmente hacia dominios o lenguas de señas diferentes. En particular, las diferencias temáticas entre *iSignos* e *id/dl*, así como el cambio de lengua de señas al predecir en *Datos Uruguay* (de LSE a LSU), parecen haber afectado negativamente la capacidad de transferencia del modelo.

Además, al comparar el rendimiento de un modelo entrenado únicamente con *id/dl* frente a otro entrenado con la combinación de *id/dl* e *iSignos*, se observa una disminución significativa en todas las métricas. Mientras que el modelo entrenado solo con *id/dl* alcanzó valores máximos de 0.5059 en CHRF, 0.3143 en BLEU-4, 0.3641 en BLEU-3 y 0.4218 en BLEU-2, el modelo combinado no superó los 0.2816 en CHRF, 0.1483 en BLEU-4, 0.1850 en BLEU-3 ni 0.2386 en BLEU-2. Esto sugiere que la inclusión de datos más heterogéneos no siempre beneficia el rendimiento y puede incluso introducir ruido o conflictos estilísticos que afectan negativamente la calidad de las predicciones.

A modo de resumen, la Tabla 4.11 presenta los mejores valores alcanzados por cada métrica en las distintas configuraciones evaluadas en estos experimentos cruzados.

**Cuadro 4.11:** Mejores resultados obtenidos para cada métrica y corpus

| Configuración | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| Train iSignos + test id/dl | 0.0008 | 0.0006 | 0.0004 | 0.0800 |
| Train iSignos + Test Datos Uruguayos | 0.0057 | 0.0039 | 0.0026 | 0.1100 |
| Train iSignos + id/dl | 0.2386 | 0.1850 | 0.1483 | 0.2816 |

## 4.5. Experimentos con Sobreciencia

### 4.5.1. Entrenar y evaluar con Sobreciencia

En este experimento, el modelo fue entrenado exclusivamente con el corpus Sobreciencia y evaluado en el conjunto de prueba correspondiente al mismo corpus. Este escenario sirve como una línea base para comparar el impacto de incorporar otros conjuntos de datos en los experimentos posteriores.

Los resultados presentados en la Tabla 4.12 permiten observar la evolución de las métricas de evaluación a lo largo del entrenamiento del modelo utilizando exclusivamente el corpus *Sobreciencia*. En la Figura 4.11, se muestra cómo las métricas BLEU-2, BLEU-3 y BLEU-4 experimentan un crecimiento inicial en las primeras etapas, con valores que se incrementan de forma sostenida hasta

**Cuadro 4.12:** Resultados de métricas BLEU y CHRF para los distintos checkpoints del experimento con el corpus del experimento entrenar y evaluar solo con Sobreciencia.

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.0028 | 0.0018 | 0.0012 | 0.1034 |
| 1000 | 0.0070 | 0.0041 | 0.0026 | 0.1262 |
| 1500 | 0.0125 | 0.0064 | 0.0037 | 0.1724 |
| 2000 | 0.0087 | 0.0049 | 0.0030 | 0.1587 |
| 2500 | 0.0184 | 0.0084 | 0.0046 | 0.1932 |
| 3000 | 0.0146 | 0.0072 | 0.0041 | 0.1890 |
| 3500 | 0.0146 | 0.0073 | 0.0041 | 0.1923 |
| 4000 | 0.0142 | 0.0071 | 0.0041 | 0.1874 |
| 4500 | 0.0142 | 0.0071 | 0.0040 | 0.1889 |
| 5000 | 0.0174 | 0.0081 | 0.0045 | 0.1973 |
| 5500 | 0.0174 | 0.0081 | 0.0045 | 0.1973 |
| 6000 | 0.0175 | 0.0081 | 0.0045 | 0.1969 |
| 6500 | 0.0182 | 0.0083 | 0.0046 | 0.2018 |
| 7000 | 0.0182 | **0.0084** | **0.0046** | **0.2026** |
| 7500 | 0.0182 | **0.0084** | **0.0046** | **0.2026** |
| 8000 | 0.0180 | 0.0083 | 0.0045 | 0.1989 |
| 8500 | 0.0176 | 0.0082 | 0.0045 | 0.1979 |
| 9000 | 0.0178 | 0.0082 | 0.0045 | 0.1954 |
| 9500 | 0.0179 | 0.0082 | 0.0045 | 0.1965 |
| 10000 | **0.0185** | **0.0084** | **0.0046** | 0.1978 |

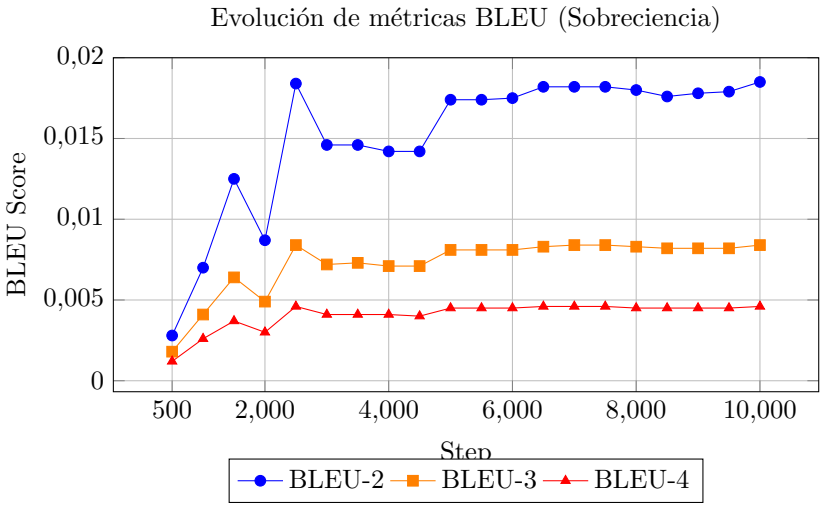


**Figura 4.11:** Evolución de las métricas BLEU-2, BLEU-3 y BLEU-4 durante el entrenamiento del modelo con el corpus *Sobreciencia*.

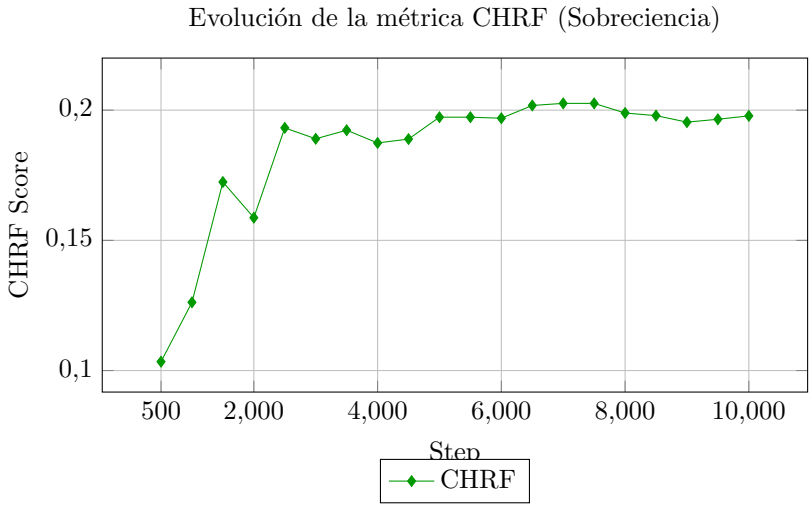


**Figura 4.12:** Evolución de la métrica CHRF durante el entrenamiento del modelo con el corpus *Sobreciencia*.

aproximadamente el paso 2500. A partir de ese punto, las curvas tienden a estabilizarse, con variaciones menores y un ritmo de mejora más moderado. BLEU-2 alcanza su valor más alto 0.0185 en el paso 10000, mientras que BLEU-3 y BLEU-4 alcanzan sus máximos, 0.0084 y 0.0046, respectivamente, en ese mismo paso. Cabe señalar que, si bien estas métricas muestran una evolución positiva a lo largo del entrenamiento, sus valores se mantienen bajos en términos absolutos. Por su parte, la Figura 4.12 ilustra la evolución de la métrica CHRF, que presenta una tendencia general de aumento, con su punto más alto (0.2026) registrado en los pasos 7000 y 7500. Al igual que con las métricas BLEU, se puede notar una cierta estabilización hacia las etapas finales del entrenamiento, con diferencias mínimas entre pasos consecutivos.

### 4.5.2. Combinación Sobreciencia e id/dl

En este experimento se combinó el corpus Sobreciencia con el corpus id/dl para el entrenamiento del modelo. La evaluación se realizó utilizando el conjunto de prueba de Sobreciencia. El objetivo fue observar si la inclusión de los datos de id/dl aporta mejoras al rendimiento del modelo al evaluar en el dominio original de Sobreciencia.

La Tabla 4.13 presenta los valores de las métricas BLEU y CHRF obtenidos en distintas etapas del entrenamiento. En la Figura 4.13, se aprecia un incremento temprano en las métricas BLEU-2, BLEU-3 y BLEU-4, particularmente hasta el paso 1500. Posteriormente, las curvas tienden a estabilizarse con leves fluctuaciones y sin grandes mejoras en las etapas finales. Los valores máximos se alcanzan en el paso 10000, con 0.0144 para BLEU-2, 0.0071 para BLEU-3

**Cuadro 4.13:** Resultados de métricas BLEU y CHRF para entrenamiento con Sobreciencia + id/dl y evaluación en Sobreciencia

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.0053 | 0.0035 | 0.0022 | 0.0850 |
| 1000 | 0.0076 | 0.0045 | 0.0028 | 0.1343 |
| 1500 | 0.0142 | 0.0070 | 0.0040 | 0.1641 |
| 2000 | 0.0136 | 0.0068 | 0.0039 | 0.1775 |
| 2500 | 0.0105 | 0.0057 | 0.0034 | 0.1528 |
| 3000 | 0.0109 | 0.0059 | 0.0035 | 0.1658 |
| 3500 | 0.0127 | 0.0066 | 0.0038 | 0.1745 |
| 4000 | 0.0131 | 0.0067 | 0.0039 | 0.1765 |
| 4500 | 0.0126 | 0.0065 | 0.0038 | 0.1732 |
| 5000 | 0.0134 | 0.0068 | 0.0039 | 0.1795 |
| 5500 | 0.0126 | 0.0065 | 0.0038 | 0.1774 |
| 6000 | 0.0126 | 0.0065 | 0.0038 | 0.1763 |
| 6500 | 0.0134 | 0.0068 | 0.0039 | 0.1783 |
| 7000 | 0.0142 | 0.0070 | **0.0040** | 0.1713 |
| 7500 | 0.0134 | 0.0067 | 0.0039 | 0.1788 |
| 8000 | 0.0135 | 0.0068 | 0.0039 | 0.1798 |
| 8500 | 0.0128 | 0.0065 | 0.0038 | 0.1792 |
| 9000 | 0.0140 | 0.0069 | **0.0040** | **0.1844** |
| 9500 | 0.0134 | 0.0067 | 0.0039 | 0.1782 |
| 10000 | **0.0144** | **0.0071** | **0.0040** | 0.1842 |

y 0.0040 para BLEU-4. Al igual que en el experimento anterior, estos valores siguen siendo bajos en términos absolutos. En cuanto al comportamiento de CHRF, ilustrado en la Figura 4.14, se observa un crecimiento inicial pronunciado hasta aproximadamente el paso 2000, momento a partir del cual la métrica alcanza una meseta. A lo largo de los pasos restantes, los valores se mantienen dentro de un rango estrecho, con su punto más alto (0.1844) registrado en el paso 9000, lo que sugiere una estabilización del rendimiento en las etapas finales del entrenamiento.

### 4.5.3. Combinación Sobreciencia y Datos Uruguay

En este experimento se combinó el corpus *Sobreciencia* con el corpus *Datos Uruguay*, ambos corpus basados en LSU, para el entrenamiento del modelo. La evaluación se llevó a cabo en el conjunto de prueba de *Sobreciencia*. Este enfoque busca analizar si la incorporación de datos de un corpus distinto pero cercano mejora la generalización del modelo en el dominio objetivo.

Los resultados que se presentan en la Tabla 4.14 corresponden al experimento en el que el modelo fue entrenado con una combinación de los corpus *Sobreciencia* y *Datos Uruguay*, y evaluado exclusivamente en el conjunto de prueba de

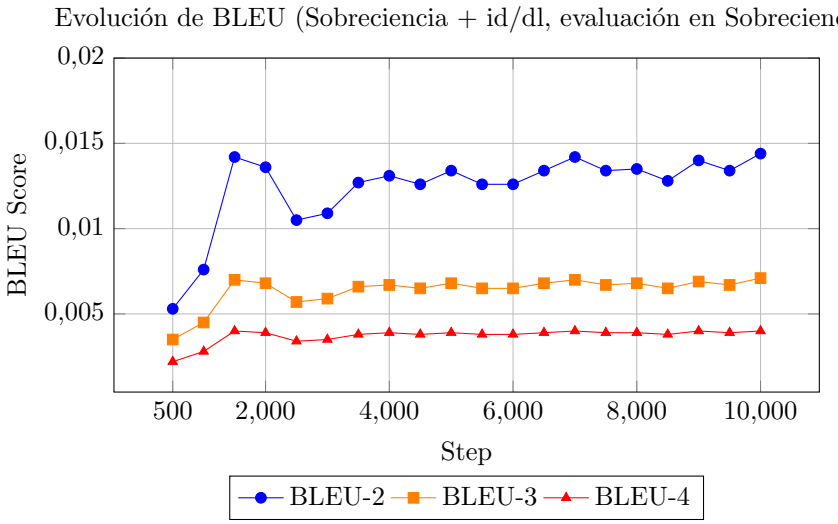


**Figura 4.13:** Evolución de las métricas BLEU-2, BLEU-3 y BLEU-4 durante el entrenamiento con Sobreciencia + id/dl y evaluación en Sobreciencia.

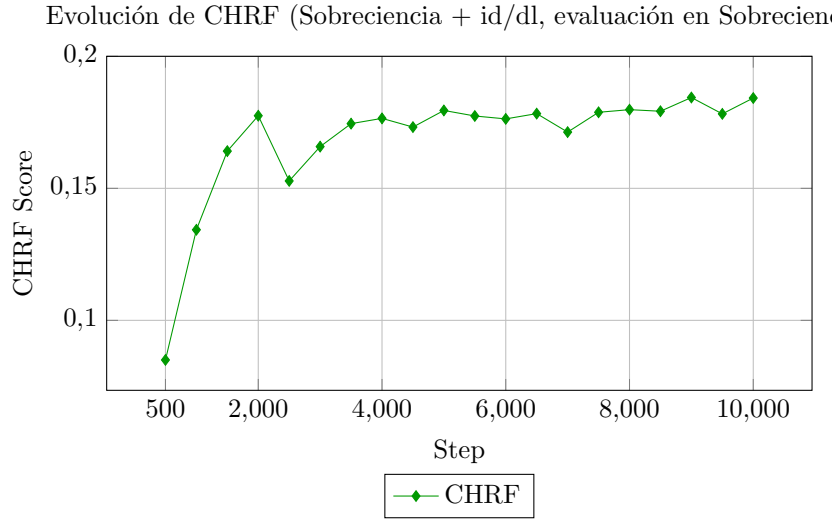


**Figura 4.14:** Evolución de la métrica CHRF durante el entrenamiento con Sobreciencia + id/dl y evaluación en Sobreciencia.

*Sobreciencia.* Las Figuras 4.15 y 4.16 ilustran la evolución de las métricas BLEU y CHRF a lo largo del entrenamiento.

En el caso de BLEU, las tres variantes alcanzan sus valores máximos al final del proceso (paso 10000), con valores de 0.0144, 0.0071 y 0.0040, respectivamente. Las curvas muestran un crecimiento inicial hasta aproximadamente el

**Cuadro 4.14:** Resultados de métricas BLEU y CHRF para entrenamiento con Sobreciencia + Datos Uruguay y evaluación en Sobreciencia

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 0.0053 | 0.0035 | 0.0022 | 0.0850 |
| 1000 | 0.0076 | 0.0045 | 0.0028 | 0.1343 |
| 1500 | **0.0142** | **0.0070** | **0.0040** | 0.1641 |
| 2000 | 0.0136 | 0.0068 | 0.0039 | 0.1775 |
| 2500 | 0.0105 | 0.0057 | 0.0034 | 0.1528 |
| 3000 | 0.0109 | 0.0059 | 0.0035 | 0.1658 |
| 3500 | 0.0127 | 0.0066 | 0.0038 | 0.1745 |
| 4000 | 0.0131 | 0.0067 | 0.0039 | 0.1765 |
| 4500 | 0.0126 | 0.0065 | 0.0038 | 0.1732 |
| 5000 | 0.0134 | 0.0068 | 0.0039 | 0.1795 |
| 5500 | 0.0126 | 0.0065 | 0.0038 | 0.1774 |
| 6000 | 0.0126 | 0.0065 | 0.0038 | 0.1763 |
| 6500 | 0.0134 | 0.0068 | 0.0039 | 0.1783 |
| 7000 | **0.0142** | **0.0070** | **0.0040** | 0.1713 |
| 7500 | 0.0134 | 0.0067 | 0.0039 | 0.1788 |
| 8000 | 0.0135 | 0.0068 | 0.0039 | 0.1798 |
| 8500 | 0.0128 | 0.0065 | 0.0038 | 0.1792 |
| 9000 | 0.0140 | 0.0069 | **0.0040** | **0.1844** |
| 9500 | 0.0134 | 0.0067 | 0.0039 | 0.1782 |
| 10000 | **0.0144** | **0.0071** | **0.0040** | 0.1842 |

paso 1500, seguido de fluctuaciones moderadas que sugieren cierta estabilización, aunque sin una tendencia de mejora sostenida. Por su parte, la métrica CHRF presenta una trayectoria ascendente más clara, con su valor más alto (0.1844) en el paso 9000, y variaciones leves hacia el final. Si bien se observa una mejora general a lo largo del entrenamiento, los valores de BLEU continúan siendo bajos en términos absolutos.

### 4.5.4. Combinación de Sobreciencia, id/dl, Datos Uruguay e iSignos

En este experimento se utilizó una combinación de los cuatro corpus disponibles: Sobreciencia, id/dl, Datos Uruguay e iSignos. El modelo fue entrenado con esta combinación heterogénea y evaluado en el conjunto de prueba de Sobreciencia. El propósito fue explorar si un entrenamiento con datos variados y de diferentes dominios contribuye a un mejor desempeño al evaluar específicamente en el contexto de Sobreciencia.

Las métricas BLEU, ilustradas en la Figura 4.17, presentan una mejora notable en las primeras etapas del entrenamiento, con un crecimiento más claro hasta el paso 2500. A partir de ese punto, las curvas muestran cierta inestabili-

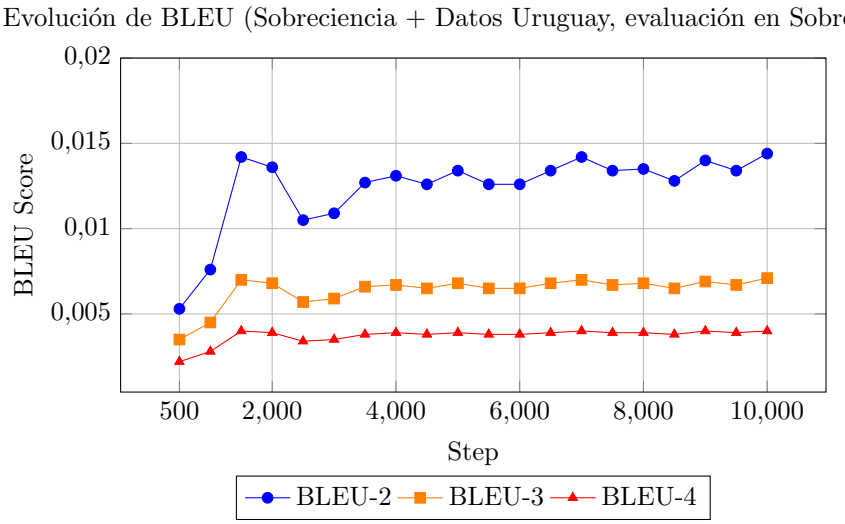


**Figura 4.15:** Evolución de las métricas BLEU-2, BLEU-3 y BLEU-4 durante el entrenamiento con *Sobreciencia + Datos Uruguay* y evaluación en *Sobreciencia*.

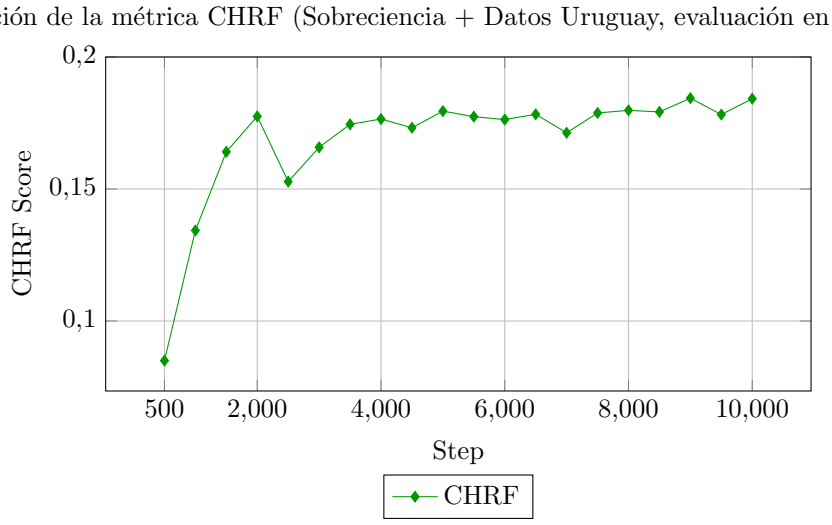


**Figura 4.16:** Evolución de CHRF durante el entrenamiento con *Sobreciencia + Datos Uruguay* y evaluación en *Sobreciencia*.

dad y oscilaciones. BLEU-2 alcanza su valor máximo (0.0172) en el paso 9500, al igual que BLEU-3 (0.0080) y BLEU-4 (0.0044), aunque este comportamiento no se sostiene hacia el final del entrenamiento. Los valores descienden en el paso 10000, indicando posibles dificultades en mantener el rendimiento. Al igual que en experimentos anteriores, los valores de BLEU continúan siendo bajos en

**Cuadro 4.15:** Resultados métricas BLEU y CHRF para entrenamiento con Sobreciencia + id/dl + Datos Uruguay + iSignos, evaluación en Sobreciencia

| Step | BLEU-2 | BLEU-3 | BLEU-4 | CHRF |
|---|---|---|---|---|
| 500 | 4.52e-07 | 3.27e-07 | 2.19e-07 | 0.0121 |
| 1000 | 2.20e-06 | 1.64e-06 | 1.12e-06 | 0.0348 |
| 1500 | 0.0040 | 0.0019 | 0.0010 | 0.1019 |
| 2000 | 0.0059 | 0.0030 | 0.0017 | 0.1344 |
| 2500 | 0.0078 | 0.0036 | 0.0019 | 0.1533 |
| 3000 | 0.0039 | 0.0020 | 0.0012 | 0.0969 |
| 3500 | 0.0077 | 0.0037 | 0.0020 | 0.1407 |
| 4000 | 0.0093 | 0.0042 | 0.0023 | 0.1559 |
| 4500 | 0.0082 | 0.0038 | 0.0021 | 0.1431 |
| 5000 | 0.0135 | 0.0068 | 0.0039 | 0.1739 |
| 5500 | 0.0126 | 0.0065 | 0.0038 | 0.1593 |
| 6000 | 0.0078 | 0.0037 | 0.0020 | 0.1505 |
| 6500 | 0.0119 | 0.0062 | 0.0036 | 0.1611 |
| 7000 | 0.0112 | 0.0060 | 0.0035 | 0.1599 |
| 7500 | 0.0086 | 0.0040 | 0.0022 | 0.1566 |
| 8000 | 0.0115 | 0.0062 | 0.0036 | 0.1607 |
| 8500 | 0.0102 | 0.0056 | 0.0033 | 0.1672 |
| 9000 | 0.0093 | 0.0042 | 0.0023 | **0.1812** |
| 9500 | **0.0172** | **0.0080** | **0.0044** | 0.1762 |
| 10000 | 0.0073 | 0.0036 | 0.0020 | 0.1360 |

términos absolutos. Estos resultados se detallan en la Tabla 4.15.

La Figura 4.18 muestra la evolución de la métrica CHRF, que exhibe un ascenso sostenido hasta aproximadamente el paso 9000, donde alcanza su valor más alto (0.1812). Sin embargo, este progreso se ve interrumpido por una caída abrupta en el paso final (10000), lo cual podría reflejar una pérdida de generalización. A lo largo del entrenamiento, la métrica también presenta oscilaciones que indican una dinámica menos estable en comparación con otros escenarios. La evolución completa de esta métrica también está incluida en la Tabla 4.15.

### 4.5.5. Conclusiones de experimentos con el corpus Sobreciencia

Los experimentos realizados con diferentes combinaciones de corpus para entrenar el modelo muestran que el rendimiento varía significativamente según la naturaleza y cantidad de los datos utilizados. Cuando el modelo fue entrenado exclusivamente con el corpus Sobreciencia, se observó una evolución positiva de las métricas durante las etapas iniciales del entrenamiento, con una posterior estabilización. El mejor resultado alcanzado fue 0.0185 en BLEU-2 y 0.2026 en CHRF. Estos valores, aunque bajos, indican que el modelo fue capaz de aprender

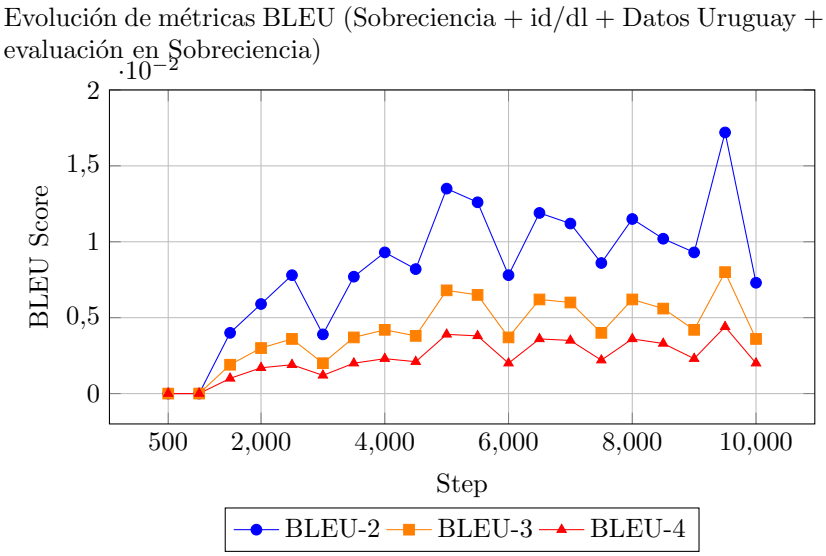


**Figura 4.17:** Evolución de BLEU-2, BLEU-3 y BLEU-4 durante el entrenamiento con Sobreciencia + id/dl + Datos Uruguay + iSignos y evaluación en Sobreciencia.

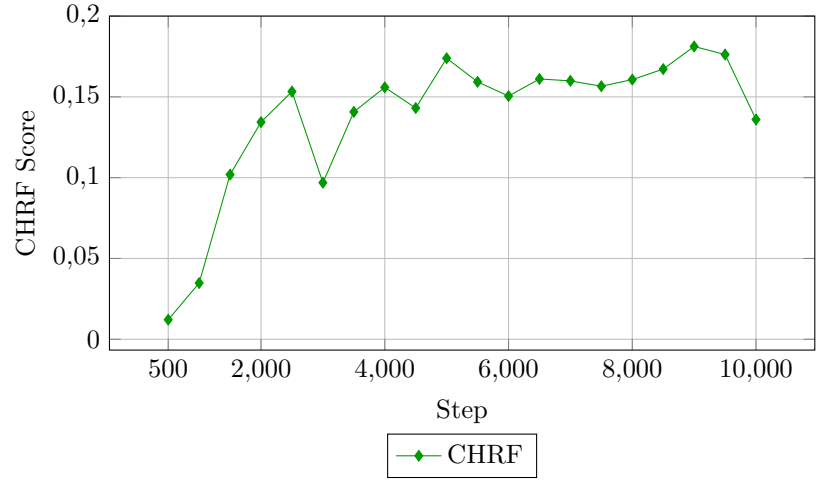


**Figura 4.18:** Evolución de la métrica CHRF durante el entrenamiento con Sobreciencia + id/dl + Datos Uruguay + iSignos y evaluación en Sobreciencia.

parcialmente el dominio específico del corpus, especialmente en lo que respecta a coincidencias léxicas a nivel de caracteres.

La incorporación del corpus id/dl al entrenamiento no logró mejoras sustan-

ciales respecto al uso exclusivo de Sobreciencia. Las métricas BLEU-2, BLEU-3 y BLEU-4 mantuvieron valores similares, y aunque CHRF alcanzó un valor máximo de 0.1844, este no superó el logrado en el primer experimento. Esto sugiere que el agregado de id/dl no aportó información significativamente útil para mejorar el rendimiento en el dominio de Sobreciencia, probablemente debido a una diferencia temática o estilística entre ambos corpus.

De forma análoga, al combinar Sobreciencia con Datos Uruguay, los resultados se mantuvieron prácticamente idénticos a los del caso anterior, sin mejoras notables. Esto refuerza la hipótesis de que, si bien estos corpus comparten características estructurales, su contenido específico no proporciona información adicional relevante para el dominio de evaluación.

Finalmente, el experimento con la combinación completa de los cuatro corpus (Sobreciencia, id/dl, Datos Uruguay e iSignos) presentó un comportamiento más errático. Si bien en ciertas etapas se observaron mejoras parciales, el rendimiento no se sostuvo hasta el final del entrenamiento, y las métricas incluso descendieron en la última etapa. Este comportamiento sugiere que la inclusión de datos más heterogéneos puede introducir ruido o desalineación respecto al dominio objetivo, dificultando la generalización específica a Sobreciencia.

En todos los escenarios evaluados, BLEU-2 fue sistemáticamente la métrica con los valores más altos dentro de las métricas BLEU, lo que confirma que el modelo logra aprender patrones cortos (bigramas) con relativa facilidad. BLEU-3 y BLEU-4, en cambio, reportaron valores significativamente menores, reflejando las dificultades del modelo para capturar dependencias más largas en secuencias de texto.

La métrica CHRF, que evalúa la similitud a nivel de caracteres, presentó un comportamiento más robusto. En general, sus valores fueron más altos y más estables que los de las métricas BLEU, especialmente hacia las etapas finales del entrenamiento. Esto sugiere que, si bien el modelo puede no generar secuencias exactas, sí logra mantener cierta coherencia léxica y estructural en sus traducciones.

La Tabla 4.16 resume los mejores resultados obtenidos para cada configuración de entrenamiento al evaluar sobre el corpus Sobreciencia. Se evidencia que ninguna de las combinaciones probadas logró mejorar los resultados obtenidos con el entrenamiento exclusivo en Sobreciencia; por el contrario, en todos los casos las métricas empeoraron levemente, especialmente en CHRF.

**Cuadro 4.16:** Mejores resultados por métrica con evaluación en Sobreciencia

| **Entrenamiento** | **BLEU-2** | **BLEU-3** | **BLEU-4** | **CHRF** |
|---|---|---|---|---|
| *Sobreciencia* | **0.0185** | **0.0084** | **0.0046** | **0.2026** |
| *Sobreciencia + id/dl* | 0.0144 | 0.0071 | 0.0040 | 0.1844 |
| *Sobreciencia + Datos Uruguay* | 0.0144 | 0.0071 | 0.0040 | 0.1844 |
| *Todos los corpus* | 0.0172 | 0.0080 | 0.0044 | 0.1812 |

En términos generales, estos resultados sugieren que, para maximizar el ren-

dimiento en un dominio específico, entrenar exclusivamente con datos de ese mismo dominio puede ser más efectivo que incorporar información de otros corpus. La inclusión de datos adicionales no garantiza mejoras, especialmente si los nuevos datos no guardan una relación estrecha con el conjunto de prueba. Además, se destaca la necesidad de explorar técnicas que permitan al modelo mejorar su rendimiento en la generación de secuencias más largas, dado que las métricas BLEU-3 y BLEU-4 siguen siendo las más difíciles de optimizar.

# Capítulo 5

# Conclusiones y Trabajo Futuro

En este capítulo se evalúan los resultados alcanzados, las dificultades encontradas y se contrastan los objetivos planteados con los logros obtenidos. Además, se destacan los aportes de esta investigación, se proponen posibles extensiones para trabajos futuros y se realiza una autocrítica sobre los aspectos que pudieron mejorarse o profundizarse.

## 5.1. Evaluación de Resultados y Aportes

El presente trabajo logró desarrollar una parte funcional del pipeline para la traducción automática del español hablado en Uruguay a glosas correspondientes a la Lengua de Señas Uruguaya (LSU). En una primera etapa, se construyó un conjunto curado manualmente de aproximadamente 238 minutos y 48 segundos , con 88 hablantes diferentes , de audio en español oral uruguayo, sobre el cual se realizaron experimentos de transcripción automática y validación de los resultados con dos sistemas distintos, complementados con correcciones manuales para garantizar la calidad de los datos.

En una segunda etapa, se llevaron a cabo diversos experimentos con combinaciones de corpus locales, evaluando el desempeño del modelo en distintos escenarios.

Por un lado, se entrenaron y evaluaron modelos sobre un mismo corpus —ya sea *id/dl*, *iSignos*, *Datos Uruguay* o *Sobreciencia*— con el fin de observar su comportamiento en contextos consistentes. Por otro lado, se diseñaron experimentos cruzados para analizar la capacidad de generalización de los modelos al enfrentarse a ejemplos fuera de su dominio de entrenamiento. En este sentido, se utilizó *iSignos* como base de entrenamiento, dado que es el corpus más numeroso, y se evaluó sobre *id/dl* y *Datos Uruguay*. También se experimentó con la combinación de *iSignos* e *id/dl* como corpus de entrenamiento.

Finalmente, dado que *Sobreciencia* constituye el corpus más extenso de glosas en LSU utilizado en este estudio, se exploraron distintas combinaciones de entrenamiento incluyendo este recurso. Se entrenaron modelos combinando *Sobreciencia* con *id/dl*, con *Datos Uruguay*, y con los tres corpus restantes (*id/dl*, *Datos Uruguay* e *iSignos*), evaluando siempre sobre el conjunto de prueba de *Sobreciencia*. Esta variedad de configuraciones permitió analizar cómo influye la procedencia y diversidad del corpus en la calidad de la traducción generada.

Sin embargo, los resultados también evidencian varias limitaciones relevantes. En primer lugar, la disponibilidad y tamaño reducido de corpus paralelos entre español y LSU representó una barrera significativa para lograr una generalización robusta. El modelo logró captar ciertos patrones útiles, pero su rendimiento sigue estando condicionado por la cantidad, calidad y coherencia de los datos de entrenamiento.

Además, si bien las métricas tradicionales como BLEU y CHRF resultan útiles para obtener una primera aproximación cuantitativa del desempeño del modelo, es posible que presenten limitaciones cuando se trabaja con lenguas de señas. Estas métricas fueron diseñadas para evaluar traducciones en lenguas orales y escritas, por lo que podrían no capturar adecuadamente aspectos semánticos, gramaticales o pragmáticos específicos de las lenguas señadas. En particular, no consideran elementos clave como la estructura visual-espacial, el orden flexible o las expresiones no manuales, que son fundamentales en la LSU. Por ello, si bien los puntajes obtenidos ofrecen cierta orientación, es necesario complementar estos análisis con métodos cualitativos o con métricas más adaptadas a este tipo de lenguas.

Otro aspecto importante es la ausencia de una evaluación cualitativa con usuarios finales, intérpretes profesionales o miembros de la comunidad sorda. Esta validación práctica es fundamental para determinar la utilidad real del sistema en contextos educativos, informativos o sociales.

En resumen, el trabajo aporta un punto de partida valioso para la exploración de traducción automática hacia LSU, especialmente en el ámbito local, pero también señala claramente los desafíos técnicos, metodológicos y de evaluación que deberán abordarse en futuras investigaciones.

## 5.2. Trabajo Futuro

Para continuar y expandir este proyecto, se proponen las siguientes líneas de trabajo:

- Realizar fine-tuning de modelos de traducción preentrenados (por ejemplo, OpenNMT) con corpus locales de LSU, incluyendo además ajustes de hiperparámetros como la cantidad y tamaño de capas, y la densidad del dropout, con el objetivo de mejorar el desempeño en este dominio específico.

- Incorporar evaluación humana mediante pruebas con hablantes e intérpretes de LSU que permitan validar y ajustar la calidad de las traducciones

generadas.

- Desarrollar un avatar animado que traduzca glosas a señas visuales utilizando tecnologías de captura y animación 3D, apoyándose en lenguajes basados en XML para describir movimientos de señas, como SiGML (Bruce, Marshall, y Johnston, 2000), y empleando herramientas como OpenPose o motores gráficos como Unity y Blender.
- Ampliar el corpus paralelo español–LSU a través de colaboraciones con comunidades y organizaciones sordas en Uruguay.
- Explorar la posibilidad de generar datos sintéticos paralelos español–LSU combinando enfoques basados en reglas y aprendizaje automático, como se propone en (Chiruzzo, McGill, Egea-Gómez, y Saggion, 2022), lo cual requeriría contar con una descripción gramatical general del LSU.
- Explorar la traducción inversa de LSU a español para facilitar una comunicación bidireccional efectiva.
- Investigar modelos multimodales que integren audio y video para mejorar la interpretación y contextualización del habla. Esto implica desarrollar sistemas que combinen la información sonora, como las palabras y características del habla capturadas por micrófonos, con datos visuales, tales como movimientos de labios, expresiones faciales y gestos, para aportar un contexto más completo. La integración de ambas modalidades permite una comprensión más precisa del mensaje, mejora el reconocimiento en ambientes ruidosos y facilita la interpretación de intenciones y emociones del hablante (Afouras, Chung, Senior, Vinyals, y Zisserman, 2018; Chung, Senior, Vinyals, y Zisserman, 2017).
- Diseñar interfaces accesibles y en tiempo real para aplicaciones prácticas en espacios públicos o educativos.
- Realizar análisis lingüísticos automáticos de la LSU para incorporar estructuras gramaticales propias en el modelo de traducción.
- Aprovechar técnicas de transferencia de aprendizaje a partir de lenguas de señas con mayor disponibilidad de datos, como la Lengua de Señas Americana o la Lengua de Señas Francesa.

En síntesis, esta investigación sienta las bases para el desarrollo de tecnologías accesibles que contribuyan a la inclusión de las personas sordas en Uruguay, abriendo el camino a futuros avances en traducción automática y representación visual de la Lengua de Señas Uruguaya.

Además se presentaron los primeros resultados obtenidos en la tarea de traducción automática del español hacia glosas de LSU. Aunque el rendimiento medido por las métricas aún no es suficientemente alto para un uso práctico, este avance constituye un paso inicial importante en una línea de investigación poco explorada. Se espera que este trabajo sirva como base para futuros estudios

que mejoren los modelos y contribuyan al desarrollo de tecnologías más precisas y accesibles para la comunidad sorda.

# Referencias


Afouras, T., Chung, J. S., Senior, A., Vinyals, O., y Zisserman, A. (2018). Audio-visual speech recognition with a hybrid ctc/attention architecture. *arXiv preprint arXiv:1809.02108*. Descargado de https://arxiv.org/abs/1809.02108

Anthony Zhang et al. (2025). *Speechrecognition — python library for performing speech recognition.* https://pypi.org/project/SpeechRecognition/. (Accessed: 2025-06-17)

Bird, S., Klein, E., y Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit.* O'Reilly Media, Inc.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bruce, S., Marshall, I., y Johnston, T. (2000). Sigml: An xml-based signing gesture markup language. En *Proceedings of the fourth international acm conference on assistive technologies (assets 2000)* (pp. 205–212). ACM. Descargado de https://doi.org/10.1145/354324.354366 doi: 10.1145/354324.354366

Caballero, M., Mariño, J. B., y Moreno, A. (2002). Multidialectal spanish modeling for asr. En *Lrec.*

Cabeza, C., y García-Miguel, J. M. (2019). *iSignos: Interfaz de datos de Lengua de Signos Española (versión 1.0).* Universidade de Vigo. Descargado de http://isignos.uvigo.es/

Chiruzzo, L., McGill, E., Egea-Gómez, S., y Saggion, H. (2022). Translating spanish into spanish sign language: Combining rules and data-driven approaches. En *Proceedings of the fifth workshop on technologies for machine translation of low-resource languages (loresmt 2022)* (pp. 75–83).

Chung, J. S., Senior, A., Vinyals, O., y Zisserman, A. (2017). Deep audio-visual speech recognition. En *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 6450–6454). Descargado de https://ieeexplore.ieee.org/document/7952154 doi: 10.1109/ICASSP.2017.7952154

Fojo, A., González, A., y Tancredi, M. (2013). Variación sintáctica de la lengua de señas uruguaya y su vinculación con los procesos de estandarización. En *Actas del vi encuentro internacional de investigadores de políticas*

*lingüísticas, núcleo de educación para la integración, augm.* Universidad Federal de Rio Grande del Sur, Brasil.

Fojo, A., y Tancredi, M. (2015). Continuación del estudio de la variación sintáctica de la lengua de señas uruguaya y su vinculación con los procesos de estandarización. En *Actas del vii encuentro internacional de investigadores de políticas lingüísticas, núcleo de educación para la integración, augm.* Universidad Nacional de Córdoba, Argentina.

Fojo, A., Tancredi, M., Etcheverry, F., Pintos, A., Pérez, B., Bianchi, J., ... Chiruzzo, L. (2024). Corpus de lengua de señas uruguaya y desarrollo de metadatos. En *Congresso internacional de pesquisas em línguística de línguas de sinais.*

Fondo de Población de las Naciones Unidas (UNFPA). (2011). *Situación de las personas sordas en uruguay.* https://uruguay.unfpa.org/es/noticias/j%C3%B3venes-sordos-prevenci%C3%B3n-en-su-propio-lenguaje. (Accedido: 2025-07-18)

Goodfellow, I., Bengio, Y., y Courville, A. (2016). *Deep learning.* MIT Press. Descargado de https://www.deeplearningbook.org/

Hochreiter, S., y Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, *15*(1), 104–129.

Jurafsky, D., y Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Pearson Prentice Hall.

Klein, G., Kim, Y., Deng, Y., Senellart, J., y Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. En *Proceedings of acl 2017, system demonstrations* (pp. 67–72). Association for Computational Linguistics. Descargado de https://aclanthology.org/P17-4012

Morris, A., Maier, V., y Green, P. (2004). From wer and ril to mer and wil: Improved evaluation measures for connected speech recognition. En *Proceedings of the international conference on spoken language processing (icslp)* (pp. 2761–2764).

Moryossef, A., Müller, M., Göhring, A., Jiang, Z., Goldberg, Y., y Ebling, S. (2023). An open-source gloss-based baseline for spoken to signed language translation. *arXiv preprint arXiv:2305.17714*. Descargado de https://arxiv.org/abs/2305.17714

Papineni, K., Roukos, S., Ward, T., y Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the association for computational linguistics (acl)* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. Descargado de https://aclanthology.org/P02-1040

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. En *Proceedings of the tenth workshop on statistical machine translation* (pp. 392–395). Lisbon, Portugal: Association for Computational Linguistics. Descargado de https://aclanthology.org/W15-3049

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., y Henning, J. (1989). *Hamnosys: The hamburg notation system for sign languages: An introductory guide*. Signum Press.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., y Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision.* Descargado de https://openai.com/research/whisper (Último acceso: 17 de junio de 2025)

San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L. F., Fernández, F., Ferreiros, J., ... Pardo, J. M. (2008). Speech to sign language translation system for spanish. *Speech Communication*, *50*(11-12), 1009–1020.

Stassi, A. E., Tancredi, M., Aguirre, R., Gómez, A., Carballido, B., Méndez, A., ... Randall, G. (2022). Lsu-ds: An uruguayan sign language public dataset for automatic recognition. En *Icpram* (pp. 697–705).

Stokoe, W. C. (2005). *Sign language structure: An outline of the visual communication systems of the american deaf.* Journal of Deaf Studies and Deaf Education. (Reedición del trabajo seminal de 1960)

Sutton, V. (1995). Signwriting: Proposed solution for a writing system for deaf sign languages. *The Sign Language Translator and Interpreter*, *1*, 59–82.

Touvron, H., Martin, L., Stone, K., y cols. (2023). *Llama 2: Open foundation and fine-tuned chat models.* https://arxiv.org/abs/2307.09288. (arXiv:2307.09288)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. En *Advances in neural information processing systems (neurips)* (pp. 5998–6008). Descargado de https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf