Sesión 1

**Explorando el campo de la lingüística de corpus: una perspectiva histórica y metodológica**

Esta clase introduce a los participantes en los conceptos fundamentales de la lingüística de corpus. Se explora la historia del campo, desde sus orígenes hasta sus aplicaciones contemporáneas, con un enfoque en cómo se construyen y utilizan los corpus. Se presentará terminología clave, como la distinción entre tipo/token, y los principios metodológicos subyacentes. La sesión también analiza cómo la lingüística de corpus trasciende los límites lingüísticos, aportando perspectivas a diversas disciplinas, incluidas los estudios culturales y las humanidades digitales. Al finalizar, los participantes tendrán una sólida base conceptual en el campo y su relevancia para múltiples áreas de investigación.

Sesión 2

**Construcción y anotación de corpus: una introducción práctica con INCEpTION**

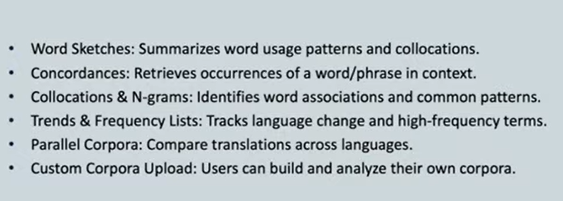
**Presentación clase**: Esta clase práctica demuestra cómo construir, anotar y analizar un corpus lingüístico, incluso para quienes no tienen conocimientos de programación. Utilizando la herramienta de anotación INCEpTION como ejemplo, los participantes aprenderán el proceso paso a paso para crear un corpus representativo, anotar manualmente características lingüísticas y generar análisis significativos. La clase incluye una demostración en vivo de INCEpTION, una introducción a la visualización básica de datos con R y orientación sobre cómo transformar anotaciones en recursos en línea. Al final, los estudiantes adquirirán habilidades prácticas e ideas para aplicar esta metodología en sus propios proyectos lingüísticos o interdisciplinarios.

Sketch Engine. Herramienta de anotación de corpus. Reúne varios corpus, incluso el corpus más grande del español: EsTenTen23.

Andrea menciona que una diferencia que considera importante respecto a INCEpTION es que Sketch Engine no permite anotaciones.

Algunas de sus key features:

* Resume patrones de palabras
* Concordancia.



Hoy veremos sobre cómo construir nuestro propio corpus.

INCEpTION: https://inception-project.github.io/

Es una herramienta de anotación.

Corpus general.

De código abierto y basado en web.

Es customizable.

Por qué usar INCEpTION?

* De código abierto y basado en web, no requiere programación.
* Anotación linguistica a varios niveles.

Características principales:

* Permite cargar texto no anotado para su posterior anotación.
* Recomienda usar texto en formato plain text. (.txt)
* Se pueden organizar perfiles en un proyecto, restringiendo las actividades a un grupo por perfil.
* Layers: Capas de anotación. Vienen algunos por defecto.
* Tagsets.
* Ejemplo: Creación de una nueva capa llamada Key Concept. Cambia su granularidad a token-level. Permite que haya overlap. Se crea el tagset relacionado a la capa Key Concept. Y lemma, part of speech y key concept. Con lemma puso las palabras en singular. En part of speech etiquetó el tipo de palabra: sustantivo, adverbio, etc. La capa *dependency* permite relacionar, por ejemplo, como un verbo se refiere a un sustantivo. Con key concept podemos anotar una frase completa. La herramienta tambien permite exportar las anotaciones.
* Con LLMs se espera llegar a anotaciones automáticas. Ejemplo con ChatGPT. Este genera texto basado en prompts. Este texto es similar al texto humano. Sobre la ingeniería de prompts. Estos son puntos de partida del modelo. Permite asistir en lematización, en taggeo de part of speech. Limitaciones de ChatGPT: Dificultad de interpretación de un contexto amplio, genera inconsistencias ya que repetir el prompt no da el mismo resultado, validación humana requerida. Nos muestra algunas de las buenas prácticas de uso como: prompts estructurados y claros, testear con muestras pequeñas del corpus, dar ejemplos de anotaciones para que el modelo los imite, combinar con verificación manual, ser especifico y dar contexto,
* Visualización de datos: Ejemplo.
* Necesitamos visualización para hacer más entendible data compleja, para reconocer patrones, para comunicar los hallazgos.
* Versus entre tablas y gráficos. Tablas para mostrar datos exactos. Gráficos para resumir puntos claves de la data.
* Algunos tipos de gráficos: Gráfico de barras. Se usa para comparar grupos mostrando sus frecuencias. Histogramas, más para datos continuos. Gráficos de líneas. Mapa de calor, visualización en matriz por intensidad de color, interesante para ver relaciones entre dos variables. Los boxplots para resumir distribución de un conjunto de datos.

Sesión 3