Universidad Nacional Autónoma de México

INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

Licenciatura en Ciencia de Datos

PLN

Tema 4

Actividad Sumativa 1

Integrante:

- Villalón Pineda Luis Enrique

¿Qué son los corpus y bajo qué criterios se pueden clasificar?

Un corpus es básicamente una colección organizada de textos o grabaciones del lenguaje real. Lo podemos pensar como una biblioteca especializada donde los lingüistas guardan ejemplos de cómo hablamos y escribimos. Hoy en día, estos corpus suelen estar en formato digital para facilitar su análisis. Son útiles porque nos permiten estudiar cómo las personas realmente usan el idioma en su día a día, no solo cómo creemos que deberían usarlo. En lugar de inventar ejemplos desde nuestro escritorio, podemos examinar conversaciones reales, artículos de periódicos, tweets, novelas... lo que sea relevante para nuestra investigación.

¿Cómo podemos clasificar los corpus?

Aunque no hay una clasificación oficial única, los lingüistas suelen organizarlos según varias características clave:

1. Según su nivel de anotación

- Corpus sin anotar: Son textos "en bruto", sin ningún tipo de análisis añadido. Básicamente, lo que ves es lo que hay: solo palabras.

- Corpus anotados: Aquí la cosa se pone interesante. Estos corpus vienen con información lingüística adicional que facilita su análisis. Pueden incluir:

    - Etiquetado gramatical: Cada palabra lleva una etiqueta que te dice si es un sustantivo, un verbo, un adjetivo, etc.
    - Análisis sintáctico: Van más allá y muestran la estructura completa de las oraciones mediante diagramas de árbol, identificando frases nominales, verbales y demás componentes.
    - Etiquetado semántico: Incluyen información sobre el significado de las palabras.
    - Anotación pragmática: Registran aspectos como los actos de habla (¿estamos preguntando, ordenando, prometiendo?) o el nivel de cortesía. Este tipo de anotación suele hacerse manualmente por su complejidad.

2. Según la modalidad del lenguaje

- Corpus escritos: Reúnen textos de libros, artículos, páginas web, correos electrónicos...
- Corpus orales: Contienen transcripciones de conversaciones, entrevistas, discursos o cualquier tipo de lenguaje hablado.

3. Según el idioma y el período histórico

Cada corpus se construye para una lengua específica y un momento histórico determinado. Por ejemplo, puedes encontrar corpus del inglés, chino mandarín, polaco o griego. Y aquí viene algo importante: un corpus del mandarín de los años 90 no te servirá para estudiar cómo se hablaba en el siglo XIX. El lenguaje cambia con el tiempo.

4. Según su tamaño

- Corpus pequeños: Aunque parezca contradictorio, pueden ser increíblemente útiles para estudiar fenómenos lingüísticos muy específicos.
- Corpus grandes: Algunos contienen cientos de millones de palabras. El *British National Corpus* (BNC), por ejemplo, es enorme. El tamaño importa, especialmente cuando haces análisis estadísticos.

5. Según su composición o diseño: Esto se refiere a cómo se eligen los textos que forman parte del corpus

- Algunos buscan ser representativos y equilibrados, intentando capturar una muestra diversa del idioma en general.
- Otros son especializados, enfocándose en un género o ámbito particular: lenguaje académico, conversaciones informales, textos legales, redes sociales... lo que necesites estudiar.

¿Para qué sirve la tokenización?

La tokenización consiste en dividir un texto en unidades más pequeñas para que la computadora pueda trabajar con ellas.

Su objetivo principal es transformar el texto en algo que se pueda analizar. Para esto sirve para:

- Hacer posible el análisis del lenguaje: Los fragmentos resultantes, llamados tokens, son los bloques básicos que permiten estudiar cómo funciona el texto. La tokenización crea una base de datos reutilizable para investigaciones futuras.
- Convertir palabras en números: Una vez separado el texto, necesitamos medir y contar esos fragmentos. Este paso es clave para entender patrones en el uso del lenguaje. Por ejemplo, después de tokenizar podemos crear listas de palabras y ver cuáles aparecen más frecuentemente.

Corpus de elección

Categoría	CORDE
Origen de los datos	Textuales (escrito)
Espontaneidad	Premeditados (textos escritos con intención comunicativa)
Codificación y anotación	Codificados (se encuentra lematizado, anotado y normalizado en parte)
Especificidad de los elementos	Corpus específico como Literarios (narrativa, poesía, teatro), Informativos (históricos, académicos, periodísticos antiguos en menor medida)
Autoría de los elementos	Canónicos y de autoría variada (autores reconocidos, textos históricos diversos)
Temporalidad de los elementos	Diacrónico (abarca desde los orígenes del español hasta el siglo XXI, con textos cronológicos e históricos)
Propósito del estudio	Multipropósito (sirve como referencia para estudios históricos, lingüísticos y literarios)
Lengua	Monolingüe (español en sus diferentes etapas y variedades históricas)

Cantidad de texto	Grande (millones de registros)
Distribución del tipo de texto	Desequilibrado (predominan ciertos géneros, épocas y autores según la disponibilidad documental)
Accesibilidad	Público, acceso libre a través de la RAE (aunque con restricciones de descarga masiva)
Documentación	Documentado (metadatos por autor, fecha, género, lugar, etc.)
Representatividad	Representativo (de la historia del español, aunque con sesgos hacia textos conservados por instituciones)

¿Cuáles son los tres tipos de información que contiene un corpus además de los textos o transcripciones?

1. Metadatos: el contexto del texto

Son los datos "de fondo" que te cuentan la historia detrás de cada texto. Piensa en ellos como la ficha técnica:

- En textos escritos: quién lo escribió, cuándo se publicó, en qué medio apareció...
- En corpus orales: información sobre los hablantes (su edad, sexo, nivel educativo, lugar de origen...)

Básicamente, todo lo que necesitas saber sobre el texto pero que no está *en* el texto mismo.

2. Marcado textual: la estructura invisible

Aquí codificamos elementos del texto que van más allá de las palabras:

- El formato: dónde empiezan y terminan las cursivas, negritas o títulos
- En transcripciones de conversaciones: quién está hablando en cada momento, cuándo cambian los turnos de palabra
- Cualquier característica estructural que ayude a entender cómo está organizado el contenido

3. Anotación lingüística: el análisis profundo

Esta es la capa más sofisticada. Consiste en hacer explícito lo que sabemos sobre el lenguaje. Por ejemplo:

- Etiquetar cada palabra con su categoría gramatical (sustantivo, verbo, adjetivo...)
- Marcar las relaciones sintácticas entre palabras
- Identificar significados o funciones pragmáticas

Básicamente, tomas el conocimiento lingüístico que está "escondido" en el texto y lo modificas de forma que las computadoras puedan trabajar con él de manera precisa y sistemática.

¿Por qué resulta tan importante la consistencia en la anotación de un corpus?

La consistencia es clave porque hace que la investigación lingüística sea transparente, verificable y replicable. Si las anotaciones están explícitas, otros investigadores pueden revisar y reproducir los hallazgos fácilmente.

El problema: la consistencia perfecta es imposible

- Anotación manual: Aunque se sigan reglas claras, los humanos cometemos errores y, además, muchas decisiones lingüísticas son subjetivas. Por ejemplo, en "su futura novia", ¿"futura" es sustantivo o adjetivo? Ambas opciones son válidas, y los analistas no siempre eligen la misma.

- Anotación automática: Las computadoras son más consistentes que nosotros, pero también fallan cuando se actualizan los programas, cambian los criterios o mejoran los algoritmos.

¿Cuáles son las principales características de las cuatro generaciones de concordancers?

- Primera generación: los pioneros

  Funcionaban en enormes computadoras centrales (*mainframes*) de universidades. Cada equipo construía su propio sistema para buscar palabras en contexto (KWIC),

pero todo era caótico: no había estándares, cada uno inventaba sus propias reglas (como escribir cafe' en lugar de *café*), y compartir investigaciones era casi imposible.

- Segunda generación: democratización... con problemas

  Llegaron las PC y cualquiera podía usar corpus lingüísticos, lo que disparó el interés en los años 80. Pero heredaron muchos defectos: seguían siendo herramientas básicas y, peor aún, solo manejaban corpus pequeños por la poca memoria de las primeras computadoras personales. Un paso adelante en acceso, pero varios pasos atrás en capacidad.

- Tercera generación: la madurez

  Programas como WordSmith, AntConc y Xaira cambiaron el juego. Ya podían procesar corpus gigantes (100 millones de palabras) en una PC común. Ofrecían herramientas estandarizadas: concordancias, listas de frecuencia, colocaciones y análisis estadísticos. Además, aprovecharon Unicode y XML, lo que facilitó trabajar con cualquier idioma. Funcionalidad sólida y confiable.

- Cuarta generación: la era web

  Las herramientas actuales funcionan en tu navegador, pero el procesamiento ocurre en servidores potentes. ¿Por qué el cambio? Para resolver tres problemas: compartir corpus sin violar derechos de autor, funcionar en cualquier sistema operativo y aprovechar el poder de procesamiento de los servidores. Ejemplos: *corpus.byu.edu*, SketchEngine y CQPweb. En cuanto a funciones, son sorprendentemente parecidas a la tercera generación, solo que más potentes y accesibles.

¿Cuáles son las tres herramientas más usadas en el análisis de corpus y qué utilidad tienen?

1. El concordancista:

¿Qué hace? Busca palabras, frases o incluso fragmentos de palabras en todo el corpus y te muestra cada resultado en su contexto, generalmente en formato KWIC (una línea por ejemplo). Es como tener un detector de patrones lingüísticos.

Ejemplos:

- Analizar cómo se usa una partícula específica
- Estudiar sufijos (todas las palabras terminadas en *-ness*)
- Investigar modismos y expresiones
- Buscar categorías gramaticales específicas (todos los verbos auxiliares)

2. Las listas de frecuencia:

¿Qué hace? Simplemente cuenta cuántas veces aparece cada elemento. Suena simple, pero es crucial porque los humanos somos pésimos estimando frecuencias de forma intuitiva. La herramienta te da dos tipos de datos:

- Frecuencia bruta: número total de apariciones
- Frecuencia normalizada: ocurrencias por millón de palabras (esencial para comparar corpus de diferentes tamaños)

Ejemplos:

- Identificar las palabras más comunes en un idioma o género textual
- Cuantificar el uso relativo de diferentes categorías gramaticales

3. El análisis de colocaciones:

¿Qué hace? Identifica qué palabras tienden a aparecer juntas de forma estadísticamente significativa. No solo te dice "estas palabras aparecen juntas", sino que calcula si es un patrón real o pura casualidad. Es fundamental para estudiar cómo se combinan las palabras en el uso natural del idioma.

Ejemplos:

- Ayudar a crear diccionarios identificando las combinaciones más típicas
- Descubrir "collostructions" (palabras que se asocian con construcciones gramaticales específicas)

McEnery & Hardy mencionan una idea alternativa a usar software open-source para el análisis de corpus. ¿Cuál es esta y cuáles son sus ventajas?

McEnery y Hardy plantean una idea interesante (inspirada en Biber y colegas): en lugar de usar programas prefabricados, ¿por qué no aprender a programar y crear tus propias herramientas personalizadas para cada proyecto?

Las ventajas son tentadoras:

- Libertad total: puedes hacer análisis que los programas convencionales simplemente no pueden realizar
- Más rápido y preciso: ajustas todo a tus necesidades exactas, sin funciones innecesarias
- Resultados a tu medida: el formato de salida es exactamente como lo necesitas
- Sin límites de tamaño: no estás restringido por las capacidades de un software comercial

En resumen, es como la diferencia entre comprar muebles de IKEA o fabricarlos tú mismo: más trabajo inicial, pero total control sobre el resultado final.

¿Cuáles son las desventajas que tiene esta alternativa?

1. La barrera de entrada es enorme

La mayoría de los lingüistas quieren estudiar el lenguaje, no convertirse en programadores. De hecho, el boom de la lingüística de corpus ocurrió precisamente cuando aparecieron herramientas que *no* requerían saber programar. Exigir que todos aprendan a codificar dejaría fuera a la mayoría de los investigadores.

2. Reinventar la rueda, una y otra vez

Trabajar solo o en equipos pequeños tiene un techo. El progreso real ocurre cuando diferentes grupos coordinan esfuerzos hacia objetivos comunes, no cuando cada uno construye sus propias soluciones desde cero.

3. El caos de los formatos incompatibles

Fue el trabajo colaborativo lo que permitió crear estándares de codificación. Sin esa coordinación, estaríamos atrapados en un mar de formatos incompatibles, con cada investigador perdiendo tiempo adaptando herramientas para diferentes corpus en lugar de hacer investigación real.

4. Adiós a la ciencia reproducible

Si desarrollas tus propias herramientas y no las compartes, otros no pueden verificar tus resultados. Esto viola un principio científico básico: la investigación debe ser replicable. Con herramientas estándar y ampliamente disponibles, cualquiera puede revisar tu trabajo.

¿Cuál es la frecuencia relativa de una palabra en un corpus y cómo se calcula?

La frecuencia relativa es una forma de medir qué tan común es una palabra, pero ajustando el cálculo al tamaño del corpus. No es solo contar cuántas veces aparece, sino ponerlo en perspectiva.

Te permite hacer comparaciones justas:

- Determinar si una palabra es realmente frecuente o no

- Comparar corpus de diferentes tamaños sin que el tamaño distorsione los resultados

Por ejemplo, encontrar una palabra 10 veces en un corpus pequeño vs. 1.103 veces en uno enorme... ¿dónde es más común? Sin normalizar, no lo sabrías.

¿Cómo se calcula?

fn = (apariciones de la palabra ÷ tamaño total del corpus) × base de normalización

La "base de normalización" suele ser 1.000 (ocurrencias por mil palabras) o 1.000.000 (por millón de palabras).

Ejemplo:

La palabra *Lancaster* en el BNC escrito:

- Apariciones: 1.103
- Tamaño del corpus: 87.903.571 palabras
- Base: 1.000.000

fn = (1.103 ÷ 87.903.571) × 1.000.000 = 12.55

Basicamente esperarías encontrar *Lancaster* unas 12.55 veces por cada millón de palabras en ese corpus.

¿Cuáles son los principales tests de significatividad y por qué resultan relevantes?

¿Por qué son importantes los tests de significatividad?

En lingüística de corpus trabajas con miles o millones de datos. Inevitablemente encontrarás patrones, pero no todos significan algo. Los tests de significatividad te permiten distinguir las diferencias que importan de las que son puro ruido estadístico. Como regla general, si hay 95% de probabilidad de que un resultado no sea coincidencia, se considera "significativo".

¿Dónde se usan más?

- Palabras clave: comparar si una palabra es significativamente más frecuente en un corpus que en otro de referencia
- Colocaciones: determinar si dos palabras aparecen juntas más de lo que esperarías por azar

Los tests más comunes:

1. Chi-cuadrado: Asume que los datos siguen una distribución normal (la famosa curva de campana), pero las palabras no se comportan así. Además, falla con datos escasos.

2. Log-verosimilitud: Lo usan mucho los lingüistas porque no asume distribución normal, lo que lo hace más apropiado para datos lingüísticos reales.

3. Prueba t

4. Prueba exacta de Fisher: Útil cuando tienes muy pocos ejemplos.

Un problema a considerar:

Cuando haces miles de tests (como al calcular palabras clave), estadísticamente *algunos* resultados falsos son inevitables. La solución es usar umbrales mucho más estrictos, como 99.9% o incluso 99.9999% en lugar del 95% estándar.