Introducción a la lingüística de corpus

Leticia Martín-Fuertes Moreno
Lingüista en Molino de ideas
@nimbusaeta
leticia@molinodeideas.es

14 de noviembre de 2015
#EOL2

Lingẅars

¿Qué es un corpus?

- ¿Un montón de palabras?
- ¿Una teoría o una metodología del lenguaje?

¿Por qué usar corpus?

- Necesitamos grandes cantidades de datos para detectar **tendencias**, lo que es normal o típico en el uso real del lenguaje
- Los corpus también revelan **casos muy raros** o excepcionales que no encontraríamos analizando textos uno a uno
- Los humanos cometemos **errores** y somos **lentos**; los ordenadores son muchísimo más rápidos y precisos

Criterios para construir un corpus

- Debe ser una **gran cantidad** de texto
- Tiene que ser **representativo** de una lengua (o de un género de esa lengua)
- Tiene que estar en un **formato legible** para una máquina (p. ej., archivos de texto .txt)
- Actúa como una **referencia estándar** sobre lo que es normal en una lengua
- A menudo, **se anota** con información lingüística adicional (p. ej., partes del discurso)

Ejemplo de texto sin anotar

“Arrest warrant out for Clowes’ partner years before collapse.”

By Daniel John

A WARRANT for the arrest of the former partner of Mr Peter Clowes was issued seven years before his Barlow Clowes investment empire collapsed, according to evidence submitted to the Parliamentary Ombudsman.

Ejemplo de texto anotado (1)

<head type=MAIN>
<s n=001><c PUQ>&bquo<w NN1>Arrest <w NN1>warrant <w AVP>out <w PRP>for <w NP0>Clowes<c PUN>' <w NN1>partner <w NN2>years <w PRP>before <w NN1>collapse <c PUQ>&equo'
<c PUN>.
</head>
<head type=BYLINE>
<s n=002><w PRP>By <w NP0>Daniel <w NP0>John
</head>

Ejemplo de texto anotado (2)

<s n=003><w AT1>A <w=NN1>WARRANT <w PRP>for <w AT0>the <w NN1>arrest <w PRF>of <w AT0>the <w DT0>former <w NN1>partner <w PRF>of <w NP0>Mr <w NP0>Peter <w NP0>Clowes <w VBD>was <w VVN>issued <w CRD>seven <w NN2>years <w CJS>before <w DPS>his <w NN1-NP0>Barlow <w NP0>Clowes <w NN1>investment <w NN1>empire <w VVD>collapsed<c PUN>, <w PRP>according to <w NN1>evidence <w VVN>submitted <w PRP>to <w AT0>the <w AJ0>Parliamentary <w NN1>Ombudsman<c PUN>.

Tipos de corpus (1)

1. **Especializado**; en, por ejemplo:

- Un género: el lenguaje periodístico
- Un momento: desde 2005 hasta ahora
- Un lugar: solo textos publicados en China

2. **General**: debe ser mucho más grande. Por ejemplo: el British National Corpus (BNC) contiene unos 100 millones de palabras del inglés hablado y escrito.

El BNC

Modo	Tipo de textos y descripción	Número de palabras
Escrito (87 284 364 palabras)	**Informativo** (8 tipos) 1) Internacional 2) Ocio 3) Arte 4) Comercio y finanzas 5) Filosofía 6) Ciencias sociales 7) Ciencias aplicadas 8) Ciencias puras y naturales	70,9 millones
	Imaginativo (1 tipo) 9) Ficción	16,4 millones
Hablado (10 341 729 palabras)	Conversaciones **informales** recogidas demográficamente de la población del Reino Unido	4,2 millones
	Lenguaje **dirigido** a realizar una tarea, grabado en lugares específicos para eventos específicos, como reuniones de trabajo, charlas en público, etc.	6,1 millones

Tipos de corpus (2)

3. Corpus **multilingüe**: inglés y español, o inglés americano y británico.
4. Corpus **paralelo**: inglés y español, exactamente los mismos textos, traducidos. Por ejemplo, el corpus CRATER.
5. Corpus **de estudiantes**: el lenguaje usado por gente que está aprendiendo una lengua. Por ejemplo, el International Corpus of Learner English.
6. Corpus histórico o **diacrónico**: como el corpus Helsinki, que recoge un millón y medio de palabras de textos escritos entre 700 y 1700.
7. Corpus **monitorizado**: está en continuo aumento. Por ejemplo, el Bank of English.