

Introducción al Procesamiento de Lenguaje Natural

Grupo de PLN - InCo
2011



Corpus



¿Qué es un corpus?

Un corpus es una colección de material lingüístico.

Es de utilidad en diferentes áreas, principalmente en lingüística computacional y lingüística teórica.



¿Cómo se construye?

Hay que definir las características deseadas:

- escrito / oral
- idioma (un idioma o multilingüe)
- tipo de texto (prensa, literario, científico, ...)
- dominio (arte, lingüística, bio-informática, ...)
- anotado / no anotado (conjunto de etiquetas)



Algunos corpus importantes

- inglés
 - Brown Corpus
 - Penn Treebank
- español
 - CREA y CORDE (RAE)
 - Corpus del español de Mark Davies



Brown corpus

- 1961
- inglés americano
- 1.000.000 palabras
- sólo material escrito (500 textos de aprox. 2000 palabras)
- anotado (*POS tags*)
- disponible pagando licencia
- contiene material clasificado en 15 categorías:
 - prensa: reportajes, editoriales, ...
 - literarios: misterio, ciencia ficción, romance, ...
 - religión
 - humor
 - ...



Penn Treebank

- 1989
- inglés americano
- 4.500.000 palabras
- material escrito y oral
- anotado (*POS tags* + información sintáctica + extensiones)
 - muy utilizado para entrenar analizadores sintácticos
- disponible pagando licencia
- contiene:
 - artículos científicos
 - noticias
 - capítulos de obras literarias
 - oraciones de manuales de computación
 - el corpus Brown completo re-etiquetado



Corpus CREA

Corpus de Referencia del Español Actual

- inicio 1993
- español actual, todas las variedades
- 160.000.000 palabras
- material escrito y oral
- no anotado
- disponible para consultas en línea (sin suscripción)
- contiene:

[ver descripción en RAE \(www.rae.es\)](http://www.rae.es), Banco de Datos



Corpus CORDE

Corpus Diacrónico del Español

- inicio 1993
- español anterior a 1975, todas las variedades
- 250.000.000 palabras
- material escrito
- no anotado
- disponible para consultas en línea (sin suscripción)
- contiene:

[ver descripción en RAE \(www.rae.es\)](http://www.rae.es), Banco de Datos



Corpus del Español

(Mark Davies, Brigham Young University)

- 2001
- español histórico y actual
- 100.000.000 palabras
- material oral y escrito
- anotado (*POS-tags, lemas*)
- disponible para consultas en línea
- contiene: ver descripción en <http://www.corpusdelespanol.org/>



Corpus ANCORA

Universidad de Barcelona

- 2008 (aprox.)
- español (500.000 pals.) y catalán (500.000 pals.)
- principalmente textos periodísticos
- anotaciones:
 - categoría morfológica
 - constituyentes y funciones sintácticas
 - estructura argumental y papeles temáticos
 - clase semántica verbal
 - sentidos de WordNet nominales
 - entidades nombradas
 - correferencia

<http://clic.ub.edu/es/ancora-es/>



Corpus CORIN

- un corpus del español de Uruguay
- proyecto desarrollado por la Licenciatura en Lingüística
- corpus anotado
- conjunto de etiquetas muy completo
- corpus pequeño
- problemas para generar anotaciones
 - > no se pudo completar



Corpus específicos

Diversos corpus anotados con información específica, relativa a ciertos fenómenos:

- TimeBank: información sobre eventos, expresiones temporales y relaciones entre ellos
- Penn Discourse TreeBank (PDTB): información sobre marcadores del discurso
- MPQA: información sobre opiniones y sentimientos



Penn Treebank: Proceso de Anotación

POS tags

asignación automática de etiquetas

+

corrección manual

Información sintáctica

generación automática de subárboles

+

generación manual de la estructura completa de cada oración



Penn Treebank: Proceso de Anotación

Conjunto de *POS tags*

- basado en las etiquetas del Brown Corpus
- conjunto más reducido
- por ejemplo, se eliminan etiquetas que aportan información recuperable a nivel léxico



Penn Treebank: Proceso de Anotación

Conjunto de *POS tags*

Brown Corpus						Penn Treebank	
VB	sing	HV	have	DO	do	VB	sing, have, do
VBD	sang	HVD	had	DOD	did	VBD	sang, had, did
VBG	singing, doing	HVG	having	--	--	VBG	singing, having, doing
VCN	sung, done	HVN	had	--	--	VCN	sung, had, done
VBZ	sings	HVZ	has	DOZ	does	VBZ	sings, has, does



Penn Treebank: Proceso de Anotación

Asignación automática de *POS tags*

- tagger PARTS (AT&T Bell Labs)
- asigna etiquetas del corpus Brown
- mapeo automático a etiquetas del Penn Treebank



Penn Treebank: Proceso de Anotación

Corrección manual de *POS tags*

- diferentes anotadores
- utilización de un software específico
- se agregan correcciones sin eliminar etiqueta del *tagger*



Penn Treebank: Proceso de Anotación

Parsing automático

- No toma decisiones cuando no hay seguridad, genera subárboles.
- Ejemplo de ambigüedad:

[[El auto [de Pedro]] [tiene [un asiento [[que se puede reclinar] y [es muy cómodo para dormir]]]].

[[El auto [de Pedro]] [tiene [un asiento [que se puede reclinar]]] y [funciona [a gas oil]]].



Texto:

Battle-tested industrial
managers here always buck up
nervous newcomers with the
tale of the first of their
countrymen to visit Mexico, a
boatload of warriors blown
ashore 375 years ago.

```
( (S
  (NP Battle-tested industrial managers
    here)
  always
  (VP buck
    up
    (NP nervous newcomers)
    (PP with
      (NP the tale
        (PP of
          (NP (NP the
            (ADJP first
              (PP of
                (NP their countrymen)))
            (S (NP *)
              to
              (VP visit
                (NP Mexico))))
          ,
          (NP (NP a boatload
            (PP of
              (NP (NP warriors)
                (VP-1 blown
                  ashore
                    (ADVP (NP 375 years)
                      ago))))
            (VP-1 *pseudo-attach*))))))
    .)
)
```

