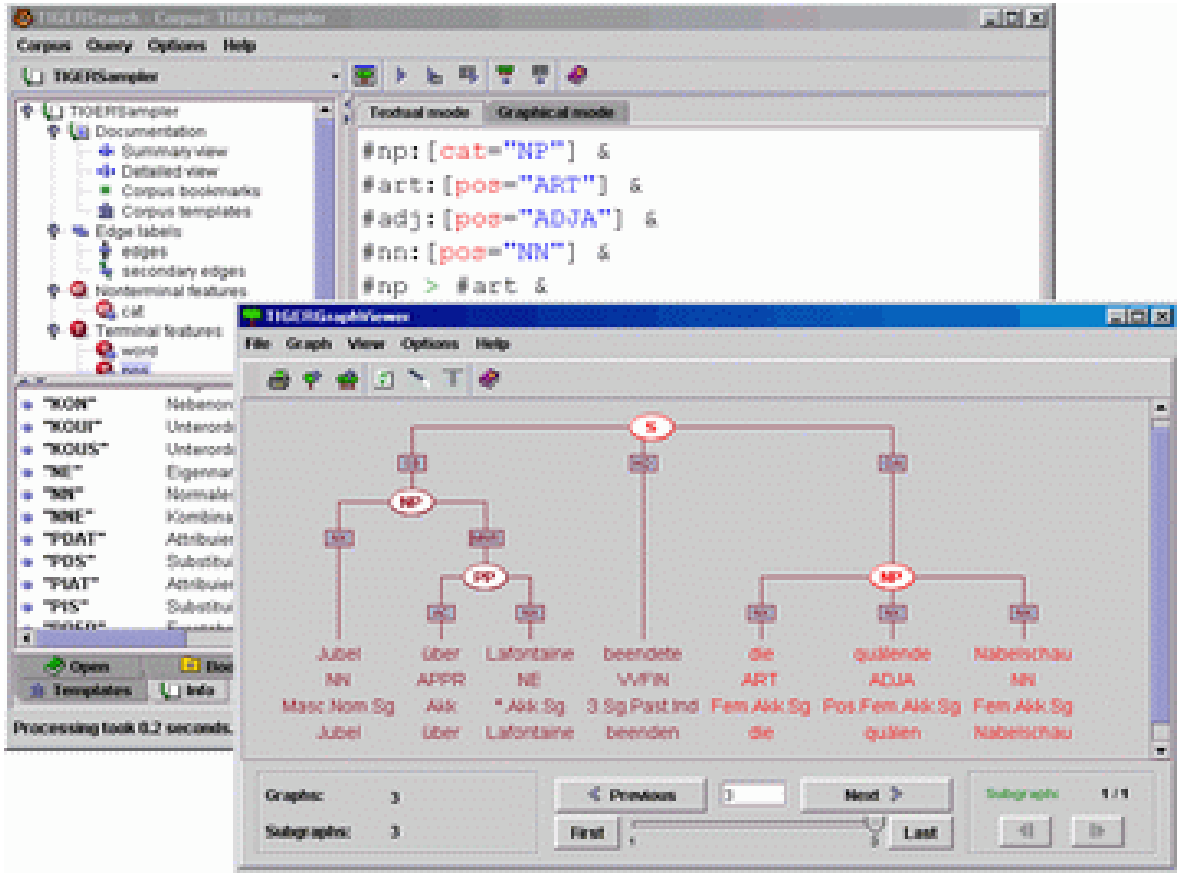
Seminario de análisis del discurso

Dr. César Antonio Aguilar
Facultad de Lenguas y Letras
13/09/2010

CAguilar@iingen.unam.mx

Búsquedas en corpus (1)



En esta clase, vamos a ver algunos métodos para hacer búsquedas en corpus textuales.

Para ello, partimos del supuesto de que tenemos una colección de documentos organizados y etiquetados, y que podemos acceder a ellos de una forma fácil y precisa.

Nota: este ejemplo está tomado del siguiente sitio WEB:

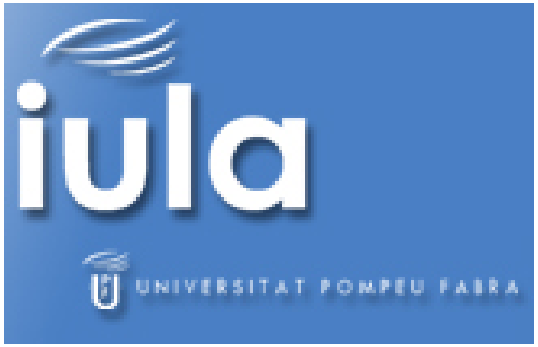
www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/oldindex.shtml

Búsquedas en corpus (2)

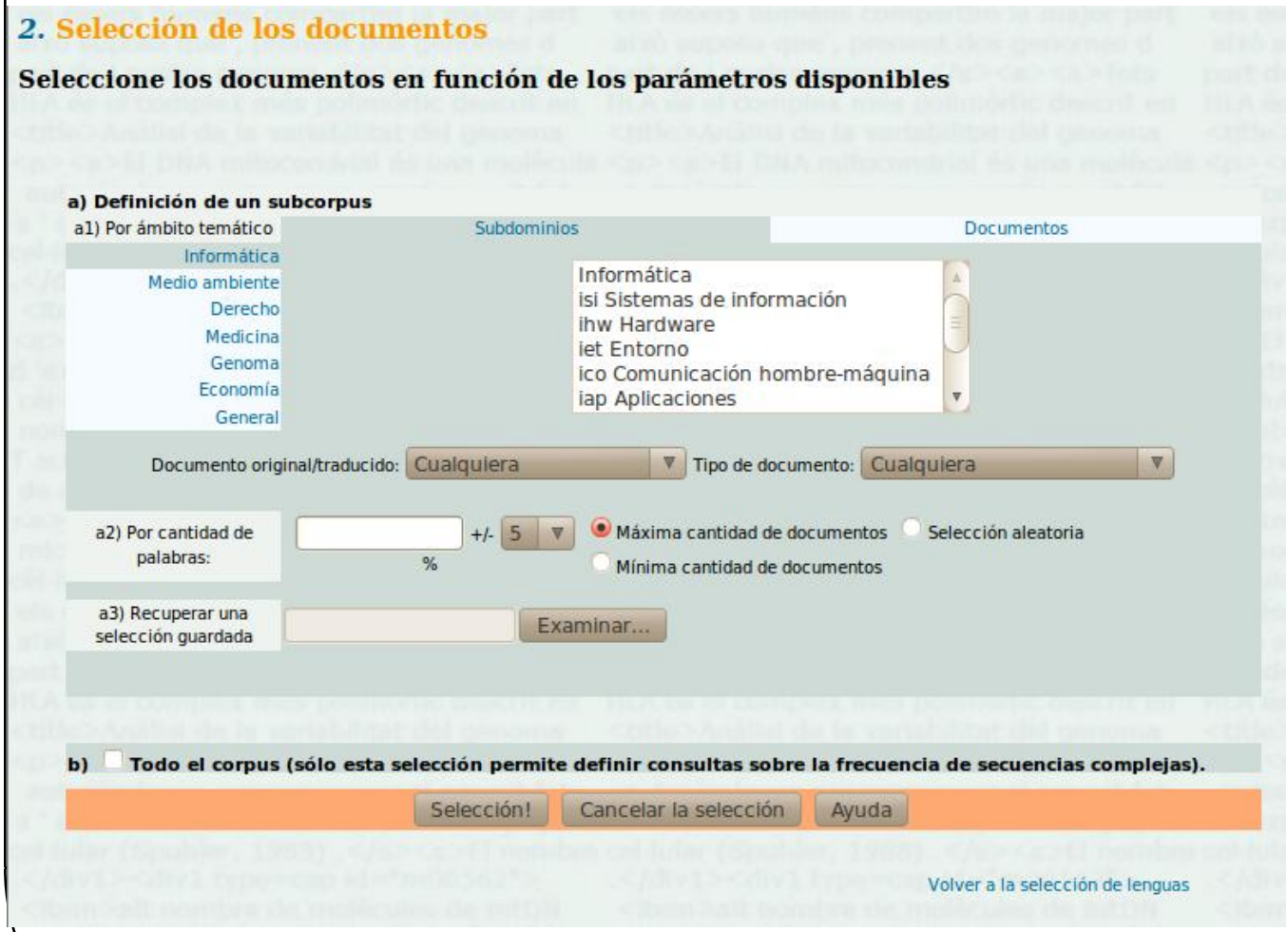
Hagamos algunas consultas al corpus **Bwananet**:

http://bwananet.iula.upf.edu/

Este corpus consiste en una colección de documentos especializados en varias áreas (informática, medio ambiente, derecho, medicina, ciencias genómicas, etc.), desarrollado por el Instituto Universitario de Lingüística Aplicada (IULA), de la Universidad Pompeu Fabra (Barcelona, España):



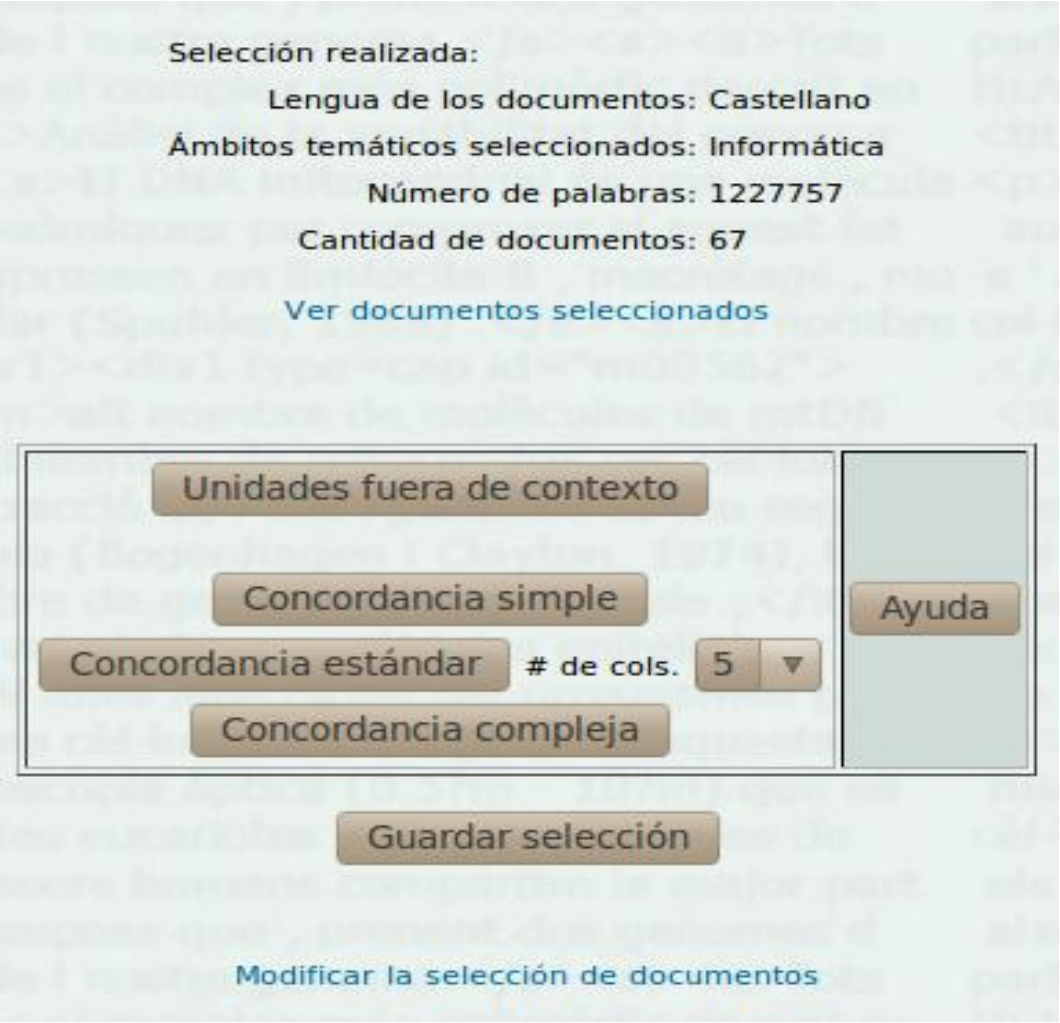
Búsquedas en corpus (3)



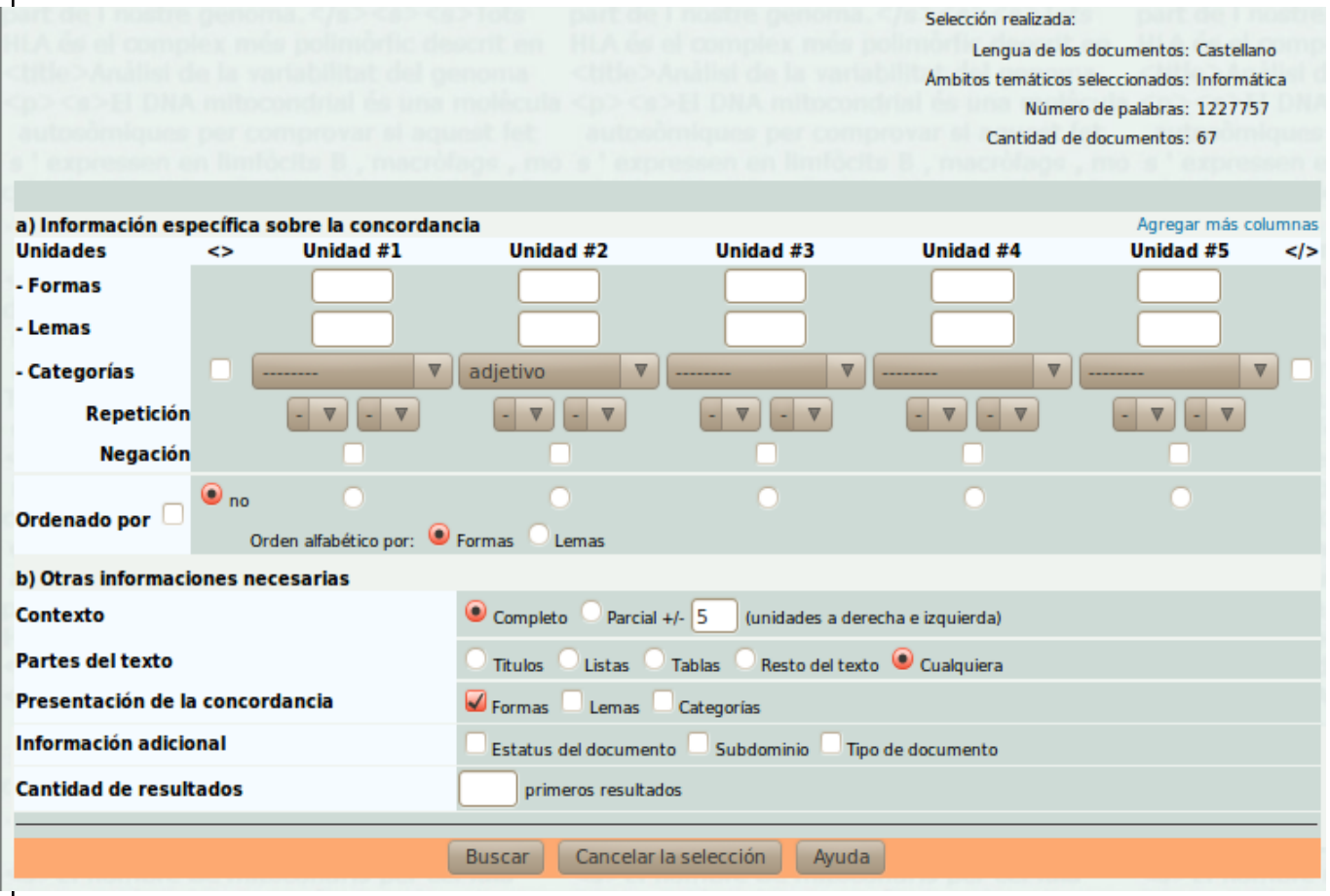
Bwananet cuenta con una ventana que nos describe primero cuáles son los temas que pueden ser buscados dentro de él, y en seguida nos permite establecer restricciones respecto a nuestra búsqueda.
Eso nos ayuda refinar qué clase de información queremos obtener, ya que nos brinda un control sobre ésta.

Búsquedas en corpus (4)

Posteriormente, podemos establecer parámetros para determinar qué cosas vamos a buscar. Bwananet es un corpus que prioriza la búsqueda de unidades léxicas, dado que fue construido para resolver tareas en terminología (en concreto, identificación de términos en textos científicos y técnicos).



Búsquedas en corpus (5)



Las opciones que tenemos para hacer búsquedas son varias: en este caso, seleccionamos la búsqueda de concordancia estándar, esto es, consideramos secuencias formadas por máximo 5 palabras.
¿Qué podemos hacer con esto? Lo veremos a continuación.

Experimento (1)

Bwananet ofrece, para cada consulta, un listado con 50 resultados. Si bien es limitado, podemos deducir algunas cosas:

1. En informática, el término *sistema*, ¿con qué adjetivos aparece asociado de manera recurrente?

2. ¿Cuál de los dos términos es más frecuente: *base de datos* o *base de conocimientos*?

3. ¿Cuántos tipos de *bases de datos* podemos reconocer en Bwananet?

4. ¿Cuántos tipos de *bases de datos relacionales* hay?

5. El adjetivo *genético*, ¿con qué términos aparece asociado en informática?

6. Los términos *secuencia* y *cadena,* ¿son sinónimos?

Experimento (2)

Tarea: haciendo búsquedas en Bwananet, en concreto en los los corpus de informática y genómica, traten de contestar lo siguiente:

1. ¿Qué términos son generados en ambas áreas asociado nombres, adjetivos o frases prepositivas al término de base *cadena*?

2. ¿Se utilizan los términos *gramática* y *sintaxis* en ambas áreas? ¿Qué términos se generan al respecto?

3. ¿Cuál de estos verbos es más productivo en ambos corpus: *definir* o *comprender*?

4. ¿Cómo pueden localizar una definición del término *gen* en el corpus de textos de genómica?

5. ¿Qué preposición es la más usada para construir términos complejos (con más de 4 palabras, por lo menos)?

Retomando la creación de un corpus (1)

Volviendo ahora a su proyecto para pensar en un conjunto de “objetos posibles de buscar” en una colección de textos (esto es, un potencial corpus), con los ejemplos de búsqueda visto en Bwananet,

Para ello, partimos del supuesto de que tenemos una colección de documentos organizados y etiquetados, y que podemos acceder a ellos de una forma fácil y precisa.

¿Qué clase de corpus tienen en mente? Considerando los 5 documentos que tienen recolectados, respondan a las siguientes preguntas:

Retomando la creación de un corpus (2)

1. De acuerdo al porcentaje de textos que tienen (p. e., si contamos su cantidad de palabras), ¿qué clase de corpus tendrían?

2. De acuerdo con la temática de sus textos, si son muy específicos o muy generales, ¿qué clase de corpus tendrían?

3. ¿Han pensado en alguna clase de etiquetas para su corpus? ¿En qué consisten estas etiquetas?

4. ¿Su corpus considera algún tipo de documentación referente al proceso de construcción?

5. ¿Qué fenómenos lingüísticos pueden ser identificados en su corpus?

6. ¿Sus corpus pueden ser analizados en todos los niveles lingüísticos, o consideran un nivel concreto (morfológico, léxico, sintáctico, semántico y/o discursivo)?

Relación entre etiquetas y palabras (1)

Ya sea que usemos etiquetas XML, o ya sea que nos animemos a usar etiquetas de partes de la oración, conviene que establezcamos qué cosas se pueden rastrear en nuestro corpus.

La idea es que tales etiquetas nos faciliten el proceso de búsqueda, considerando una alternancia entre el uso éstas, o de los ítems léxicos a los que van asociadas.

Figure 1. An Example of Tagged Text (excerpted from the Brown Corpus)

```
The/at grand/jj jury/nn commented/vbd on/in a/at number/nn of/in
other/ap topics/nns ,/, among/in them/ppo the/at Atlanta/np and/cc
Fulton/np-tl County/nn-tl purchasing/vbg departments/nns which/wdt
it/pps said/vbd "/" are/ber well/ql operated/vbn and/cc follow/vb
generally/rb accepted/vbn practices/nns which/wdt inure/vb to/in
the/at best/jjt interest/nn of/in both/abx governments/nns "/" ./.
```

Relación entre etiquetas y palabras (2)

Veamos un ejemplo: si accedemos al siguiente sitio, tenemos un prototipo de un etiquetador automático desarrollado por el Centro de Tecnologías del Lenguaje de la Universidad de Copenhagen.

http://cst.dk/online/pos_tagger/uk/index.html

Relación entre etiquetas y palabras (3)

Este prototipo es capaz de asociar a una palabra un tipo de etiqueta morfosintáctica, usando un algoritmo diseñado por un computólogo llamado Eric Brill, actualmente asociado a Microsoft Research.

http://research.microsoft.com/%7Ebrill/

Sin entrar en demasiados detalles, la idea de este algoritmo diseñado por Brill es:

1. Si quieres etiquetar un documento, primero necesitas saber si ese documento tiene palabras identificables.

2. Una vez que reconoces estas palabras identificables, les asignas las primeras etiquetas.

3. Si hay palabras a las que no sabes qué etiqueta ponerle, toma como guía la palabra etiquetada más cercana.

Relación entre etiquetas y palabras (4)

Volviendo con el etiquetador, veamos si es capaz de etiquetar cualquier clase de texto (en este caso, en inglés). El siguiente ejemplo es una entrada de Wikipedia:

A computer is a programmable machine that receives input, stores and manipulates data//information, and provides output in a useful format. While a computer can, in theory, be made out of almost anything (see misconceptions section), and mechanical examples of computers have existed through much of recorded human history, the first electronic computers were developed in the mid-20th century (1940–1945). Originally, they were the size of a large room, consuming as much power as several hundred modern personal computers (PCs).[1] Modern computers based on integrated circuits are millions to billions of times more capable than the early machines, and occupy a fraction of the space.[2] Simple computers are small enough to fit into mobile devices, and can be powered by a small battery. Personal computers in their various forms are icons of the Information Age and are what most people think of as "computers". However, the embedded computers found in many devices from MP3 players to fighter aircraft and from toys to industrial robots are the most numerous.

Relación entre etiquetas y palabras (5)

Supongamos que este etiquetador pudo hacer la tarea porque a la mejor tiene un lexicón con palabras relacionadas con cuestiones de computación. ¿Y si lo probamos con una nota deportiva?

were very confusing," she said, "because they were happiness but then alsoIt was over in less than an hour, but no matter. Kim Clijsters will savor her 6-2, 6-1 victory against Vera Zvonareva in the United States Open final much longer than she did last year's championship.
Clijsters's 2009 title run was the first major moment in her life that she was unable to share with her father, Leo, whose death from lung cancer that January had driven her back to the tennis court and out of a 27-month retirement.
"Last year my emotions sad at the same time."
After Clijsters dispatched Zvonareva in 59 minutes Saturday night, her joy was unfettered. She scampered into the Arthur Ashe Stadium seats to give her husband, Brian Lynch, the kind of kiss generally bestowed upon soldiers returning from the battlefield.

Relación entre etiquetas y palabras (6)

Finalmente, ¿qué resultados nos arroja el etiquetador cuando buscamos textos en un documento de medicina? Veamos:

Impairments in social cognition in early medicated and unmedicated Parkinson disease.

Roca M, Torralva T, Gleichgerrcht E, Chade A, Arévalo GG, Gershanik O, Manes F.

*Department of Neuropsychology §Congnitive Neuroscience Laboratory ‖Department of Neurology ♯Congnitive Neurology Department, Institute of Cognitive Neurology (INECO) †Department of Neuropsychology ¶Department of Neurology **Cognitive Neurology Department, Institute of Neuroscience, Favaloro University, Buenos Aires, Argentina ‡Laboratory of Neuroscience, Universidad Diego Portales, Chile.
Abstract

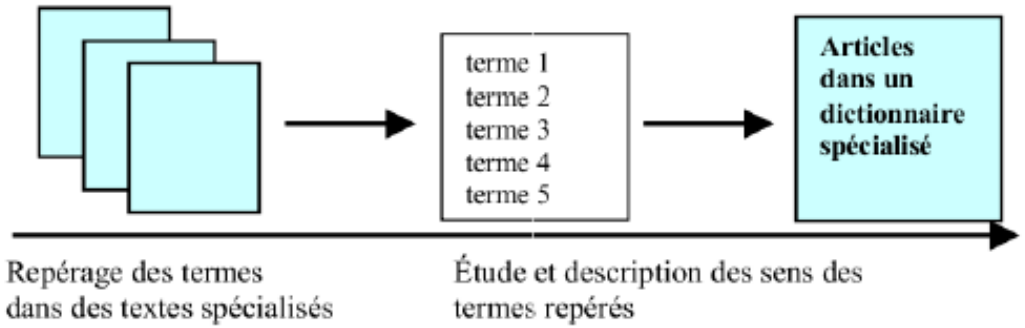
BACKGROUND: Theory of mind (ToM) refers to the ability to infer others' mental states, including intentions and feelings, and is considered to be a critical part of social cognition. Earlier studies in individuals with Parkinson disease (PD) have shown ToM deficits in the more advanced stages of the disease. There is currently no evidence of social cognition deficits in patients in the early stages of PD.

Algunas conclusiones (1)

Cuando busco algo en un corpus, ¿lo que encuentro son palabras o secuencias de etiquetas?
Piensen por ejemplo en tareas como buscar términos en Bwananet: en realidad, conviene que los textos tengan etiquetas porque nos permiten buscar **patrones constitutivos** de palabras o frases (que posteriormente pueden ser consideradas como buenos candidatos a términos):

$$\text{Adj}_1\ \text{Noun}_2 \rightarrow \text{Noun}_1\ ((\text{CC Det}^?)^?\ \text{Prep Det}^?\ (\text{Adj}|\text{N}|\text{Part})^{0-3})\ \text{Noun}_2' \quad (19.7)$$

Esto nos ayuda a explorar otros textos, en aras de reconocer estos patrones, y así enriquecer nuestro listado de términos.



Algunas conclusiones (2)

Lo interesante de hacer este tipo de análisis en un corpus es que nos ayuda a confirmar (o a refutar) supuestos que planteamos como hipótesis, y que ahora podemos validar a partir del cálculo de frecuencias.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

En este ejemplo, lo que se trata de determinar es saber la probabilidad de identificar y asignar una categoría gramatical a una palabra, dada la información que nos brinda una palabra precedente. Así, por ejemplo, podemos decir que es altamente probable que a un pronombre personal le siga un verbo conjugado, y que la combinación de ambos elementos condicione la introducción de una palabra (o secuencia de palabras) concreta.

Algunas conclusiones (3)

El uso de **probabilidades condicionales** es muy útil a la hora de hacer etiquetados, justo porque permiten asignar etiquetas a palabras, considerando cuán frecuente es la asociación entre dos o más palabras dentro de un contexto dado.

<s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>

Estas combinaciones de palabras reciben el nombre de **n-gramas**. Un grama es un ítem léxico que potencialmente va a aparecer ligado a otros gramas. Dada la recurrencia de su asociación, podemos considerar que forman estructuras más complejas, basándonos en las frecuencias que observamos cada vez que aparecen juntas.

Algunas conclusiones (4)

Finalmente, unas preguntas para reflexionar:

1. ¿Piensan buscar en sus corpus patrones de constitución de palabras complejas y/u oraciones?

2. ¿Cómo construirían una *gramática* para determinar tales patrones?

3. Si estos patrones aparecen de forma recurrente en su corpus, ¿esto ayuda a validar su gramática?

4. Si tales patrones no son recurrentes, ¿cuál sería el problema que hay de por medio?

5. ¿Cómo lo solucionarían?

Gracias por su atención

Blog del curso: http://discurso-uaq.weebly.com/