LA CIENCIA DE LA GRAMÁTICA: MÉTODO Y PRÁCTICA
GUÍA DE CLASE (24-02-2014)

Las fuentes de datos del gramático I: corpus

Carlos Muñoz Pérez
cmunozperez@filo.uba.ar
Universidad de Buenos Aires

1. Introducción: posibles fuentes de datos

El lenguaje presenta dificultades prácticamente únicas para su estudio: desde un punto de vista estrictamente platónico se puede decir que "no existe", que es una mera abstracción. ¿Cómo se estudia aquello que no existe, que no podemos ver o tocar?

El problema en términos más teóricos es: **¿cómo accedemos a la *competencia lingüística* que queremos investigar y definir?**

Una manera de hacerlo es a través de la *actuación*. El estudio de los datos basados en la actuación lingüística podría permitirnos inferir algunas de las propiedades de la competencia. Una idea más osada es intentar acceder a la *competencia* a partir de métodos y técnicas especialmente diseñados para ello.

En este sentido, hay tres tipos de datos que pueden utilizarse para contrastar teorías gramaticales.

- Datos que existen independientemente del «deseo» del lingüista por estudiar el lenguaje (i.e. corpus).
- Conducta verbal inducida con el fin específico de estudiar la competencia lingüística (i.e. juicios de gramaticalidad).
- Observación de los procesos mentales y fisiológicos que se dan durante la utilización inconsciente de la competencia lingüística (e.g., un experimento psicolingüístico o neurofisiológico).

Desde ya les voy adelantando la moraleja que pretendo que se lleven a partir de las discusiones de esta unidad:

El estudio de la gramática es algo complejo e indirecto (por lo dicho en el primer párrafo, por ejemplo). Por tanto, no es buena idea descartar ningún tipo posible de fuente de datos. Lo mejor que puede hacer el gramático es complementar las ventajas y desventajas de estas fuentes, e intentar realizar un relevamiento lo más adecuado posible a partir de cada uno de ellas siempre que sea posible.

2. La noción moderna de corpus

En esta primera parte, nos vamos a dedicar a los *corpora*. Empecemos por una definición de corpus.

(1) *Corpus (Crystal 2006)*
Una colección de datos lingüísticos, ya sean textos escritos o transcripciones de habla, que puede ser usada (i) *como punto de partida de la descripción lingüística* o (ii) *como medio de verificar hipótesis sobre el lenguaje*.

La noción de corpus nos lleva a hablar de la llamada *lingüística de corpus*. La lingüística de corpus no es el estudio de ningún componente particular del sistema lingüístico.

(2) *Lingüística de corpus*
Área de la lingüística que se aboca al desarrollo de procedimientos y métodos para estudiar el lenguaje.

Es una caracterización bastante amplia, lo sé. McEnery & Hardie (2012) observan algunas características particulares del trabajo moderno con corpus:

> We could reasonably define corpus linguistics as dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions.
>
> Corpora are invariably exploited using tools which allow users to search through them rapidly and reliably
>
> The corpus data we select to explore a research question must be well matched to that research question. [...] we cannot (or can only with some caution) make general claims about the nature of a given language based on a corpus containing only one type of text or a limited number of types of text.

A su vez, hay varios aspectos en los que puede variar un corpus.

MODO DE COMUNICACIÓN: un corpus puede estar formado a partir de textos escritos, orales o gestuales.

CORPUS-BASED VERSUS CORPUS-DRIVEN: muchos lingüistas de corpus no parecen estar muy contentos con la caracterización de su área como "mera metodología". De hecho, existe una dicotomía entre los enfoques que se denominan *corpus-based* y *corpus-driven*.

Los *estudios corpus-based* típicamente usan datos de corpus para explorar una teoría o hipótesis establecida en la bibliografía con el fin de validarla, refutarla o refinarla.

La definición de la lingüística de corpus en tanto método subyace a este uso de los datos de corpus en lingüística.

Los *estudios corpus-driven* rechazan la caracterización de la lingüística de corpus como método, y asumen, en cambio, que el corpus en sí mismo debería ser la única fuente de hipótesis sobre el lenguaje. Involucran, entonces, que el corpus encarna una teoría del lenguaje.

RÉGIMEN DE RECOLECCIÓN DE DATOS:

> Two broad approaches to the issue of choosing what data to collect have emerged: the *monitor corpus approach* (see Sinclair1991: 24–6), where the corpus continually expands to include more and more texts over time; and the *balanced corpus or sample corpus approach* (see Biber1993and Leech2007), where a careful sample corpus, reflecting the language as it exists at a given point in time, is constructed according to a specific sampling frame. (McEnery & Hardie 2012)

CORPUS ANOTADOS Y NO ANOTADOS: Si, por ejemplo, se incluye la categoría gramatical de las palabras que aparecen en el corpus.

REGISTRO TOTAL DE DATOS VERSUS SELECCIÓN:

> The principle of total accountability is, simply, that we must not select a favourable subset of the data in this way. When approaching the corpus with a hypothesis, one way of satisfying falsifiability is to use the entire corpus – and all relevant evidence emerging from analysis of the corpus –to test the hypothesis. This principle is the reason for the quantitative nature of many corpus-based methods. Minimally, however, where there is too much evidence for using the entire corpus to be practical, the analyst must at least, as Leech suggests, avoid conscious selection of data. Short of using the corpus in its totality, total accountability can in principle be preserved by using an unbiased (e.g. randomised) subsample of the examples in the corpus.

3. Pros y cons de los corpora

Primero vamos con las ventajas.

(3) *Ventaja N° 1*
Los corpus contienen muestras de lengua que fueron producidas cuando el hablante no estaba concentrado en cómo se expresaba (a nivel de forma).

Muchas veces, los hablantes consideran inaceptables expresiones que sin embargo utilizan sin darse cuenta. Este tipo de fenómeno no puede captarse a partir del uso exclusivo de juicios gramaticales. Uno de los primeros en observar esto fue Labov a mediados de los 70s con respecto a la utilización de *anymore*. En la mayoría de los dialectos del inglés, *anymore* se utiliza negativamente.

(4) Trains don't stop here anymore.

Sin embargo, en algunos dialectos se lo utiliza en contextos gramaticales positivos para dar cuenta de cosas que son negativas en su valor.

(5) John is smoking a lot anymore.
'John empezó a fumar mucho, lamentablemente'.

Labov y sus colaboradores encontraron que la gente que suele producir este tipo de expresiones lo hacía sin darse cuenta.

> Cuando se les presentaba una oración como *John is smoking a lot anymore* decían que nunca habían escuchado algo parecido, no lo reconocían como inglés, pensaban que significaba 'John no está fumando' y mostraban los mismos signos de desconcierto que cualquiera tiene ante hablantes de otras variedades de la propia lengua. Esto describe la conducta de Jack Greenberg, un constructor de 58 años. Sus juicios de introspección eran tan convincentes que tuvimos que aceptarlos por descripciones válidas de su gramática. Pero dos semanas más tarde se escuchó que le decía a un plomero: *Do you know what's a lousy show anymore? Johnny Carson*. (Labov 1975).

(6) *Ventaja N° 2*
Hay fenómenos gramaticales que se dan en contextos particulares, sólo detectables en un corpus.

(7) a. Saw no one.
b. Hurt myself when trying to cut the roses.
c. Left the party exhausted.

Haegeman (1990) observa que oraciones con sujeto nulo como (3-5) son recurrentes en un tipo de texto particular: el diario íntimo. Sin embargo, dichas oraciones son sistemáticamente agramaticales para los hablantes de inglés. Se trata, en definitiva, de datos que no podrían haber sido capturados sin mirar un corpus específico.

(8) *Ventaja N° 3*
Sólo a través de corpus podemos estudiar lenguas ya extintas o en peligro de desaparecer.

Hay un estudio muy interesante de Golston (1995) con respecto a restricciones de anti-homofonía en palabras adyacentes (i.e. dos palabras iguales no pueden estar una al lado de la otra. El estudio se hizo sobre el griego antiguo. Es normalmente observado

que en esta lengua un modificador genitivo podía aparecer a la izquierda o a la derecha del núcleo nominal al que modifica.

(9) [[hee tólma] [tóon legóntoon]]
el-NOM valor-NOM ellos-GEN hablantes-GEN
'la valentía de los que hablan'

(10) [[hee [tóu himatíoon] ergasías]
el-NOM el-GEN multitud-GEN ley-NOM
'*La ley de la multitud*'

Dada la recurrencia de esta alternancia, sería esperable que existiera también el siguiente par.

(11) [[[tóon oikeíoon] tinàs] [tóon ekeínoon]]
los-GEN esclavos-GEN algunos-GEN los-GEN aquellos-GEN
'Algunos de los esclavos de aquellos'.

(12) *[[tóon [tóon ekeínoon] oikeíoon] tinàs]
los-GEN los-GEN aquellos-GEN esclavos-GEN algunos-GEN
'Algunos de los esclavos de aquellos'.

Golston (1995) realizó una búsqueda en un corpus que reunía más de 500 años de literatura griega antigua, pero no encontró ningún caso en el que se diera un patrón sintáctico como (12).

(13) *Ventaja N° 4*
Es posible automatizar la búsqueda de datos y la generación de estadísticas.

Ahora veamos las desventajas:

(14) *Desventaja N° 1*
Los corpus son útiles para encontrar patrones léxicos o morfológicos, pero no permiten buscar patrones sintácticos o semánticos.

Es decir, podemos buscar los mil usos de la palabra "banco". O podemos buscar las palabras en las que aparece el morfo *–ería*. Pero no podemos buscar cláusulas relativas de objeto con sujeto léxico u oraciones pasivas con verbos psicológicos. Es cierto que existen corpus de estructuras sintácticas (*treebanks*), pero obviamente dependen de análisis sintácticos particulares. Demás está decir, además, que no podemos buscar patrones de ambigüedad semántica, o casos de uso de estructuras con determinada interpretación (e.g. alcance amplio de cuantificador existencial).

(15) *Desventaja N° 2*
¿Cómo se interpreta la ausencia de un dato en el corpus?
Ejemplo absurdo pero claro: si en el corpus no aparece la palabra *mamá,* ¿significa que esa palabra no existe?

(16) *Desventaja N° 3*
El acceso a muchos corpus es pago.

Si, acceder el *Penn Treebank* (un corpus particularmente rico) nos costaría U$3150.

(17) *Desventaja N° 4*
Ningún corpus posee evidencia negativa.

Supongamos que vamos a evaluar las predicciones de esta gramática con respecto a un corpus de oraciones. ¿La gramática es adecuada a los datos?

(18) <u>Gramática 1</u>
O → N SV
SV → V
SV → V N
N → *Homero*
N → *Marge*
N → *Bart*
V → *duerme*
V → ve
…

Sin evidencia negativa no se pueden hacer generalizaciones negativas.

(19) a. [La tortuga ninja]$_i$ bailaba mientras *pro*$_i$ comía pizza.
b. Mientras *pro*$_i$ comía pizza, [la tortuga ninja]$_i$ bailaba
c. **pro*$_i$ bailaba mientras [la tortuga ninja]$_i$ comía pizza.

Este patrón se suele explicar a partir del llamado *Principio C* de la Teoría de Ligamiento.

(20) *Principio C*
Una Expresión Referencial no puede estar coindizada con un elemento que la mande-c.

¿Es posible postular o contrastar este principio sin contar con (19c) como dato?

4. Apéndice: algunos corpus para consultar

1. Corpus de la Real Academia Española
   http://www.rae.es/

2. Corpus de Mark Davis
   http://www.corpusdelespanol.org/

3. Corpus de la Universidad de Barcelona.
   http://clic.fil.ub.es/

4. Corpus de la Universidad de Santiago de Compostela.
   http://www.bds.usc.es/

5. Corpus de la Universidad autónoma de Madrid
   http://www.lllf.uam.es/esprincipal.html

6. Corpus de la Universidad de Pompeu Fabra.
   http://www.iula.upf.es/corpus/corpusuk.htm