IAC: Una interfaz dinámica de acceso a corpus
IAC: A Dynamic Corpora Access Interface

Toni Badia
Universitat Pompeu Fabra
C/ Roc Boronat, 138, 08018 Barcelona
toni.badia@upf.edu

Judith Domingo
Barcelona Media - Centre d'Innovació
Av. Diagonal, 177, planta 9, 08018 Barcelona
judith.domingo@barcelonamedia.org

Resumen: En esta demostración presentamos IAC (Interfaz de Acceso a Corpus), una herramienta on-line desarrollada por Barcelona Media - Centro de Innovación y la Universidad Pompeu Fabra que permite crear interfaces dinámicas para hacer búsquedas en corpus.

Palabras clave: interfaz, corpus, corpus alineado, indexación, corpus monolingüe

Abstract: In this demo we present IAC (Corpus Access Interface), an on-line tool developed by Barcelona Media - Innovation Center and the Pompeu Fabra University to create dynamic interfaces to search in corpora.

Keywords: interface, corpus, aligned corpora, indexation, monolingual corpora

1 Introducción

En los últimos años, el desarrollo de corpus ha experimentado un aumento importante gracias al auge de las nuevas tecnologías, que han facilitado el acceso a los datos lingüísticos y a su tratamiento automatizado.

Este hecho ha supuesto la necesidad de la creación de interfaces para facilitar su consulta. Muestra de ello son las interfaces del Corpus Diacrónico del Español (CORDE), o la del Corpus Textual Informatizado del Catalán (CTILC).

A pesar de la utilidad de las interfaces para consultar corpus, la creación de una interfaz supone un esfuerzo personal y económico importante. Además, cada interfaz está diseñada específicamente para un único corpus, con lo cual la creación de un nuevo corpus implica, de nuevo, invertir dinero y esfuerzos en la creación de una interfaz de consulta.

Conscientes de esta situación, Barcelona Media – Centro de Innovación y la Universidad Pompeu Fabra han creado IAC (Interfaz de Acceso a Corpus), una herramienta que permite diseñar una interfaz de consulta e indexar un corpus sin conocimientos previos de diseño gráfico, programación o indexación de corpus.

2 Características de IAC

A continuación se detallan las principales características de IAC:

- Es una herramienta multilingüe.
- Acepta corpus tanto monolingües como alineados.
- Tiene control de usuarios: cada propietario de corpus decide si su corpus es público o privado, y si debe pertenecer a algún grupo concreto.
- Contiene una herramienta on-line de diseño de interfaces.
- Utiliza Corpus WorkBench[1] (CWB) como sistema de indexación y búsqueda.
- Permite 3 tipos de búsqueda: simple, avanzada y estadística.

2.1 Formato del corpus

Un corpus que se quiera incorporar a IAC debe estar verticalizado (un token por línea). Además, los atributos a nivel de palabra deben tener formato tabular, es el caso del lema o la categoría morfológica, y los atributos que

1 Corpus Workbench (CWB) ha sido desarrollado por el Institut für Maschinelle Sprachverarbeitung (http://cwb.sourceforge.net/)

afectan a un grupo de palabras, como la función sintáctica o los metadatos del texto, deben estar etiquetados en formato xml (Véase Figura 1).

```
<metadatos titulo = “demo” año=”2010”>
<sujeto>
el              D
chico           N
</sujeto>
canta           V
</metadatos>
```

Figura 1: Ejemplo de formato de un corpus

2.2 Generación de interfaz e indexación de corpus

IAC incorpora una herramienta *user-friendly* para diseñar las interfaces de búsqueda en el corpus. Mediante esta herramienta se introducen los atributos del corpus en IAC para personalizar la interfaz de consulta, sin necesidad de tener conocimientos de diseño ni programación (Véase Fig. 2).

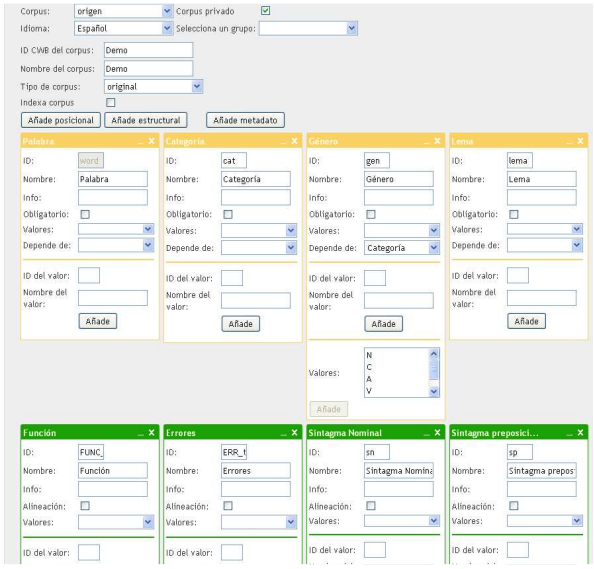

Figura 2: Herramienta para diseñar interfaces

2.3 Interfaces y tipos de búsqueda

Todos los corpus de IAC tienen tres tipos de búsqueda: búsqueda simple, avanzada y estadística.

En la interfaz de búsqueda simple se pueden hacer búsquedas de palabras aisladas (KWOC), determinar el contexto de búsqueda (15, 30, 50 palabras u oración) y filtrar por metadatos. Los resultados que se obtienen son oraciones con el resultado de la búsqueda destacado.

En la interfaz de búsqueda avanzada (Véase Figura 3) también se pueden filtrar las búsquedas por metadatos y determinar el contexto de búsqueda, pero además permite hacer búsquedas contextuales (KWIC). Normalmente, también es posible hacer búsquedas por rasgos lingüísticos, p. ej. categoría principal, género, número, etc., pero esto depende de qué información contenga el corpus y de cómo se haya diseñado la interfaz. Los resultados, al igual que la búsqueda simple, son oraciones con el resultado de la búsqueda destacado.

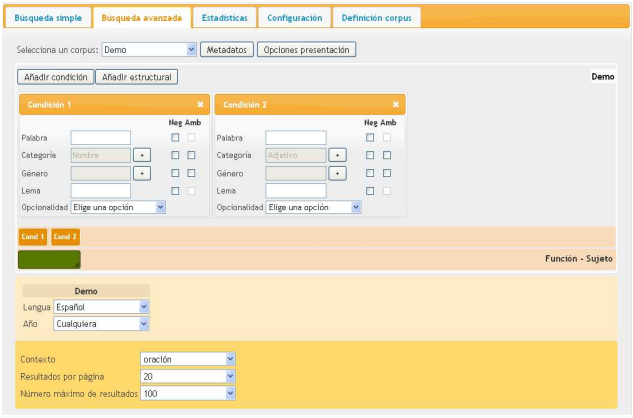

Figura 3: Interfaz de búsqueda avanzada

En la interfaz de estadísticas se pueden hacer búsquedas contextuales. Sin embargo, los resultados son porcentajes sobre cualquiera de los atributos del corpus. Esta opción sólo está disponible para los corpus monolingües.

3 Conclusiones

En este artículo, se ha presentado una herramienta que permite crear interfaces de corpus de una manera fácil e intuitiva. En la demostración se mostrará el proceso de creación de interfaces y de consulta de corpus.

4 Bibliografía

Colominas, Carme and Badia, Toni (2008), The real use of corpora in teaching and research contexts. In: Elia Yuste (ed.), *Topics in Language Resources for Translation and Localisation*, John Benjamins Publishing Company. ISBN: 978 90 272 1688 5. pp. 71-88.

Christ, Oliver (1994) “A modular and flexible architecture for an integrated corpus query system", COMPLEX'94, Budapest.