Introducción a la extracción terminológica bilingüe y bidireccional basada en bitextos para la documentación de intérpretes

Miriam Seghiri
Universidad de Málaga
seghiri@uma.es

Miriam Seghiri Seghiri Universidad

1.Introducción

- Las **fuentes de información** de las que puede beber un intérprete son múltiples y variadas, pues van desde la consulta oral a un experto hasta la utilización de diccionarios y glosarios especializados.

- Sin embargo, existe una **preferencia contundente de los usuarios por uso el glosario/diccionario bilingüe** (Seghiri et al, 2017), seguido muy de lejos por el monolingüe, frente a cualquier otro tipo de recurso, según un estudio realizado a alumnos de la Licenciatura y Grado de Traducción e Interpretación de la UMA.
- Estos resultados son análogos a los de Atkins y Knowles, en la Universidad de Támpere (Finlandia) o el de Meyer y Roberts, en la Universidad de Ottawa.

- Este recurso *ideal* es, a día de hoy, el denominado ***corpus virtual*** (cfr. Laviosa, 1998; Bowker, 2002; Bowker y Pearson, 2002; Zanettin et al. 2003).

- Predilección por el uso del glosario o diccionario por parte de intérpretes.
- El recurso ideal para muchos investigadores es el corpus virtual pues permite ver los términos *in vivo.*

Corpus + glosario= glosario basado en corpus

- Por consiguiente, presentaremos una **metodología protocolizada de extracción de glosarios basados en corpus virtuales bilingües paralelos** para la **documentación en interpretación.**

2. Corpus en Interpretación

¿Qué es un corpus?

- **corpus**, pl. **corpus** (esp.)/ **corpus**, pl. **corpora** (ing.), proviene de la palabra latina *corpus*, i.e. "cuerpo".

*A **collection of texts** assumed to be **representative** of a given language, dialect, or other subset of a language, to be used for linguistic analysis* (Francis, 1982)

A large collection of **authentic text** that have been gathered in **electronic form** according to a **specific set of criteria** (Bowker & Pearson, 2002)

3. Metodología para la creación de corpus virtuales para la labor documental previa en interpretación

3.1.Criterios de diseño

Supuesto:

- Interpretación (inglés->español) en conferencia en la que el ponente va a lanzar al mercado nuevos televisores de la marca Sony.

Diseño del corpus…

- **Tipo textual:** manuales, especificaciones técnicas, etc...
- **Lengua/s:**

inglés (subcorpus 1) y español (subcorpus 2)

- **Original /traducción:** inglés (original) y español (traducción)
- **Texto completo /parcial:** completo

3.2. Protocolo de compilación

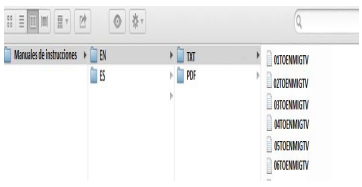
El protocolo de compilación se compone de cuatro pasos:

I. Búsqueda
II. Descarga
III. Formato
IV. Almacenamiento

- **¿Cómo se convierte un documento PDF a TXT?**

http://www.ensode.net/pdf-crack.jsf

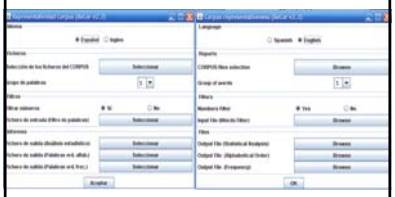
IV. Almacenamiento



- El resultado ha sido la compilación de un corpus **bilingüe** (inglés-español), **paralelo** y **virtual** que se integra por:
  - 1 subcorpus en inglés (20 documentos)
  - 1 subcorpus en español (20 documentos)

¿Son suficientes textos?

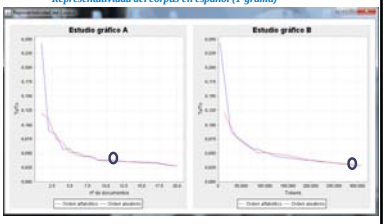
There is no general agreement as to what the size of a corpus should ideally be. In practice, however, the size of a corpus tends to reflect the **ease or difficulty of acquiring the material**. *Usually, the* ***availability*** *of material in the particular field of study* ***determines the final size of the corpus*** (Giouli y Piperidis, 2002).



ReCor

REPRESENTACIONES GRÁFICAS

Representatividad del corpus en español (1-grama)



11 documentos 300 000 palabras

- ReCor (y su algoritmo N-Cor) se encuentra patentado (ref. P200695657) a través de la Oficina Española de Patentes y Marcas del Ministerio de Industria, Turismo y Comercio.
  - Premio extraordinario de doctorado (2005-2011).
  - Premio de Tecnologías de la Traducción de España (2007).
  - Premio en Humanidades "María Zambrana" (2013) .
  - URL (OTRI, Málaga): http://umapatent.uma.es/es/patent/metodo-para-la-determinacion-de-la-representa4b0/ (**alinares@uma.es**)

3.3. Gestión del corpus para la interpretación

Gestión de corpus paralelo (TO+TM)

Paraconc:

http://www.athel.com/para.html

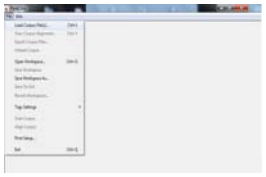


Ilustración 9. Interfaz de ParaConc para cargar los subcorpus

Ilustración 11. Subcorpus (inglés, a la izquierda y en español, a la derecha) subidos a ParaConc

STOP WORD LIST

EN: ranks.nl/stopwords

ES: ranks.nl/stopwords/spanish

Antilock system	Sistema antibloqueo
Assembly conditions	Condiciones de Montaje
Attachment	Accesorio
Audio decoder	Descodificador de audio
Audio Digital Out Socket	Enchufe de salida de audio digital
Audio in Jack	Toma de entrada de audio
Audio output specifications	Especificaciones para la fuente de audio
Audio return channel	Canal de retorno de audio

Glosario de manuales de televisores inglés-español.

Opcional: Transcriptor fonético
EN: photransedit.com ES: aucel.com

4. Conclusiones

- Los corpus **paralelos** son particularmente útiles para cubrir las necesidades documentales del intérprete.
- Un corpus representativo, bien gestionado, es una herramienta de gran utilidad para **identificar, extraer** y **traducir unidades terminológicas, en forma de glosario bilingüe**, que ayuden al intérprete en el **proceso de documentación previo** a una interpretación y **durante la propia** interpretación.

Introducción a la extracción terminológica bilingüe y bidireccional basada en bitextos para la documentación de intérpretes

Miriam Seghiri
Universidad de Málaga
seghiri@uma.es

Miriam Seghiri Seghiri Universidad