LA ALINEACIÓN DE UN CORPUS PARALELO MULTILINGÜE: PROPUESTA DE FASES PARA LA DIDÁCTICA DE TRADUCCIÓN ESPECIALIZADA INVERSA[1]

Cristina Castillo Rodríguez
Universidad de Málaga
ccastillor@uma.es

Resumen: La alineación de corpus paralelos multilingües puede parecer una tarea trivial. Sin embargo, se trata de uno de los procesos más importantes a la hora de gestionar con éxito los textos contenidos en un corpus paralelo multilingüe y así extraer el máximo provecho del análisis contrastivo de traducciones a otra lengua diferente. Generalmente, los textos de un corpus *ad hoc* paralelo multilingüe, extraídos de la red Internet, suelen ofrecer una estructura similar —tanto en el texto original (TO) como en los distintos textos meta (TM) que presenten—. No obstante, en muchas ocasiones los TM ofrecen una estructura diferente con respecto a su homólogo en lengua original, como ocurre, por ejemplo, con textos turísticos españoles del segmento de salud y belleza con versiones traducidas a otras lenguas. En este caso, para que los alumnos, y futuros traductores profesionales, puedan evaluar la calidad de dichas traducciones, no basta con la simple recopilación de los textos y posterior explotación con un programa de gestión de corpus paralelos. El objetivo de este artículo es ofrecer una serie de pautas que deben seguirse para alinear TO con sus respectivos TM cuando éstos últimos presentan una estructura dispar a partir de un corpus de textos *ad hoc* paralelo multilingüe (español-inglés/francés/italiano).
Palabras clave: Lingüística de corpus, alineación, corpus paralelo, traducción especializada.

Abstract: Alignment of multilingual parallel corpora can be seen as a trivial task. However, it is one of the most important process before managing successfully all the texts contained in a multilingual parallel corpus in

order to extract all the benefits from the contrastive analysis of translations into another language. Texts from an *ad hoc* multilingual parallel corpus, compiled from the Internet, usually presents a similar text structure —both in source text (ST) and in target texts (TT)—. Nevertheless, TT may show a text structure different from their ST, as it is the case, for instance, of Spanish tourist texts of wellness and beauty subdomain with their translated versions into other languages. Taking this as a premise, and with the aim of teaching our students of Translation Studies, and future professional translators, how to evaluate the quality of these translations, it would be not enough the simple compilation of the texts and their subsequent exploitation with a parallel corpus management software. The purpose of this paper is to provide certain guidelines that have to be followed in order to align ST with their TT, most of which show a different text structure in an *ad hoc* multilingual parallel corpus (Spanish-English/French/Italian).
Keywords: Corpus linguistics, alignment, parallel corpora, specialised translation.

1. Introducción

El uso de corpus paralelo es muy útil en el campo de la traducción, sobre todo, en lo que respecta al análisis contrastivo de traducciones. No obstante, para gestionar con éxito todos los textos contenidos en un corpus paralelo, los textos meta (TM) deben estar perfectamente alineados con sus textos originales (TO). Normalmente, los TM suelen presentar una estructura similar a la de sus TO, aunque, a veces, encontramos textos que, aunque no dejan de ser traducciones, tienen estructura muy diferente. Es el caso, por ejemplo, de textos turísticos publicados en la red, especialmente, los textos del subdominio del turismo de salud y belleza.

Para poder llevar a cabo un análisis contrastivo de traducciones inversas en una clase de traducción especializada, se deben serguir ciertas pautas previas que conciernen, sobre todo, al proceso de alineación de los textos contenidos en un corpus paralelo multilingüe. En el caso que nos ocupa, hemos compilado, a modo de ejemplo, un corpus *ad hoc* paralelo multilingüe, integrado por TO en lengua española y TM en inglés, francés e italiano. De esta forma, y antes

de la propia gestión de todos los textos, los alumnos deberán ceñirse a una serie de pautas que proponemos en aras de poder evaluar de forma satisfactoria la calidad de las traducciones publicadas en la red, ya que de acuerdo con Barlow (1999), es necesario que los textos de un corpus paralelo estén alineados de forma tal que facilite al usuario una rápida búsqueda de los equivalentes de traducción y una mejor gestión de los textos mediante un sistema informático que se emplee a tal efecto.

Podría, incluso, afirmarse que el proceso de alineación constituye una de las fases más importantes a la hora de compilar y gestionar con éxito un corpus paralelo multilingüe, ya que, según Abaitua (2002: 6), es «el proceso que mayor valor añadido aporta a un corpus multilingüe». Además, este proceso genera lo que se conoce como corpus paralelo alineado y los textos que se alinean pasan a denominarse bitextos, siguiendo la terminología propuesta por Harris (1988: 8), quien, por su parte, además, considera que un TO y su traducción no son en realidad dos textos sino que, en realidad, conforman «a single text in two dimensions, each of which is a language».

El principal problema en la alineación de textos de un corpus paralelo reside, según Rabadán y Fernández Nistal (2002: 76-77), en conseguir preparar los textos del corpus paralelo, es decir, que tanto el TO como el TM se puedan analizar en segmentos, por lo que es necesario que los segmentos de ambos se correspondan de una manera explícita. Aunque las autoras afirman que esta labor puede realizarse de forma totalmente manual (en el caso de corpus de pequeña extensión), la mayoría de las veces es necesario recurrir a sistemas informáticos que permitan la alineación semiautomática de los textos contenidos en un corpus paralelo, como, por ejemplo, con la herramienta *WinAlign*, contenida en el paquete de herramientas de *Trados*, y con el programa *ParaConc*, que contiene un módulo de análisis de alineación sencilla y básica de un TO y sus traducciones, que son las dos herramientas que empleamos para establecer las pautas de alineación de bitextos.

Para la compilación del corpus *ad hoc* paralelo multilingüe del segmento de turismo de salud y belleza, que engloba TO en lengua española y TM en las lenguas inglesa, francesa e italiana, seguimos la metodología protocolizada de compilación de corpus establecida por Seghiri Domínguez (2006)[2]. Si bien la autora establece esta metodología de compilación para corpus comparables, es perfectamente aplicable para el propósito que este artículo, esto es, compilar un corpus paralelo multilingüe, ya que la autora incide en que una de las fases más importantes en la recopilación de los textos que integrarán el corpus es el establecimiento de los criterios de diseño y la búsqueda de la información, que son dos tareas comunes a la compilación de corpus comparables y paralelos.

No obstante, la mayor particularidad de la compilación de un corpus de estas características reside en que los textos deben estar identificados con un código unívoco para cada uno de los registros que se compilen. Además, se deben establecer correspondencias entre los TO y los TM de forma tal que el usuario pueda recuperarlos y gestionarlos fácilmente. Así, si para un registro en lengua española se le da, por ejemplo, el código 1001TOES (donde 1 millar se corresponde con el país, esto es, España; TO es texto original; y ES lengua española), los registros de textos traducidos llevarán los siguientes códigos: 1001TMEN, 1001TMFR y 1001TMIT (donde 1 millar sigue siendo el código de identificación del país, es decir, España, puesto que los textos están publicados en páginas web españolas, aunque estén publicados en diferentes lenguas; TM es texto meta; y EN lengua inglesa, FR lengua francesa e IT lengua italiana).

Una vez que se almacenan correctamente todos los registros con sus códigos de identificación unívocos, los usuarios, en nuestro caso, los alumnos de traducción, podrán comenzar con las fases de alineación que se proponen en este artículo.

2. Fase I: Alineación con WinAlign

El módulo de análisis *WinAlign* permite al usuario llevar a cabo un proceso de alineación de forma muy versátil, ya que presenta un editor interactivo muy fácil de manejar, por lo que resulta bastante útil a la hora de alinear corpus paralelos cuyos textos presentan una estructura diferente. Así, este módulo se ha empleado para la alineación de los textos contenidos en nuestra corpus paralelo multilingüe (español-inglés/francés/italiano) del dominio de especialidad del turismo de salud y belleza.

Como ya se ha mencionado, los textos que integran el corpus compilado presentan la particularidad de que algunos de los segmentos de los textos traducidos (francés, inglés e italiano) muestran una macroestructura diferente con respecto a sus homólogos en lengua original (español). Sin embargo, aunque la estructura final sea bastante diferente, los textos no dejan de ser traducciones, ya que, obviamente, la información contenida en ellos será similar a la ofrecida en los TO. El problema estriba principalmente en que los segmentos traducidos no se encuentran perfectamente alineados con los segmentos del TO y viceversa. Por ello, se requiere llevar a cabo el proceso de alineación con esta herramienta para poder realizar, posteriormente, la explotación y análisis contrastivo.

2.1. Configuración de criterios de alineación

Una vez que se ejecuta *WinAlign*, se debe abrir un proyecto nuevo mediante la opción *New Project* del menú *File* de la barra de herramientas. Para configurar los requisitos de la alineación, dentro de este cuadro de diálogo, en la pestaña *General*, se deben definir: a) la lengua de origen y la lengua meta de los bitextos; b) el tipo de segmentación que se desea, en nuestro caso, los saltos de carro[3]; y c) el tipo de formato de los textos, por ejemplo, texto enriquecido (o formato .rtf) como tipo de archivo[4].

En la siguiente captura de pantalla se refleja el cuadro de diálogo en el que debemos configurar los requisitos previos al proceso de alineación:

Ilustración 1. Cuadro de diálogo para la configuración de la alineación de bitextos

2.2. Alineación de bitextos

La siguiente tarea dentro de esta primera fase de alineación con la herramienta *WinAlign* implica añadir los bitextos desde la pestaña denominada *Files*. De esta forma, los bitextos se alinean como archivos mediante la opción *Align File Names* para que la herramienta pueda comenzar a ejecutar el proceso de alineación de los

segmentos candidatos a ser equivalentes de traducción de los bitextos seleccionados.

A continuación, se abre automáticamente el editor de alineación, proponiendo alineaciones entre pares de segmentos (del TO y del TM). Estas alineaciones de los segmentos aparecen representadas mediante líneas de puntos, como bien se muestra en la siguiente ilustración:

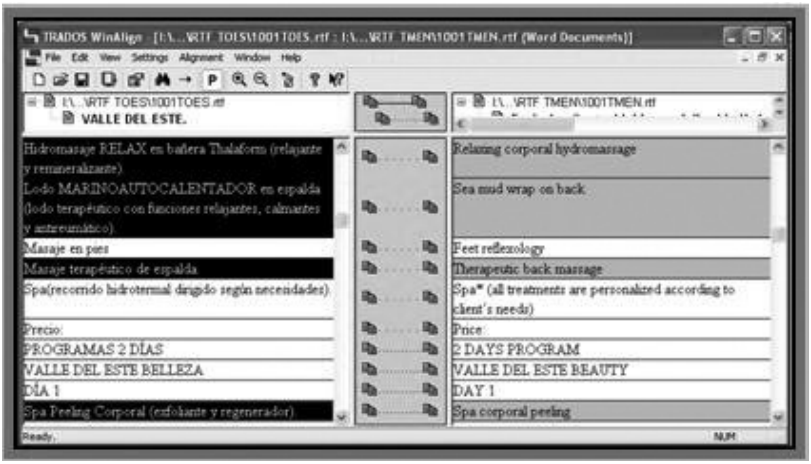


Ilustración 2. Ejemplo de segmentos alineados automáticamente por *WinAlign*

Algunos de los segmentos de los TM del corpus aparecen en una posición distinta a la de los segmentos del TO. No obstante, el usuario siempre puede modificar los pares alineados propuestos automáticamente por el sistema mediante una serie de opciones que se ofrecen pulsando el botón derecho del ratón, como se explica en el subapartado posterior.

2.3. Edición en el proceso de alineación con *WinAlign*

Las principales opciones que ofrece *WinAlign* para poder editar la información arrojada en el proceso de alineación automática son las siguientes que pasamos a explicar a continuación.

La opción *Edit Segment* permite al usuario editar la información contenida en el segmento que se haya seleccionado. Para unir dos

segmentos el usuario debe seleccionar la opción *Join Segments*, mientras que para dividirlos en dos debe seleccionar *Split Segment*. Por otro lado, *Insert Segment* sirve para habilitar un segmento vacío entre dos segmentos dados. La opción *Disconnect* desconecta dos segmentos alineados, mientras que la opción *Commit* confirma la alineación propuesta entre dos segmentos de un bitexto.

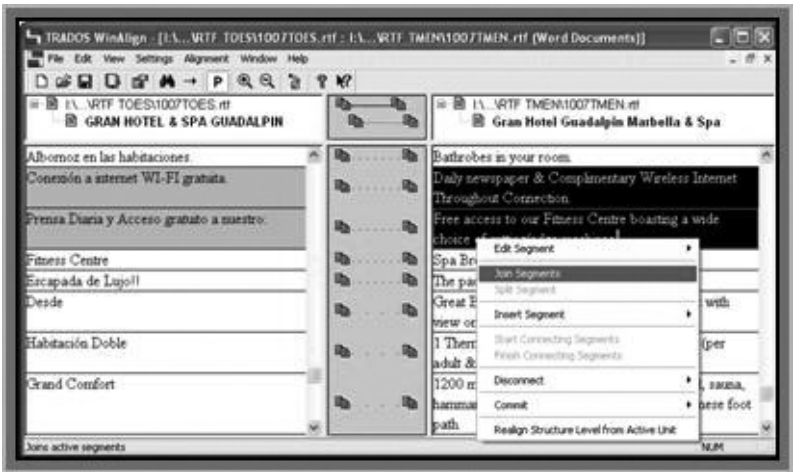


Ilustración 3. Opciones de edición de la alineación en *WinAlign*

La principal ventaja de utilizar este alineador reside en que el usuario puede alinear segmentos independientemente de la posición en que se encuentren tanto en el TO como en el TM, una vez que se han desconectado otros segmentos[5], como bien ejemplificamos en la siguiente ilustración:

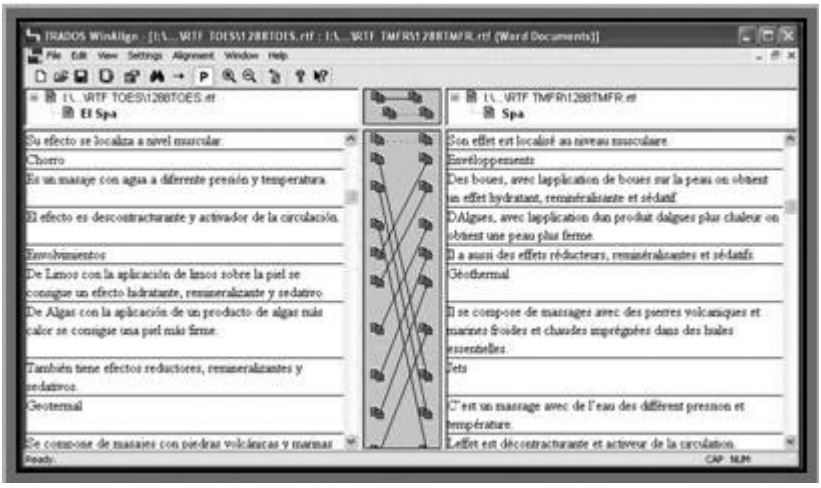


Ilustración 4. Ejemplo de segmentos alineados por el sistema y por el usuario

Cuando se comprueban todos los segmentos del bitexto, se procede a su validación mediante la opción *Commit All Units* del menú *Alignment* situado en la barra de herramientas[6]. A continuación, se guarda el proyecto del bitexto alineado[7] y se exporta como archivo de texto plano (formato .txt o ASCII), que es el formato requerido por el programa de gestión de corpus paralelos *ParaConc*, mediante el cual se procederá a completar el proceso de alineación de los bitextos y continuar así con el análisis contrastivo de los mismos. No obstante, los archivos alineados con *WinAlign* necesitan ser segmentados para que puedan integrarse en *ParaConc*, como explicamos en el paso siguiente.

3. Fase II: Segmentación de archivos alineados

El programa *ParaConc* requiere de archivos de textos individuales para poder realizar análisis contrastivos con los módulos de análisis que ofrece. Por otro lado, el programa utilizado para

la alineación de segmentos nos exporta un único archivo donde se encuentran todos los segmentos alineados, los cuales aparecen identificados con etiquetas descriptivas de la lengua correspondiente, como bien se observa a continuación:

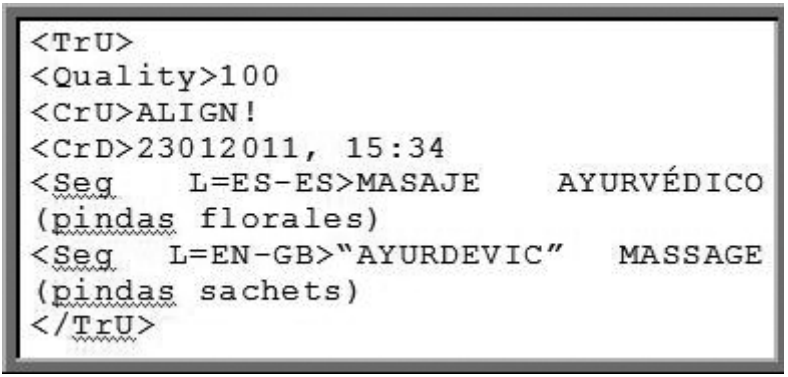


Ilustración 5. Muestra de segmento de bitexto alineado exportado desde *WinAlign*

Para poder segmentar los archivos únicos en formato .txt, exportado desde *WinAlign*, hemos empleado una distribución de Linux denominada Ubuntu (versión 8.10)[8]. Este sistema operativo permite al usuario aplicar un guión (*script*) que automatiza el proceso de segmentación de los textos alineados del corpus paralelo. Este guión se ha aplicado a cada par de lenguas de los componentes del corpus en cuestión; se trata de un guión específico para un intérprete de comando del sistema operativo UNIX, el cual permite reunir la ejecución de diferentes comandos de forma que sea posible realizar una ejecución secuencial sin tener la necesidad de invocar individualmente cada uno de estos comandos.

En el caso de nuestro corpus, se ha ejecutado el comando de segmentación de cada uno de los subcorpus en cada lengua —por un lado, los textos del subcorpus español, que constituyen la lengua original, y, por otro, los textos de los subcorpus francés, inglés e

italiano, que representan las lenguas meta objeto de estudio—, generando así los elementos segmentados en archivos diferentes. Por tanto, estos ficheros contienen los textos segmentados adaptados a las exigencias del *input* de *ParaConc*. Además, también se ha realizado la adaptación del formato del fichero (.txt) para compatibilizar los resultados de la herramienta ejecutada en UNIX con el sistema operativo *host*[9], es decir, Windows.

Las siguientes tres ilustraciones muestran los tres guiones con los algoritmos de separación en cada uno de los bitextos:

```
separawinesfr.sh
#! /bin/bash
for i in *.txt; do
echo "procesando fichero:" $i
echo "Realizando segmentación de la
lengua origen: Español"
awk -f spanish.awk $i > $i".es"
sed -e 's/$/\r/' $i".es" > $i".es.wn"
echo "segmentación realizada en el
fichero" $i".es.wn"
echo "Realizando segmentación de la
lengua meta: Francés"
awk -f french.awk $i > $i".fr"
sed -e 's/$/\r/' $i".fr" > $i".fr.wn"
echo "segmentación realizada en el
fichero" $i".fr.wn"
Done
```

Ilustración 6. Guión que representa el algoritmo de separación del bitexto español-francés

```
separawinesen.sh
#! /bin/bash
for i in *.txt; do
echo "procesando fichero:" $i
echo "Realizando segmentación de la
lengua origen: Español"
awk -f spanish.awk $i > $i".es"
sed -e 's/$/\r/' $i".es" > $i".es.wn"
echo "segmentación realizada en el
fichero" $i".es.wn"
echo "Realizando segmentación de la
lengua meta: Inglés"
awk -f english.awk $i > $i".en"
sed -e 's/$/\r/' $i".en" > $i".en.wn"
echo "segmentación realizada en el
fichero" $i".en.wn"
Done
```

Ilustración 7. Guión que representa el algoritmo de separación del bitexto español-inglés

```
separawinesit.sh
#! /bin/bash
for i in *.txt; do
echo "procesando fichero:" $i
echo "Realizando segmentación de la lengua
origen: Español"
awk -f spanish.awk $i > $i".es"
sed -e 's/$/\r/' $i".es" > $i".es.wn"
echo "segmentación realizada en el
fichero" $i".es.wn"
echo "Realizando segmentación de la lengua
meta: Italiano"
awk -f italian.awk $i > $i".it"
sed -e 's/$/\r/' $i".it" > $i".it.wn"
echo "segmentación realizada en el
fichero" $i".it.wn"
Done
```

Ilustración 8. Guión que representa el algoritmo de separación del bitexto español-italiano

La explicación del algoritmo de separación de estos guiones es la siguiente: 1) `for` selecciona de forma secuencial por orden lexicográfico el archivo con el bitexto de dos archivos que que hay que segmentar; 2) `awk` realiza el proceso de segmentación de cada una de las lenguas, generando un fichero `$i.es`[10], donde `$i` representa las secuencias que se van a segmentar de una determinada lengua; 3) `sed` expresa la adaptación de ficheros entre los dos sistemas operativos; y por último, 4) `echo` imprime en pantalla la información sobre el análisis y la segmentación realizada en los pasos anteriores.

En el sistema operativo Windows, el salto de carro viene representado por dos caracteres, a saber, *CR* y *LF*, que equivalen a los números 13 y 10, respectivamente[11]. En cambio, en los sistemas UNIX, los saltos de carro vienen representados sólo por el carácter *LF*, lo cual puede llevar a incompatibilidad al migrar los ficheros generados de un sistema a otro. En este caso, los archivos segmentados se han realizado en UNIX, mientras que el software gestor de corpus *ParaConc* está albergado en el sistema operativo Windows. Por ello, el comando `sed` permite realizar dicha transformación mediante la aplicación de la regla `'s/$/\r/'`, que sustituye el fin de línea representado por el símbolo del dólar por la secuencia *CR LF*, representado por `\r`. Esta regla se aplica sobre el fichero `$i.es` generado anteriormente por `awk`, obteniéndose así el fichero `$i.es.wn`, el cual representa la adaptación descrita.

Como acabamos de explicar, la herramienta `awk` presenta una utilidad diseñada para el procesamiento de datos basado en texto. El funcionamiento de `awk` conlleva la definición de una serie de reglas que determinan, por ejemplo, qué partes del texto deben ser procesados. Además, cada regla lleva asociada una serie de instrucciones para poder llevar a cabo la manipulación del texto. Partiendo de esta base y del objetivo que nos proponemos, esto es, la extracción de todas las secuencias en cada una de las lenguas de los ficheros generados por el proceso de alineación realizado anteriormente con *WinAlign*, se definirán las reglas con el conjunto de instrucciones asociado a ellas para `awk`, únicas para cada idioma, que se deben ejecutar por cada lengua que se desea extraer y que

consiste en cuatro líneas fundamentales que mostramos, a continuación, en las siguientes ilustraciones:

```
'<Seg L=ES-ES>'
        {sub(/\r$/,"")
        gsub(/\\endash/,"",$0)
        $(NF+1)
="</Seg>";print $0}
```

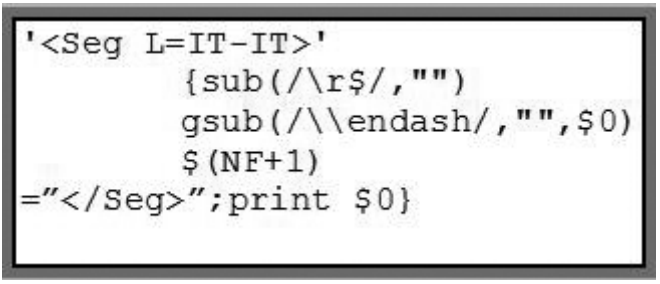
Ilustración 9. Instrucción generada por la herramienta *awk* para los textos en lengua española

```
'<Seg L=EN-GB>'
        {sub(/\r$/,"")
        gsub(/\\endash/,"",$0)
        $(NF+1)
="</Seg>";print $0}
```

Ilustración 10. Instrucción generada por la herramienta *awk* para los textos en lengua inglesa

```
'<Seg L=FR-FR>'
        {sub(/\r$/,"")
        gsub(/\\endash/,"",$0)
        $(NF+1)
="</Seg>";print $0}
```

Ilustración 11. Instrucción generada por la herramienta *awk* para los textos en lengua francesa



```
'<Seg L=IT-IT>'
        {sub(/\r$/,"")
        gsub(/\\endash/,"",$0)
        $(NF+1)
="</Seg>";print $0}
```

Ilustración 12. Instrucción generada por la herramienta *awk* para los textos en lengua italiana

La primera parte, donde se expresa `'<Seg L=ES-ES>'`, `'<Seg L=EN-EN>'`, `'<Seg L=FR-FR>'`, y `'<Seg L=IT--IT>'`, representa la regla encargada de seleccionar únicamente el idioma determinado por la etiqueta L. La diferencia entre las distintas herramientas para `awk` viene determinada precisamente por la especificación de la regla, lo cual revela que por cada idioma existe una regla diferente. Las instrucciones de procesamiento son únicamente cuatro. Las sentencias, descritas a continuación, se aplican sobre la línea actual procesada (representada por `$0`, en la herramienta `awk`), es decir, una frase correspondiente al idioma dado por la etiqueta L.

La primera instrucción, esto es, `sub(/\r$, "")`, elimina los saltos de carro que aparecen en los ficheros del sistema operativo Windows. Para ello, se sustituye la cadena `\r` correspondiente a *CR LF* por la cadena vacía, indicada en las comillas. La segunda instrucción, `gsub(/\\endash/,"", $0)`, indica que todas las apariciones de `\endash` deben ser reemplazadas por la cadena vacía sobre `$0`. Esta instrucción es necesaria ya que el alineador que hemos utilizado, esto es, *WinAlign*, introduce esa secuencia para la representación textual del símbolo de la flecha.

La tercera instrucción añade al final de cada frase el cierre de la etiqueta Seg (representado por `</Seg>`) para lo cual se necesita añadir el campo nuevo a la frase, representado por NF+1[12]. Por

último, la cuarta instrucción simplemente imprime en pantalla la nueva sentencia actual, modificada por las anteriores instrucciones, representada por la regla `print $0`.

De esta forma, todos los archivos de nuestro corpus se encuentran listos para ser cargados dentro del programa *ParaConc*.

4. Fase III: Alineación con ParaConc

ParaConc[13] es un programa de concordancias multilingüe que permite realizar análisis contrastivos mediante corpus paralelos. Está integrado por un conjunto de módulos de análisis que facilita la gestión de corpus paralelos y, además, ofrece una interfaz muy sencilla, lo cual posibilita al usuario su fácil manejo a la hora de empezar a ejecutar el programa, una vez que se ha instalado previamente.

Por otro lado, el programa requiere que los textos contengan segmentos alineados en cada uno de los textos escritos en diferentes lenguas, por lo que la alineación es crucial para llevar a cabo con éxito las distintas tareas que ofrecen los módulos de análisis contenidos en *ParaConc*, puesto que cuando el programa busca, por ejemplo, en los TO, la única información que tiene sobre los enlaces entre las diferentes lenguas cargadas es la ofrecida por la propia alineación previa.

A pesar de estas limitaciones, el programa contiene un módulo de análisis por el cual se permite una alineación sencilla de los bitextos integrados.

4.1. Carga de textos del corpus paralelo

Para que el programa pueda alinear todas las secuencias contenidas en los textos es necesario cargarlos dentro del programa. Para ello, se debe seleccionar la opción *Load Corpus File(s)*[14] del menú *File* de la barra de herramientas inicial. A continuación, se abre un

cuadro de diálogo que permite cargar de dos a cuatro archivos de textos paralelos[15], como mostramos en la siguiente ilustración, con ejemplos de carga de archivos del corpus paralelo:

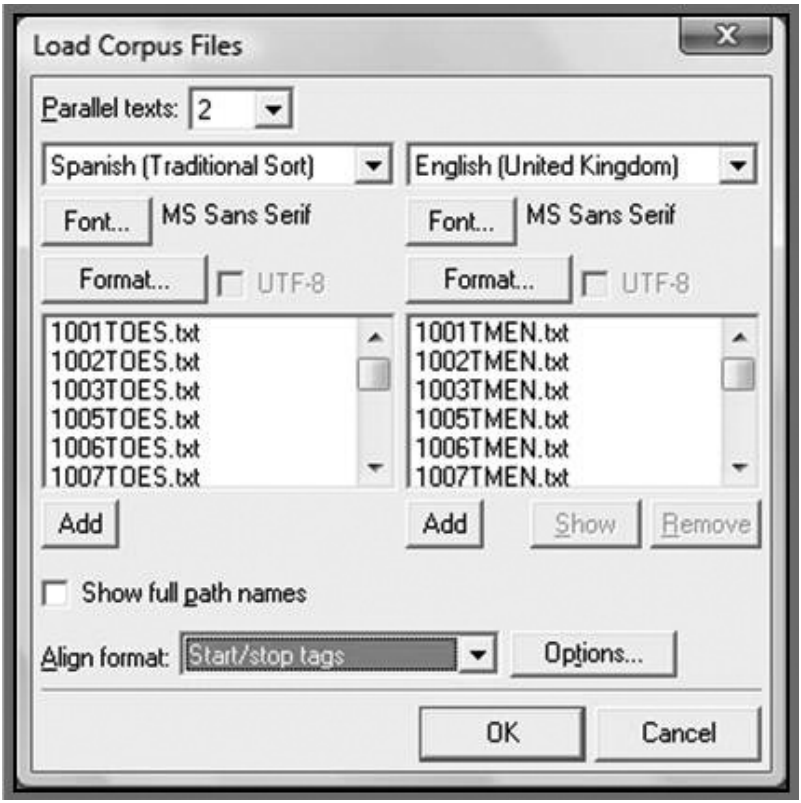


Ilustración 13. Proceso de alineación de archivos paralelos en *ParaConc*

Si alguno de los archivos no estuviera perfectamente alineado con los demás, el programa permite colocarlo en su sitio correspondiente, simplemente seleccionando el archivo en cuestión para «arrastrarlo» hasta su lugar en la lista de los archivos cargados (Barlow, 2003: 19).

4.2. Alineación de bitextos

Una vez que se han cargado todos los TO y los TM del corpus paralelo que se va a analizar, el usuario debe seleccionar el indicador de alineación adecuado. El formato de alineación que ofrece el programa por defecto es *New line delimiter*, esto es, la delimitación mediante saltos de carro, aunque para este tipo de corpus se debe seleccionar otro tipo de delimitación, en concreto, *Start/Stop tags*, que se encuentra en el desplegable *Align format*.

Cuando se valida la alineación, los textos del corpus se encuentran perfectamente cargados dentro del programa. Para comprobar la alineación de cada uno de los textos, simplemente se selecciona la opción *View Corpus Alignment* del menú *File* de la barra de herramientas; a continuación, se pincha en los pares de textos alineados y se selecciona la opción *Alignment*, como bien ilustra la siguiente captura de pantalla con ejemplos de selección de archivos de nuestro corpus:

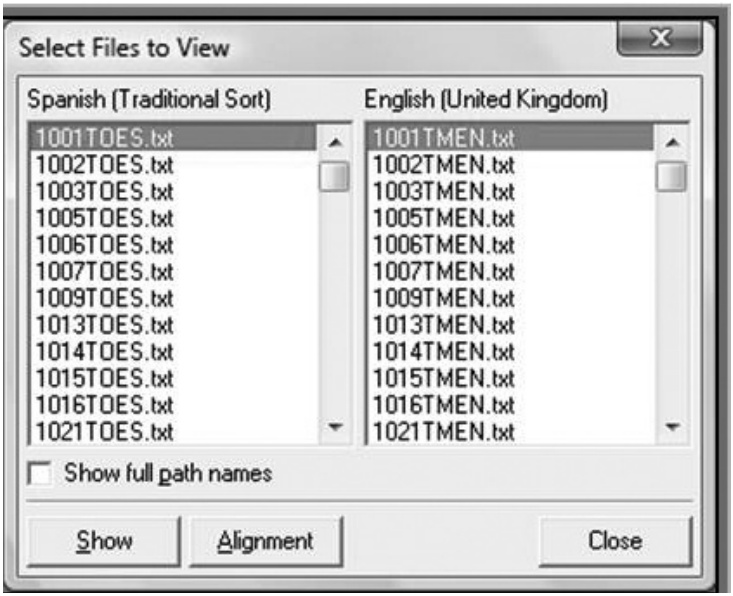


Ilustración 14. Selección de archivos para comprobar su alineación en *ParaConc*

En este cuadro de diálogo, el usuario debe comprobar que los archivos se encuentran alineados correctamente, es decir, que el archivo 1001TOES en español esté alineado con su correspondiente 1001TMEN en inglés[16]. Tras esta comprobación, el usuario puede proceder a la alineación del contenido de cada uno de los bitextos.

4.3. Edición en el proceso de alineación con *ParaConc*

A pesar de que la alineación se realiza de forma correcta dentro del programa, es posible que éste desconecte algunas palabras o, incluso, frases, ya que las identifica como secuencias diferentes. Es el caso, por ejemplo, de determinadas abreviaturas que incluyen puntos. Para poder integrar cada una de las letras pertenecientes a la abreviatura, el programa permite conectarlas de nuevo activando la opción de *Merge with Next Sentence* con el botón derecho del ratón[17], como bien ilustramos a continuación:

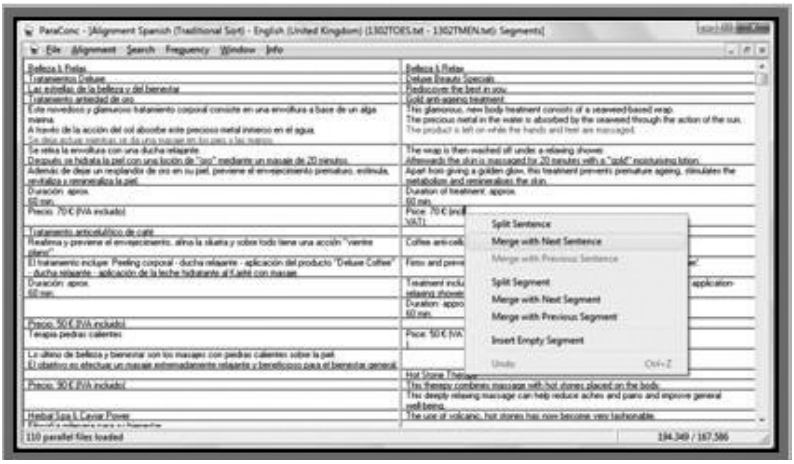


Ilustración 15. Ejemplo de la opción *Merge with Next Sentence* de *ParaConc*

Además de esta opción, el alineador de *ParaConc* permite otras ediciones en la alineación de los bitextos, como puede observarse

en el desplegable de la ilustración anterior. Las opciones de edición de este programa son las siguientes: *Split Sentence*, que permite dividir una oración en dos; *Split Segment*, para separar un párrafo de otro; *Merge with Next Segment*, para unir un párrafo con otro anterior[18]; y la opción *Insert Empty Segment*, para introducir un párrafo en blanco, siempre y cuando en alguno de los textos (original o traducido) se muestre el mismo párrafo vacío. Además, se ofrece la opción *undo*, mediante la que el usuario puede deshacer una tarea ejecutada.

Una vez que se han alineado todos los del TO y del TM en cuestión, se puede proceder al análisis contrastivo de los textos a través de otros módulos de análisis contenidos en el programa *ParaConc*.

5. Conclusiones

El proceso de alineación de corpus paralelos constituye una de las tareas más importantes para la correcta gestión, análisis y evaluación de la calidad de los textos contenidos en estos tipos de corpus, convirtiéndose, incluso, en una tarea crucial e imprescindible, sin la cual no se podría llevar a cabo con éxito un análisis contrastivo de TO y TM.

Generalmente, basta con utilizar un programa de gestión de corpus paralelos que contenga un módulo de análisis de alineación sencillo, como, por ejemplo, el programa *ParaConc*. Sin embargo, en algunas ocasiones los TM contenidos en los corpus paralelos pueden presentar una estructura diferente a la de los TO. En estos casos, es preciso utilizar otro tipo de alineadores más versátiles que permitan llevar a cabo el proceso de alineación de los bitextos, como es el caso del módulo de análisis *WinAlign* del programa de herramientas *TRADOS*.

En este artículo hemos ofrecido una serie de pautas que deben seguirse cuando la estructura y, por tanto, las secuencias de los TO y TM están organizadas de forma muy diferente, como, por ejem-

plo, en el caso de un corpus paralelo multilingüe del segmento del turismo de salud y belleza, integrado por TO en lengua española y TM en inglés, francés e italiano.

La primera fase consiste en alinear los bitextos, original y traducido, con la herramienta *WinAlign*. A continuación, la segunda fase implica la segmentación de los archivos generados por esta herramienta mediante la aplicación de un guión para poder automatizar esta segmentación. Y, por último, en la tercera fase, los textos segmentados se alinean, de nuevo, con el módulo de alineación del programa *ParaConc* para que éste pueda comenzar a ejecutarse y se pueda llevar a cabo el análisis contrastivo de los bitextos con éxito y evaluar la calidad de los textos traducidos a otras lenguas diferentes de la lengua materna.

Note

1. El presente trabajo ha sido realizado en el seno del proyecto *Ecosistema* (FF/2008-06080-C03-03/FILO, Ministerio de Ciencia y Tecnología).

2. A este respecto remitimos a la publicación de Corpas y Seghiri (2009), donde también ejemplifican esta metodología protocolizada de compilación de corpus comparables.

3. Los saltos de carro vienen identificados mediante el símbolo del calderón.

4. Otros posibles formatos son: .doc, .ppt, .pps, .pot, .xls, .xlt, .xml, .html, entre otros.

5. Se trata de un proceso muy sencillo mediante el cual se unen segmentos utilizando el ratón, es decir, haciendo doble clic en el segmento del TO y arrastrando una flecha hasta el segmento del TM candidato a ser alineado.

6. Aunque el usuario puede ir validando los segmentos alineados mediante la opción *Commit*, como hemos descrito anteriormente, la opción de *Commit All Units* valida todos los segmentos alineados, tanto por el alineador como por el propio usuario, una vez que se ha procedido a la revisión de todos los segmentos de un bitexto dado.

7. Los proyectos se guardan a través de la opción *Save Project As*, del menú *File* situado en la barra de herramientas. Además, todos los proyectos se guardan en formato .pjt que, aunque no es compatible con el formato requerido por el programa *ParaConc*, sí es útil para volver a recuperar los bitextos alineados en el caso de requerir una revisión posterior de los mismos.

8. Para llevar a cabo esta segmentación de bitextos nos hemos basado en el algoritmo de separación propuesto en Corpas Pastor (2008). No obstante, debido a las exigencias del programa utilizado para esta investigación, se ha realizado una serie de adaptaciones que nos han llevado a la descripción del algoritmo empleado a tal efecto.

9. El sistema operativo bajo el cual funciona el programa *ParaConc*.

10. Este fichero indica el texto que integra las secuencias en lengua española. Para la lengua inglesa, francesa e italiana los ficheros generados se denominan $i.en, $i.fr y $i.it, respectivamente.

11. Códigos ASCII que representan a esos dos caracteres.

12. Advertimos aquí que la herramienta *WinAlign* no lo introduce, aunque para seguir los estándares del lenguaje XML es recomendable añadir este campo nuevo.

13. El programa ofrece una versión de prueba que puede descargarse directamente desde su página web, aunque el número de resultados se restringe a 150, los cuales no pueden ni guardarse ni imprimirse. Para descargar la versión demo, remitimos a la siguiente URL, donde, además, se puede adquirir el producto: <http://www.athel.com/para.html>.

14. En el manual del programa *ParaConc* (cf. Barlow, 2003), se indica que esta opción puede activarse directamente mediante el *shortcut* Ctrl+L.

15. Aunque el corpus paralelo contiene cuatro componentes en distintas lenguas, no se procede a la alineación de cuatro archivos, debido a que no todos los textos originales presentan traducciones en las tres lenguas restantes y, si las muestran, en muchas ocasiones la información contenida en dichas traducciones está organizada de forma diferente.

16. Los archivos alineados pueden guardarse mediante la opción *Save Workspace As* en un archivo con extensión .pws, de forma que estarán disponibles cuando el usuario desee recuperarlos, sin tener la necesidad de volver a cargar de nuevo cada uno de los textos del corpus multilingüe.

17. Otra forma posible es realizar el proceso a la inversa, esto es, situando el cursor del ratón en la secuencia posterior de la palabra cortada y pulsando sobre la opción *Merge with Previous Sentence*.

18. Al igual que en el caso de unión de oraciones, se puede proceder a llevar a cabo este proceso a la inversa, es decir, colocando el cursor en el párrafo posterior para unirlo con el anterior mediante la opción *Merge with Previous Segment*.

Bibliografía

ABAITUA ODRIOZOLA, J. (2002). Tratamiento de corpora bilingües. En M.A. Martí Antonín y J. Llisterri Boix (eds). *Tratamiento del lenguaje natural*. Barcelona: Universidad Autónoma de Barcelona. 61-90.

BARLOW, M. (1999). MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics*. 4 (1): 319-27.

BARLOW, M. (2003). *ParaConc: A Concordancer for Parallel Texts (Draft 3/03)*. Houston. Athelstan.

CORPAS PASTOR, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt: Peter Lang.

CORPAS PASTOR, G. y SEGHIRI DOMÍNGUEZ, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). En A. Beeby, P. Rodríguez Inés y P. Sánchez-Gijón (eds.). *Corpus Use and Translating*. Ámsterdam y Filadelfia. John Benjamins Publishing Company: 75-107.

HARRIS, B. (1988). Bi-text, a New Concept in Translation Theory. *Language Monthly*. 54: 8-10.

RABADÁN ÁLVAREZ, R. y FERNÁNDEZ NISTAL, P. (2002). *La traducción inglés-español: fundamentos, herramientas, aplicaciones*. León. Servicio de publicaciones de la Universidad de León.

SEGHIRI DOMÍNGUEZ, M. (2006). *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. Tesis doctoral. Málaga: Universidad de Málaga.