ALINEAMIENTO DE FRASES Y TRADUCCIÓN: ALFRACOVALT Y EL PROCESAMIENTO DE CORPUS*

Josep R. Guzmán
Àlvar Serrano
Universitat Jaume I

Resumen

La aparición de los corpus en el campo de la traductología ha motivado la necesidad de generar instrumentos que permitan manejar toda esta información de forma rápida y efectiva. En este orden de cosas, este trabajo se ocupa de la presentación de una herramienta de alineamiento de textos paralelos integrados en el corpus COVALT (*Corpus Valencià de Literatura Traduïda*). Así pues, tras la realización de un pequeño repaso de los diversos métodos y instrumentos de alineamiento, se analizan las características del programa, AlfraCOVALT, especialmente por lo que se refiere a la utilidad para el investigador de la traducción y su necesidad de alineamientos ajustados.

Palabras clave: corpus, traducción, alineamiento, textos paralelos.

Abstract

The use of corpus in the field of translation has motivated the need to create new instruments to deal with a great deal of information in a short period of time. From this perspective, the present article explains how an aligning tool integrated in the COVALT (*Corpus Valencià de Literatura Traduïda*) corpus works. In particular, we review the different methods and aligning tools and illustrate the characteristics of the application, AlfraCOVALT, mainly those highly relevant for translation research.

Keywords: corpus, translation, aligning, parallel texts.

1. Introducción

Dependiendo de la tipología textual, el alineamiento de frases a partir de un determinado segmento de texto puede resultar en numerosas ocasiones un hecho bastante arduo. Esta dificultad resulta todavía más significativa cuando se trata de textos narrativos, en los cuales, no sólo las omisiones, reordenamientos o inserciones tienen una mayor relevancia, sino que todo aquello que englobamos dentro del término creatividad o estilo particular tienen una importancia capital. Podemos decir que los

* Este trabajo se ha realizado en el marco del proyecto financiado por el MCT –BFF 2003-05422 y Bancaja P1-1B2003-25

hapax legomena, los distintos grados de lexicalización de determinados conceptos, objetos, acciones, etc., en lenguas distintas, o simplemente los referentes, generan traducciones que se alejan de forma significativa de lo que resultaría la traducción literal.

Así mismo, las diversas necesidades en campo de la enseñanza y el estudio de la traducción han motivado el desarrollo de diferentes clases de aplicaciones informáticas que intentan solucionar el problema del alineamiento de textos enunciado en la párrafo anterior. En este trabajo presentamos nuestras investigaciones para colaborar en esta tarea y construir un instrumento de ayuda para el estudio de traducciones editadas en el contexto valenciano durante el periodo de 1990-2000. Este periodo es el que engloba el corpus COVALT (*Corpus Valencià de Literatura Traduïda*) (Lawick, 2005; Marco, 2006; Burdeus y Verdegal, en prensa; Guzmán, en prensa). El corpus COVALT es un corpus paralelo formado por un conjunto de pares de textos escritos en diversas lenguas que van desde el inglés hasta el árabe, pasando por el griego moderno, o el vasco, entre otras. Sin embargo, exclusivamente por la significación de la muestra, únicamente se han considerado los textos originales en alemán, francés o inglés y traducidos al catalán: en total más de cuatro millones de palabras, correspondientes a 42 obras en inglés, 27 obras en francés, 10 obras en alemán, y sus traducciones correspondientes al catalán. A fin de operar más fácilmente con este corpus paralelo se elaboró un programa gestor, AlfraCOVALT, que, por una parte, permite búsquedas entre las diversas entradas según el autor, traductor o la lengua de los textos y, por otra, realiza el alineamiento de frases entre el texto original (TO) y el texto traducido (TT). Así pues, este trabajo hace un somero repaso de los diversos sistemas de alineamiento de textos para posteriormente analizar tanto las características de AlfraCOVALT como su eficacia en esta labor.

2. Sistemas de alineación

La mayoría de los métodos existentes de alineamiento se basan en dos corrientes: la iniciada por Kay y Röscheisen (1988) y la de Brown et al. (1991) y Gale y Church (1991). Estos dos planteamientos parten del uso, o no, de información léxica. El método de Kay y Röscheisen (1988) considera que las palabras que forman los textos se corresponden, es decir, que la información necesaria para la alineación se encuentra en los propios textos de manera que lo que hay que hacer es comparar la distribución que tienen las palabras en cada uno de los textos. Por su parte, en el segundo método, aunque también se utiliza información interna, ésta no es léxica. Lo que se utiliza es la longitud del texto y de las frases (según el número de caracteres), ya que se considera que existe una relación constante de estas longitudes entre un texto y otro. De ambos métodos se han ido produciendo modificaciones, según se han introducido otros factores tales como: la búsqueda de pares de secuencias de caracteres iguales en ambos textos (Church 1993; Simard et al. 1992, 1998), la búsqueda

de las distancias entre las ocurrencias de una palabra (Fung y McKeon 1996), la construcción de un modelo estadístico de traducción palabra a palabra (Chen 1993); la adaptación a lenguas muy alejadas (Haruno y Yamazaki, 1997); etc.

Actualmente lo que se ha producido es una combinación de diferentes elementos como: la longitud del texto y la longitud de las frases, la información léxica, reconocimiento de cognados, introducción de métodos de filtro (*stoplist*) y reducción de espacio de búsqueda, etc. (Macklovitch y Hannan 1998; Melamed 2000; etc.). En este sentido encontramos diversos trabajos más o menos recientes que han realizado un estudio de las diversas aportaciones a los métodos de alineamiento de textos paralelos (Och y Ney 2003), o concretamente al alineamiento a nivel de palabra (Mihalcea y Pedersen 2003; Hao y Gildea, 2004).

Como colofón de todo este proceso han aparecido numerosos programas e instrumentos para el trabajo con textos paralelos. Estos van desde herramientas que permiten la búsqueda de textos paralelos en internet (*Strand data*), a la creación de programas específicos como los elaborados para el corpus *MULTEXT-East* —*MtRecode*: de conversión de caracteres, *MtSgmlQL*: de búsquedas, *MtSeg*: segmentador de textos, *MtLex*: de acceso léxico o el *MtTag*: programa que realiza también desambiguaciones—, o el *Uplug* que reúne diversas herramientas que podemos encontrar por separado —*Giza++*, diversos etiquetadores como *TreeTagger*, *TnTtagger*—, etc.

En general estos programas presentan variaciones significativas algunas de las cuales pasan por: la interface de comunicación con el usuario, el tipo de lenguaje con el que esta escrito, o los objetivos que se buscan, como por ejemplo, la concordancia de palabras o la de frases. En definitiva, encontramos programas escritos en Perl como *Alpaco* (*Aligner for Parallel Corpora*), en Java, como *Cairo*, o en C++, *Align* (A.Berger). Igualmente cabe destacar algunos que son específicos para pares de lenguas, como sueco e inglés —*PLUG Word Aligner* (PWA)—, o húngaro e inglés, como *Hunalign*. Por una parte, algunos de estos programas se concretan en el alineamiento a nivel de palabra: *Lingua-AlignmentSet*, *UMIACS Word Alignment Interface*, el *Natura Alignment Tools* (*NATools*), *Twente*, *Kvec++* —en algun caso con la necesidad de que los textos hayan sido previamente alineados a nivel de frase (*I*Link*)—, mientras que otros se ciñen al alineamiento a nivel de frase, como el *Geometric Mapping and Alignment* (GMA) —para plataformas como la Linux o Solaris; o como la propuesta de *Bilingual sentence aligner* (Microsoft). Finalmente, cabe señalar que mientras unos siguen la implementación del algoritmo de Gale y Church, como el *Vanilla*, otros tienen una dependencia total de la información léxica como el *CTK*: *Champollion Tool Kit*.

3. AlfraCOVALT: estructura y funcionamiento

Lo que se desprende del apartado anterior es la gran abundancia de propuestas que se han realizado para alinear palabras o frases de textos paralelos. Sin embargo,

existen una serie de lagunas que quedan en buena medida sin cubrir, entre las que habría, entre otras: las lenguas alineadas (lenguas no suficientemente representadas), la preparación de los textos para ser alineados y una interface amigable. Por lo que respecta a las lenguas implicadas, los investigadores y docentes en traducción al catalán y del catalán se encuentran que, tal como se ha visto, aunque algunos de los algoritmos de los programas mencionados anteriormente pretenden poder ser utilizados con cualquier lengua, o incluso pueden operar con lenguas muy diversas —que van desde la mayoría de las europeas (con presencia actual o futura en el Parlamento Europeo), a lenguas asiáticas, como el japonés o el chino, entre otras—, la inmensa mayoría de ellos no tienen el catalán como una de las lenguas específicas en los pares de lenguas. Obviamente las causas que motivan este hecho son muy diversas, y no es éste el momento de explicitarlas.

La segunda de las lagunas mencionadas se refiere al hecho de que en muchos programas, para que la búsqueda obtenga una precisión significativa necesitan recurrir a ajustes manuales en los alineamientos de las frases —cuando el alineamiento no es simplemente manual como en el programa *Alpaco*. Nuestro interés radica en que esta acción quede reducida a la mínima expresión, ya que el investigador necesita eliminar la mayor cantidad posible de procesos manuales o tediosos y que estas tareas estén lo más automatizadas posible. Junto a esto habría que añadir la necesidad de aumentar también la precisión en el alineamiento de las frases cruzadas, es decir aquellas que son producto de inversiones, o de la segmentación de otras.

Un tercer aspecto que habíamos apuntado era la necesidad de una interface amigable. Muchos de los programas mencionados resultan de acceso engorroso, tanto por lo que se refiere a la introducción de los ficheros a alinear, o la generación de corpus, como del acceso a los ficheros de resultados. Y esto en muchos casos con la utilización de formatos más o menos específicos, como por ejemplo COAL, CES, etc., y con formatos de visualización como HTML, etc.

A fin de paliar, aunque sea en parte las carencias anteriores, surgió la necesidad de crear un instrumento que realizara dos funciones: 1) gestión del corpus COVALT y 2) alineamiento de los textos paralelos en un formato generalizado y que facilitara la comprobación de los resultados obtenidos: AlfraCOVALT.

3.1 Gestión del corpus COVALT

Para esta primera función el programa gestiona una base de datos externa en formato *Access* con diversas tablas: una que incluye todos los datos del corpus COVALT (Tabla1) y otras que incluyen códigos de búsqueda y el campo que ha de ser seleccionado. La tabla1 incluye:

—autor (apellidos y nombre)
—Título original

—Lengua de partida (inglés, francés y alemán)
—texto original
—Título traducción
—Lengua de llegada (catalán, español)
—texto traducción
—traductor (apellidos y nombre)
—y diversos campos de código de autores, de traductores y de obras.

También incluye otros datos como por ejemplo el ISBN, editorial, etc., pero que actualmente no tienen ninguna relevancia. En total la base dispone de 143 registros que incluyen referencias a textos en otras lenguas, como ya se ha indicado. Las otras tablas contienen únicamente dos campos: el código (que puede ser de autor, de traductor, o de título de la obra —original o traducción—, y el segundo campo, que recopila los nombres correspondientes. Esta gestión del corpus se realiza con una interface amigable como se ve en la Figura 1:

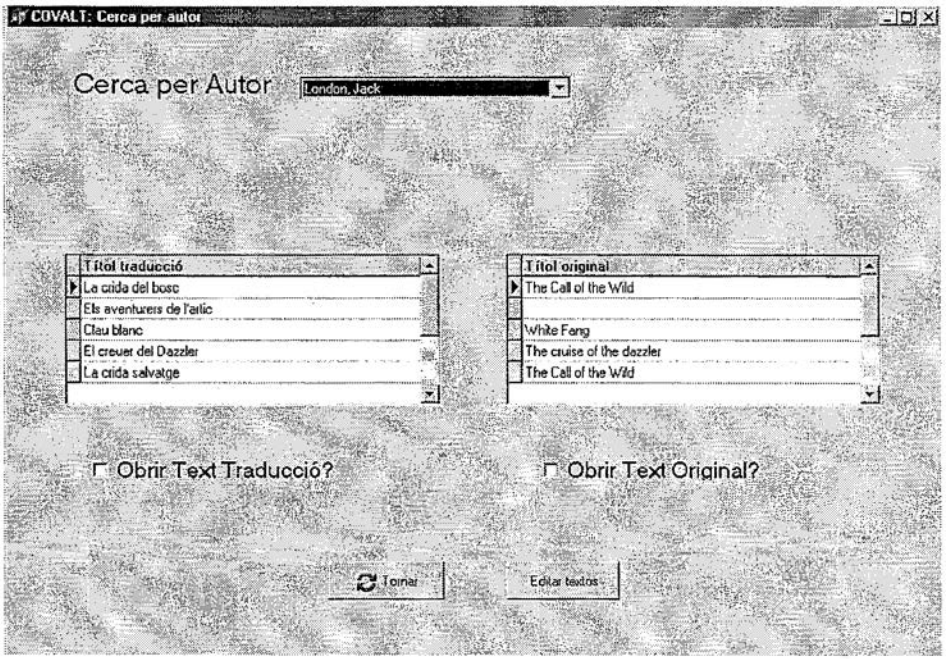

Figura 1. Búsqueda por autor en el corpus COVALT

En esta figura se presenta la búsqueda por autor (en este caso la de Jack London), y donde aparecen los títulos registrados en el corpus tanto por lo que se refiere a los originales como a las traducciones correspondientes. Igualmente, existen

otros botones que sirven para seleccionar los textos o para editarlos, o para volver a la ventana anterior de la aplicación.

3.2 AlfraCOVALT y el alineamiento de los textos paralelos

El programa de alineamiento opera con una serie de fuentes externas y internas. Las fuentes externas están formadas por tres bases de datos *Access* correspondientes a otros tantos diccionarios con dos campos, uno con una lista de lemas en catalán y otro según la lengua de origen del texto (inglés, francés, o alemán), todos ellos indexados y con duplicados. La base de alemán tiene 54.581 entradas, la de inglés 45.399 y la de francés 73.474.

Por lo que respecta a las fuentes internas, éstas corresponden a bases de datos *Paradox* que recogen la segmentación de las frases de los textos que han de ser alineados. Esta segmentación se realiza a partir del componente *TRegExp* realizado por Andrey Sorokin ([htpp://anso.virutalave.net/delphi_stuff.htm]) para trabajar las expresiones regulares en el entorno de Delphi (que es en el que está escrita la aplicación). La expresión regular es la siguiente:

```
[.\s(?)¿¡!;]*[\w\s),\[ \]«»ÇœÄÖËÜäöüïëßâêôûîàéèíòóúùç\('’·-]*
```

La expresión esta compuesta por dos clases de caracteres separadas ambas por los corchetes. La primera clase de caracteres se encuentra marcada por el asterisco que nos indica que esta clase se puede repetir las veces que sea necesario. Esta clase está constituida por los límites de la frase: el punto, como signo de puntuación ortográfico (que en este caso no es un metacaracter porque se halla dentro de una clase de caracteres); «\s» que es un metacaracter que representa cualquier espacio acompañando al punto; los signos de interrogación, o admiración, y el punto y coma. En este sentido cabe señalar que aunque la definición de frase siempre ha sido dentro de la lingüística un concepto controvertido, hemos optado por una serie de signos que caracterizan estos límites, excluyendo algunos otros que también podían ser claramente considerados como tales, como por ejemplo los dos puntos.

La segunda clase de caracteres comienza por «\w» que indica que la expresión puede ser cualquier carácter alfanumérico incluido «_». La clase incluye todo tipo de caracteres específicos en las diversas lenguas, como «œ» del francés, la «ß» de la lengua alemana o los diversos tipos de vocales acentuadas tanto en francés, catalán, español o alemán.

Cada frase se ubica en la tabla según la posición de ocupa en el texto el primer carácter válido, es decir tras eliminar los posibles caracteres iniciales de frase como por ejemplo los espacios en blanco, y los posibles retornos de carro.

Una vez seleccionados los textos, las bases correspondientes a cada texto son almacenadas en vectores (en principio de valor 100, pero que según las necesidades

se van redimensionando) que agrupan, por una parte, las frases del texto traducido y su posición y, por otra, las del texto original y su posición en el texto. El siguiente paso es el de determinar la cadena lingüística de búsqueda, la cual define a su vez si el sentido de esta búsqueda es a partir del texto traducido o del texto original, o lo que es lo mismo: el texto de referencia depende de la dirección de la búsqueda. Esta es una diferencia sensible con otras aplicaciones ya que en la mayoría de los casos el texto de referencia es el texto original.

Dependiendo de las lenguas implicadas se activan también: 1) las fuentes externas (diccionario de alemán, francés, o inglés); 2) la lista de palabras omitidas (*stoplist*), es decir, aquellas palabras que si no fuesen eliminadas podrían falsear los datos; y 3) la lista de variaciones morfológicas (género, número, flexión verbal, etc.), ya que en las fuentes externas las entradas aparecen lematizadas.

Una vez localizada la cadena deseada en el Texto 1 (T1) partimos, en parte, de los métodos basados en la longitud del texto (Gale y Church, 1991). Según este trabajo, existe una correlación entre la posición de la frase de T1 y la de T2. Igualmente, también consideramos que el alineamiento biunívoco —una frase de T1 tiene su correspondencia en una frase de T2— es el más frecuente porcentualmente (como se desprende también de nuestros resultados en el apartado 4).

Para determinar si la expresión que buscamos se encuentra en la frase posible del T2, en primer lugar extraemos las palabras de la frase del T1 teniendo en cuenta, en primer lugar, las variaciones morfosintácticas de género y número, de forma que se descartan algunos sufijos y, en segundo lugar, se excluyen aquellas que por su frecuencia distorsionarían la correcta alineación. A éstas añadimos aquellas palabras con similitud en la superficie, como por ejemplo antropónimos y topónimos, siguiendo el criterio establecido, entre otros, por Simard et al. (1992) y Kraif (1999).

Seguidamente se realiza una búsqueda en SQL de los lemas encontrados y la equivalencia correspondiente (según la lengua) en el diccionario externo. En este punto se busca en T2 la frase que tiene más palabras coincidentes con la lista de lemas dentro de la ventana delimitada. En el caso que el número de palabras encontradas sea superior a un parámetro determinado (5) y que el porcentaje de estas supere el 20% de las palabras existentes, la frase se acepta como concordante y se considera alineada con la del T1. En caso que esto no suceda se almacena como candidata y se continua la búsqueda en las frases anteriores y posteriores a las que nos encontramos, siempre que sean significativas. Ahora, mientras por una parte procedemos con parámetros menos restrictivos respecto al número de palabras encontradas, introducimos también la variable correspondiente a la distancia respecto de la posición inicial, de forma que vamos aumentando las posibles candidatas, hasta obtener la mejor correlación posible.

Un esquema general del proceso se muestra en la Figura 2:

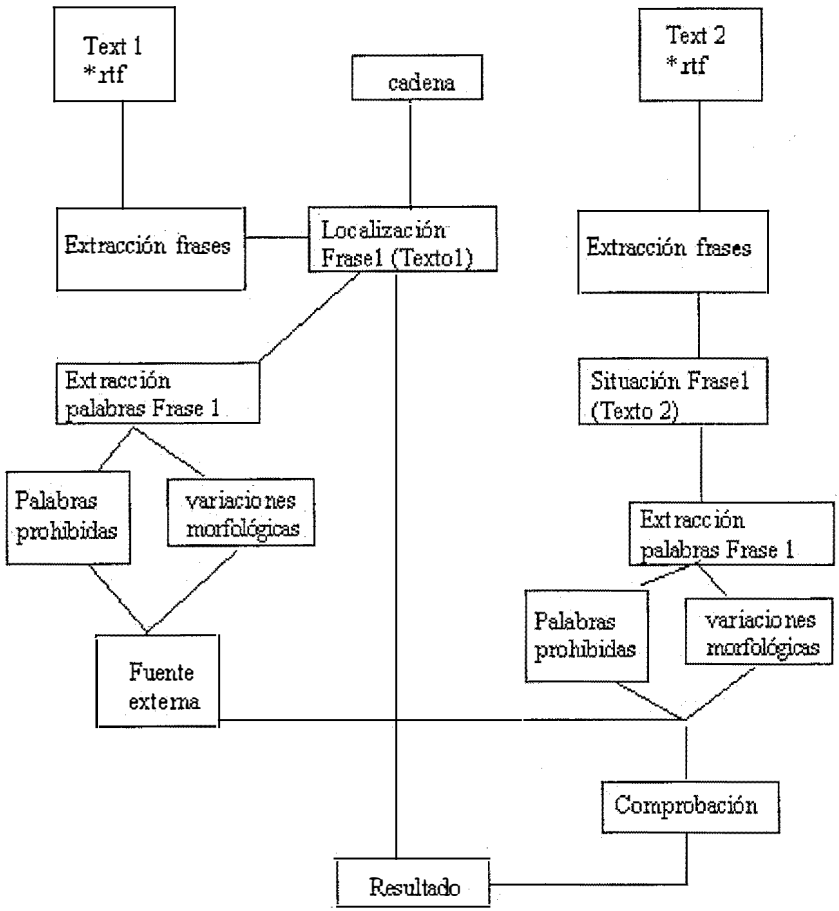

Figura 2. Esquema del proceso de alineamiento de frases con AlfraCOVALT

A la hora de establecer la relación exacta entre las dos frases, la del T1 y la del T2, interviene un nuevo criterio que cabe considerar: la simetria de la relación. Es bien sabido que entre la frase donde se encuentra la expresión buscada en el T1 y la frase correspondiente del T2, existe generalmente una simetría perfecta, es decir una relación 1-1. Esta relación se puede dar en el lenguaje literario incluso en aquellos casos en los cuales la traducción se aleja de forma muy significativa de las equivalencias literales de las palabras, como se observa en la concordancia entre el T1 y el T2, tal y como aparece en el *Cuadro 1* de la búsqueda de la palabra inglesa *feet* en la obra de L. Frank Baum, *The Wonderful Wizard of Oz*:

encara pots veure-li les sabates There are her two **feet**, still sticking out from under a block of wood

Cuadro 1. Concordancia entre TO y TT (feet en la obra de L. Frank Baum, *El meravellos màgic d'Oz (The Wonderful Wizard of Oz)*

Ahora bien, por el contrario, también puede darse una gran disimetría. No hace falta recordar que ello se debe al carácter intrínsecamente creativo de la traducción, que huye por tanto de formas en muchos casos extremadamente literales como se desprende también del ejemplo presentado en el Cuadro 1. Todas estas asimetrías han sido suficientemente analizadas; no obstante, indicaremos que a la luz de los numerosos ejemplos de que disponemos, podemos hablar de que la frase de T1 exista en T2 o que no exista y por tanto nos encontremos ante una omisión, o un añadido, según el texto que utilicemos como referente. Estaríamos hablando de una relación 0-1 o 1-0 según el caso. Pero esta no es la única asimetría existente, si no que las combinaciones complejas donde son diversas las frases que se unen están a la orden del día. Una de las combinaciones típicas es la de 2-1 o 1-2. Así, una frase de T1 puede aparecer dividida en diversas oraciones en T2, tal y como vemos en el ejemplo del Cuadro 2:

«¡Anem-hi!», digué. *I mamprengué el viatge cap al nord amb l'orfenet de **cara** de color canyella als talons.* "Come on!" he said, and resumed his journey into the north. And close at his heels followed the motherless little tan-faced cub.

Cuadro 2. Concordancia entre TO y TT (*cara* en la obra de James Oliver Curwood, *El rei dels óssos. (The grizzly king)*

Así pues, nos encontramos en el TT la frase 1 (FTT1): «*Anem-hi*», *digué*; y la frase 2 (FTT2): *I mamprengué el viatge cap al nord amb l'orfenet de cara de color canyella als talons.* Por su parte en el TO nos encontramos la frase 1 (FTO1): "*Come on*!" *he said, and resumed his journey into the north*, y la frase 2 (FTO2): *And close at his heels followed the motherless little tan-faced cub*. Con la distribución como aparece en la Figura 1:

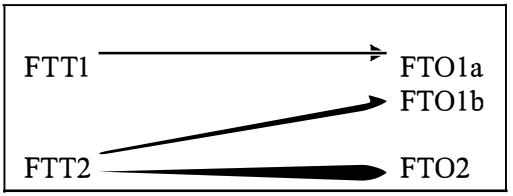

Figura 3. Esquema del alineamiento entre TO y TT (*cara* en la obra de James Oliver Curwood, *El rei dels óssos. (The grizzly king)*

Para establecer una mayor similitud entre la frase de T1 y la del T2, la aplicación realiza una serie de correcciones que son de segmentación o de unión. En primer lugar, si la frase de T2 es superior en número de caracteres al 35% de la de T1, se articula una nueva búsqueda dentro de la frase de T2 segmentándola a partir de las comas, o los dos puntos, existentes en dicha frase. Así en la búsqueda de la palabra *feet* nos encontramos con la concordancia de frases siguiente: T1: *But my **feet** would not touch the ground, and I was forced to stay on that pole* y T2: *No em va agradar gens que m'abandonaren d'aquella forma, per això vaig intentar seguir-los, però com els peus no em tocaven terra, em vaig haver de quedar clavat en aquella canya.* Como se puede observar la longitud de la frase de T2 es mucho más larga que la de T1 por tanto se produce el recorte tal y como se observa en el Cuadro 3:

però com els peus no em tocaven terra, em vaig haver de quedar clavat en aquella canya
But my **feet** would not touch the ground, and I was forced to stay on that pole

Cuadro 3. Concordancia entre TO y TT (*feet* en la obra de L. Frank Baum, *El meravellos màgic d'Oz (The Wonderful Wizard of Oz)*

En segundo lugar, puede darse el caso contrario, es decir que la frase del T1 sea un 35% más grande que la de T2. En este caso establecemos la posición de la expresión buscada en la frase de T1 y añadimos la anterior o la posterior a la frase de T2.

Ella portava una preciosa túnica que li queia des dels muscles en airosos plecs i arribava fins als ***peus****, estava tota coberta d'estrelletes que brillaven al sol com si foren diamants*
the little woman's hat was white, and she wore a white gown that hung in pleats from her shoulders. Over it were sprinkled little stars that glistened in the sun like diamonds

Cuadro 4. Concordancia entre TO y TT (*peu* en la obra de L. Frank Baum, *El meravellos màgic d'Oz* *(The Wonderful Wizard of Oz)*

Los resultados obtenidos en la búsqueda son presentados de dos formas: en un fichero adjunto que agrupa todos los resultados, y también con la interface que se presenta en la Figura 4:

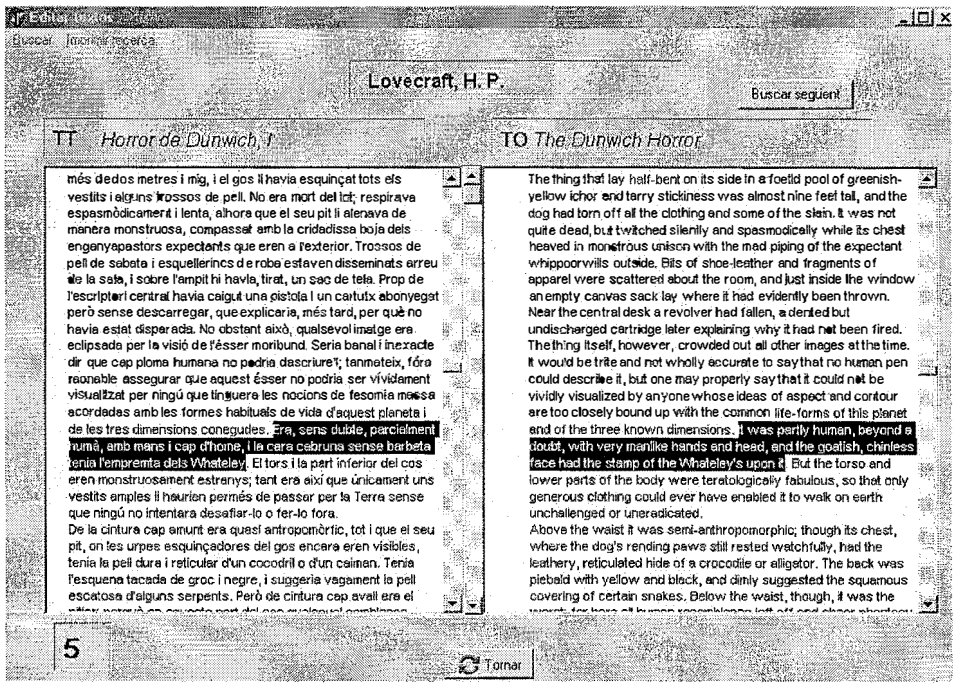

Figura 4. Interface de resultado de la búsqueda de la cadena *head* en el texto original y la traducción de The Dunwich Horror de H. P. Lovecraft

En la Figura 4, se observan los textos paralelos de la obra de H. P. Lovecraft, *The Dunwich Horror*, y el resultado número 5 de la búsqueda de la cadena lingüística *head* en TO y la frase alineada correspondiente en el TT.

4. Evaluación de AlfraCOVALT

A fin de establecer la validez, fiabilidad y precisión de las alineaciones entre T1 y T2 realizadas por la aplicación hemos realizado un total de 24 búsquedas correspondientes a un total de catorce obras y sus correspondientes traducciones. La suma total de palabras de los textos era de 768.070 con un promedio de 28.764 palabras por texto. El criterio para seleccionar las búsquedas y evaluarlas consistió en que el total de ocurrencias a concordar fuera igual o superior a 15. Una vez realizadas las búsquedas, tres evaluadores realizaron el análisis de las concordancias. Todos ellos con competencia lingüística en las lenguas de los textos, inglés y catalán, coincidieron en todos los casos analizados y los resultados obtenidos se presentan en la Tabla 1.

AUTOR	EXPRES	TOTAL	ACIERTOS	UNIR	SEPARAR	PRECISIÓN
Baum	Peu	28	28	3	1	100
Baum	neck	16	15	1	1	93,75
Chesterton	Ulls	34	34			100
Chesterton	hand	27	27		1	100
Conrad	Pit	15	13			86,7
Conrad	head	108	99	1	4	91,7
Curwood,	Cara	36	33	1	6	91,7
Curwood,	Eyes	54	50	3	1	92,6
Doyle	Cos	19	19		1	100
Doyle	Face	15	15			100
Faville	Braç	42	40			95,2
Faville	Mouth	37	36			97,3
Joyce	Cara	37	36		1	97,3
Joyce	Eyes	40	35			87,5
Le Guin	Peu	43	42			97,7
Le Guin	Heart	17	17			100
London	Mans	18	17	2		94,4
London	Ear	17	16			94,1
Lovecraft	Head	21	20		3	95,2
Melville	Mà	15	15			100
Melville	Cara	17	17			100
Saki	Ulls	19	19			100
Stevenson	Mans	15	14			93,3
Stevenson	Face	15	15			100

Tabla 1. Resultados de las búsquedas en el corpus COVALT con la herramienta AlfraCOVALT

La primera columna de la tabla, *autor*, indica los autores de las obras en las cuales se ha realizado la búsqueda; en la segunda columna, *expres*, aparece la expresión o palabra utilizada en la búsqueda; la tercera columna, *total*, recoge el número total de ocurrencias de la expresión buscada entre T1 y T2; la cuarta columna, *aciertos*, expresa el número de alineaciones correctas; en la quinta columna, *unir*, aparece el número de casos en que se ha producido en T2 la unión de una segunda frase, anterior o posterior, a la concordante para ajustar los dos textos; la sexta columna, *separar*, representa el número de casos en que se ha producido la separación de la frase concordante de T2; y finalmente la última indica los porcentajes de aciertos.

Analizando los resultados de la tabla, comprobamos que la media de los porcentajes de cada una de las 24 búsquedas supera el 96,7% de aciertos (la media ponderada: 95,37). Estos aciertos en el alineamiento incluyen también aquellos que implican o el necesario recorte de la frase alineada o la correspondiente unión con otra frase precedente, o posterior, según el caso y, a su vez, los errores incluyen también aquellos en los que la unión o la separación no se ha realizado de forma precisa. De los 33 errores observados, el 21 %, tienen relación con esta última circunstancia. De ellos, el 12,1% responden a frases que no han sido adecuadamente recortadas, o lo que es lo mismo, que la segmentación de la frase de T2 no ha sido suficientemente ajustada para correlacionarla exactamente con la frase seleccionada de T1. Por su parte, el 9 % lo representan frases de T2 incorrectamente unidas, generalmente porque se ha unido la frase anterior a la frase relacionada de T2, cuando había de unirse la posterior o a la inversa. El resto de los errores obedecen a que la aplicación detecta la frase anterior o la posterior a la correspondiente en T2, el 36,36%, en lugar de la que se corresponde realmente, o bien unas frases más alejadas, con el 39,39%. El caso que AlfraCOVALT no detecte ninguna frase (sin que sea una omisión o inserción) representa únicamente el 3% de los errores. Respecto a la dirección de la búsqueda cabe indicar que los errores cuando la búsqueda se realiza sobre el TO, el número de errores es sensiblemente superior que cuando se hace esta misma búsqueda sobre el TT (64.5% frente al 35,5%).

Las razones que motivan la mayor imprecisión en la búsquedas obedecen principalmente a dos factores: la limpieza de los textos, y la creatividad. Por limpieza de los textos cabe entender el hecho que los textos no están pre-editados lingüísticamente, de forma que pueden generar, por ejemplo y como es el caso, errores en el momento de detectar el límite de la frase. Las razones se hallan especialmente en las convenciones del lenguaje literario (como por ejemplo las entradas de los parlamentos de los personajes), que en unos casos son fruto de las convenciones generales de la propia lengua, pero que, en la mayoría de ellos, lo son más bien de los criterios de edición de cada texto. Por lo que respecta al segundo factor, la creatividad de la obra literaria, éste es uno de los que resultan también determinantes ya que se manifiesta en diversos aspectos lingüísticos, como son las convenciones tipográficas, los usos metafóricos, el empleo de fraseologismos, los diversos grados de lexicalización de las acciones, los conceptos, etc., de cada lengua, y de forma particular en el modelo de lengua de cada traductor.

5. Conclusiones y perspectivas de futuro

En este trabajo hemos presentado un programa de alineamiento de textos, AlfraCOVALT, que, aunque está destinado a la gestión de un determinado corpus de textos narrativos, COVALT, podría servir para gestionar otros corpus cuyo objetivo se halle en la investigación para la traducción y los estudios contrastivos o de elaboración de materiales docentes. Igualmente hemos comprobado el programa y tras observar la fiabilidad significativa de su empleo, hemos analizado también las causas de las posibles deficiencias en la correlación atribuyéndolas a dos factores claramente determinantes: los problemas suscitados por las convenciones tipográficas y la creatividad en la ficción, en la medida que las decisiones (elección de las equivalencias léxicas, construcciones, fraseologismos, etc.) del traductor, cuando estas se alejan de las formas más transitadas, la labor de la herramienta de alineamiento resulta mucho más compleja.

A la luz de los resultados, la búsqueda de una mayor optimización de los alineamientos obtenidos surge comos una tarea para el trabajo futuro. En este sentido, nos encontramos con dos posibles vías de trabajo. Por una parte la introducción de filtros más efectivos tanto por lo que respecta a los lemas, como también por lo que respecta a las variaciones morfológicas, ya que ambos se perciben como generadores de distorsión. Y por otra parte, la introducción de métodos estadísticos que ayuden en esta misma labor.

Agradecimientos

Mi más profundo agradecimiento a los profesores Juan Carlos Ruiz, Heike van Lawick y Josep Marco por la lectura atenta de este trabajo y los comentarios realizados.

Bibliografía

1. Biblografía primaria

Baum, Lyman Frank (1900). *The Wonderful Wizard of Oz* [documento en internet http://www.gutenberg.org/etext/419].

— (2000). *El meravellós màgic d'Oz*, Josep Franco Martínez (trad.). Alzira: Bromera.

Chesterton, Gilbert Keith (1911). *The Secret Garden* [documento en internet http://www.dur.ac.uk/martin.ward/gkc/books/Complete_Father_Brown/chapter2.html]

— (1997). *El jardí secret i altres contes*, Salvador Montaner (trad.). Alzira: Bromera.

Conrad, Joseph (1903). *Typhoon and other stories* [documento en internet http://etext.library.adelaide.edu.au/c/conrad/joseph/c75ty/]

— (1991). *Tifó*, Remei Bataller Martínez (trad.). Alzira: Bromera.

Curwood, James Oliver (1916). *The Grizzly King* [documento en internet http://www.gutenberg.org/etext/10977]

— (1994). *El rei dels óssos,* Remei Bataller Ferrer (trad.). Alzira: Bromera.

Doyle, Arthur Conan (1908). *The Bruce Partington Plans* [documento en internet http://www.gutenberg.org/etext/2346]

— (1996). *Sherlock Holmes i els plànols del Bruce Partington,* Víctor Oroval Martí (trad.). Alzira: Bromera.

Faville, B. 1986 *The Keeper*. Oxford: Oxford University Press.

— (1999) *El supervivent,* Víctor Oroval Martí trad. Alzira: Bromera.

Joyce, James (1907). *The Dead* [documento en internet http://www.gutenberg.org/etext/2814]

— (1992). *Els morts,* Joan Talens (trad.). València: Eliseu Climent.

Le Guin, Ursula Kroeber (1966). *Worlds of Exile and Illusion.* New York : Orb Omnibus Edition.

— (1998). *El món de Rocannon,* Carles Ayuso (trad.). Alzira: Bromera.

London, Jack. (1902). *The Cruise of the Dazzler* [documento en internet http://www.gutenberg.org/etext/11051]

— (1995). *El creuer del Dazzler,* Remei Bataller Ferrer (trad.). Alzira: Bromera.

Lovecraft, Howard Phillips (1929). *The Dunwich Horror* [documento en internet http://www.dagonbytes.com/thelibrary/lovecraft/thedunwichhorror.htm]

— (1992). *L'horror de Dunwich,* Elisabeth Mateo (trad.). València: Tabarca.

Melville, Herman (1853). *Bartleby, the Scrivener* [documento en internet http://www.bartleby.com/129/index.html]

— (1995). *Bartleby, l'escrivent*, Pilar Aguilar Cortina (trad.). Alzira: Germania.

— (1924). *Billy Budd, sailor* [documento en internet http://www.bibliomania.com/0/0/36/1006/frameset.html]

— (2000). *Billy Budd, el mariner,* Jesús Cortés (trad.). Alzira: Bromera.

Saki (1912). *Tobermory* [documento en internet http://www.sff.net/people/DoyleMacdonald/l_tober.htm]

— (1994). *Tobermory,* Domènec Ardit Climent (trad.). Alzira: Bromera.

Stevenson, Robert Louis (1893) *The bottle imp* [documento en internet http://gaslight.mtroyal.ab.ca/bottlimp.htm]

— (1997). *El diable de la botella,* Joan E. Pellicer Borrás (trad.). Alzira: Bromera.

2. Bibliografía secundaria

Brown, Peter, Lai Jenifer y Mercer, Robert L. (1991) Aligning Sentences in Parallel Corpora. En *Proceedings of the 29th conference on Association for Computational Linguistics*, 169-176, Berkeley, CA: University of California.

Burdeus, María Dolores y Verdegal, Joan (En prensa). Claves para una sociología de la traducción de narrativa a partir de COVALT (1990-2000). *META*.

Chen, Stanley (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information, *ACL-93*, 9-16.

Church, Kenneth (1993). Char-align: A Program for Aligning Parallel Texts at the Character Level, *ACL-93*, 1-8.

Fung, Pascale y McKeown, Kathleen. (1996). A Technical Word and Term Translation Aid using Noisy Parallel Corpora Across Language Groups, *The Machine Translation Journal,* Special Issue on New Tools for Human Translators, 30, 53-87.

Gale, William A. y Church, Kenneth (1991). A Program for Aligning Sentences in Bilingual Corpora. En *Proceedings of the 29th conference on Association for Computational Linguistics*, 177-184. Berkeley, CA: University of California.

Guzmán, Josep R. (en prensa) El uso de COVALT y AlfraCOVALT en el aprendizaje traductor, *Actas del XXIV Congreso de AESLA*, Madrid, marzo 2006.

Haruno, Masahiko. y Yamazaki, Takefumi (1997). High Performance Bilingual Text Alignment Using Statistical and Dictionary Information. *Natural Language Engineering* 3(1), 1-14.

Kay, Martin y Röscheisen, Martin. (1988). Text-Translation Alignment. *Computational Linguistics*, 19(1), 121-142.

Kraif, Olivier (1999). Identification des cognats et alignement bi-textuel : une étude empirique. En *Actes de la 6ème conférence annuelle sur le Traitement Automatique des Langues Naturelles,* 205-214. Cargèse: Institut d'Etudes Scientifiques.

Lawick, Heike. v. (2005). El corpus paralelo bitextual en la enseñanza de traducción: identificación y soluciones para doch. En *II Congreso Internacional AIETI, Formación, investigación y profesión*, Cd-Rom.

Macklovitch Elliott y Hannan, Marie-Louise (1998). Line 'Em Up: Advances in Alignment Technology and their Impact on Translation Support Tools. *Machine Translation* 13, 41–57.

Marco Borillo, J. (2006). A Corpus-Based Approach to the Translation of Evaluativo Adjectives as Modality Markers. En *Corpus Linguistics. Applications for the Study of English*, Ana María Hornero, María José Luzón y Silvia Murillo (eds.), 241-253. Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Oxford, Wien: Peter Lang.

— (En prensa). Estudios sobre la traducción basados en corpus y análisis estilístico: el caso de las unidades fraseológicas. En *I Congreso Internacional de la Asociación de traductores peruanos*.

Melamed, I. Dan (2000). Pattern recognition for mapping bitext correspondence. En *Parallel Text Processing*. Véronis, Jean (Ed.), 25-48. Dordrecht: Kluwer Academic Publishers.

Och, Franz Josef y Ney, Hermann (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29:1,

Rada, Mihalcea y Pedersen, Ted. (2004). An Evaluation Exercise for Word Alignment, *Proceedings of the HLT-NAACL 2004 Workshop on "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*. Edmonton, Canada [documento en internet http://www.cs.unt.edu/~rada/wpt/papers/pdf/Mihalcea.pdf]

sentences in bilingual corpora. In *Meeting of the Association for Computational*

Simard, Michel, Foster, George y Isabelle, Pierre (1992). Using Cognates to Align Sentences in Parallel Corpora. En *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal, Canada, 67-81.

Hao, Zhang y Gildea, Daniel (2004). Syntax-Based Alignment: Supervised or Unsupervised? En *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*. Ginebra [documento en internet http://141.3.25.88/MTCourse/Paper/zhang-gildea-coling04.pdf].

3. Programas y herramientas de alineamiento

Align (A.Berger) [http://www.cs.unt.edu/~rada/wa/tools/aberger]

Alpaco (Aligner for Parallel Corpora) [http://www.d.umn.edu/~tpederse/parallel.html], Cairo [http://www.clsp.jhu.edu/ws99/ projects/mt/toolkit],

Bilingual sentence aligner (Microsoft) [http://research.microsoft.com/research/downloads/default.aspx]

CTK: Champollion Tool Kit [http://champollion.sourceforge.net].

Geometric Mapping and Alignment (GMA) [http://nlp.cs.nyu.edu/GMA]

Hunalign [http://mokk.bme.hu/resources/hunalign]

I*Link [http://www.ida.liu.se/~nlplab/ILink/]

Kvec++, http://www.d.umn.edu/~tpederse/parallel.html

Lingua-AlignmentSet [http://www.lsi.upc.es/~lambert/software/AlignmentSet.html]

MULTEXT-East [http://nl.ijs.si/ME/CD/mte-home.html]

Natura Alignment Tools (NATools) [http://natura.di.uminho.pt/natura/natura?&topic=NATools]

PLUG Word Aligner, PWA [http://numerus.ling.uu.se/~corpora/plug/pwa/]

Strand data [http://www.umiacs.umd.edu/~resnik/strand/]

Twente [http://wwwhome.cs.utwente.nl/~irgroup/align/download.html]

UMIACS Word Alignment Interface [http://www.umiacs.umd.edu/~nmadnani/alignment/forclip.htm]

Uplug [http://stp.ling.uu.se/cgi-bin/joerg/Uplug]

Vanilla [http://nl.ijs.si/telri/Vanilla/]