Bifid: un alineador de corpus paralelo a nivel de documento, oración y vocabulario

Rogelio Nazar
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
rogelio.nazar@upf.edu

Resumen

Este artículo presenta un algoritmo que integra distintos aspectos del procesamiento de corpus paralelo y que ha sido implementado como una aplicación web. El trabajo se enmarca en la lingüística computacional pero puede interesar a terminólogos, traductores y estudiantes de lenguas extranjeras. El sistema está diseñado para operar con cualquier par de lenguas ya que es exclusivamente estadístico. Acepta como entrada un corpus paralelo definido como un conjunto de documentos en una lengua A y sus traducciones en una lengua B. Sin requerir más especificaciones, el sistema puede separar el conjunto de documentos en las dos lenguas, alinear cada documento con su traducción y luego alinear los segmentos dentro de cada par de documentos para producir finalmente un vocabulario bilingüe que incluye unidades poliléxicas.

Palabras clave

Alineación de corpus paralelo, extracción de vocabularios bilingües, lexicografía computacional

Abstract

This paper presents an algorithm that integrates different aspects of parallel corpus processing, which is now implemented as a web application. It is a computational linguistics project but can also be of interest to translators, terminologists and foreign language learners. The system is designed to operate with any pair of languages since it is exclusively based on statistical techniques. It takes a parallel corpus as input, defined as a set of documents in a language A and their translations into a language B. Without any further specification, the system separates the set of documents in the two languages, aligns each document with its translation and then aligns the segments within each pair of documents to finally produce a bilingual vocabulary that includes multiword units.

Keywords

Bilingual lexicon acquisition, computational lexicography, parallel corpus alignment

1 Introducción

Si bien en los últimos años ha aparecido una gran cantidad de publicaciones en las que se describen metodologías para la obtención de terminología bilingüe desde corpus comparables (Gaussier et al., 2004; Daille y Morin, 2005; Morin et al., 2008) e incluso corpus no relacionados (Fung, 1995; Rapp, 1999; Nazar, Wanner, y Vivaldi, 2008), el procesamiento de corpus paralelo continúa siendo el método más consolidado para obtener vocabularios bilingües o como herramienta de apoyo a la traducción, particularmente en el caso de la traducción técnica o especializada (Kübler, 2011). Esta tendencia se ha visto sin duda favorecida por la masificación de la World Wide Web, la principal fuente de corpus paralelos en nuestra época (Almeida, Simões, y Castro, 2002; Resnik y Smith, 2003).

El presente artículo describe el sistema Bifid, un proyecto en curso de herramienta de alineación de corpus paralelo. Se implementa como aplicación web y opera de forma independiente de lengua, llevando a cabo los siguientes pasos en forma secuencial: 1) separar el corpus en las dos lenguas que lo componen; 2) alinear a nivel de documento; 3) alinear a nivel de oración; 4) extraer un vocabulario bilingüe con unidades poliléxicas; 5) comenzar de nuevo el proceso desde el paso 2 introduciendo como parámetro adicional el resultado obtenido en el paso 4.

El sistema se encuentra actualmente implementado como una demo online en forma de CGI Perl[1], pero debe advertirse que el propósito de este demostrador no es el de funcionar ya como un producto informático sino el de permitir al lector reproducir el experimento con otro corpus. El artículo no incluye capturas de pantalla ni instrucciones de uso porque no pretende poner el foco de atención en un programa informático concreto sino en la metodología que se

[1] La dirección URL del proyecto es la siguiente: http://www.bifidalign.com/

propone para resolver el problema de la alineación, lo cual representa un nivel mayor de abstracción. En un programa informático, por ejemplo, se busca el mejor rendimiento posible, por lo que no habría razón para no incorporar conocimiento explícito de las lenguas analizadas (en la forma de diccionarios, lematizadores, analizadores morfosintácticos, etc.). En el caso de esta propuesta, en cambio, lo que se pretende es averiguar qué resultado es posible conseguir sin la ayuda de estos recursos.

Además de obedecer al principio de parsimonia, el enfoque "pobre" en conocimiento lingüístico tiene una doble motivación, teórica y práctica. Desde un punto de vista teórico, se puede afirmar que un algoritmo capaz de resolver el problema de la alineación sin recurrir a conocimiento explícito sobre el par de lenguas analizadas tiene un mayor poder de generalización y permite poner de relieve fenómenos que trascienden las lenguas particulares. Desde el punto de vista práctico, es útil porque la gran mayoría de las lenguas del mundo no goza de los recursos lingüísticos comunes a las lenguas mayoritarias, y la adaptación de los recursos lingüísticos a una lengua minoritaria, incluso los de nivel más elemental como el etiquetado morfosintáctico, implica un coste en tiempo, esfuerzo y presupuesto que en muchos casos resulta inasumible. También desde el punto de vista práctico, la capacidad de este algoritmo para adaptarse a distintos tipos de datos facilita su aplicación a una diversidad de escenarios. Por mencionar un ejemplo, organismos internacionales como la Organización de las Naciones Unidas, el Fondo Monetario Internacional o la Unión Europea, por nombrar algunos de los más importantes, disponen de cantidades ingentes de documentos traducidos a una gran diversidad de lenguas y, si todo ese material se pudiera reaprovechar para la generación de recursos léxicos o terminológicos sin tener que analizar las particularidades de cada una de esas lenguas o la codificación específica que cada organismo aplica a sus documentos, estaríamos ante la posibilidad de producir recursos lexicográficos o terminológicos de gran envergadura y con presupuesto mínimo.

El artículo se organiza de la siguiente manera: la sección 2 presenta un breve repaso de los trabajos más importantes realizados en el área de la alineación de corpus paralelos en los distintos niveles. La sección 3 presenta los detalles del presente algoritmo en cada uno de los pasos sucesivos de la alineación. La sección 4 presenta una evaluación de los resultados de la aplicación en distintos corpus. La sección 5 describe cómo funciona la implementación en forma de aplicación web y, finalmente, la sección 6 enuncia las conclusiones de este trabajo.

2 Antecedentes de la alineación de corpus paralelo

Como señala Véronis (2000), el primer gran hito en la historia de la alineación de corpus paralelo se da en 1822 con el desciframiento de la Piedra Rosetta llevado a cabo por Jean François Champollion. Algo más de un siglo más tarde, en el trabajo de Weaver (1955) se puede comprobar que existe lo esencial de la idea, aunque sea de forma embrionaria, cuando describe los primeros métodos estadísticos de traducción automática. A pesar de este aporte visionario, la alineación de corpus paralelo no prosperó hasta después de la segunda mitad de la década de los años ochenta, cuando las computadoras fueron capaces de procesar grandes matrices numéricas. Hasta entonces, la mentalidad que primaba en lo que se conocía como traducción automática era ajena a los corpus paralelos o a los métodos estadísticos en general. El pensamiento normal de la época era el de los sistemas basados en reglas, como por ejemplo en el proyecto llamado justamente Rosetta Stone (Appelo y Landsbergen, 1986).

A comienzos de los años noventa, sin embargo, se produce la explosión de publicaciones sobre alineación de corpus paralelos con métodos estadísticos y el ámbito se constituye como un campo de estudio específico, diferenciado respecto de la traducción automática. Los primeros trabajos se centraron en la alineación de oraciones en función de su extensión medida en caracteres o palabras (Gale y Church, 1991a; Brown, Lai, y Mercer, 1991), bajo el supuesto de que existe una correlación entre la extensión de las oraciones del texto de origen con las del texto meta. Al mismo tiempo, se exploró también la posibilidad de extraer vocabularios bilingües mediante el cálculo de la coocurrencia de las unidades léxicas en los pares de oraciones ya alineadas (Gale y Church, 1991b).

Posteriormente se incorporaron nuevas pistas, como la detección de cognados como señaladores de la correspondencia entre oraciones (Church, 1993; Simard, Foster e Isabelle, 1993; McEnery y Oakes, 1995). Se estudió también la posibilidad de establecer una retroalimentación entre las salidas de los dos procesos de alineación oracional y a nivel de vocabulario, según la idea de que el vocabulario bilingüe que resulta de la alineación oracional puede servir para refinar la misma alineación oracional, creando así un círculo virtuoso (Kay y Röscheisen, 1993; Moore, 2002; Braune y

Fraser, 2010). Posiblemente el trabajo más influyente sea el de Brown et al. (1993), ya que en él se inspiran algunos de los alineadores más conocidos en la actualidad, como Hunalign (Varga et al., 2005), Giza++ (Och y Ney, 2000; Och y Ney, 2003) o Champollion (Ma, 2006), entre otros.

Algunas propuestas de alineadores incorporan conocimiento lingüístico explícito, ya sea en la forma de vocabularios bilingües (Hofland y Johansson, 1998) o información morfosintáctica (De Yzaguirre et al., 2000; Gammallo, 2005; Gómez Guinovart y Simões, 2009), pero prevalecen las visiones puramente estadísticas o incluso geométricas, como en el caso de Melamed (2000), que representa los textos paralelos en un plano (dos ejes) de manera que la mejor alineación entre las oraciones se selecciona calculando la distancia que tienen con la diagonal.

Queda aún trabajo por hacer en la integración de las distintas fases de alineación, y es sobre todo en ese sentido en el que este artículo presenta una nueva contribución. Han aparecido diversas herramientas que integran diferentes niveles de alineación, como Twente Aligner (Hiemstra, 1998), NATools (Simões y Almeida, 2003), o Uplug (Tiedemann, 2006), pero hasta la fecha no se había intentado un enfoque que integrara la totalidad de los procesos, como en una herramienta que, partiendo de cero, sin ningún tipo de información sobre las lenguas del corpus ni intervención humana, produjera resultados de alineación en todos los niveles, desde el documento hasta el léxico.

3 El algoritmo

El proceso de este algoritmo comienza por un conjunto de documentos escritos en dos lenguas desconocidas y la tarea consiste en separar los documentos en estas dos lenguas (subsección 3.1.), alinear cada documento con su correspondiente traducción u original (subsección 3.2.), alinear las oraciones del texto de una lengua con las oraciones del texto en la otra lengua (subsección 3.3.) y, finalmente, extraer un vocabulario bilingüe (subsección 3.4.).

En todos estos pasos la intervención humana es posible, ya que como el resultado del proceso descrito en cada subsección alimenta el proceso siguiente, un usuario siempre puede controlar y corregir eventuales errores producidos en cada paso con el objeto de mejorar el resultado posterior. En ningún caso, sin embargo, esta intervención humana es contemplada como una condición *sine qua non* para la tarea del algoritmo ni se ha tenido en cuenta esta posibilidad en la evaluación de los resultados descrita en la sección 4.

3.1 Separación de los documentos en lenguas

Partiendo de un conjunto de documentos escritos en una lengua con su correspondiente traducción a otra, la primera operación consiste en separar los documentos en dos subconjuntos correspondientes a cada una de las lenguas. Para ello, en esta operación el algoritmo asume la universalidad de la distribución de frecuencias del vocabulario en los textos. De acuerdo con el sencillo principio de que todos los documentos escritos en una misma lengua tendrán al menos una parte del vocabulario en común consistente en las unidades léxicas más frecuentes, podemos agrupar los documentos en función de la similitud que tengan con respecto a las n palabras más frecuentes de cada documento. Para ello se llevan a cabo los siguientes dos pasos:

1. Ordenar por frecuencia decreciente el vocabulario del documento más extenso del corpus, al que llamaremos documento D_a.
2. Ubicar en un conjunto A todos los documentos del corpus que entre sus 10 palabras más frecuentes tengan al menos 3 palabras en común con las 10 más frecuentes del documento D_a.

Si esta operación consigue dividir el corpus en las dos lenguas, ya es posible pasar a la fase siguiente. Si, en cambio, todos los documentos han quedado en un mismo conjunto, entonces esto quiere decir que estamos trabajando con lenguas muy similares. En tal caso, se asume que las lenguas de los documentos se distribuyen en mitades iguales del corpus, lo cual sería de esperar en un corpus paralelo. Para ello reutilizamos el rango dado a los documentos por la puntuación otorgada en el cálculo recién descrito, es decir, en la similitud que tienen con el documento más largo calculada en función de las palabras altamente frecuentes que tienen en común.

3.2 Alineación a nivel de documento

Tomando el resultado del proceso anterior, en este paso se construirá una tabla de correspondencias entre los documentos[2] de la lengua A con los documentos de la lengua B, que representa los coeficientes aplicados a cada una de las parejas de documentos obtenidas por el producto cartesiano

[2]Para los fines de este proceso, no se tiene interés en saber cuál documento es el original y cuál la traducción.

de ambos conjuntos. Las variables que aparecen en el conjunto de coeficientes (1) son definidas a continuación en esta sección.

$$C = \{l, ln, sim, voc, num, bvoc\} \quad (1)$$

$$w(i,j) = \prod_{n=1}^{|c|} (1 + c_n(i,j)) \quad (2)$$

Una vez calculados los coeficientes, cada pareja de documentos *i–j* recibe una puntuación final que es el producto de los coeficientes (2), a los que sumamos 1 para no perder toda la puntuación en caso de que con alguno de los coeficientes se obtenga un valor 0. Esta puntuación permite ordenar todas las parejas de documentos de mayor a menor y así ir "sacándolas" una a una. Por una decisión metodológica, no se permite a un mismo documento estar en más de dos parejas, sólo se le permite estar en la que tiene el valor más alto, pero esto se podría considerar un parámetro de ejecución. Una vez explicada la lógica general del proceso de selección, pasamos a definir cada uno de los coeficientes.

Coeficiente *l*: Este coeficiente simplemente compara el tamaño de los documentos en número de caracteres. Se fundamenta en un criterio similar al que utilizan Gale y Church (1991a) para la alineación a nivel oracional, bajo el supuesto de que el documento original y su traducción deben tener un tamaño similar, tal como se define en la ecuación 1, donde la expresión *length* refiere a la extensión en caracteres de un documento.

$$l(i,j) = \frac{\text{mín}(length(i), length(j))}{\text{máx}(length(i), length(j))} \quad (3)$$

Como una forma de atender a las diferencias de tamaño que se pueden producir como consecuencia de la distinta redundancia natural de una y otra lengua, fenómeno conocido como *Language Proportion Coefficient* o *LPC* (Choueka, Conley, y Dagan, 2000), se lleva a valor 1 todo par de documentos que al ser comparados arrojen un valor de similitud superior o igual a 0.7 (o cualquier otro umbral arbitrario que se ajuste como parámetro en la implementación).

Coeficiente *ln*: Aunque también está basado en la comparación de extensión en caracteres, a diferencia del anterior este otro coeficiente recurre a un criterio metatextual que es comparar el largo de los nombres de los documentos, suponiendo que las parejas de documentos original-traducción tendrán nombres de extensión similar. Este coeficiente queda sin efecto en los experimentos descritos en la sección 4 ya que los nombres de los ficheros que designan los documentos obedecen a códigos que no guardan relación con el contenido. Se define de la misma forma que el coeficiente anterior (ecuación 3), por lo tanto puede decirse que se trata del mismo coeficiente pero aplicado de forma distinta.

Coeficiente *sim*: Este coeficiente, al igual que el anterior, analiza el nombre de los documentos. Su función es tratar de encontrar una similitud ortográfica entre los nombres de los ficheros de una y otra lengua, bajo el supuesto de que los nombres del fichero original y su traducción pueden tener elementos en común en un nivel inferior a la palabra.

Tal como se mencionó en la sección 2, la idea de la aplicación de coeficientes de similitud ortográfica para la alineación de corpus paralelos ha sido utilizada ya con el objeto de encontrar cognados (por ejemplo, en McEnery & Oakes, 1995). El coeficiente aplicado en el presente algoritmo compara formas como vectores binarios cuyas dimensiones son bigramas de caracteres (secuencias de dos letras) y la similitud entre vectores se calcula utilizando el coeficiente de Dice, expuesto en la ecuación 4.

$$sim(I,J) = \frac{2|I \cap J|}{|I| + |J|} \quad (4)$$

Tanto este coeficiente como el anterior, al estar aplicados según criterios metatextuales, son de menor interés teórico. Sin embargo, en casos reales es muy frecuente encontrar este tipo de similitud ortográfica entre nombres de ficheros y, por lo tanto, el coeficiente puede tener un potencial práctico importante de cara a un usuario final del sistema. Como en el caso anterior, en el experimento llevado a cabo en este artículo (sección 4) este coeficiente también queda sin efecto ya que, por las particularidades de los corpus utilizados en los experimentos, los nombres de los ficheros están compuestos por símbolos arbitrarios que no guardan relación con el contenido de los textos.

Coeficiente *voc*: Como el primero de los coeficientes, este mide características propias de los documentos. Se fundamenta en la probabilidad de que una pareja de documentos original-traducción tenga elementos del vocabulario en común. Uno puede pensar, por ejemplo, en diversos símbolos, siglas y nombres propios que puedan escribirse de la misma forma aunque se trate de lenguas distintas. El coeficiente, definido en la ecuación 5, normaliza la cantidad de unidades del vocabulario en común por la cantidad de unidades de vocabulario distintas encontradas en el

más extenso de los dos documentos.

$$voc(I, J) = \frac{|I \cap J|}{\text{máx}(|I|, |J|)} \quad (5)$$

Coeficiente *num*: El coeficiente *num* funciona y se define de la misma manera que el coeficiente *voc*, solamente que se restringe a la detección de números, y el objeto de mantenerlos como coeficientes distintos tiene una utilidad práctica ya que de esta manera es más grande su contribución al peso final de la comparación.

Coeficiente *bvoc*: Este último coeficiente es opcional, ya que hace referencia a la aplicación de un vocabulario bilingüe tal como el que resulta del producto final del algoritmo, y que puede ser incorporado de nuevo a una segunda ejecución. Cualquier léxico bilingüe puede servir a este coeficiente pero, naturalmente, su calidad afectará la precisión del resultado final. La forma en que se calcula este coeficiente es exactamente igual a la que se expone en la ecuación 5, solo que esta vez, en lugar de comparar las mismas palabras, se comparan palabras equivalentes.

Es evidente cuál puede ser el servicio que puede prestar un vocabulario bilingüe para la alineación a nivel de documento: aquel par de documentos que contenga más palabras equivalentes será probablemente la alineación correcta.

3.3 Alineación a nivel de oración

En la alineación a nivel de oración, el algoritmo elabora una matriz similar a la descrita en la subsección anterior. Presupone, sin embargo, la existencia de una división en el texto (oraciones o segmentos de otra extensión) mediante el carácter de final de línea, un signo universal, presente por definición en todo archivo de texto. La situación en la que se encuentra el algoritmo en la alineación del corpus a nivel de la oración conserva una serie de similitudes con la alineación a nivel de documento, pero dispone de nuevas pistas, como la información posicional de las oraciones, definida a continuación:

Coeficiente *pos*: No era posible, en la subsección 3.2., hacer alguna suposición respecto a la situación posicional de los documentos en el corpus. A nivel oracional, en cambio, es legítimo suponer que existirá un orden que debe ser respetado al menos en parte por quien hizo la traducción a la otra lengua. Por lo tanto, es altamente probable que la primera oración del documento original se corresponda con la primera oración de la traducción y, de la misma forma, que la última oración del original se corresponda también con la última oración de la traducción. De esta manera, se observará una correlación entre la posición de cada oración en el original con respecto a la posición de su traducción en el texto meta. Esta correlación permite definir un coeficiente posicional tal como se indica en la ecuación 6, donde los símbolos $P_{i,a}$ y $P_{j,b}$ representan la posición relativa de cada oración a en el documento original i y la posición relativa de cada oración b en la traducción j.

$$pos(a, b) = \frac{\text{mín}(P_{i,a}, P_{j,b})}{\text{máx}(P_{i,a}, P_{j,b})} \quad (6)$$

El coeficiente *sim*, también incluido en la matriz para la alineación oracional, se aplica en este caso al contenido de las oraciones candidatas a alineación en lugar de al nombre de los documentos, tal como se hizo en la subsección 3.2., esta vez con el objeto de detectar la presencia de cognados. Funciona de forma paralela a otro coeficiente que mide los cognados, *cogn*, y la única diferencia es que en el caso de este último no se comparan las dos oraciones como una única cadena de caracteres sino que se comparan las palabras de las oraciones entre sí. La comparación se realiza de la misma manera que el coeficiente *sim* pero la lógica de su aplicación es ligeramente distinta. La idea sería que a mayor cantidad de cognados tenga un par de oraciones, más probable será que se trate de una alineación correcta. En este punto, además, la detección de números queda nuevo constituida como una variable independiente, con el coeficiente *num*, debido a la vital importancia que en algunos casos tienen los números para la alineación. Así, en la ecuación 7 queda definido un nuevo conjunto C prima de coeficientes.

$$C' = \{l, ln, pos, sim, num, cogn, voc, bvoc\} \quad (7)$$

La manera de seleccionar las mejores alineaciones entre oraciones es muy similar a la que se describe en la alineación a nivel de documento, con la diferencia de que en este caso se debe respetar el orden de las oraciones en los textos. El primer paso para la alineación es la búsqueda de puntos de anclaje a lo largo de los documentos alineados. Eso se consigue alineando primero todos aquellos pares de oraciones que sean más seguros. Concretamente, en esta implementación se toman como puntos de anclaje los pares de oraciones cuya puntación esté por encima del percentil 80, pero de nuevo esto puede funcionar como un parámetro más de la ejecución.

Una vez encontrados los puntos de anclaje resulta más fácil distribuir las oraciones dentro de

cada fragmento entre anclajes, aunque sin llegar a forzar una alineación uno-a-uno ya que la alineación oracional no es una función biyectiva. Con relativa frecuencia, una misma oración del texto meta puede alinearse con más de una oración en el texto original y viceversa, por lo tanto es necesario flexibilizar el criterio y permitir que una misma oración de un texto sea alineada con más de una oración de su contraparte.

3.4 Alineación a nivel de vocabulario

Una vez que el corpus ha sido separado en lenguas y alineado a nivel de documento y de oración, ya es posible elaborar una primera versión del vocabulario bilingüe, que podrá servir, posteriormente, como un parámetro para iterar el algoritmo. Es necesario advertir que no todos los autores consideran que la alineación a nivel de vocabulario equivale a la extracción de un vocabulario bilingüe, ya que en el primer caso se trata de alinear las unidades léxicas en el contexto mismo en el que ocurren (Santos y Simões, 2008). En este artículo se ha preferido la extracción del vocabulario bilingüe con independencia del contexto de aparición porque parece lo más útil desde el punto de vista práctico.

Tal como se advirtió ya en la introducción, el resultado de esta alineación a nivel de vocabulario no se limita únicamente a la palabra ortográfica, ya que la alineación a nivel de las unidades sintagmáticas es de una importancia capital en particular para usuarios interesados en la terminología especializada, que muy a menudo se presenta en forma de unidades poliléxicas. Por este motivo, tomamos como vocabulario para analizar no solamente las palabras ortográficas sino también todas las secuencias de hasta cinco palabras siempre y cuando no tengan como primer o último componente una palabra con menos de cuatro caracteres. La expansión del vocabulario no llega a saturar la memoria porque se descartan los hapax legomena y dis legomena.

$$C'' = \{l, sim, coo\} \tag{8}$$

Como en la alineación a nivel de documento y oración, en el caso de la alineación a nivel del léxico definimos nuevamente un conjunto biprima de coeficientes (8), en este caso con tres. Dos de ellos son comunes a las instancias anteriores. Se añade, sin embargo, un coeficiente de asociación basado en la coocurrencia de las palabras, el Coeficiente *coo*, descrito a continuación.

Coeficiente *coo*: Este coeficiente mide el grado de asociación estadística entre dos palabras a través de la coocurrencia en una misma alineación oracional. La gran mayoría de los autores que han realizado alineación de corpus paralelo a nivel de léxico han calculado de una forma u otra la coocurrencia de las unidades, utilizando para ello distintos coeficientes de asociación. El que se define en la ecuación 9 pone en relación la frecuencia de coocurrencia de las unidades léxicas i y j, candidatas a ser alineadas, normalizada por la frecuencia total de las unidades i y j en el corpus.

$$coo(i,j) = \frac{f(i,j)}{\sqrt{f(i)}.\sqrt{f(j)}} \tag{9}$$

Tal como se describió antes en la ecuación 2, la puntuación final de cada pareja de unidades léxicas potencialmente equivalentes se calcula nuevamente como el producto de los coeficientes.

4 Evaluación de los resultados

Esta sección describe los resultados obtenidos con la aplicación de la metodología descrita en la sección 3 sobre tres corpus paralelos en los pares de lenguas inglés-castellano, inglés-francés y gallego-inglés. La subsección 4.1. describe las características de los corpus utilizados. Posteriormente, las subsecciones 4.2., 4.3., 4.4. y 4.5 describen, respectivamente, los resultados obtenidos durante las fases de reconocimiento de lengua, de alineación a nivel de documento, alineación a nivel de oración y alineación a nivel de vocabulario en cada uno de estos corpus. Es preciso tener en cuenta que en las subsecciones en las que se describen estos resultados aparecen distintos valores de precisión que corresponden a distintas ejecuciones del algoritmo, ya que, como se dijo en la introducción, se trata de un algoritmo iterativo: su resultado final es un vocabulario bilingüe que luego puede funcionar como un parámetro de entrada a una nueva ejecución, en un proceso de retroalimentación que le permite mejorar el desempeño. Este artículo describe los resultados obtenidos después de tres iteraciones, aunque no hay una razón teórica para limitar este número. De cualquier modo, es de esperar que la mejora en el desempeño sea menor con cada iteración.

4.1 Corpus utilizados en los experimentos

Con el objetivo de evaluar este sistema se llevaron a cabo experimentos de alineación en tres corpus paralelos diferentes. El primero es el corpus CLUVI-TECTRA, un conjunto de obras literarias en inglés con sus correspondientes traducciones al gallego (Gómez Guinovart y Sacau Fontenla, 2004), con un tamaño total aproximado de

CLUVI	JRC-ACQUIS	HANSARDS
100 %	95,4 %	100 %

Cuadro 1: Precisión de la separación de documentos por lengua

1.500.000 palabras. El segundo es una muestra del corpus JRC-Acquis (Steinberger et al., 2006) consistente en textos paralelos de la Unión Europea, de naturaleza legal. En el caso de este último corpus, se utilizaron sólo los documentos en castellano y en inglés del año 1990, lo que totaliza 800.000 palabras en las dos lenguas. El tercer corpus está constituido por una parte de las actas del parlamento canadiense (Canadian Hansards) del año 2000 publicadas por Ulrich Germann (2001), para el par de lenguas inglés-francés, también con un total de 800.000 palabras. El corpus CLUVI se divide en 600 ficheros, el JRC-Acquis en 332 y el Hansards en 42. Naturalmente, antes de utilizar estos corpus para la evaluación se eliminó toda la metainformación contenida en las etiquetas XML, dejando todos los documentos como texto plano.

4.2 Resultado de la separación de documentos por lengua

La precisión del resultado de la división del corpus por lenguas aparece detallada en la tabla 1 para los distintos corpus que han servido para el experimento. En este caso se expresan los resultados en una única vez ya que estos no se modifican con las distintas iteraciones como en los demás procesos de alineación. La discriminación por lengua resulta 100 % correcta excepto en el caso del corpus JRC-Acquis, y una de las razones que pueden explicar los errores es que los documentos en castellano de este corpus incluyen también largas leyendas en inglés, y cuando los documentos son muy cortos, lógicamente esta leyenda puede dificultar la detección.

4.3 Resultado de la alineación a nivel de documento

El cuadro 2 expone los resultados de la alineación a nivel de documento para cada uno de los corpus. Debido a que la alineación es afectada por las sucesivas iteraciones del algoritmo, en este cuadro se expone el resultado de las mismas muestras en tres iteraciones sucesivas. Como se puede apreciar, también en este caso la alineación a nivel de documento es óptima en el caso del CLUVI y los Hansards pero no en el JRC-Acquis, como consecuencia de los errores cometidos por el algoritmo en la instancia anterior.

iteración	CLUVI	JRC-ACQUIS	HANSARDS
1º	98,3 %	89,7 %	100 %
2º	100 %	90,9 %	100 %
3º	100 %	90,9 %	100 %

Cuadro 2: Precisión de la alineación a nivel de documento

iteración	CLUVI	JRC-ACQUIS	HANSARDS
1º	85,0 %	97,4 %	93,1 %
2º	86,8 %	97,9 %	94,3 %
3º	86,8 %	98,3 %	94,5 %

Cuadro 3: Precisión de la alineación oracional

4.4 Resultado de la alineación a nivel de oración

Para llevar a cabo la evaluación de la alineación a nivel de oración hacemos un muestreo aleatorio estratificado de cinco documentos por corpus (dos en el caso del Hansards, ya que son documentos mucho más largos), con el objeto de obtener una muestra representativa de los documentos de distinto tamaño, ya que es de esperar que la alineación oracional sea más difícil en el caso de los documentos más largos. Como en el caso anterior, en el cuadro 3 también se expone el rendimiento de las alineaciones en tres iteraciones.

La evaluación se lleva a cabo revisando manualmente los documentos de la muestra, y el porcentaje simplemente representa la proporción entre alineaciones correctas e incorrectas. Para evaluar se toma como medida el segmento, que coincide en general con una oración pero puede ocurrir que contenga más de una. La forma de evaluar es simplemente controlar que a cada oración del texto de origen le corresponda una alineación correcta en el texto meta.

En la alineación a nivel oracional, el corpus CLUVI es el que ha mostrado el peor rendimiento, lo cual no es del todo sorprendente dada la naturaleza del corpus. Un corpus literario es siempre más complejo, porque los traductores sienten mayor libertad y los segmentos se encuentran menos estructurados. No es infrecuente que se sustraigan o se inserten segmentos, tal como se puede apreciar en el ejemplo del cuadro 4. Además, en un corpus literario existe menor cantidad de símbolos y cognados que son frecuentes en un corpus técnico y que ayudan a un alineador de estas características.

En el caso del corpus JRC-Acquis los errores de alineación son infrecuentes, lo cual se puede explicar por la extrema pulcritud con que se ha llevado a cabo la traducción. Las traducciones son muy similares al original, son pocos los casos

Original	**Traducción al gallego**
Then old Luce ordered another martini and told the bartender to make it a lot drier.	Logo o Luce pediu outro martini, e aínda máis seco.
Listen.	
How long you been going around with her, this sculpture babe?' I asked him.	-¿E canto tempo levas saíndo con esa escultura?
I was really interested.	
Did you know her when you were at Whooton?'	¿Coñecíala xa cando estabas en Whooton?
Hardly.	-¿Como a ía coñecer?
She just arrived in this country a few months ago.'	Hai só uns meses que chegou a este país.
She did?	-¿Si?
Where's she from?'	¿De onde é?
She happens to be from Shanghai.'	-Pois é de Shanghai.
No kidding!	-¿En serio?
She Chinese, for Chrissake?'	¿É chinesa?
Obviously.'	-Dende logo.
No kidding!	
Do you like that?	-¿E gústache iso?
Her being Chinese?'	¿Que sexa chinesa?

Cuadro 4: Ejemplo de alineación oracional (fragmento de "The Catcher in the Rye", de J.D. Salinger)

en los que una oración se traduce por más de una y solo en muy contadas ocasiones los traductores han eliminado o insertado pasajes. En el caso del Hansards la proporción de errores es ligeramente mayor, pero se trata también de una traducción sumamente fiel al original, se podría decir incluso "ideal" para una alineación.

Hay que reconocer que los usuarios del sistema aquí presentado no siempre utilizarán corpus paralelos de una calidad comparable a estos, lo cual es un factor de riesgo para la calidad del resultado. En cualquier caso, también hay que destacar que gracias a la estrategia de los puntos de anclaje, cuando se produce un error en una alineación no hay una reproducción en cadena de ese error, ya que rápidamente se recupera la alineación correcta en las oraciones siguientes. Además, los errores de alineación casi siempre se dan en oraciones contiguas, la oración correspondiente en la traducción está a una o dos posiciones antes o después de la que se seleccionó erróneamente.

4.5 Resultado de la alineación a nivel de léxico

El último paso de esta evaluación produce un vocabulario bilingüe que incluye unidades poliléxicas. Como en los casos previos, en esta subsección se exponen los resultados de tres iteraciones. Se evaluaron manualmente los primeros 2.000 pares de equivalentes resultantes de cada experimento. Naturalmente, es de esperar que la calidad de los resultados decaiga progresivamente al seguir elementos que aparecen más abajo en la lista. En el cuadro 5 se exponen los porcentajes de precisión como la proporción de alineaciones correctas para los tres corpus en las tres iteraciones. Tal como se puede apreciar, en

iteración	CLUVI	JRC-ACQUIS	HANSARDS
1º	99,8 %	96,8 %	97,8 %
2º	99,9 %	98,4 %	98,6 %
3º	99,9 %	98,4 %	98,8 %

Cuadro 5: Resultados de la alineación a nivel de vocabulario

todos los casos el aumento de la precisión de alineaciones se da fundamentalmente de la primera ejecución a la segunda. Prácticamente no hay diferencias entre la segunda y la tercera ejecución en el caso de los 2.000 pares mejor posicionados. El cuadro 6 muestra algunos ejemplos de la alineación léxica en el corpus JRC-Acquis, en las posiciones 357-371 de la lista. Como se puede apreciar, el algoritmo es capaz de resolver alineaciones léxicas complejas como las de unidades de tres componentes (*microorganismos modificados genéticamente*) con otro de cuatro componentes (*genetically modified micro organisms*).

En cuanto a los errores que se producen en la alineación a nivel de vocabulario, estos se dan casi exclusivamente en el caso de la alineación de términos poliléxicos, y se perciben en mayor proporción en el caso del corpus JRC-Acquis debido a que allí los términos poliléxicos son mucho más frecuentes por la naturaleza más técnica de ese corpus. Los siguientes son algunos ejemplos de alineaciones incorrectas:

- *consecutivos durante* ≠ *consecutive years during*
- *prueba suficiente* ≠ *being sufficient proof*
- *monetaria internacional* ≠ *competent international monetary*
- *república democrática* ≠ *German Democratic Republic*

Rango	Término castellano	Término inglés
...	...	...
357	autoridades administrativas	administrative authorities
358	provisionales	provisional
359	microorganismos modificados genéticamente	genetically modified micro organisms
360	iniciativa	initiative
361	expertos	experts
362	recomendaciones	recommendations
363	portugal	portugal
364	microorganismos	micro organisms
365	integrado	integrated
366	programa	programme
367	contacto	contact
368	interpretación uniforme	uniform interpretation
369	racional	rational
370	diferencia	difference
371	secretario general	secretary general
...	...	...

Cuadro 6: Ejemplos de alineación léxica en el corpus JRC-Acquis

Estos ejemplos, que son representativos de los errores que se encuentran, llevan a pensar que se podrían resolver con un mínimo grado de conocimiento lingüístico, tal como un modelo de sintaxis derivado de un etiquetador morfosintáctico, que se puede conseguir con facilidad en el caso de la mayoría de las lenguas europeas.

5 La interfaz

Esta sección ofrece una breve descripción del funcionamiento de la interfaz web que es la forma que se propone como implementación del algoritmo. Tal como se advierte en la introducción, el artículo no pretende ofrecer una descripción pormenorizada de los aspectos técnicos de la aplicación informática en sí, ya que el interés del trabajo está más en el método que en el programa.

El programa en sí es menos importante porque un programador podría preferir implementar el mismo algoritmo en C en lugar de Perl por cuestiones de eficiencia, o hacerlo como una aplicación de escritorio en lugar de una aplicación web. Los aspectos técnicos de un producto informático son complejos y motivarían un artículo diferente. El desarrollo de productos informáticos requiere estudios de usuario y el cuidado de una serie de aspectos relacionados con la usabilidad de las interfaces. Por ejemplo, la demo online solo acepta como entrada ficheros de texto plano. Un terminólogo o traductor no tiene por qué saber cuál es la diferencia entre un archivo binario y un archivo de texto, por tanto, si se le solicita un archivo de texto probablemente proporcionará un documento de Word o un PDF. Algo en apariencia trivial como la conversión de formatos PDF a texto puede resultar muy complejo en algunos casos (eliminación o tratamiento de símbolos, tablas, fórmulas y epígrafes de las figuras o imágenes, reconocimiento y conversión de codificación de caracteres, reconstrucción de texto dividido en columnas, de palabras que se cortan a final de línea, restauración de diacríticos, de errores de reconocimiento óptico de caracteres y toda una serie de temas que no tienen relación intrínseca con la alineación de corpus paralelo), por lo que se ha preferido dejar de lado ese tipo de cuestiones técnicas en favor de un modelo básico y una argumentación más abstracta.

Dicho esto, también es verdad que el mismo diseño de Bifid facilita su utilización por parte de usuarios no informáticos, lo cual no deja de ser importante ya que, si bien se han presentado distintas herramientas para la alineación de corpus paralelo, como se comentó en la sección 2, en general estas no se caracterizan por un diseño amigable para un usuario sin conocimientos avanzados de informática. Ejecutar una aplicación en línea de comando, como muchas de ellas requieren, no es algo que esté en el horizonte de posibilidades de la mayoría de los usuarios no expertos en informática en la actualidad. Es, por tanto, un valor práctico que el sistema tiene ya en su estado actual de desarrollo: el poder ser operado con facilidad por un usuario que al menos sea capaz de proporcionar un corpus en forma de archivos de texto.

La particularidad de una implementación de este algoritmo en forma de aplicación web tiene obvias ventajas como el poder ser ejecutada en cualquier plataforma sin necesidad de llevar a cabo una instalación. Sin embargo, la decisión también acarrea algunos inconvenientes, principalmente que el coste computacional del sistema y las limitaciones de infraestructura hacen que no sea posible una respuesta instantánea del servidor a las solicitudes de los usuarios. El diseño del algoritmo tiene aún que mejorar para funcio-

nar más rápido, pero por el momento la potencia del hardware es un factor clave, y ello exige que el programa funcione en varios servidores en red para que sea viable.

En este momento el sistema se encuentra instalado en un solo servidor, y en lugar de devolver los resultados de inmediato, envía una dirección URL al correo electrónico que el usuario indica en la solicitud. En esta URL es posible observar los resultados a medida que se van generando, pero se debe esperar que estos resultados "maduren" para obtener la calidad óptima, lo que se consigue después de dos o tres iteraciones. El envío de los datos se da por medio de un formulario web a través del cual el usuario sube un archivo comprimido que contiene los documentos de su corpus paralelo. Existe la posibilidad de realizar correcciones en la salida de cada proceso para evitar que los errores se propaguen a las instancias posteriores. A la salida de cada proceso el usuario puede descargar la información en ficheros de texto con nombres que corresponden a cada proceso ("lang.txt" para la separación de lenguas, "doc.txt" para la alineación a nivel de documento, etc.) que podrá modificar con un editor de texto para luego repetir el experimento incluyendo estos ficheros en el archivo del corpus.

En su estado actual, el tiempo de respuesta del sistema varía en función del tamaño del corpus y de la carga de usuarios, pero como referencia general, el procesamiento con tres iteraciones de un corpus de un millón de palabras puede tardar 24 horas, a lo que hay que sumar el tiempo que lleve el proceso en lista de espera. Una leyenda en la interfaz informa en todo momento sobre el número de trabajos pendientes.

6 Conclusiones

Este artículo ha presentado un sistema integral de alineación de corpus paralelo que no incorpora ningún tipo de conocimiento lingüístico y que puede por tanto ser utilizado en cualquier par de lenguas. La datos generados muestran que los resultados son competitivos y que ya es posible su aplicación a casos reales para obtener material de una calidad suficiente como para que sea rentable el posterior procesamiento y corrección por parte del usuario, tarea que puede ser llevada a cabo a lo largo de cada uno de los pasos del proceso (alineación a nivel de documento, de oración y de léxico).

En general, la calidad de los resultados tanto en el nivel de la alineación oracional como la del vocabulario, que son los dos niveles que han sido ya explorados en la bibliografía, se encuentran en el mismo orden que los mejores resultados de las publicaciones consultadas aunque, como es sabido, las comparaciones son siempre meramente orientativas por las diferentes características de los corpus y las lenguas analizadas por cada autor. Para conseguir una comparación rigurosa sería preciso llevar a cabo un experimento con distintos algoritmos trabajando sobre un mismo corpus, que se deja para trabajo futuro ya que no era el objetivo principal de este artículo.

Del diseño del algoritmo se puede decir que es original en lo teórico y económico en lo práctico, ya que permite una alta portabilidad a otras lenguas sin necesidad de organizar o procesar lingüísticamente el corpus. Hay que decir, con todo, que aún es necesario seguir experimentando con distintos pares de lenguas antes de poder afirmar categóricamente que es independiente de lengua, ya que, como algunos autores advierten (Choueka, Conley, y Dagan, 2000), estos algoritmos deben ser evaluados en pares como hebreo-inglés, árabe-inglés, chino-inglés, etc. En el caso de las lenguas altamente aglutinantes, como el turco, es de esperar que los resultados sean peores. Es también el caso del swahili, por ejemplo, en el que una secuencia en inglés como "I have turned him down" se puede traducir con una sola unidad léxica, "Nimemkatalia" (Pauw, Wagacha, y Schryver, 2009).

En cuanto a trabajo futuro, existe toda una serie de mejoras que se están llevando a cabo de cara a una nueva versión de este alineador que será de mayor complejidad. Estas mejoras son el objeto de un nuevo artículo que se encuentra en preparación y que se dedica por entero a medir cómo cambia la calidad de los resultados en cada uno de los niveles de alineación con cada modificación que se hace sobre la versión básica presentada en este artículo.

Agradecimientos

Este proyecto es posible gracias a un contrato del autor como técnico en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra. Agradezco a Alberto Simões, Alberto Lugrís y Diana Santos por leer una versión previa del artículo y contribuir con sus comentarios a mejorarlo sustancialmente.

Bibliografía

Almeida, J., A. Simões, y J. Castro. 2002. Grabbing parallel corpora from the web. *Procesamiento del Lenguaje Natural*, (29):13–20.

Appelo, L. y J. Landsbergen. 1986. The Machine Translation Project Rosetta. En *International Conference on the State of Machine Translation in America, Asia and Europe. Proceedings of IAI-MT86, 20-22 August*, Bürgerhaus, Dudweiler.

Braune, F. y A. Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. En *Proceedings of the 23th International Conference on Computational Linguistics Coling*, páginas 81–89.

Brown, P., V. DellaPietra, S. DellaPietra, y R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Brown, P.F., J. C. Lai, y R. L. Mercer. 1991. Aligning sentences in parallel corpora. En *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 169–176, Berkeley.

Choueka, Y., E. Conley, y I. Dagan. 2000. A comprehensive bilingual word alignment system. application to disparate languages: Hebrew and English. En *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, páginas 69–96.

Church, K.W. 1993. Charalign: a program for aligning parallel texts at the character level. En *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 1–8, Columbus, Ohio.

Daille, B. y E. Morin. 2005. French-English terminology extraction from comparable corpora. En *Proceedings of the Second international joint conference on Natural Language Processing*, IJCNLP'05, páginas 707–718, Berlin, Heidelberg. Springer-Verlag.

De Yzaguirre, Ll., M. Ribas, J. Vivaldi, y M. T. Cabré. 2000. Some technical aspects about aligning near languages. En *Proceedings of the 2nd International Conference on Language Resources and Evaluation, (LREC'2000)*, páginas 545–548, Athens, Greece.

Fung, P. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. En *Proceedings of the Third Workshop on Very Large Corpora*, páginas 173–183.

Gale, W. y K. Church. 1991b. Identifying word correspondence in parallel texts. En *Proceedings of the workshop on Speech and Natural Language*, HLT '91, páginas 152–157, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gale, W. A. y K.W. Church. 1991a. A program for aligning sentences in bilingual corpora. En *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, páginas 177–184, Berkeley.

Gammallo, P. 2005. Extraction of translation equivalents from parallel corpora using sense-sensitive context. En *Proceedings of Conference of the European Association for Machine Translation (EAMT'05)*, Budapest, Hungary.

Gaussier, E., J.M. Renders, I. Matveeva, C. Goutte, y H. Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. En *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, páginas 526–533, Barcelona, Spain, July.

Germann, U. 2001. Aligned Hansards of the. 36th. Parliament of Canada - release 2001-1a. Informe técnico, http://www.isi.edu/natural-language/download/hansard/.

Gómez Guinovart, X. y E. Sacau Fontenla. 2004. Parallel corpora for the Galician language: building and processing of the CLUVI (Linguistic Corpus of the University of Vigo). En *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, páginas 1179–1182, Lisboa (Portugal), May.

Gómez Guinovart, X. y A. Simões. 2009. Parallel corpus-based bilingual terminology extraction. En *Proceedings of the 8th International Conference on Terminology and Artificial Intelligence*, IRIT (Institut de recherche en Informatique de Toulouse), Université Paul Sabatier, Toulouse.

Hiemstra, D. 1998. Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus. En *Computational linguistics in the Netherlands 1997*, volumen 25 de *Language and computers*, páginas 41–58, Amsterdam, the Netherlands. Rodopi.

Hofland, K. y S. Johansson. 1998. The translation corpus aligner: A program for automatic alignment of parallel texts. En *Corpora and Cross-linguistic research. Theory, Method, and Case Studies*. Rodopi, Amsterdam/Atlanta, páginas 87–100.

Kay, M. y M. Röscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

Kübler, N. 2011. Working with corpora for translation teaching in a French-speaking setting. En *New Trends in Corpora and Language Learning*. London, UK, páginas 62–80.

Ma, Xiaoyi. 2006. Champollion: A robust parallel text sentence aligner. En *Proceedings of the 5th International Conference on Language Resources and Evaluation, (LREC'2006)*, Genova, Italy.

McEnery, A. M. y M. P. Oakes. 1995. Sentence and word alignment in the CRATER project: methods and assessment. En *Proceedings of the EACL-SIGDAT Workshop: from texts to tags, Issues in Multilingual Language Analysis(ACL)*, páginas 77–86, Dublin, Ireland.

Melamed, D. 2000. Pattern recognition for mapping bitext correspondence. En *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, páginas 25–47.

Moore, R. 2002. Fast and accurate sentence alignment of bilingual corpora. En *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users, AMTA '02*, páginas 135–144, London, UK. Springer-Verlag.

Morin, E., B. Daille, K. Takeuchi, y K. Kageura. 2008. Brains, not brawn: The use of "smart" comparable corpora in bilingual terminology mining. *ACM Trans. Speech Lang. Process.*, 7(1):1–23, Octubre.

Nazar, R., L. Wanner, y J. Vivaldi. 2008. Two step flow in bilingual lexicon extraction from unrelated corpora. En *Proceedings of the 12th conference of the European Association for Machine Translation*, páginas 138–147, Hamburg: HITEC.

Och, F. y H. Ney. 2000. Improved statistical alignment models. En *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, páginas 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Och, F. y H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Pauw, G. De, P. Wagacha, y G. De Schryver. 2009. The SAWA Corpus: a parallel corpus English-Swahili. En *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009)*. Association for Computational Linguistics.

Rapp, R. 1999. Automatic identification of word translations from unrelated English and German corpora. En *Proceedings of 37th Annual Meeting of the ACL*, páginas 519–526.

Resnik, P. y N. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, Septiembre.

Santos, D. y A. Simões. 2008. Portuguese-English word alignment: some experiments. En *Proceedings of LREC 2008 Workshop on Comparable Corpora*, páginas 2988–2995, Marrakech, Marroco.

Simões, A. y J. Almeida. 2003. Natools - a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, (31):217–226.

Simard, M., G. Foster, y P. Isabelle. 1993. Using cognates to align sentences in bilingual corpora. En *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing - Volume 2*, CASCON '93, páginas 1071–1082. IBM Press.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, y D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. En *In Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Tiedemann, J. 2006. ISA & ICA - two web interfaces for interactive alignment of bitexts. En *Proceedings of the 5th International Conference on Language Resources and Evaluation, (LREC'2006)*, Genova, Italy.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, y V. Nagy. 2005. Parallel corpora for medium density languages. En *Proceedings of the RANLP 2005*, páginas 590–596.

Véronis, J. 2000. From the Rosetta stone to the information society: A survey of parallel text processing. En *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, páginas 1–24.

Weaver, W. 1955. Translation. En *Machine Translation of Languages*. MIT Press, Cambridge, Massachusetts, páginas 15–23.