Generación automática de diccionarios para la traducción especializada a partir de corpus alineados

II Congreso Internacional de Traducción Especializada
La traducción científica
Barcelona, Universitat Pompeu Fabra

Maria Teresa Cabré Castellví
Carles Tebé i Soriano
Lluís de Yzaguirre Maura
Araceli Alonso Campo

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
de_yza@upf.es

Palabras clave: *generación automática de diccionarios, alineación de textos, validación de traducciones*

1.- Presentación

El alineador elaborado en el Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra consiste en una herramienta que permite alinear frases de un documento con su traducción.

En la actualidad, el alineador da resultados óptimos en la alineación a nivel de frase y unos resultados bastante satisfactorios a nivel de alineación léxica. En este punto, se ha desarrollado una técnica de retroalimentación del alineador a partir de diccionarios extraídos automáticamente de los corpus alineados.

Esta técnica ha demostrado ser especialmente beneficiosa tanto para la validación de traducciones, como para la sistematización de criterios de marcaje y herramientas de procesamiento de corpus paralelos y para producir microdiccionarios especializados según el dominio del corpus al cual se aplique.

Los objetivos que se pretenden alcanzar en esta comunicación son los siguientes:

- Presentar el estado actual del desarrollo del alineador y su aplicación enla producción de diccionarios a medida del especialista.
- Mostrar la técnica de retroalimentación del alineador a partir de diccionarios extraídos automáticamente de los corpus alineados.
- Valorar los resultados obtenidos hasta el momento para la validación de traducciones especializadas.

2.- El alineador del IULA

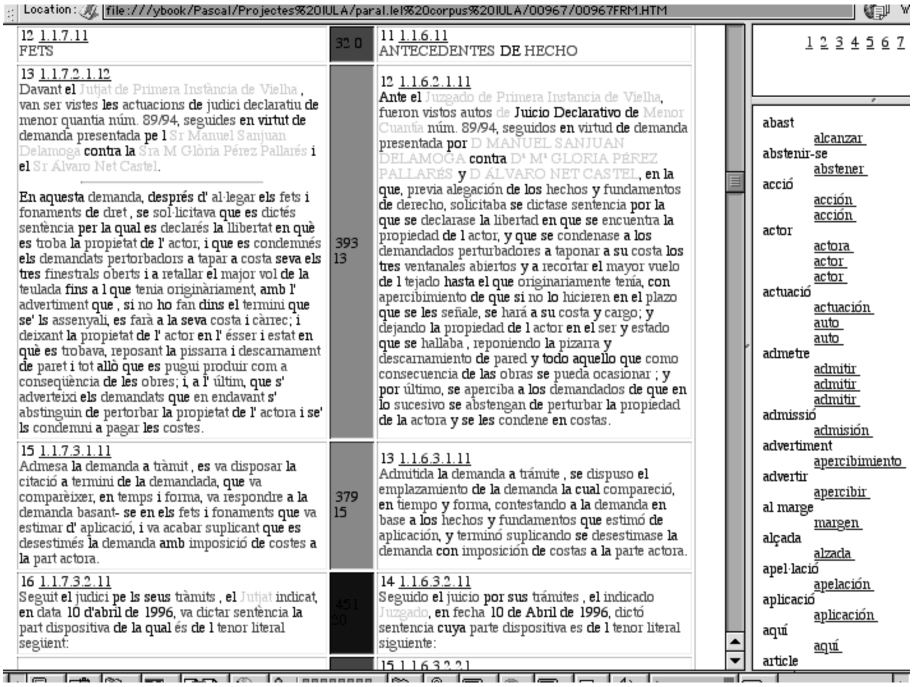
Concebimos la alineación automática de traducciones como la aplicación de un sistema capaz de alinear dos textos, uno de los cuales es el original y el otro, su traducción a otra lengua (o dos traducciones del mismo original a dos lenguas distintas o a dos variantes geográficas o cronológicas de la misma lengua), con el objetivo de vincular cada frase del texto original con la frase correspondiente de la traducción, basándose en el grado de similitud entre ellas.

Siguiendo a Baker (1995), consideramos un corpus *paralelo* como un conjunto de textos *"originally written in a language A alongside their translations into a language B"*. Hablamos de *corpus alineado* cuando éstos disponen de vínculos explícitos entre cada frase del texto origen y cada frase del texto traducción (alineación oracional), o bien entre cada palabra del texto origen y cada palabra del texto traducción (alineación léxica).

El alineador que se ha desarrollado en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra permite tanto alinear frases como también llevar a cabo una alineación léxica. A diferencia de otras herramientas de alineación, basadas en criterios estadísticos, el alineador del IULA toma las decisiones basándose en la utilización de información lingüística añadida (lemas y etiquetas), es decir, es un programa dependiente de herramientas de marcaje lingüístico para las lenguas que se quieran alinear (De Yzaguirre, L. *et al.* 2000a; Morel, J. *et al.* 1998; Vivaldi, J. *et al.* 1996). Esto implica que el alineador no compara únicamente la forma de la lengua A con la forma de la lengua B, sino también, sus lemas y etiquetas morfológicas. Así pues, para establecer un par léxico, se da prioridad primero a la coincidencia de lemas, después a la coincidencia de etiquetas y finalmente, al parecido ortográfico. Por lo tanto, funciona a partir de textos procesados con la cadena de herramientas del Corpus Técnico del IULA (Bach, C. *et al*. 1997), en catalán y castellano. También puede alinear textos analizados por otras herramientas (p. ej. Constraint Grammar, en inglés) si se homogeneizan el formato de salida y la información morfológica.

El programa trabaja en tres niveles: a) nivel de palabra, b) nivel de frase, y c) nivel de documento. En el primer nivel, se mide el grado de similitud de dos palabras; en el segundo nivel, se globalizan los resultados del nivel de palabra para cada pareja de frases; finalmente, en el tercer nivel, se establece una estrategia para decidir qué frase de la lengua A es comparada con qué frase de la lengua B, siguiendo, entre otros, los modelos clásicos de cálculo de longitud de frase, de número de palabras y de posición en el documento.

El resultado es una versión hipertextual de la alineación de dos textos, tal y como se puede observar en la siguiente pantalla:



Otros resultados son un fichero SGML que formaliza los vínculos oracionales entre ambos documentos y, opcionalmente, un diccionario derivado de un par de documentos o de un conjunto de pares de documentos de una determinada especialidad. También puede exportar los pares de frases alineadas al formato de importación de Translator's Workbench. A continuación, presentamos un fragmento del fichero de vínculos en SGML, donde se puede observar que además de los vínculos oracionales también se puede obtener el grado de certeza a que ha llegado el alineador:

```
<xprt targType="S" id=d1id0001 from='1' to='1'>
<xprt targType="S" id=d2id0001 from='1.1' to='1.1'>
<link targets="d1id0001 d2id0001" certainty=33>
<xprt targType="S" id=d1id0002 from='2' to='2'>
<xprt targType="S" id=d2id0002 from='1.2' to='1.2'>
<link targets="d1id0002 d2id0002" certainty=77>
```

De todos estos resultados, el que más interesa aquí es el de la producción de diccionarios generadosa partir de corpus paralelos de textos de especialidad. La técnica de generación de diccionarios se ha incorporado al proceso por necesidades del mismo y de alguna de sus finalidades, pero se puede aplicar íntegramente a la posibilidad que estamos comentando.

Las situaciones en que puede interesar generar un diccionario especializado a partir de corpus alineados se nos antojan múltiples:

- estandarización de las pautas de traducción de un organismo o agencia a partir de la generación automática de un diccionario bilingüe de un conjunto de textos (original y traducción) de un mismo ámbito temático;
- extracción monolingüe de candidatos a término en textos paralelos formación de traductores en un determinado perfil profesional muy especializado, especialmente si se trata de un dominio muy innovador que no esté bien reflejado en los diccionarios convencionales
- alimentación de sistemas de traducción automática focalizados en una determinada especialidad
- producción de diccionarios "papel" que cubran distintas especialidades
- preparación de tesauros y ontologías y el establecimiento de criterios de clasificación documental

En todos estos escenarios, considerando que hoy toda la documentación profesional se produce informáticamente y que se traduce más que nunca, los corpus están virtualmente disponibles, esperando ser utilizados (y reutilizados) con la ayuda de las técnicas y procesamientos adecuados.

3.- La técnica de generación de diccionarios

En el curso de desarrollo del alineador se ha demostrado que el programa mejoraba notablemente su eficiencia si era capaz de aprender de sus alineaciones, materializándose el aprendizaje en un diccionario de pares léxicos prealineados generado a partir de la alineación léxica que llamamos diccionario de inclusiones, y en un diccionario de exclusiones, complementario del anterior, que incluye todos los pares léxicos que queremos rechazar sistemáticamente, especialmente errores debidos a la similitud, como "set" inglés frente a "sed, ser, sea" en castellano o "set, ser, seu, sot, net, etc." en catalán.

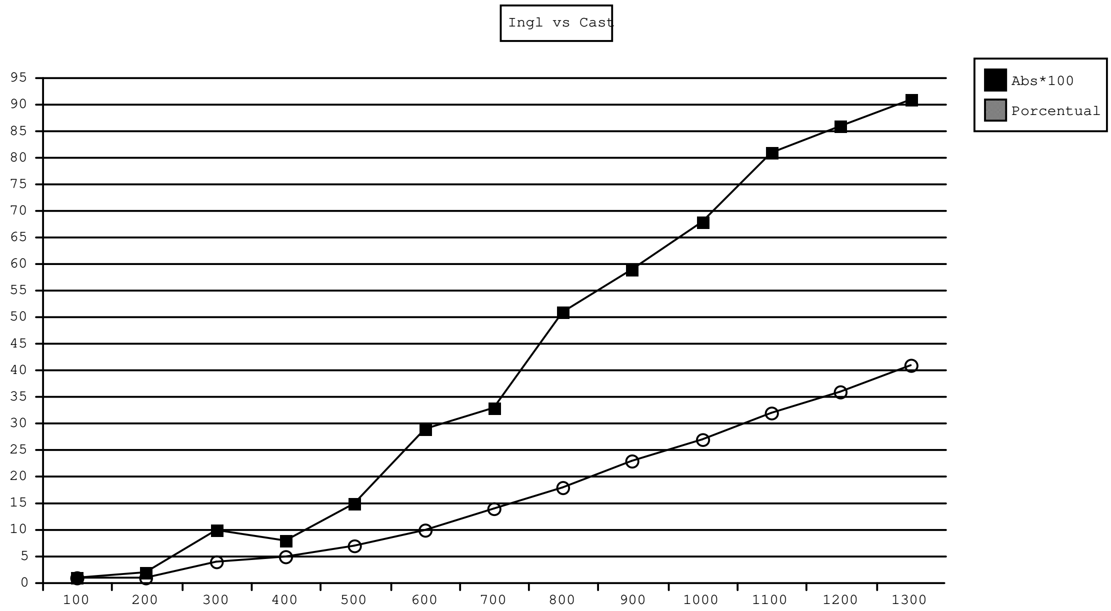
En principio, el programa, al alinear dos textos, genera un diccionario automáticamente. Este diccionario se crea a partir de la indexación y acumulación de todas las alineaciones léxicas hechas por el ordenador, por lo que incluye también entre un 10% y un 40% de errores, según el par de lenguas y la conflictividad de la traducción (calidad, errores de segmentación en frases, de procesamiento, etc.).

Para evitar todos los errores o una parte significativa de ellos, se puede hacer una revisión manual de dicho diccionario generado automáticamente, o se pueden usar criterios numéricos para aislar los errores de los aciertos. Sin embargo, la alineación automática de textos interesa, en general, si se trata de grandes volúmenes de datos, por lo que los costes de una revisión humana exhaustiva pueden llegar a ser prohibitivos.

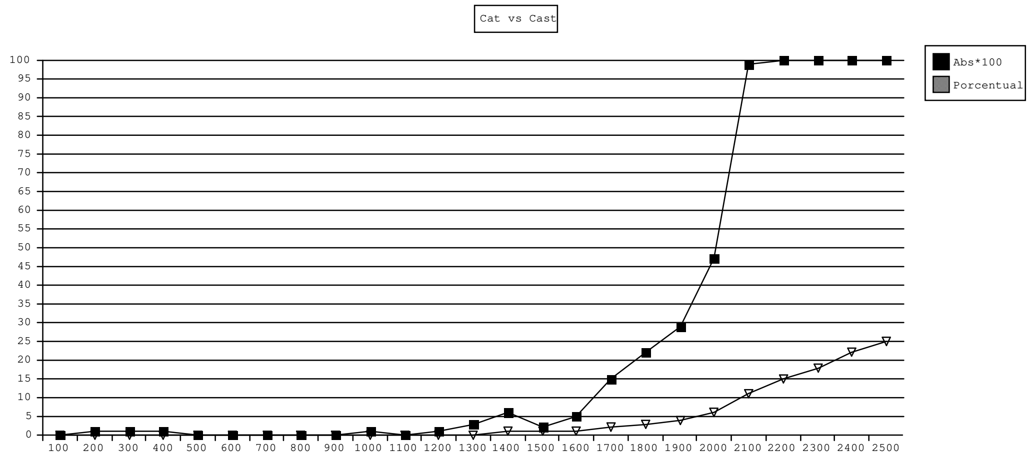
Por ello nos propusimos elaborar un método automático de ponderación de la probabilidad de que un par léxico fuera correcto, basándonos en el número de veces que el par ocurre, en cuantos documentos distintos, el índice de confianza que el alineador atribuye a cada par y a la frase en que lo encuentra y otros criterios numéricos que el ordenador puede manipular.

A este coeficiente lo hemos llamado Probabilidad Biléxica Decreciente (PBD) y puede ser usado para retroalimentar selectivamente el alineador reduciendo la proporción de errores consolidados o para optimizar la intervención de un validador humano, haciendo que revise solamente una parte de los pares léxicos, elegida con el coeficiente PBD de manera que contenga el mínimo de errores.

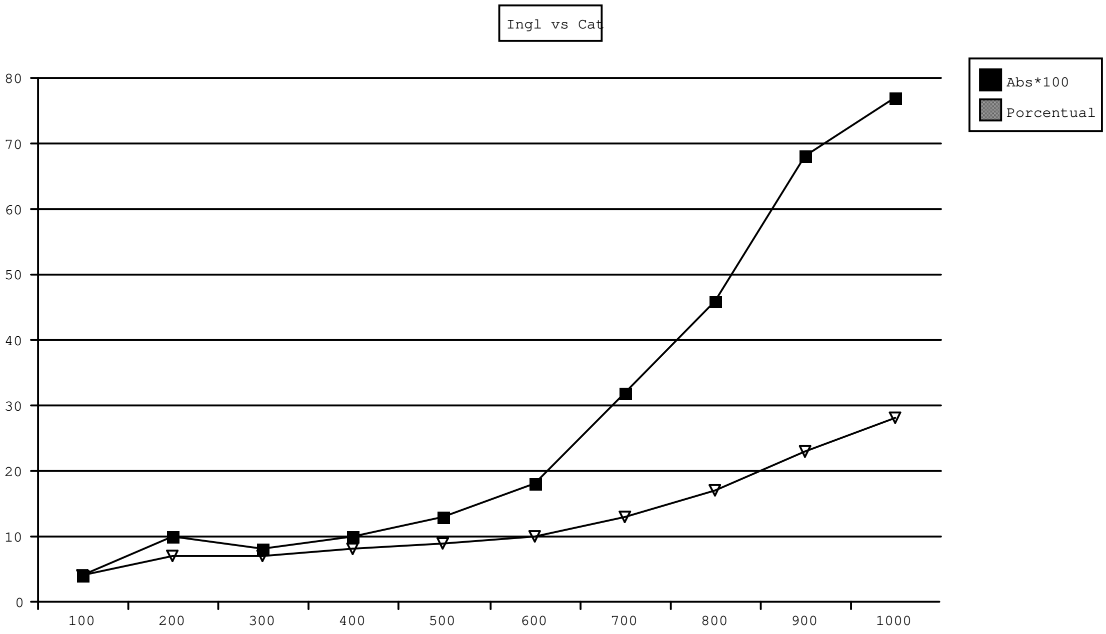
En un experimento destinado a comprobar la fiabilidad del PBD (Alonso, A. *et al.* 2001) pudimos verificar que, de entre las entradas de un diccionario generado automáticamente (de casi 1400 pares) a partir de la alineación automática del texto en inglés y su traducción al castellano y ordenado según el PBD, los 400 primeros pares acumulan hasta un 5% de errores, o sea que el 95% de los pares son correctos. La gráfica siguiente muestra el número de pares erróneos por cada cien pares en términos absolutos (número de errores en cada centenar de pares) y acumulación porcentual (por ejemplo, en el sexto centenar se comenten 29 errores que, sumados a los 36 acumulados de los cinco centenares precedentes dan un 10% porcentual acumulativo. Los pares están ordenados según su PBD y el trazado ascendente de la línea de errores absolutos por cada centenar de pares evidencia que el PBD ha logrado su objetivo de concentrar los aciertos en el inicio de la lista y los errores en el final:



Para el par inglés-castellano podríamos retroalimentar automáticamente el alineador con los 400 primeros pares con un margen de error del 5%. Hay que tener en cuenta que la consolidación de errores de alineación léxica tiene relativamente poca importancia, porque la mayoría de los errores son fortuítos y difícilmente se repetirán en otro texto diferente o incluso en el mismo si se reprocesa después de retroalimentar el alineador, mientras que con los aciertos pasa exactamente lo contrario, la probabilidad de que se repita el mismo par es muy elevada. Repetir el proceso de alinear con el vocabulario bilingüe generado con el PBD y volver a generar nuevos diccionarios recursivamente permite reducir los errores sin intervención humana.



La gráfica resultante del mismo experimento para el par de lenguas Catalán-Castellano permite ver que para dos lenguas más próximas el proceso en general de alineación y en particular de generación de diccionarios es mejor por cuanto a) se generan muchos más pares léxicos y b) la concentración de errores al final de la lista gracias al PBD es mucho más contundente. En concreto, en los 1300 primeros pares hay menos del 1% de errores acumulados y hemos de subir hasta el par 1900 para llegar al 5% que en el caso de la alineación inglés-castellano aparecía con el par 400. Aparte de la proximidad entre ambas lenguas, influye también en las diferencias el hecho de que en el par catalán-castellano empezamos a trabajar dos años antes que en los otros pares.



La gráfica del par de lenguas Inglés-Catalán se parece más a la del par Inglés-Castellano que al par Catalán-Castellano, por las razones que ya hemos comentado.

El material generado para el experimento de validación del PBD está almacenado en una base de datos, junto con la alineación de los textos, y resultó ser útil no sólo para la validación del funcionamiento del programa sino también como material para usar en la docencia de la traducción y de los lenguajes de especialidad, ya que permite al alumno detectar problemas y errores de traducción frente a errores propios del programa, observar las diferencias estructurales y fraseológicas entre la lengua de partida y la lengua de llegada, etc. (Alonso, A. *et al*. 2001).

4.- Cálculo del coeficiente de probabilidad biléxica decreciente

Hasta aquí hemos visto que la utilización de un coeficiente generado computacionalmente puede permitir tratar grandes volúmenes de pares léxicos obtenidos de corpus alineados, que tendrán diversas utilidades (De Yzaguirre, L. *et al.* 2001). A continuación, presentaremos el método de obtención del PBD.

La Probabilidad Biléxica Decreciente se basa conceptualmente en la siguiente premisa: "*cuantas más veces decida el ordenador aparear un determinado par de palabras concretas, más probable es que dicho par sea correcto*", lo cual es lo mismo que decir que es muy difícil que pueda el ordenador repetir el mismo par erróneo múltiples veces. De hecho, el ordenador retiene y rechaza los pares léxicos con un fundamento numérico que se parece conceptualmente en gran manera al procedimiento que se sigue en una estrategia de detección de concurrencias (en inglés, *collocations*) al discriminar entre secuencias casuales de palabras de aquellas que se basan en alguna regularidad.

Además, el PBD se basa en otros factores numéricos que completan su solidez, como mostraremos a continuación; para calcular el PBD de cada par léxico hay que calcular una serie de coeficientes parciales y combinarlos al final:

a) Calcular el coeficiente parcial X5.

El coeficiente X5 es el número de veces que se repite cada lema en la L1. Cada lema de la L1 tiene su X5.

b) Calcular el coeficiente parcial X6.

El coeficiente X6 es el número de veces que se repite cada lema en la L2. Cada lema de la L2 tiene su X6.

c) Calcular el coeficiente parcial X2.

El coeficiente X2 es el número de veces que el lema de la L1 está relacionado con un lema de la L2. Cada par L1-L2 tiene su X2.

d) Calcular el coeficiente parcial X3.

El coeficiente X3 interrelaciona X2, X5 y X6: es el promedio del porcentaje de X2 sobre X5 con el de X2 sobre X6. Cada par L1-L2 tiene su X3.

$$x3=((((x2*100) / x5) + ((x2*100) / x6))/2)$$

e) Establecer el coeficiente por frase X7, coeficiente léxico X8 y número de documentos X4.

El cómputo del PBD aprovecha información generada por el alineador (De Yzaguirre, L. *et al*. 2000c), la cual expresa el nivel de confianza que dicho programa tiene sobre cada par léxico (Coeficiente de Similitud Léxica, CSL) y sobre cada frase (Coeficiente de Similitud de Frase, CSF, siendo éste último el promedio de los CSL de sus pares decrementado por el número de palabras desaparejadas). Por lo tanto, los coeficientes parciales X7 y X8 representan un procedimiento para transferir al PBD los conocimientos lingüísticos en que se basa el alineador y su grado de satisfacción respecto a los pares léxicos y oracionales.

Para cada par léxico, X8 es el promedio del CSL de todas sus ocurrencias (coeficiente que puede variar por cuanto depende del grado de conflictividad de su contexto); y X7 es el promedio del CSF de cada una de las frases en que aparece dicho par léxico.

Finalmente, para cada par léxico, X4 es el número de documentos diferentes en que se ha encontrado.

Los coeficientes X7, X8 y X4 se transforman en un porcentaje sobre el valor máximo observado.

f) Combinar todos los coeficientes parciales.

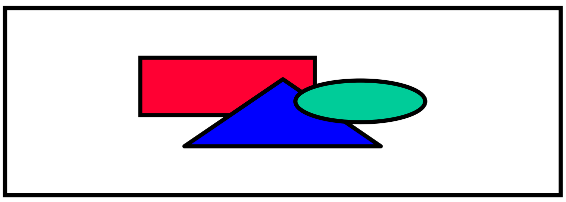
PBD = x8 + x7 + x4 + exp(x2)*5 + x3*2, donde **exp** es una función exponencial que incrementa la preponderancia de los pares muy repetidos.

Como puede verse, hay unos factores de corrección que ponderan X2 y X3. Una de las mejoras posibles del procedimiento de cómputo de PBD es la de comprobar si estos factores de ponderación se pueden sintonizar (*fine tunning*) según el par de lenguas, el volumen del corpus o el ámbito de especialidad.

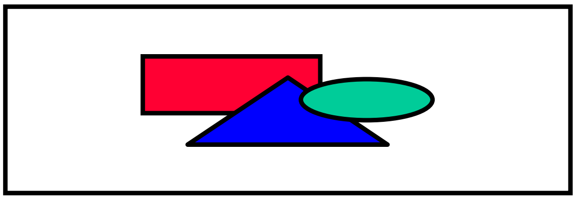
g) Ejemplificación.

De los diversos integrantes del PBD, el más complejo es X3, por lo que pasamos a presentar algunos ejemplos de cómo se calcula.

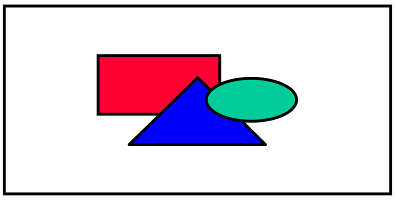
El par *politician-más* ha ocurrido 2 veces (X2), sobre un total de 7 (X5) *politician* y 170 (X6) *más*. Por tanto, ese par representa aproximadamente el 28% de los pares de *politician* y menos del 1% de los pares de *más*, dándonos un redondeado 14% como X3.



En el segundo caso, el par *political-político* agrupa la mayoría de las ocurrencias de ambos términos: 61 ocurrencias de *political* (sobre un total de 73), o sea el 83%, están apareadas con *político* y el 73% a la inversa (61 ocurrencias sobre 83 casos), dando un X3 de 78.



En el tercer caso tenemos un ejemplo de polisemia, donde la palabra *política* tiene repartidos sus apareamientos con *policy* y con *politics*. En el primer caso, X3=27 y en el segundo X3=30. Es evidente que este caso es más conflictivo que el anterior, pero resulta significativo **a)** que la suma de ambos X3 sea 57, superior al 50% y **b)** que cada uno de ellos doble un par erróneo, como *politician-más*.



5.- La validación de traducciones

Uno de los terrenos novedosos en que se pueden aplicar los alineadores como el que presentamos es la validación de traducciones, basándose en la localización de discrepancias entre ambas lenguas (que no siempre serán errores). Esta tarea exige robustez a la herramienta usada, que en nuestro caso se obtiene retroalimentándola con sus propios diccionarios, tal como hemos explicado en los apartados anteriores.

El principio básico del funcionamiento del alineador como validador de traducciones es que cuanto peor sea una traducción, peor resultado obtendrá el alineador. Aunque asumimos el principio, en nuestro caso hay que matizarlo en el sentido de que el alineador es una herramienta en desarrollo con varias limitaciones (por ejemplo, sólo alinea léxicamente palabras aisladas, no grupos de palabras); casi siempre responde mejor ante una traducción correcta que ante una errónea.

En cualquier caso, hay que destacar que los alineadores de base estadística no permiten esta utilización y están limitados a traducciones altamente literales. En cambio, los alineadores de base lingüística son más tolerantes a discrepancias y errores, sean permutaciones de frases, distribución desigual de las frases entre ambos documentos o simples adiciones o supresiones de frases enteras por criterio del traductor.

Un caso donde, en principio, el alineador produciría lo contrario de lo previsto sería ante los errores conocidos como *falsos amigos*, donde una estrategia de comparación como la nuestra es especialmente vulnerable.

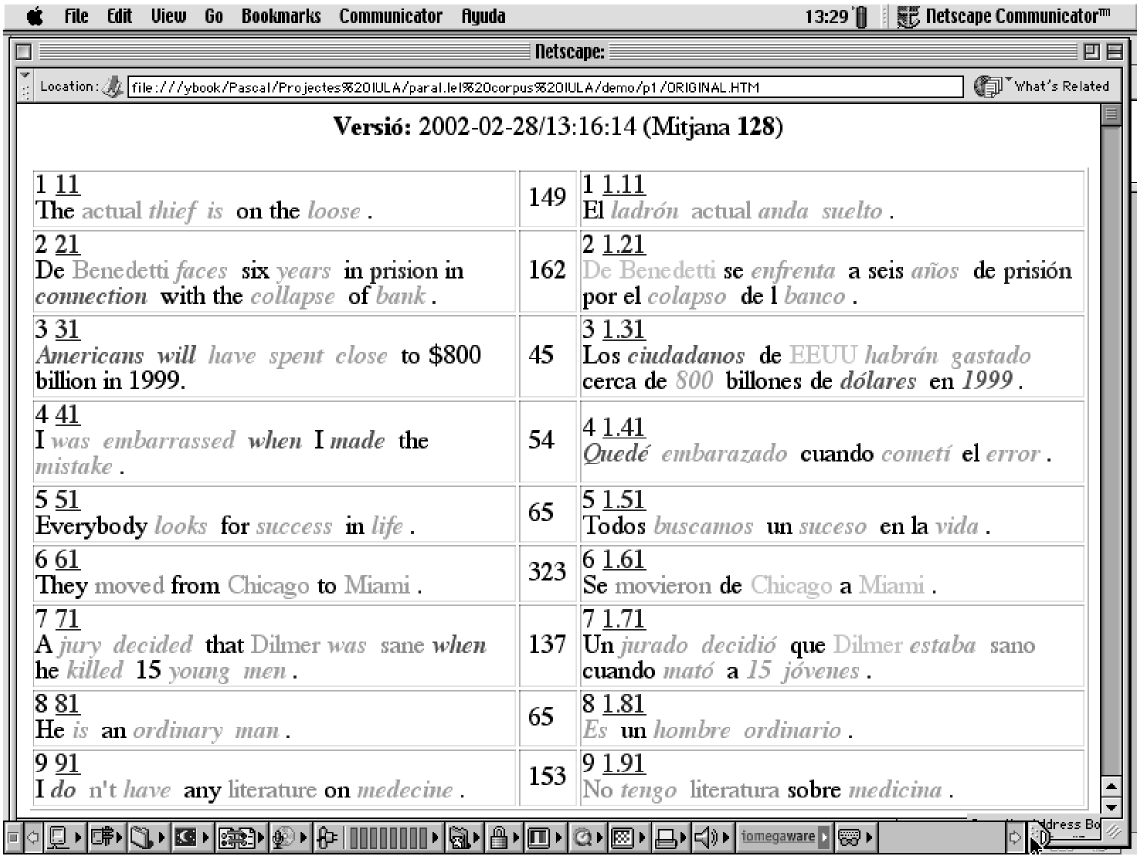
Este es un terreno donde se demuestra el interés de realimentar el alineador con diccionarios generados automáticamente (de hecho, los generados manualmente también podrían interesar, si no fueran prohibitivos sus costes de producción).

Nuestra experiencia usando el alineador del IULA nos permite afirmar que si usamos el alineador sin diccionarios da unos resultados peores con traducciones que contengan errores del tipo *falsos amigos*, del mismo modo que si la traducción no respeta el orden original de las frases. Pero en ambos casos, disponiendo de los diccionarios de inclusiones y de exclusiones que generamos con la ayuda del coeficiente PBD, se neutralizan estos problemas y los resultados pasan a ser ampliamente satisfactorios.

Para mostrar la influencia del diccionario en la robustez del alineador, hemos procesado seis veces el mismo par de documentos con algunas variantes. El original es siempre el mismo, mientras que la traducción presenta dos variaciones: traducción correcta vs. traducción con falsos amigos y traducción con las frases en el mismo orden que el original vs. traducción desordenada. La tercera variable es el uso o no de diccionarios de inclusiones y exclusiones.

versión	calidad	falsos amigos	in/exclusiones	traducción
1	128	sin	no	ordenada
2	98	sin	no	ordenada
3	295	con	sí	ordenada
4	365	con	sí	ordenada
5	91	sin	no	desordenada
6	361	con	sí	desordenada

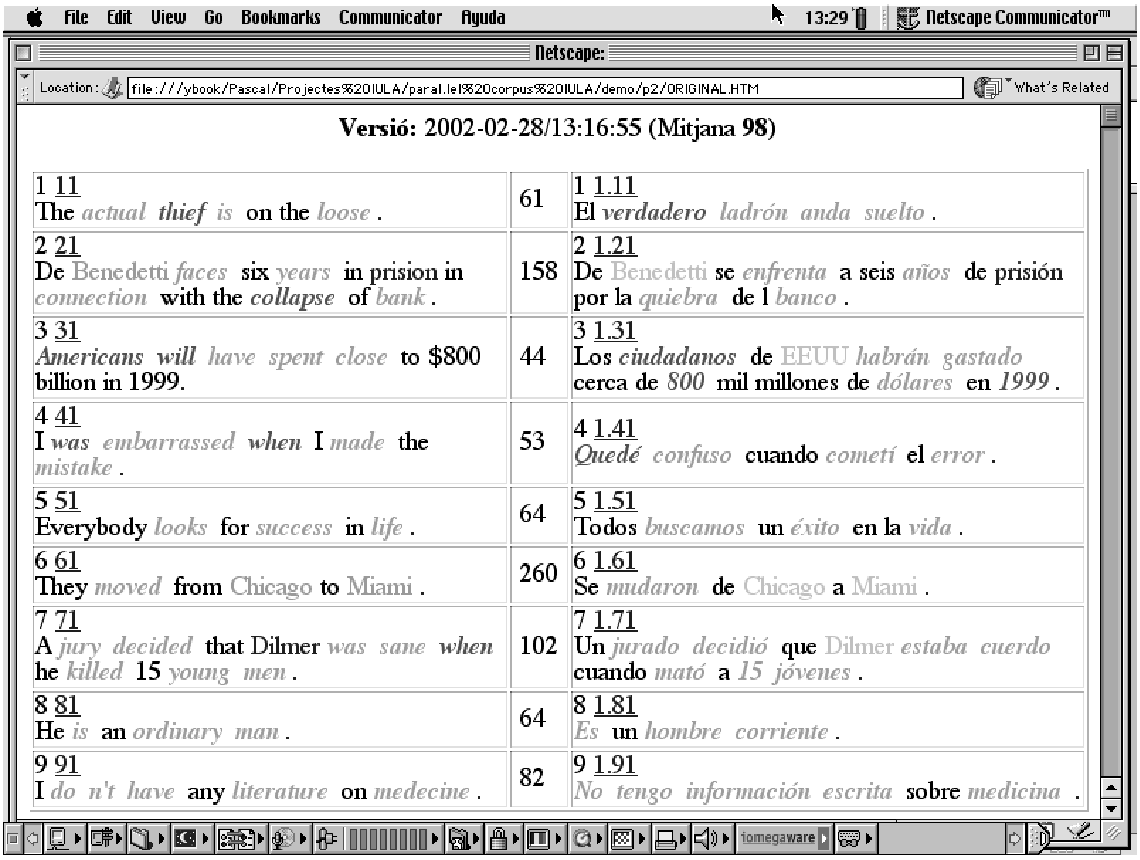
La tabla precedente muestra las combinaciones con las que hemos experimentado para esta presentación, cuyos resultados se recogen en las seis capturas de pantalla que siguen, donde hemos destacado en cursiva las palabras que obtienen un Coeficiente de Similitud Léxica inferior al umbral de confianza:



Versió: 2002-02-28/13:16:14 (Mitjana **128**)

1 11 The actual thief is on the loose .	149	1 1.11 El ladrón actual anda suelto .
2 21 De Benedetti faces six years in prision in connection with the collapse of bank .	162	2 1.21 De Benedetti se enfrenta a seis años de prisión por el colapso de l banco .
3 31 Americans will have spent close to $800 billion in 1999.	45	3 1.31 Los ciudadanos de EEUU habrán gastado cerca de 800 billones de dólares en 1999 .
4 41 I was embarrassed when I made the mistake .	54	4 1.41 Quedé embarazado cuando cometí el error .
5 51 Everybody looks for success in life .	65	5 1.51 Todos buscamos un suceso en la vida .
6 61 They moved from Chicago to Miami .	323	6 1.61 Se movieron de Chicago a Miami .
7 71 A jury decided that Dilmer was sane when he killed 15 young men .	137	7 1.71 Un jurado decidió que Dilmer estaba sano cuando mató a 15 jóvenes .
8 81 He is an ordinary man .	65	8 1.81 Es un hombre ordinario .
9 91 I do n't have any literature on medecine .	153	9 1.91 No tengo literatura sobre medicina .

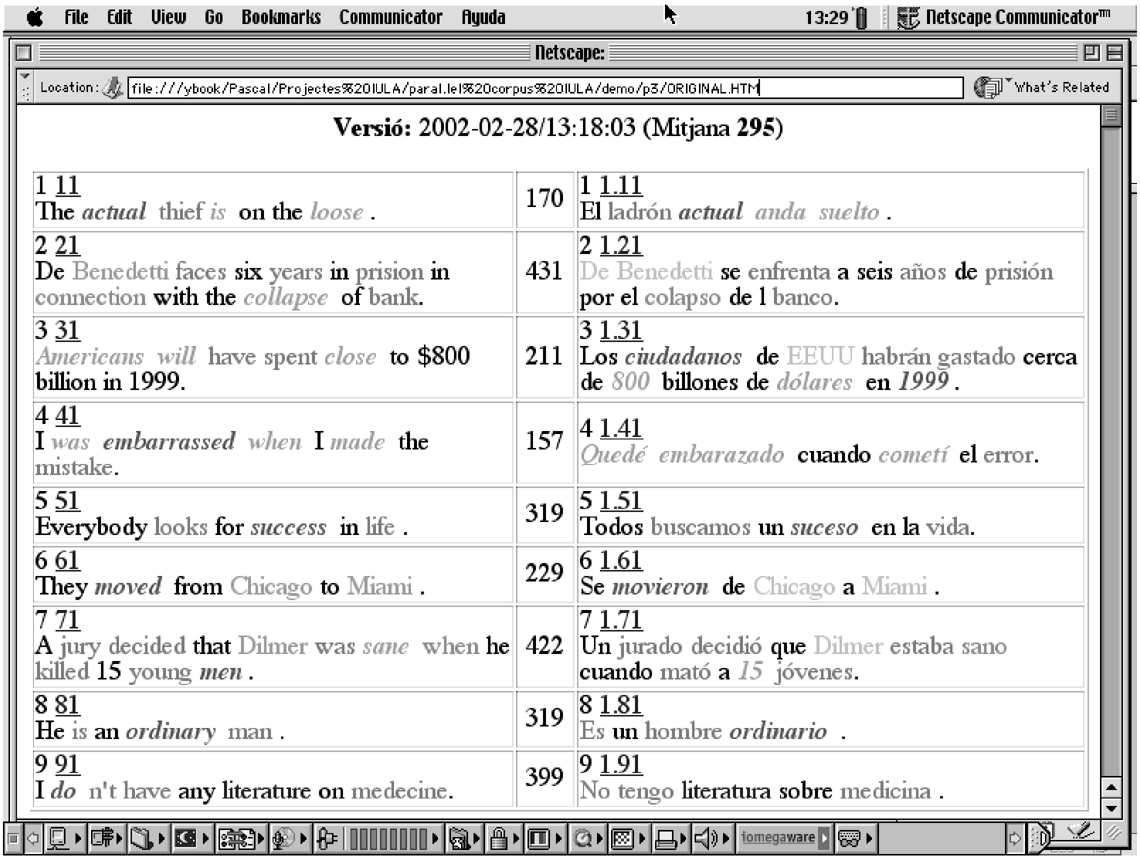
versión 1, con falsos amigos y sin diccionario



Versió: 2002-02-28/13:16:55 (Mitjana **98**)

1 11 The actual thief is on the loose .	61	1 1.11 El verdadero ladrón anda suelto .
2 21 De Benedetti faces six years in prision in connection with the collapse of bank .	158	2 1.21 De Benedetti se enfrenta a seis años de prisión por la quiebra de l banco .
3 31 Americans will have spent close to $800 billion in 1999.	44	3 1.31 Los ciudadanos de EEUU habrán gastado cerca de 800 mil millones de dólares en 1999 .
4 41 I was embarrassed when I made the mistake .	53	4 1.41 Quedé confuso cuando cometí el error .
5 51 Everybody looks for success in life .	64	5 1.51 Todos buscamos un éxito en la vida .
6 61 They moved from Chicago to Miami .	260	6 1.61 Se mudaron de Chicago a Miami .
7 71 A jury decided that Dilmer was sane when he killed 15 young men .	102	7 1.71 Un jurado decidió que Dilmer estaba cuerdo cuando mató a 15 jóvenes .
8 81 He is an ordinary man .	64	8 1.81 Es un hombre corriente .
9 91 I do n't have any literature on medecine .	82	9 1.91 No tengo información escrita sobre medicina .

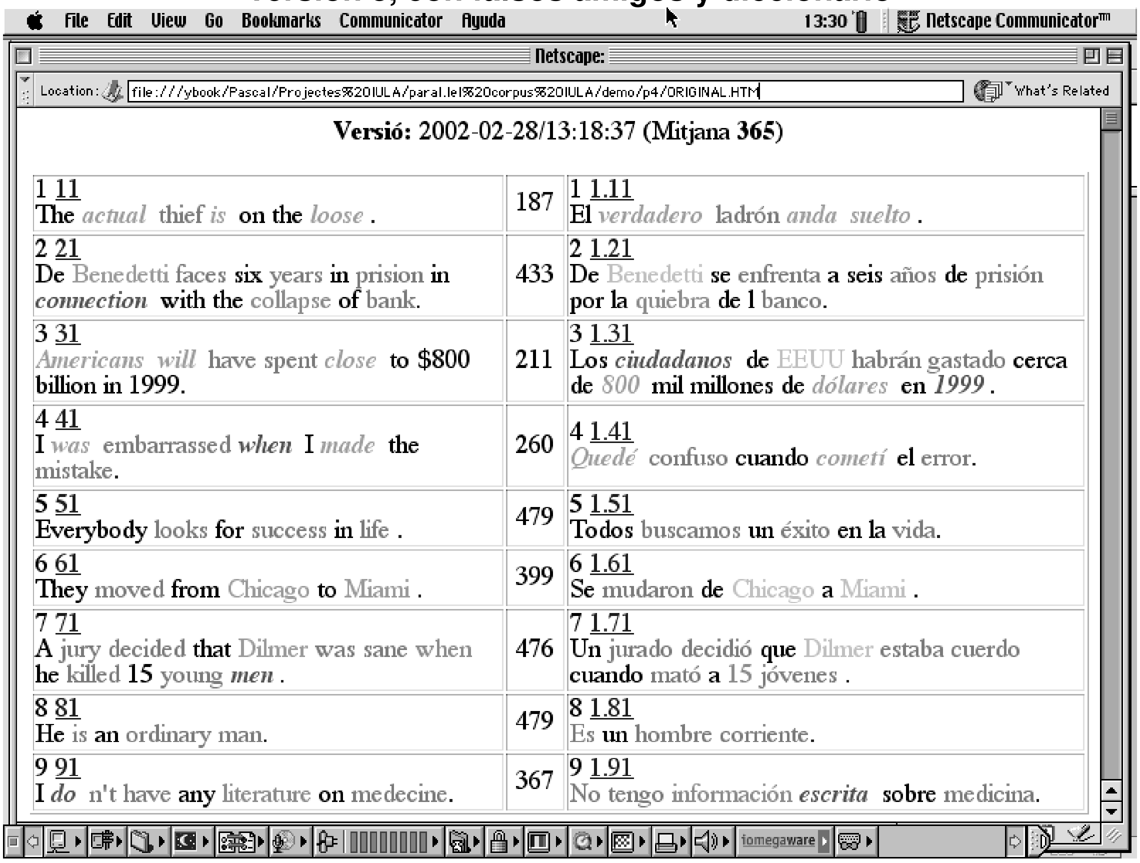
versión 2, sin errores ni diccionario



Versió: 2002-02-28/13:18:03 (Mitjana 295)

1 11 The actual thief is on the loose .	170	1 1.11 El ladrón actual anda suelto .
2 21 De Benedetti faces six years in prision in connection with the collapse of bank.	431	2 1.21 De Benedetti se enfrenta a seis años de prisión por el colapso de l banco.
3 31 Americans will have spent close to $800 billion in 1999.	211	3 1.31 Los ciudadanos de EEUU habrán gastado cerca de 800 billones de dólares en 1999 .
4 41 I was embarrassed when I made the mistake.	157	4 1.41 Quedé embarazado cuando cometí el error.
5 51 Everybody looks for success in life .	319	5 1.51 Todos buscamos un suceso en la vida.
6 61 They moved from Chicago to Miami .	229	6 1.61 Se movieron de Chicago a Miami .
7 71 A jury decided that Dilmer was sane when he killed 15 young men .	422	7 1.71 Un jurado decidió que Dilmer estaba sano cuando mató a 15 jóvenes.
8 81 He is an ordinary man .	319	8 1.81 Es un hombre ordinario .
9 91 I do n't have any literature on medecine.	399	9 1.91 No tengo literatura sobre medicina .

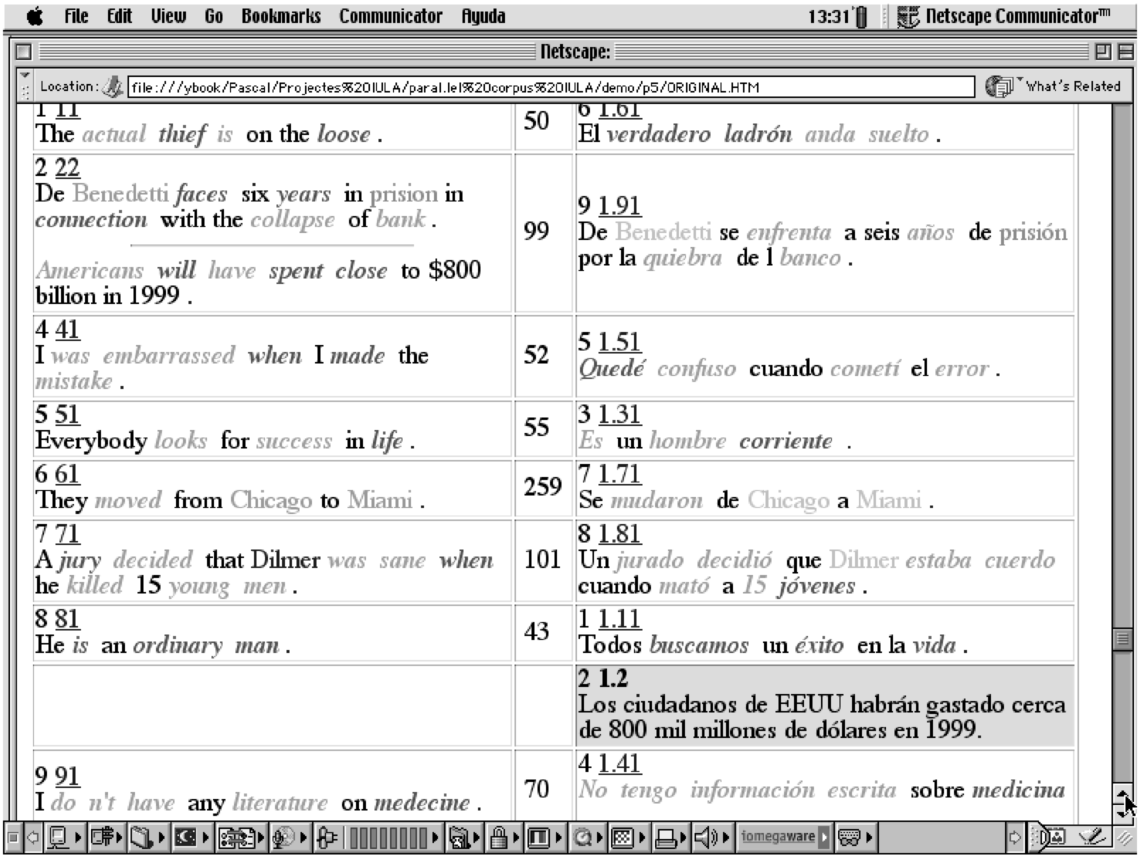
versión 3, con falsos amigos y diccionario



Versió: 2002-02-28/13:18:37 (Mitjana 365)

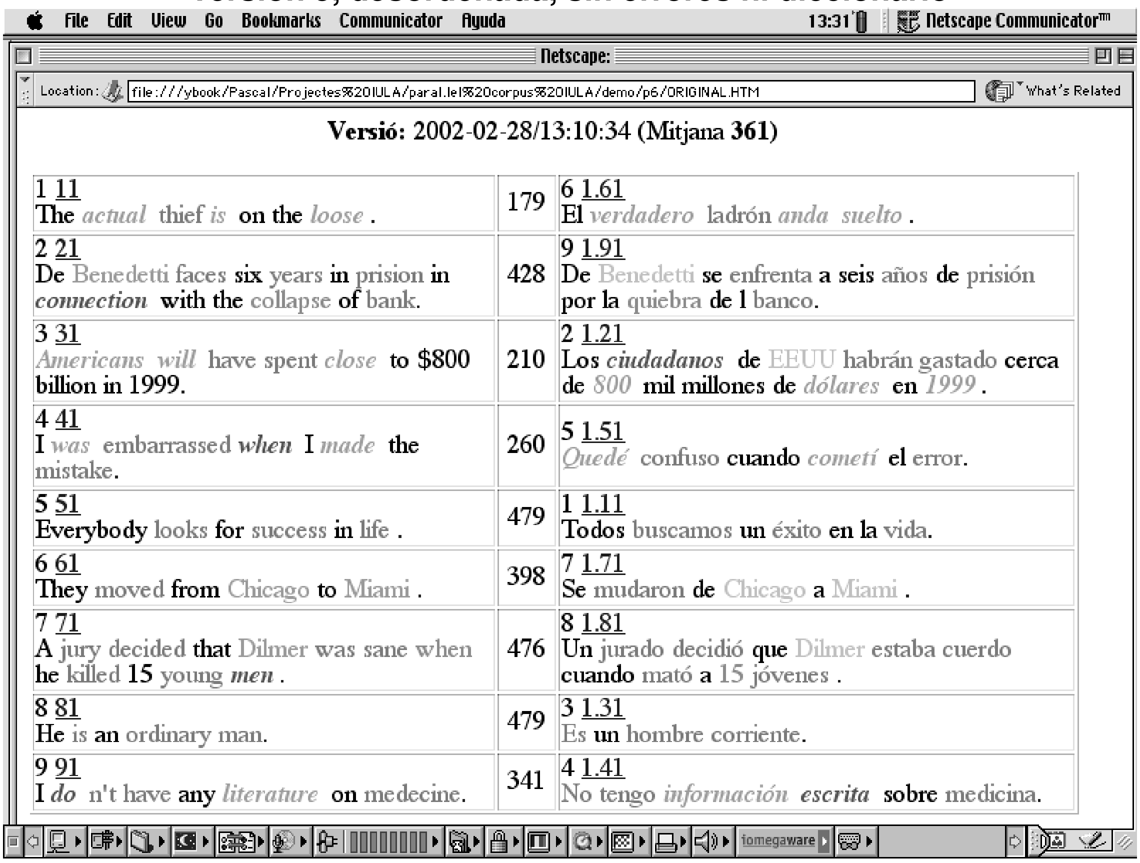
1 11 The actual thief is on the loose .	187	1 1.11 El verdadero ladrón anda suelto .
2 21 De Benedetti faces six years in prision in connection with the collapse of bank.	433	2 1.21 De Benedetti se enfrenta a seis años de prisión por la quiebra de l banco.
3 31 Americans will have spent close to $800 billion in 1999.	211	3 1.31 Los ciudadanos de EEUU habrán gastado cerca de 800 mil millones de dólares en 1999 .
4 41 I was embarrassed when I made the mistake.	260	4 1.41 Quedé confuso cuando cometí el error.
5 51 Everybody looks for success in life .	479	5 1.51 Todos buscamos un éxito en la vida.
6 61 They moved from Chicago to Miami .	399	6 1.61 Se mudaron de Chicago a Miami .
7 71 A jury decided that Dilmer was sane when he killed 15 young men .	476	7 1.71 Un jurado decidió que Dilmer estaba cuerdo cuando mató a 15 jóvenes .
8 81 He is an ordinary man.	479	8 1.81 Es un hombre corriente.
9 91 I do n't have any literature on medecine.	367	9 1.91 No tengo información escrita sobre medicina.

versión 4, con diccionario y sin errores



1 11 The actual thief is on the loose .	50	6 1.61 El verdadero ladrón anda suelto .
2 22 De Benedetti faces six years in prision in connection with the collapse of bank . Americans will have spent close to $800 billion in 1999 .	99	9 1.91 De Benedetti se enfrenta a seis años de prisión por la quiebra de l banco .
4 41 I was embarrassed when I made the mistake .	52	5 1.51 Quedé confuso cuando cometí el error .
5 51 Everybody looks for success in life .	55	3 1.31 Es un hombre corriente .
6 61 They moved from Chicago to Miami .	259	7 1.71 Se mudaron de Chicago a Miami .
7 71 A jury decided that Dilmer was sane when he killed 15 young men .	101	8 1.81 Un jurado decidió que Dilmer estaba cuerdo cuando mató a 15 jóvenes .
8 81 He is an ordinary man .	43	1 1.11 Todos buscamos un éxito en la vida .
		2 1.2 Los ciudadanos de EEUU habrán gastado cerca de 800 mil millones de dólares en 1999.
9 91 I do n't have any literature on medecine .	70	4 1.41 No tengo información escrita sobre medicina

versión 5, desordenada, sin errores ni diccionario



Versió: 2002-02-28/13:10:34 (Mitjana **361**)

1 11 The actual thief is on the loose .	179	6 1.61 El verdadero ladrón anda suelto .
2 21 De Benedetti faces six years in prision in connection with the collapse of bank.	428	9 1.91 De Benedetti se enfrenta a seis años de prisión por la quiebra de l banco.
3 31 Americans will have spent close to $800 billion in 1999.	210	2 1.21 Los ciudadanos de EEUU habrán gastado cerca de 800 mil millones de dólares en 1999 .
4 41 I was embarrassed when I made the mistake.	260	5 1.51 Quedé confuso cuando cometí el error.
5 51 Everybody looks for success in life .	479	1 1.11 Todos buscamos un éxito en la vida.
6 61 They moved from Chicago to Miami .	398	7 1.71 Se mudaron de Chicago a Miami .
7 71 A jury decided that Dilmer was sane when he killed 15 young men .	476	8 1.81 Un jurado decidió que Dilmer estaba cuerdo cuando mató a 15 jóvenes .
8 81 He is an ordinary man.	479	3 1.31 Es un hombre corriente.
9 91 I do n't have any literature on medecine.	341	4 1.41 No tengo información escrita sobre medicina.

versión 6, desordenada, sin errores, con diccionario

Las dos versiones más satisfactorias, la 4 y la 6, prácticamente se confunden: el alineador ha sido capaz de detectar y resolver el desorden en la versión 6 y dar un resultado parecido al de la versión 4 gracias a los diccionarios que hemos generado a partir del procesamiento inicial. La versión 3 da un resultado levemente inferior al de la versión 4, precisamente porque hemos rechazado la mayoría de los falsos amigos gracias al diccionario de exclusiones. En cambio, las versiones 1, 2 y 5 dan resultados muy insatisfactorios, con ligeras diferencias, pero manifestando globalmente que, sin diccionarios, el rendimiento del alineador no llega a niveles útiles.

En resumen, un alineador de base lingüística, en la medida que es mucho más robusto ante los errores que uno de base estadística, permite ser utilizado como validador de traducciones, en cuyo caso mejorará notablemente sus prestaciones a medida que vaya retroalimentando sus diccionarios.

6.- Conclusiones

En esta comunicación, hemos presentado el estado de desarrollo del alineador, aún no terminado pero plenamente operativo y hemos procurado exponer que el funcionamiento del alineador mejora en calidad y en robustez si dispone de diccionarios. Hemos mostrado también que el propio alineador puede contribuir a la constitución de los diccionarios que necesita, especialmente gracias al cálculo del coeficiente de probabilidad biléxica decreciente (PBD), cuyo método de obtención hemos presentado exhaustivamente. Finalmente, hemos aportado nuestra experiencia de utilización del alineador para validar traducciones, papel que desarrolla con buenos resultados siempre que sea retroalimentado con sus propios diccionarios.

En cuanto a las perspectivas de mejora de la herramienta, contemplamos distintas posibilidades. En la medida en que el IULA disponga de lematizador para otras lenguas, intentaremos incorporar dichas lenguas a los pares que el alineador acepta; cuando disponga de herramientas de enriquecimiento lingüístico a otros niveles (analizadores sintácticos, semánticos...) para un determinado par de lenguas, sería deseable que dichas informaciones fueran utilizadas por el alineador. A más corto plazo, hay que pensar en la posibilidad de que los diccionarios del alineador permitan relacionar pares léxicos dispares (una sola palabra de la lengua A con varias de la lengua B o viceversa: terminología, fraseología, locuciones...), así como la formulación de reglas metatácticas que resuelvan diferencias en el número de palabras por razones gramaticales sistemáticas (v.g. formas verbales perifrásticas en la lengua A contra formas simples en la lengua B o activa vs. pasiva). Un reto especialmente atractivo es el de conseguir que el alineador detecte automáticamente los candidatos a pares dispares.

7.- Otras informaciones

Para obtener más información sobre el funcionamiento del alineador, así como ejemplos de textos alineados consúltense las páginas siguientes:

http://terminotica.upf.es/academic/
http://terminotica.upf.es/demo_alinea/
http://terminotica.upf.es/crel/alinea.htm
FAQ: http://www.iula.upf.es/faq/cache/149.html

BIBLIOGRAFÍA

Alonso, A; Martínez, M.; Polo, N.; Puertas, G.; de Yzaguirre, L.; Tebé, C.; Cabré, M. T. (en prensa). "La utilización de corpus paralelos alineados en la docencia de la traducción y de los lenguajes de especialidad", comunicación presentada en la *2nd International Contrastive Linguistics Conference*, Depto. de Filoloxía Inglesa, Facultade de Filoloxía, Universidad de Santiago de Compostela, 25-27 d'octubre del 2001.

Bach, C.; Saurí, R.; Vivaldi, J.; Cabré, M. T. 1997. *El Corpus de l'IULA: descripció*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, (Papers de l'IULA. Sèrie Informes, 17).

Baker, M. 1995. "Corpora in Translation Studies: An Overview and Some Suggestions for the Future Research". *Target* 7(2):223-43.

Brown, P.F.; Lai, J.C.; Mercer, R.L. 1991. "Aligning Sentences in Parallel Corpora". En: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Morriston, NJ.: University of California.

Cabré, M. T. 1999. "Elementos para una teoría de la terminología: hacia un paradigma alternativo". En: *La Terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*, 69-92. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Chang, J.S.; Chen, M.H. 1997. "An Alignment Method for Noisy Parallel Corpora based on Image Processing Techniques". En: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference- of the European Chapter of the Association for Computational Linguistics*. San Francisco, Madrid: UNED.

De Yzaguirre, Ll. 2000a. "Validador de traduccions, una eina de paral·lelització", comunicación presentada en las *Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*. IEC. Barcelona, 4-5 de abril del 2000.

De Yzaguirre, Ll. 2000b. "L'etiquetador PALIC i el desambiguador AMBILIC", comunicación presentada en las *Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*. IEC. Barcelona, 4-5 de abril del 2000.

De Yzaguirre, Ll.; Ribas, M.; Vivaldi, J.; M.T. Cabré. 2000c. "Some Technical Aspects About Aligning Near Languages", comunicación presentada en la LREC-2000, Atenas, 31 mayo-2 junio del 2000. En: *Second International Conference on Language Resources and Evaluation. Proceedings*, M. Gabrilidou **et al**. (eds.), vol I, 545-548. Atenas: National Technical University of Athens Press.

De Yzaguirre, Ll.; Ribas, M.; Vivaldi, J.; Cabré, M. T. 2001. "Alineación automática de traducciones: descripción y usos en los ámbitos de la profesión, de la docencia y de la investigación traductológica", comunicación presentada en los IV Encuentros Alcalaínos de Traducción, Alcalá de Henares, 17-18 de febrero del 2000. En: *Traducción y nuevas tecnologías. Herramientas auxiliares del traductor. Encuentro en* *torno a la traducción 4.*, C. Valero y C. De la Cruz (eds), 391-398. Alcalá de Henares: Universidad de Alcalá.

De Yzaguirre, Ll.; Matamala, A.; Cabré, M.T. (en prensa) "El lematizador PALIC del IULA (UPF)", comunicación presentada en *el XVIII Congreso de AESLA*. Barcelona, 4-6 de mayo del 2000.

De Yzaguirre, Ll.; Matamala, A.; Bach, C.; Castillo, N.; Ustrell, E. (en prensa) "AMBILIC, el desambiguador lingüístico del Corpus del IULA", comunicación presentada en el *XVIII Congreso de AESLA*. Barcelona, 4-6 de mayo del 2000.

Gale, W.A.; Church, K.W. 1991. "A Program for Aligning Sentences in Bilingual Corpora". En: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. Morriston, NJ.: University of California.

Johansson, S.; Oksefjell, S. (eds.) 1998. *Corpora and Cross-Linguistic Research. Theory, Method and Case Studies*. Amsterdam y Atlanta (GA): Rodopi.

Kenny, D. 1998. "Corpora in translation studies". En: *Routledge Encyclopedia of Translation Studies*, M. Baker et al. (eds). London: Routledge.

Melamed, D. 1997. "A Portable Algorithm for Mapping Bitext Correspondence". En: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. San Francisco, Madrid: UNED.

Morel, J.; Torner, S.; Vivaldi, J.; De Yzaguirre, Ll.; Cabré, M.T. 1998. *El corpus de l'IULA: Etiquetaris*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Papers de l'IULA. Sèrie Informes, 18). [Segunda edición, revisada y corregida].

Prado, M. 2001. *Diccionario de falsos amigos Inglés-Español*. Madrid:Gredos.

Vivaldi, J.; De Yzaguirre, Ll.; Solé, X.; Cabré M. T. 1996. *Marcatge estructural i morfosintàctic del corpus tècnic amb l'estàndard SGML.* Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. (Papers de l'IULA. Sèrie Informes, 1).