ALINEACIÓN AUTOMÁTICA DE TRADUCCIONES: DESCRIPCIÓN Y USOS EN LOS ÁMBITOS DE LA PROFESIÓN, DE LA DOCENCIA Y DE LA INVESTIGACIÓN TRADUCTOLÓGICA

Lluís de Yzaguirre, Marta Ribas, Jordi Vivaldi y M.T.Cabré
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra, Barcelona
E-mail: de_yza@upf.es

1. EL ALINEADOR DE TRADUCCIONES - DESCRIPCIÓN

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, 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. El porcentaje de aciertos en el nivel léxico es más bajo que en el de la frase pero las alineaciones léxicas erróneas suelen poner de relieve diferencias estructurales o fraseológicas entre las dos lenguas analizadas, que constituyen un material de estudio valioso.

2. POSIBILIDADES DE USO DEL PROGRAMA

2.1. USO PARA LA DOCENCIA

En el marco de la docencia, la utilización de este programa posibilita la *creación de corpus alineados*, en cualquier ámbito de especialidad, que pueden resultar de interés para situaciones diversas, especialmente:

a. Desde la perspectiva del profesor de traducción, para aproximarse a las especificidades propias del sublenguaje que esté estudiando, localizar dificultades de traducción entre dos lenguas y poder preparar, de esta manera, un material de base real para sus clases.

b. Desde la perspectiva del alumno, para -a partir de la utilización de corpus alineados- sensibilizarse de manera más profunda sobre diferencias léxicas y estructurales entre lenguas distintas en un mismo ámbito de especialidad.

2.2. USO PARA LA TRADUCCIÓN PROFESIONAL

En el terreno profesional, el uso de corpus alineados puede permitir al traductor analizar posibles problemas de traducción y encontrar ideas para su solución. Además, los corpus generados a través de un programa de estas características pueden constituir una ayuda importantísima para la generación de memoria de traducción.

Para más detalles sobre la utilitzación de corpus en la traducción profesional y en la docencia, remitimos a los trabajos de otros autores que también se presentan en estos Encuentros.

2.3. USO PARA LA INVESTIGACIÓN

En otros usos, la generación de memoria de traducción puede ser beneficiosa tanto para la traducción asistida como para mejorar los sistemas de traducción mecánica. Además, los

---

[1] véanse Johanson (1998) y Kenny (1998)

[2] Véanse: Brown (1991), Gale & Church (1991), Chang (1997) y Melamed (1997).

[3] En inglés “*Machine Translation*”.

corpus alineados pueden resultar de interés para el desarrollo de estrategias de detección multilingüe de candidatos a términos. El alineador también puede servir para la validación de traducciones mecánicas.

Por otro lado, en el terreno de la lexicografía bilingüe, la explotación de corpus alineados puede constituir una fuente básica para la elaboración de nuevos productos.

3. FUNCIONAMIENTO GENERAL DEL PROGRAMA

Respecto a trabajos precedentes, la especificidad de nuestra estrategia consiste en la utilización de información lingüística añadida (lemas y etiquetas morfológicas), lo que hace que nuestro programa sea dependiente de la disponibilidad de herramientas de marcaje lingüístico para las dos lenguas que se quieran alinear, pero, en compensación, el programa es mucho más robusto ante paralelizaciones ruidosas.

Esto implica que el alineador no compara únicamente la forma de la lengua A con la forma de la lengua B, sino también - y principalmente- sus lemas y sus etiquetas. Considérense los ejemplos siguientes:

Forma	***Lema***	***Etiqueta***	***Sentido***
Seguíssim	Seguir	VJA1P	Siguiéramos
Siguiéramos	Seguir	VJA1P	Siguiéramos
Con	Con	N5MS	Cono
Con	Con	P	Con
Fumadores	Fumador	JQMP/ N5FP	Fumadoras
Fumadores	Fumador	JQFP/ N5MP	Fumadores
Jueves	Jueu	JQFP	Judías
Jueves	Jueves	N5M6	Jueves

Así, pues, para establecer un par léxico se da prioridad a la coincidencia de lemas, después a la coincidencia de etiquetas, y finalmente a la coincidencia ortográfica.

El programa trabaja en 3 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 los modelos clásicos de cálculo de longitud de frase, de número de palabras y de posición en el documento.

El punto de partida es la lectura de dos ficheros: por un lado, el texto original (escrito en la lengua A) y, por otro, el texto correspondiente a su traducción (escrito en la lengua B). Se trata de dos ficheros cuyo contenido ha sido enriquecido previamente con marcaje estructural y con marcaje lingüístico.

El marcaje estructural explicita la fronteras sintácticas, lo cual evita posibles ambigüedades sintácticas, como por ejemplo la utilitzación del signo "." (punto) en funciones tan distintas como indicar el final de una frase, delimitar una abreviatura o separar, en el caso de los números, los millares o los decimales.

En cuanto al marcaje lingüístico, éste se realiza en tres etapas: a) el pre-proceso; b) la lematización, y c) la desambiguación.

El hecho de que el alineador use la información lingüística añadida implica que no sólo compara la forma de la lengua A con la forma de la lengua B, sino también – y principalmente- sus lemas y sus etiquetas. Para establecer un par léxico, se da prioridad primero a la coincidencia de lemas, después a la coincidencia de etiquetas y, finalmente, a la coincidencia ortográfica.

[4] Véanse De Yzaguirre (2000a, 2000b y 2000c), Morel (1998) y Vivaldi (1996).

El resultado de este análisis ofrece tres posibilidades:

- Obtener una versión hipertextual de la alineación de estos dos textos.
- Obtener un documento en formato SGML que especifica la vinculación entre los párrafos de una lengua a otra.
- Crear un listado de pares léxicos (equivalencias de una lengua a otra).

El programa ofrece al usuario la posibilidad de revisar, posteriormente, este listado de pares léxicos, eliminar los que crea conveniente y dejar únicamente los que son pertinentes según el ámbito de especialidad en que se trabaja.

El listado de pares léxicos validados sirve para realimentar el alineador, lo cual permitirá llegar a unos resultados posteriores más satisfactorios, tanto en el mismo texto como en otros del mismo ámbito de especialidad.

4. LENGUAS DE TRABAJO

El programa desarrollado en el IULA trabaja sobre dos lenguas: el catalán y el castellana, ya que son las lenguas de las que se dispone actualmente de marcaje lingüístico. Ello no impide, sin embargo, que pueda ser aplicado para textos de otras lenguas, siempre que previamente hayan sido tratados de la misma forma (esto es, con etiquetarios morfológicos similares).

El programa no dispone de etiquetario sintáctico; por el momento únicamente funciona con lenguas que presentan estructuras sintácticas próximas (como es el caso del castellano y el catalán). Cuando se disponga de este marcaje sintáctico, será posible elaborar una versión del programa que lo tenga en cuenta y, por lo tanto, evaluar si aporta mejoras en alinear también otras lenguas sintácticamente diferenciadas, como es el caso del castellano-inglés.

5. MUESTRAS

A continuación presentamos, a modo de ejemplo, una muestra del tipo de información que

aporta la utilització del alineador. Se trata de un pequeño fragmento de un documento jurídico, en concreto, una sentencia del Tribunal Superior de Justícia de Catalunya en su versión catalana y castellana. La muestra presenta, en la columna de la izquierda, la versión catalana del documento; en la columna de la derecha, el original castellano y, en la columna central, la "supuesta" retro-versión literal de la traducción teniendo en cuenta la estructura del documento original.

12 **1.1.9.2.1.3+**		11 **1.1.9.2.1.2**
A ls efectes que prevé l ' art. 1798 de la Llei d ' enjudiciament civil, feia constar que aquest recurs s ' interposa abans d ' haver transcorregut el termini assenyalat des de l descobriment de l nou document que acompanyava de data 6 d'abril de 1992 . < tree="1.1.9.2.1.4" Després d ' al·legar els requisits processals de l recurs, constituir el dipòsit necessari per recórrer i establir la competència d ' aquesta Sala per tramitar l ' esmentat recurs, al·legant els motius en què es basa aquest recurs i els fonaments de dret que va estimar pertinents, acabava demanant que un cop tramitat aquest recurs d'acord amb el dret, es dictés sentència que estimés aquest recurs i rescindís	A los efectos **que prevé** el art. 1798 de la Ley de **e**njuiciamiento **c**ivil, hacía constar que **este** recurso se interpone antes de **haber** trascurrido el plazo señalado desde el descubrimiento del nuevo documento que acompañaba de fecha 6 de abril de 1992. **Después de** alegar los requisitos procesales del recurso, constituir el depósito necesario para recurrir y establecer la competencia de esta Sala para **tramitar el citado** recurso, alegando los motivos en que se basa **este recurso** y los fundamentos de derecho que estimó pertinentes**, acababa pidiendo** que **una vez** tramitado este recurso **de acuerdo con el** derecho, se dictara sentencia **que**	A los efectos **prevenidos** en el art. 1798 de la Ley de **E**njuiciamiento **C**ivil , hacía constar que **el presente** recuso se interpone antes de transcurrido el plazo señalado desde el descubrimiento de l nuevo documento que acompañaba de fecha 6 de abril de 1992 **y tras** alegar los requisitos procesales de l recurso, constituir el depósito necesario para recurrir y establecer la competencia de esta Sala para **conocer de dicho** recurso alegando los motivos en que se basa **el mismo** y los fundamentos de derecho que estimó pertinentes **terminaba suplicando** que tramitado este recurso **con arreglo a** derecho, se dictara sentencia **dando lugar a l mismo y rescindiendo en todo** la sentencia

[5] Esta muestra forma parte del corpus de la futura tesis doctoral de Marta Ribas, titulada *Perspectiva traductològica de l'alineació de textos jurídics paral·lels català-castellà*, que pretende llevar a cabo un análisis de la diferencias discursivas de las sentencias jurídicas del Tribunal Superior de Justícia de Catalunya, en catalán y en castellano. Dicho análisis se propone detectar las principales diferencias en las especificidades de cada sublenguaje que constituyen ruído en el proceso de alineación mecánica, en beneficio de este proceso y, en general, del tratamiento automático del lenguaje jurídico en este par de lenguas.

totalment la sentència que s ' ha impugnat.	**estimase este recurso y rescindiese totalmente** la sentencia **que se ha impugnado**.	**impugnada**.

6. OTRAS INFORMACIONES

Para obtener otras muestras sobre cómo trabaja el alineador, véase: http://traductica.upf.es/alinea/frm2.htm

Una descripción técnica del producto ha sido aceptada en ELRA 2000 (mayo/junio).

7. BIBLIOGRAFÍA

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

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

Chang, J.S.; Chen, M.H. "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* (Madrid, UNED). San Francisco, 1997.

De Yzaguirre, L. , A. Matamala y T. Cabré (2000a): “El lematizador "PALIC" del IULA (UPF)”, comunicación aceptada para el XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

De Yzaguirre, L., A. Matamala, C. Bach, N. Castillo y E. Ustrell (2000b): “AMBILIC, el desambiguador lingüístico del corpus del IULA (UPF)”, comunicación aceptada para el

[6] Véase De Yzaguirre (2000d).

XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

De Yzaguirre, L., S. Torner y A. Matamala (2000c) “El tratamiento automático de las ambigüedades segmentales del castellano”, comunicación aceptada para el XVII Congreso AESLA, panel sobre Lingüística de Corpus y Computacional.

De Yzaguirre, L., M. Ribas, J. Vivaldi y T. Cabré (2000d) “Some technical aspects about aligning near languages”, póster aceptado ELRA 2000 (mayo/junio).

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

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

Kenny, D. “Corpora in translation studies”. En: Baker, M. (ed.) y Malmkjaer, K. (asist.). 1998. *Routledge Encyclopedia of Translation Studies*. London: Routledge.

Melamed, D. "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* (Madrid, UNED). San Francisco, 1997.

Morel, J.; Torner, S.; Vivaldi, J.; De Yzaguirre, Ll.; Cabré, M.T. 1998. *El corpus de l’IULA: Etiquetaris*. Barcelona: Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada. Papers de l’IULA. Sèrie Informes, 18, [Segona edició, revisada i corregida].

Vivaldi, J.; De Yzaguirre, Ll.; Solé, X.; Cabré, M.T. 1996. *Marcatge estructural i morfosintàctic del corpus tècnic amb l’estàndard SGML*. Barcelona: Universitat Pompeu

Fabra. Institut Universitari de Lingüística Aplicada. Papers de l’IULA. Serie Informes, 1.