SEGMENTACIÓN DE CORPUS PARALELOS PARA MEMORIAS DE TRADUCCIÓN

Joseba K. Abaitua Odriozola
Universidad de Deusto
abaitua@fil.deusto.es

Arantza Casillas Rubio
Universidad de Alcalá de Henares
arantza@aut.alcala.es

Raquel Martínez Unanue
Universidad Complutense de Madrid
raquel@eucmos.sim.ucm.es

Palabras clave:

Unidad de traducción, traducción automática, memorias de traducción, análisis de corpus, alineamiento de corpus paralelos, traductología, tipología textual, lexicografía computacional, documentación estructurada, SGML, TEI.

Resumen

Esta comunicación plantea las ventajas de tratar un corpus de textos paralelos bilingües (el bitexto) mediante procedimientos que reconocen y etiquetan segmentos variables de equivalencias de traducción. Estos segmentos constituyen unidades de traducción sensibles al contexto en el que se reconocen, es decir, al tipo de documento, a la sección o división interna del documento y al registro lingüístico (lengua común, lenguajes de especialidad, etc.). Esta forma de tratar el corpus supone un avance sobre otros métodos conocidos basados en palabras y oraciones.

1 Introducción

La comunicación presenta la metodología aplicada a la explotación de un corpus[1] bilingüe de textos administrativos o *bitexto*, en la terminología de Harris 1988:8-9, como fuente de datos para la creación de entornos de edición y traducción en el proyecto LEGEBIDUNA[2]. El corpus se ha tratado por medios atomáticos que introducen etiquetas descriptivas cuyo principal cometido es identificar en las dos versiones lo que

denominamos *unidades de traducción variables*. Mediante algoritmos de alineamiento se obtienen listas de pares de equivalencias. Además, a partir del texto etiquetado se generan *definiciones de tipo de documentos* (DTDs de SGML, *Standard Generalized Markup Language*, ISO 8879:1986), que equivalen a gramáticas de estados finitos capaces de reproducir la estructura de los textos. La presente comunicación discute las ventajas de segmentar el corpus en unidades de traducción en lugar de en palabras u oraciones.

2 Estado actual del proyecto *LEGEBIDUNA*

Se están siguiendo en la actualidad cuatro líneas complementarias de actuación:

* **Creación de un corpus**. El corpus está compuesto por los boletines oficiales de tres administraciones: de las Diputaciones de Álava (BOA 1990-92) y Bizkaia (BOB 1989-95) y del Gobierno Vasco (BOPV 1995). Esto hace un corpus bastante considerable, de aproximadamente 7 millones de palabras en cada lengua (130 Mb). No tenemos previsto, de momento, ampliar más el corpus antes de tratar convenientemente el que ya disponemos.

* Etiquetado del corpus. Nuestro esfuerzo ahora se centra en el tratamiento de los formatos y en la conversión de los textos a versiones adaptadas de SGML, en la línea de las propuestas de TEI (*Text Encoding Iniciative*, McKelvie y Thompson 1994) y MULTEXT (*Multilingual Text Tools and Corpora*, Ide y Véronis 1994). A partir de un análisis detallado de las distintas clases de documentos en una parte del corpus (Órdenes Forales del BOB, que constituye aproximadamente una quinta parte del corpus, 1,5 millones de palabras), se ha realizado un inventario de etiquetas descriptivas. Se están definido distintos modelos de DTDs, para cada tipo de documento, aproximadamente 900 documentos de 22 tipos distintos, con una muestra representativa de 40 textos de cada uno). Se han

implementado rutinas en lenguaje AWK (próximamente en PERL) que reconocen y segmentan automáticamente un 41% del corpus objeto de estudio, con unos resultados de cobertura (*recall*) y precisión excepcionales.

* **Creación de memorias de traducción.** Los textos paralelos se someten a un cotejo automático que tiene como objeto la identificación de equivalencias en las dos versiones mediante la aplicación de diversos algoritmos de alineamiento (basados en Gale y Church 1991, Brown y otros 1991). Una vez reconocidas, estas unidades se catalogan formando memorias de traducción (Satò y Nagao 1990, Sumita y H.Iida. 1991.). La particularidad del alineamiento es que se realiza sobre unidades de traducción, en lugar de oraciones enteras, como hasta ahora realizan la mayoría de los sistemas de alineamiento (Gale y Church 1991, Winarske y otros 1992, Brown y otros 1990, 1991).
* **Diseño de un editor de textos administrativos bilingües.** A partir de los textos etiquetados en SGML es posible deducir gramáticas en la forma de DTDs. Se está diseñando un editor que muestra las posibilidades de simultanear en un mismo proceso la edición de las dos versiones (castellano y euskara) de la documentación analizada. El editor aprovecha la información procedente de la DTDs para proponer plantillas con la estructura inicial en las dos versiones del documento. Estas plantillas se rellenan automáticamente en un 30%, con la información disponible en la memoria de traducción. El 70% restante se resuelve de dos maneras. Un 45% lo componen unidades terminológicas, la mayoría propias del lenguaje administrativo, cuya resolución también es dada por la máquina (diccionario de términos), aunque su distribución en el texto se decide en modo interactivo, entre el usuario y las opciones que ofrece la DTD del editor. El 25% restante son unidades lexicológicas, por el momento no recogidas en la memoria de traducción, por lo que deben ser resueltas íntegramente por el usuario.

La originalidad del proyecto LEGEBIDUNA reside en haber centrado la investigación primero en el estudio estructural y tipológico de los textos, tal y como recomiendan los traductólogos, Snell-Hornby 1988:30-33, Sager 1993:170, entre otros; segundo, en la identificación de los distintos registros lingüísticos que coaparecen en un mismo texto; y tercero, en llevar a cabo la segmentación en unidades de traducción sensibles a la tipología, estructura y registro de los textos. Este enfoque supone una aproximación al tratamiento computacional del lenguaje natural desde el polo opuesto al tradicional que parte de unidades lexicológicas. Por otro lado, la adopción de unidades de traducción en lugar de oraciones para alimentar las memorias de traducción supone una considerable mejora sobre los métodos hasta ahora publicados.

La fase inicial del proyecto (elección de la metodología y selección y creación del corpus) comenzó en 1993. Desde finales de 1995, el proyecto se ha integrado en un marco de actuación más amplio (Proyecto ITEM: **Recuperación de información textual en un entorno multilingüe con técnicas de lenguaje natural**, TIC96-1243-C03-03) que contempla la integración de un conjunto de herramientas en una plataforma común para la gestión y extracción de información en bases documentales. Nuestro corpus se ha adoptado como campo de pruebas de este proyecto.

3 Unidades lexicológicas, terminológicas y traductológicas

Nuestro corpus está compuesto por textos administrativos y por ello el registro lingüístico preponderante viene marcado por las características del lenguaje de esta especialidad, o *sublenguaje* (MAP 1991, Prieto de Pedro 1989, Calvo Ramos 1980). Sin embargo, además de unidades del lenguaje administrativo, en los textos aparece terminología de una considerable variedad de especialidades (arquitectura, medicina, botánica, etc.), además de otras unidades no catalogables y que consideramos pertenecen a la lengua común.

Melby 1995:59–69 recomienda tratar de manera diferenciada las unidades de los lenguajes de especialidad (**unidades terminológicas**, UTTs, o términos) y las unidades de la lengua común (**unidades lexicológicas**, UTLs, o palabras). Nuestra hipótesis de trabajo se fundamenta en esta sugerencia de Melby, ya que pretendemos reconocer y separar los fragmentos del corpus que pertenecen al lenguaje administrativo de los que no. Según nuestras estimaciones, la mayor parte del corpus (en torno al 60%) se compone exclusivamente de unidades del lenguaje administrativo, un 65% de las cuales (41% del total del corpus) se reconoce mediante sencillas técnicas de cotejo de patrones (*pattern matching*). El 35% (19% del total) restante requiere un tratamiento más pormenorizado, que integra estadísticas de coaparición de palabras, con confrontación en glosarios terminológicos y revisión humana. Otro 15% del corpus está formado por terminología de disciplinas diversas y el 25% restante lo forman unidades de la lengua común.

Los términos de especialidad tienen la ventaja de que son estáticos (en el sentido de que su significado se mantiene fijo) y, una vez reconocidos e identificados, se traducen inequívocamente. Las palabras de la lengua común son fundamentalmente ambiguas y dinámicas, su uso e interpretación está en permanente cambio y son muy difíciles de tratar en un entorno de traducción (Melby 1995:45–50). Nuestro objetivo es resolver adecuadamente ese 60% del corpus que pertenece al lenguaje administrativo y que constituye la base de nuestras memorias de traducción. Téngase en cuenta que por estos medios se obtienen ratios de acierto óptimos (cercanos al 100%), ya que se generan traducciones previamente contrastadas. La aparición en los nuevos textos de ese 15% aproximado de unidades terminológicas ajenas al lenguaje administrativo no se puede preveer a priori, por lo que nuestro esfuerzo se limita a la detección y creación de glosarios parciales y multidisciplinares de terminología especializada que se integran en el editor. El 25% restante, compuesto por unidades de la lengua común, será tratado en el marco del proyecto ITEM mediante otras técnicas de análisis lexicológico (véase

Ezeiza y otros 1996). En el caso concreto de nuestro editor, el tratamiento de estas unidades lexicológicas se deja en manos del traductor humano.

Consideramos que el diseño de un sistema sensible a la unidad de traducción, que reconoce los casos y los contextos en los que la traducción se puede resolver automáticamente con un 100% de garantía (lenguaje administrativo), supone una contribución metodológica de la máxima importancia. La utilización de memorias de unidades de traducción en un entorno de edición basado en SGML, con ayudas a la redacción, diccionarios terminológicos y otras herramientas lingüísitcas más genéricas constituye en nuestra opinión un enfoque óptimo nunca hasta ahora ensayado.

4 Unidades de traducción variables

Pese a la relevancia del concepto de unidad de traducción en los estudios traductológicos, apenas se ha abordado esta cuestión en el campo de la traducción automática. Bennett 1994 es la primera referencia importante que intenta trasladar las conclusiones de la traducción humana a la realizada por medios mecánicos, pero limita su indagación a la técnica de transferecia, equiparando la unidad lexicologica con la unidad de transferencia. Omitie de esta manera la posibilidad de encontrar otras correspondencias en métodos como, por ejemplo, las memorias de traducción o la traducción basada en analogias.

En nuestro proyecto realizamos en primer lugar una diferenciación entre unidad lexicológica y terminológica. Durante años en los estudios de traducción, la unidad de traducción se ha equiparado con la unidad lexicológica (Vinay y Darbelnet 1958, Vázquez-Ayora 1977, etc.). Nosotros, sin embargo, siguiendo a Melby 1995, hemos decidido distinguir entre unidades lexicológicas y terminológicas, siendo la principal diferencia en que las primeras pertenecen a un área de especialidad determinada y las segundas a la lengua común. La principal ventaja, como se ha señalado,

es que la resolución de la terminología es inequívoca. Pero esta división no nos ha parecido suficiente. Los textos que manejamos dan la razón a quienes en traductología proponen considerar unidades más amplias, que trascienden el marco lingüístico y que configuran lo que Hatim y Mason 1990:105-113 denominan unidades semiológicas y pragmáticas.

Tanto Hatim y Mason 1990, como también Sager 1993 o Toury 1995, hablan de la conveniencia en algunos casos de reconocer unidades de traducción que abarcan el texto completo. En nuestro caso, hemos identificado en la documentación que estudiamos lo que denominamos **unidades formulaicas** (UTFs) y que suponen casi una tercera parte del corpus. Estas unidades pueden corresponderse con secciones completas de un documento y se componen con frecuencia de secuencias multiclausales. Estos son algunos ejemplos:

<seg type=utf#1_es>Mediante la Orden Foral de referencia se ha dispuesto lo siguiente:</seg>

<seg type=utf#1_eu>Aipameneko Foru Aginduaren bidez honako hau xedatu da:</seg>

<seg type=utf#2_es>Contra dicha Orden Foral, que agota la vía administrativa, podrá interponerse recurso de reposición ante (el Diputado Foral de Urbanismo), como trámite previo a la impugnación ante la Jurisdicción Contencioso-Administrativa, en el plazo de un mes, contado desde el día siguiente a esta notificación, sin perjuicio de la utilización de otros medios de defensa que estima oportunos.</seg>

<seg type=utf#2_eu>Administrazio bidea agortzen duen aipaturiko Foru Aginduaren aurka, jakinerazpen honen biharamunetik zenbatu beharreko hilabeteko epearen barruan, birjarpenezko errekurtsoa jarri ahal izango da (Hirigintzako Foru Diputatuaren) aurrean, Administrazioarekiko Auzien Jurisdikzio aurrean egiteko aurkapenaren aldez aurretiko tramite gisa, komeniesten diren beste defentsabideak erabil daitezkeelako kalterik gabe.</seg>

<seg type=utf#3_es>Durante el referido plazo el expediente <nº expediente> quedará de manifiesto para su examen en las dependencias situadas en <lugar>.</seg>

<seg type=utf#3_eu>Adierazi den epearen barruan, <nº expediente> espedientea <lugar> egongo da ageriko, azter dadin.</seg>

Las unidades formulaicas configuran en la documentación administrativa el esqueleto que vertebra los textos [<encabezado>,...,<interposición>, <pie>]. Su reconocimiento es una tarea relativamente trivial, sin embargo, como indicaremos más adelante, suponen no solo una descarga importante para el algoritmo de alineamiento, ya que su identificación es inmediata, sino que además sirven de "puntos ancla" para el alineamieto de otros segmentos.

La identificación de las unidades terminológicas (UTTs) es más compleja y constituye el núcleo central de nuestra investigación en estos momentos. Este es un ejemplo típico de UTT, que denominamos "término propio" y cuya aparición en el corpus (10%) está resuelta en su totalidad:

<seg type=utt#1_es>**Modificación de las Normas Subsidiarias**</seg>
<seg type=utt#1_eu>**Sorospidezko Arauen aladarazpena**</seg>

El siguiente ejemplo ilustra la complejidad de la traducción si se llevara a cabo de manera indiscriminada mediante métodos como el de transferencia:

"**para la inclusión del** <seg type=utt#1_es>***Corredor del Cadagua***</seg>**, en este** <seg type=utt#2_es>**término municipal**</seg>"

"<seg type=utt2#_eu>**udalerritik**</seg> **igarotean** <seg type=utt#1_eu>***Cadaguako pasabidea***</seg> **barru sartzeko**"

Se observa que la versión en euskara realizada por un traductor humano no es en absoluto literal, como demuestra la retraducción al castellano *"al pasar por el término municipal el corredor del Cadagua se incluya [en él]"*. Esto da idea de la dificultad que entrañaría el tratamiento de estos textos mediante técnicas tradicionales basadas únicamente e indiscriminadamente en, por ejemplo, la traducción por transferencia. Creemos que es mejor construir un sistema que sea capaz de modular el método de resolución según la naturaleza de las unidades encontradas.

En el caso del ejemplo señalado, el sistema puede resolver sin dificultad la traducción de las unidades terminológicas. Sin embargo, se abstiene de tratar de resolver casos de trasformación estructural tan inesperados e imprevisibles como los que se producen con las unidades lexicológicas. En estos casos, consideramos que es mejor dejar la resolución en manos del redactor-traductor humano. A lo más que aspiramos es a ofrecer el corpus como una fuente de modelos de traducciones, en los que el fragmento en su conjunto se ofrece al traductor como resultado de una búsqueda a partir de, por ejemplo, las palabras "inclusión" y "término municipal". El hecho de que la búsqueda se realice únicamente sobre el 60% del corpus, el que no está reconocido como formulaico, además de la posibilidad de disponer de fragmentos indizados por unidades terminológicas, como es el caso de "término municipal", permite agilizar considerablemente la precisión y rapidez de la búsqueda.

5 El alineamiento

Las principales propuestas de alineamiento de corpus paralelo se basan en métodos estadísticos similares a los aplicados por Gale y Chuch 1991, por un lado, y Brown y otros 1991, por otro. La propuesta de Gale y Church se basa en un modelo estadístico fundamentado en la longitud, medida en caracteres, de las oraciones de uno y otro corpus y no utiliza conocimiento lingüístico. Dicho modelo utiliza la idea de que las oraciones largas en una lengua tienden a ser traducidas a oraciones largas en la otra lengua, y que las oraciones cortas tienden a ser traducidas a oraciones cortas. El algoritmo que proponen los autores asigna a cada par de oraciones candidatas al emparejamiento una razón probabilística. Esta razón se basa en el ratio de las longitudes de las dos oraciones en caracteres, en una y otra lengua, y en la varianza de ese ratio. Mediante programación dinámica se determina, a partir de esas razones probabilísticas, el alineamiento con mayor probabilidad.

El alineamiento de oraciones no es un problema trivial ya que se puede

dar el caso de que el emparejamiento no sea de una a una, sino que una oración en una lengua puede ser traducida a ninguna o a dos o más oraciones en la otra, lo que complica severamente el alineamiento. Otra estrategia de alineamiento es la planteada por Brown y otros. que está basada en la utilización de "puntos ancla" para ayudar al alineamiento, además de servirse de la estrategia de Gale y Church. La longitud de las oraciones no se mide en caracteres sino en palabras.

En nuestro caso, las etiquetas incorporadas a los textos sirven de puntos ancla que delimitan e identifican, en ambos corpus, una parte considerable de las unidades variables de traducción, todas las del tipo UTF, y gran parte de las del tipo UTT. El resultado del alineamiento consiste en añadir a dichas etiquetas el atributo que establece la correspondencia entre segmentos en ambas lenguas. Aún a pesar de tener que salvar la complejidad añadida, consideramos que las unidades de traducción variables representan un modelo considerablemente más adecuado que la oración. La memoria de traducción que se obteniene mediante este procedimiento es más eficaz y económica que las basadas en segmentos del tamaño de oraciones. La utilización de SGML como sistema de etiquetado permite, además, que se utilicen las DTDs como gramáticas que dan cuenta de la estructura y distribución de una parte importante de las unidades alineadas.

6 Algunos datos estadísticos

Se ha realizado un muestreo sobre un total de 65.303 palabras, que se corresponde con el tipo de documento Orden Foral del año 1995 del BOB (la muestra es pequeña en relación con todo el corpus, pero hemos adoptado criterios de representatividad por lo que creemos que los resultados son fiables y extrapolables al conjunto de Órdenes Forales en el BOB).

Total de palabras del muestreo:	65.303	100%
Palabras reconocidas en UTF "encabezamiento"	11.760	18%
Otras UTF	7.955	12,24%
Total palabras en UTFs	19.715	30,24%
UTT (términos propios)	7.063	10,81%
Total reconocido mediante pattern matching	26.818	41,06%

Para el reconocimiento de estas unidades se han utilizado técnicas de cotejo de patrones (*pattern matching*) y heurísticas, con un conocimiento lingüístico mínimo. Han resultado ser técnicas eficientes con un coste computacional aceptable. Los valores de precisión y cobertura, referidos al conjunto de UTTs y UTFs, son para el castellano:

	Precisión	Cobertura
UTT (términos propios)	98%	93%
UTF	100%	100%

Estos valores resultan sorprendentemente altos, pero hay que tener en cuenta que las unidades de traducción identificadas como unidades formulaicas responden a patrones y reglas de fácil detección. Mediante reglas heurísticas, con un conocimiento lingüístico muy sencillo se ha

reconocido un 10,81% del corpus, lo que denominamos términos propios (nombre de instituciones, cargos públicos, titularidades, disposiciones, órdenes, regulaciones, etc.). Para el reconociemiento de las UTTs restantes se están utilizando técnicas estadísticas aplicadas a la búsqueda de coaparición de palabras, que se cotejan con un glosario de 13.000 términos del lenguaje administrativo y sobre el que posteriormente se realizan revisiones manuales. Es un trabajo de reconocimiento de unidades terminológicas simples y compuestas, diferenciándolas de las palabras sin especialización referencial. De acuerdo con el muestreo realizado sobre las 38.485 palabras sin reconocer (aproximadamente el 60% del total), una tercera parte (20% del total) se corresponde con terminología administrativa; una cuarta parte (un 15% del total), es terminología de otras disciplinas, y el resto (solo un 25% del total) está formado por unidades de la lengua común.

Estas cifras confirman que la metodología adoptada permitirá automatizar la composición y traducción de documentación administrativa en tres cuartas partes del proceso con unos índices de precisión elevadísimos. El diseño de un sistema sensible a la unidad de traducción, que reconoce los casos y los contextos en los que la traducción se puede resolver automáticamente con un 100% de garantía (lenguaje administrativo), junto a la utilización de memorias de unidades de traducción en un entorno de edición basado en SGML, con ayudas a la redacción, diccionarios terminológicos y otras herramientas lingüísitcas más genéricas constituyen en nuestra opinión un enfoque óptimo para automatizar la traducción de documentación especializada.

---

[1]Los autores quieren expresar su gratitud al Departamento de Presidencia de la Diputación de Bizkaia, al Departamento de Diputado General de la Diputación de Álava y al Servicio de Publicación del Boletín Oficial del Gobierno Vasco por la cesión de los textos que componen el corpus.

[2]Más información sobre el proyecto LEGEBIDUNA, con muestras del corpus etiquetado, y sobre unidades de traducción en la URL: <http://www.deusto.es/~abaitua/konzeptu/lege2dun.htm>

Citas

P. Bennett. 1994. "Translation Units in Human and Machine", *Babel* 40 12–20.

P. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer y P.S. Roossin. 1990. "A Statistical Approach to Machine Translation", *Computational Linguistics* 16:79–85.

P. Brown, J. Lai y R.L. Mercer. 1991. "Aligning sentences in Parallel Corpora", *Proceedings of the Association for Computational Linguistics*, 169–176. Berkeley.

L. Calvo Ramos. 1980. *Introducción al estudio del lenguaje administrativo* Gredos, Madrid.

N. Ezeiza, I. Aldezabal, R. Urizar, I. Alegría, y I. Aduriz I. 1996. "Del analizador morfológico al etiquetador/lematizador: Unidades léxicas complejas y desambiguación", *Procesamiento del Lenguaje Natural* 19:90–100.

W. Gale y K. Church. 1991. "A Program for Aligning Sentences in Bilingual Corpora", *Proceedings of the Association for Computational Linguistics*, 177–184. Berkeley.

B. Harris. 1988. "Bi-text: A New Concept in Translation Theory", *Language Monthly*, 54:8–10.

B. Hatim y I. Mason. 1990. *Discourse and the Translator*. Longman.

N. Ide y J. Véronis. 1994. "MULTEXT (Multilingual Text Tools and Corpora)", *Proceedings of the International Wordshop on Sharable Natural Language Resources*, 90–96.

IVAP (Instituto Vasco de Administración Pública. 1994. *Hizkera argiaren bidetik*, Vitoria-Gasteiz.

MAP (Ministerio para las Administraciones Públicas). 1991. *Manual de estilo del lenguaje administrativo*, Madrid.

D. McKelvie y H.S. Thompson. 1994. "TEI-Conformant Structural of a Trilingual Parallel Corpus in the ECI Multilingual Corpus 1", *Proceedings of the International Wordshop on Sharable Natural Language Resources*, 108–112.

A. K. Melby. 1995. *The Possibility of Language. A discussion of the nature of language with implications for human and machine translation*, John Benjamins.

J. Prieto de Pedro. 1989. "Los vicios del lenguaje legal. Propuestas de estilo", *La calidad de las leyes*, Gobierno Vasco, Vitoria-Gasteiz.

J. C. Sager. 1993. *Language Engineering and Translation. Consequences of automation.* John Benjamins.

M. Snell-Hornby. 1988. *Translation Studies.* John Benjamins.

E. Sumita y H.Iida. 1991S. Sato y M. Nagao. 1990. "Toward Memory-Based Translation" *COLING-90: Papers presented to the 13th International Conference on Computational Linguistics* 3:247-252, Helsinki.

E. Sumita y H.Iida. 1991. "Experiments and Prospects of Example-Based Machine Translation", *Proceedings of the Association for Computational Linguistics*, 185-192, Berkeley.

G. Toury. 1995. *Descriptive Translation Studies and beyond.* John Benjamins.

G. Vázquez-Ayora. 1977. *Introducción a la Traductología.* Georgetown University Press, Whasington.

J.P. Vinay y J. Darbelnet. 1958. *Stylistique comparée du français et l'anglais*, Didier, Paris.

A. Winarske, S. Warwick-Armstrong, J. Hajic. 1992. "Tagging and Alignment of Parallel Texts: Current Status of BCP", *Proceedings of the Third Conference on ANLP*, 227-228, Terento.