http://dx.doi.org/10.12795/PH.1997.v11.i02.09

TRATAMIENTO DE TEXTOS ADMINISTRATIVOS BILINGÜES: EL PROYECTO *LEGEBIDUN*[*]

Joseba K. Abaitua Odriozola
Arantza Casillas Rubio
Raquel Martínez Unanue

This paper introduces the methodology of the LEGEBIDUN project that concentrates on the explotation possibilities of a bitext corpus of administrative documents in both Basque and Spanish as a source for the development of simultaneous editing and translating software. The paper includes tokens of tagged text where *variable translations units* are identified.

"Over the past thirty years, what types of computer software that deal with language in any way have been the most successful? Clearly, it has been word processing and indexing software."
(Melby 1995:73)

Resumen

Este artículo presenta la metodología del proyecto LEGEBIDUN que trata de estudiar las posibilidades de explotación de un corpus bilingüe de textos administrativos como fuente de datos para la creación de entornos de edición y traducción. Se acompaña la presentación con muestras de texto etiquetado en el que se reconocen lo que denominamos *unidades de traducción variables*.

1. Introducción

La cita de Alan K. Melby, que procede de una obra reciente en torno a las posibilidades de la traducción automática, ilustra muy adecuadamente las premisas sobre las que se asienta la metodología del proyecto *LEGEBIDUN*, tal y como vamos a exponer en este artículo.

---

[*] Los autores quieren expresar su gratitud al Servicio de Traducciones 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.

El objetivo principal del proyecto *LEGEBIDUN* es demostrar de qué manera se puede optimizar la edición de documentación administrativa bilingüe mediante la aplicación de distintas tecnologías de la lingüística computacional. Concretamente, estudiamos las posibilidades de explotación de un corpus bilingüe como fuente de datos para la creación de entornos de procesamiento de textos administrativos con ayudas para la composición y traducción simultánea. El corpus se ha tratado por medios atomáticos para introducir etiquetas descriptivas cuyo principal cometido es identificar en las dos versiones lo que denominamos *unidades de traducción variables*. Mediante algoritmos de alineamiento se están construyendo catálogos de pares de equivalencias. Además, como resultado del etiquetado, se han generado *definiciones de tipo de documentos* (DTDs del estándar SGML), que equivalen a gramáticas de estados finitos capaces de reproducir la estructura de los textos. En este artículo se defiende la idoneidad de la metodología empleada y se presentan muestras de textos etiquetados y de unidades de traducción variables.

2. Antecedentes

A partir del establecimiento de la cooficialidad de las dos lenguas (euskara y castellano) en todos los ámbitos institucionales de la Comunidad Autónoma Vasca, se ha incrementado notablemente la traducción especializada en euskara, hasta el punto de que la traducción de textos administrativos supone en torno al 80% del total de traducciones al euskara. La oficialidad implica la incorporación de nuevos ámbitos de uso para la lengua vasca y su adaptación a nuevas funciones. Todo ello entraña un importante cúmulo de obstáculos debido sobre todo a la necesidad de desarrollar nuevos registros y expresiones inexistentes hasta el momento de su promoción a lengua oficial.

Se pueden definir dos situaciones extremas en la traducción especializada. La situación idónea para el traductor se da cuando la lengua de llegada dispone de expresiones y registros equivalentes a los empleados en la lengua de partida. La misión del traductor consistirá en conocer tales equivalencias y en aplicarlas adecuadamente. La situación opuesta se produce cuando la lengua de llegada no ha sido desarrollada en el ámbito del texto que se desea traducir. En una circunstancia así el traductor se ve obligado a adoptar una delicada responsabilidad de desarrollo lingüístico para la que no siempre está capacitado. A este problema suele añadirse un segundo factor que agrava todavía más la situación, es el aislamiento en el que muy a menudo trabajan los traductores. Esta situación conduce a que las soluciones terminológicas y registros de nueva acuñación sean muchas veces dispares. En el ámbito de la Administración Vasca, se han tomado medidas de diversa índole encaminadas a paliar el problema, de forma destacada por parte de un instituto a quien atañe especialmente este cometido, el **Instituto Vasco para la Administración Pública** (IVAP 1993). Sin embargo el problema dista mucho de estar resuelto. El proyecto LEGEBIDUN se concibió en 1993, a partir de una iniciativa del Master en Traducción de la Universidad de Deusto, con la pretensión de demostrar la validez de las técnicas de la lingüística computacional ante este tipo de situaciones. Aunque los organismos públicos han prestado una estimable colaboración para la creación del corpus, por el momento el proyecto no está vinculado ni oficial ni financieramente con

ningún organismo. Se trata de un trabajo experimental desarrollado en el marco de la investigación de dos tesis doctorales.

3. El proyecto *LEGEBIDUN*

En el proceso manual, tanto los redactores de los documentos originales como los traductores de los distintos organismos se dedican a la reutilización y reciclado permanente de los textos. Primero localizan los documentos según el tipo (actas, diligencias, disposiciones, decretos, órdenes, normas, etc.) y después someten el documento a las operaciones habituales de cortar y pegar los fragmentos que pueden reutilizarse. Los traductores sulen acometer este trabajo en ámbitos cerrados, es decir, dentro de un organismo determinado (sea Ayuntamiento, Diputación Foral, Gobierno Vasco o Parlamento), y no siempre cuentan con la colaboración de los redactores de los documentos originales. Tampoco existe una red que relacione a los distintos organismos entre sí, aunque sí se han dado casos aislados de coordinación, como fue hace unos años la comisión para la normalización de la terminología de los impresos de la declaración de la renta en las Diputaciones Forales. Ante este estado de cosas, LEGEBIDUN pretende definir una metodología para la configuración futura de un entorno telemático óptimo que permita la reutilización y el reciclado permanente de la documentación bilingüe en cualquier ámbito de la Administración Pública Vasca.

El proyecto en la actualidad consta de los siguientes apartados:

* **Creación de un corpus**. El corpus en la actualidad está compuesto por boletines de tres administraciones: de las Diputaciones de Álava (BOA 1990-91) 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 concentra 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 y MULTEXT.

* **Estudio estructural**. A partir de un análisis detallado de las distintas clases de documentos en una parte del corpus (Órdenes Forales del BOB), se ha realizado un inventario de etiquetas descriptivas. Se ha definido un modelo (DTD) normalizado para cada tipo de documento, aproximadamente 900 documentos de 22 tipos distintos, con una muestra representativa de 40 textos de cada uno).

* **Creación de memorias de traducción.** Los textos paralelos se someten a un cotejo automático que tiene como objeto la identificación de unidades de traducción equivalentes en las dos versiones mediante la aplicación de diversos algoritmos de alineamiento. Una vez reconocidas, estas unidades se catalogan formando memorias de traducción.

4. El lenguaje administrativo

El lenguaje administrativo es un claro exponente de los llamados lenguajes de especialidad y, en este sentido, exhibe de manera generalizada todas las propiedades de dichos lenguajes, tal y como ponen de manifiesto diversos estudios (MAP 1991, Prieto de Pedro 1989, Calvo Ramos 1980). Los lenguajes de especialidad tienen una ventaja para su estudio sobre la lengua común y es que sus unidades se corresponden con unidades conceptuales cuya referencia pertenece a un mundo extralingüístico acotado. En consecuencia, los lenguajes de especialidad requieren que cada concepto se exprese mediante un único término y, recíprocamente, que cada término exprese un solo concepto, de manera que se eviten imprecisiones y ambigüedades (Cabré 1993, Sager 1993). Por contra, las unidades de la lengua común son unidades *léxicas* que por su propia naturaleza son *dinámicas*, en el sentido de que su significado es fundamentalmente variable y ambiguo.

Una característica que comparten los lenguajes de especialidad es que suelen etar sujetos a importantes medidas de sistematización. Sin embargo, no siempre los colectivos de científicos, técnicos o profesionales obtienen resultados por igual. En el contexto de lenguas en proceso de normalización, como es el caso del que nos ocupamos, las dificultades para llevar a la práctica estas medidas se incrementan. El proyecto LEGEBIDUN aporta, en este sentido, el valor añadido de permitir evaluar la dispersión de variantes terminológicas en las traducciones del lenguaje administrativo en euskara.

Una vez subsanado el problema de la normalización terminológica, los lenguajes de especialidad son, como afirma Melby, el dominio más adecuado para ensayar la traducción automática. Esta opinión generalizada y bien contrastada alienta considerablemente las expectativas de nuestro proyecto, más si cabe cuando se considera la relevancia de la traducción administrativa en euskara. No por ello el tratamiento de los textos administrativos está exento de los problemas de la lengua común. Los textos de especialidad no se componen exclusivamente de expresiones terminológicas, referencialmente acotadas; también contienen elementos de la lengua común, que se utilizan como mecanismo aglutinador y cohesionador del texto. Por ello, una parte central de nuestro proyecto consiste en esayar métodos para el reconocimiento de las unidades especificas del lenguaje administrativo y su distinción de las unidades de la lengua común.

En opinión de Melby, el fracaso de muchos proyectos de traducción automática recae precisamente en la falta de discernimiento entre lo que son unidades terminológicas y lexicológicas. Dicho de otra manera, el fracaso se debe al desconocimiento del dominio en el que se va a realizar la traducción. Por el contrario, los proyectos que mayores éxitos han cosechado se han circunscrito siempre al ámbito de los mal llamados *sublenguajes* (Somers 1993). Decimos "mal llamados" porque durante algún tiempo se ha considerado que los lenguajes de especialidad forman subconjuntos de la lengua común. Los estudiosos de la materia en la actualidad prefieren considerar que existen relaciones de interseción y no de inclusión (Bergenhotz y Tarp 1995:16-19, Cabré 1993:139-141).

5. Las memorias de traducción

Hasta fechas recientes, la TA aspiraba a resolver los problemas de la traducción sin limitaciones. Las técnicas simbólicas utilizadas, preponderantes en la lingüística computacional hasta la década de los noventa, se han centrado en el estudio de las gramáticas sintagmáticas de las distintas lenguas. Generalmente tales gramáticas representan un modelo ideal del lenguaje humano, separado, en exceso, del uso real. En los últimos años hemos asistido a un retorno hacia postulados más pragmáticos, similares a los utilizados en la década de los cincuenta. Los más conocidos son los siguientes: *memoria de traducción* (Sato y Nagao 1990), *traducción basada en el corpus* (Winarske *et al.* 1992), *traducción por ejemplos* (Sumita *et al.* 1991) y *traducción basada en la estadística* (Brown *et al.* 1990, 1991). La metodología del proyecto LEGEBIDUN se apoya fundamentalmente en las aportaciones de Winarske *et al.* 1992 y Brown *et al.* 1991 sobre los textos bilingües en inglés y francés de las *Actas del Parlamento Canadiense*, recopilados en el **Hansard Corpus**. Nuestro trabajo consiste en etiquetar los textos bilingües por segmentos que se corresponden con unidades de traducción variables. Estos textos etiquetados se procesan mediante técnicas de alineamiento y los resultados se catalogan a la manera de listas ordenadas de segmentos de equivalencias. Estas listas son las que conforman la *memoria* de traducción.

6. El etiquetado descriptivo

El tratamiento de los textos mediante sistemas de etiquetado descriptivo tiene importantes ventajas. Un etiquetado descriptivo es un código de identificación de los elementos que concurren en el texto. Nuestro proyecto estudia diversas propuestas de etiquetado, fundamentalmente las aportadas por los consorcios internacionales TEI y MULTEXT. TEI (*Text Encoding Iniciative*) es una iniciativa internacional que se encarga de la extracción y codificación de textos tratables por ordenador (McKelvie y Thompson 1994). TEI utiliza SGML (*Standard Generalized Markup Language*, ISO 8879:1986) que es un estándar para la codificación documental. SGML abarca cuestiones tan diversas como la disposición del texto, su estructura, contenido, etc. Por su parte, el proyecto MULTEXT (*Multilingual Text Tools and Corpora*) contribuye al desarrollo de software para manipular y analizar corpus textuales y a la creación de corpus plurilingües con etiquetado estructural y lingüístico. Sus propuestas amplían y complementan las directrices del TEI (Ide y Véronis 1994).

Se presentan a continuación dos ejemplos en cada lengua de textos etiquetados en los que se puede comprobar la estructura de los documentos (las etiquetas se han adaptado para permitir una mejor interpretación):

Ejemplo de documento 1 en castellano:

```
<documento>
<encabezado>
<tipo>OF_T2<\tipo>
```

```
<idioma>castellano<\idioma>
<localizacion>BOB junio 1994<\localizacion>
<\encabezado>
<encabezamiento>Orden Foral
 <numero>número 313/1994,<\numero>
 <fecha> de fecha 10 de mayo.<\fecha>
Mediante la Orden Foral de referencia se ha dispuesto lo
siguiente:
<\encabezamiento>
<desarrollo>
 <enumeracion1>1.-Aprobar definitivamente la Modificación de
las Normas Subsidiarias de Barakaldo en el ámbito de Alonsótegui
para la inclusión del Corredor del Cadagua, en este término
municipal.
 <\enumeracion1>
 <enumeracion2>2.-El Ayuntamiento de Alonsótegui, deberá
remitir un ejemplar del documento, en el plazo de un mes, para
proceder a su autentificación.
 <\enumeracion2>
 <interposicion>Contra dicha Orden Foral, que agota la vía
administrativa, podrá interponerse recurso de reposición ante
        <emisor-orden> el Diputado Foral de
Urbanismo,<\emisor-orden>
 como trámite previo a la impugnación ante la Jurisdicción
Contencioso-Administrativa,
        <plazo>en el plazo de un mes,<\plazo>
 contado desde el día siguiente a esta notificación, sin
perjuicio de la utilización de otros medios de defensa que estima
oportunos.
 <\interposicion>
<\desarrollo>
<pie>Durante el referido plazo
 <numero-expediente>el                    expediente
BHI-004/94-P05-A,<\numero-expediente>
quedará de manifiesto para su examen en las dependencias
situadas en
 <lugar>Bilbao, Alameda Rekalde, 30, 5.a y 6.a
plantas.<\lugar>
<lugar-fecha>Bilbao, 10 de mayo de 1994.<\lugar-fecha>
<emisor-orden>-El Diputado Foral de Urbanismo,
Pedro Hernández González<\emisor-orden>
<\pie>
<\documento>
```

Ejemplo de documento 1 en euskara:

```
<documento>
<encabezado>
<tipo>OF_T2<\tipo>
<idioma>euskara<\idioma>
<localizacion>BOB junio 1994<\localizacion>
<\encabezado>
<encabezamiento>Foru Agindua,
 <numero>313/1994 zk.eko<\numero>
 <fecha> maiatzak 10<\fehca>
Aipameneko Foru Aginduaren bidez honako hau xedatu da:
<\encabezamineto>
<desarrollo>
 <enumeracion1>1.- Alonsotegiko eremuan Barakaldoko
Sorospidezko Arauen aladarazpena, udalerritik igarotean Cadaguako
pasabidea barru sartzeko, behin betiko onestea.
 <\enumeracion1>
 <enumeracion2>2.- Alonsotegiko udalak, hilabeteko epearen
barruan agiriaren ale bat bidaliko du kautotua izan dadin.
 <\enumeracio2>
 <interposicion>Administrazio bidea agortzen duen aipaturiko
Foru Aginduaren aurka, jakinerazpen honen biharamunetik zenbatu
beharreko
        <plazo>hilabeteko epearen barruan,<\plazo>
 birjarpenezko errekurtsoa jarri ahal izango da
 <emisor-orden>Hirigintzako Foru Diputatuaren<\emisor-orden>
 aurrean, Administrazioarekiko Auzien Jurisdikzio aurrean
egiteko aurkapenaren aldez aurretiko tramite gisa, komeniesten
diren beste defentsabideak erabil daitezkeelako kalterik
gabe.<\interposicion><\desarrollo>
<pie>Adierazi den epearen barruan,
 <numero-expediente>BHI-004/94-P05-A espedientea
<\numero-expediente>
 <lugar>Bilbaoko Rekalde zumarkaleko 30.eko 5 eta 6. solairuko
bulegoetan<\lugar>
egongo da ageriko, azter dadin.
<lugar-fecha>Bilbon, 1994.eko maiatzaren 10ean.<\lugar-fecha>
<emisor-orden> Hirigintzako foru diputatua. Pedro Hernández
González<\emisor-orden>
<\pie>
<\documento>
```

Ejemplo de documento 2 en castellano:

```
<documento>
<encabezado>
<tipo>OF_T2<\tipo>
<idioma>castellano<\idioma>
<localizacion>BOB junio 1994<\localizacion>
<\encabezado>
<encabezamiento>Orden Foral
 <numero>número 316/1994,<\numero>
 <fecha> de fecha 10 de mayo.<\fecha>
Mediante la Orden Foral de referencia se ha dispuesto lo
siguiente:
<\encabezamiento>
<desarrollo>
 <enumeracion1>1.-Aprobar definitivamente la Modificación de
las Normas Subsidiarias en la Insula 2 de la localidad de
Arantzatzu, en la que deberá incorporarse lo que a continuación
se expone:
        1.1. Deberá completarse la normativa de vuelos
aportada, en el sentido de permitir una mayor flexibilidad de las
alineaciones de la edificación.
        1.2. Se suprimirá la acera que circunda el jardín
privado de las edificaciones consolidadas que pasará a formar
parte de esta calificación.
 <\enumeracion1>
 <enumeracion2>2.-El Ayuntamiento de Arantzatzu deberá remitir
a este Departamento tres ejemplares del proyecto refundido para
proceder a su autentificación.
 <\enumeracion2>
 <interposicon>Contra dicha Orden Foral, que agota la vía
administrativa, podrá interponerse recurso de reposición ante
        <emisor-orden>  el  Diputado  Foral  de
Urbanismo,<\emisor-orden>
 como trámite previo a la impugnación ante la Jurisdicción
Contencioso-Administrativa,
        <plazo>en el plazo de un mes,<\plazo>
 contado desde el día siguiente a esta notificación, sin
perjuicio de la utilización de otros medios de defensa que estima
oportunos.
 <\interposicon>
<\desarrollo>
```

```
<pie>Durante el referido plazo el expediente
 <numero-expediente>BHI-229/93-P05-A,<numero-expediente>
quedará de manifiesto para su examen en las dependencias
situadas
 <lugar>en Bilbao, Alameda Rekalde, 30, 5.a y 6.a
plantas.</lugar>
<lugar-fecha>Bilbao, 10 de mayo de 1994.<\lugar-fecha>
<emisor-orden>-El Diputado Foral de Urbanismo,
Pedro Hernández González<\emisor-orden>
<\pie>
<\documento>
```

Ejemplo de documento 2 en euskara:

```
<documento>
<encabezado>
<tipo>OF_T2<\tipo>
<idioma>euskara<\idioma>
<localizacion>BOB junio 1994<\localizacion>
<\encabezado>
<encabezamiento>Foru Agindua,
 <numero> 316/1994 zk.,<\numero>
 <fecha> maiatzaren 10ekoa.<\fecha>
Aipameneko Foru Aginduaren bidez honako hau xedatu da:
<\encabezamiento>
<desarrollo>
 <enumeracion1>1.-Arantzazu udalerriko 2. Insulan Sorospidezko
Arauen aldarazpena behin betiko onestea, bertan jarraian
adierazten dena barru sartu beharko dela.
        1.1. Ekarritako hegalduren arautegia osotu beharko da,
erakikinaren lerrokaduran malgutasun handiagoa uzteko.
        1.2. Eraikin sendotuetako lorategi pribatuari bira
ematen dion espaloia kenduko da, kalifikazio honetako atal egina
izango dela.
 <\enumeracion1>
 <enumeracion2>2.-Arantzazuko udalak, Sail honi proiektu
bateginaren hiru ale bidaliko dizkio kautotua izan dadin.
 <\enumeracion2>
 <interposicion>Administrazio bidea agortzen duen aipaturiko
Foru Aginduaren aurka, jakinerazpen honen biharamunetik zenbatu
beharreko
        <plazo>hilabeteko epearen barruan,<\plazo>
 birjarpenezko errekurtsoa jarri ahal izango da
```

```
          <emisor-orden>Hirigintzako                    Foru
Diputatuaren<\emisor-orden>
    aurrean, Administrazioarekiko Auzien Jurisdikzio aurrean
egiteko aurkapenaren aldez aurretiko tramite gisa, komeniesten
diren beste defentsabideak erabil daitezkeelako kalterik gabe.
   <\interposicion>
   <\desarrollo>
   <pie>Adierazi den epearen barruan,
    <numero-expediente>BHI-229/93-PO5-A<\expediente> espedientea
    <lugar>Bilbaoko Rekalde zumarkaleko 30.eko 5 eta 6. solairuko
bulegoetan<\lugar>
   egongo da ageriko, azter dadin.
   <luagr-fecha>Bilbon, 1994.eko maiatzaren 10ean.<\lugar-fecha>
   <emisor-orden> Hirigintzako foru diputatua. Pedro Hernández
González<\emisor-orden>
   <\pie>
   <\documento>
```

7. Unidades de traducción variables

LEGEBIDUN adopta técnicas de la lingüística de corpus, aplicadas al estudio exhaustivo y a la constatación estadística de los datos. Para la traducción, además, una fuente importantísima de datos son los textos paralelos. Estos textos, que contienen una versión bilingüe del mismo documento, permiten establecer correspondencias no limitadas a la palabra, ni siquiera a la expresión multipalabra, o al giro terminológico. Existen, de hecho, correspondencias entre fragmentos tan extensos como el párrafo. En el ámbito jurídico y administrativo este tipo de correspondiencias son frecuentes.

La discusión en torno a la naturaleza de la unidad de traducción (UT) se remonta a los estudios de Vinay y Darbelnet 1958. Desde entonces la bibliografía ha aumentado considerablemente, pero salvo contadas excepciones se ha ocupado de estudiar la UT en el terreno de la traducción manual o humana. Bennett 1994 es la primera referencia importante que traslada estos estudios al terreno de la traducción automática, adaptando las propuestas a los métodos de traducción por transferencia y equiparando la unidad de traducción lexicológica con la unidad de transferencia. LEGEBIDUN añade, a la consideración tradicional de unidades lexicológicas (UTL) y terminológicas (UTT), una tercera, la *unidad de traducción de fórmulas cliché* (UTC). Esta unidad abarca toda la colección de coletillas, retailas, fórmulas y clichés jurídicos tan frequentes en los textos administrativos y que se corresponden de manera sistemática en las dos versiones. Estos son algunos ejemplos extraídos de los textos que han servido de muestra:

<UTC_1>Mediante la Orden Foral de referencia se ha dispuesto lo siguiente:

<UTC_2>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.

<UTC_3>Durante el referido plazo el expediente <n1 expediente> quedará de manifiesto para su examen en las dependencias situadas en <lugar>.

<UTC_1>Aipameneko Foru Aginduaren bidez honako hau xedatu da:

<UTC_2>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.

<UTC_3>Adierazi den epearen barruan, <n1 expediente> espedientea <lugar> egongo da ageriko, azter dadin.

Las UTC autónomas (las que ocupan todo un párrafo) 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, supone 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 resto del proceso.

La identificación de las otras unidades de traducción es más compleja y constituye el núcleo central de nuestra investigación. Estos son algunos ejemplos de párrafos segmentados en UTs variables (<UT_num> identifica el orden de distribución en castellano y sus equivalencias en euskara):

```
<UT_1>Aprobar definitivamente/
<UT_2>la Modificación de las Normas Subsidiarias de ()/
<UT_3>en el ámbito de ()/
<UT_4>para la inclusión del Corredor del Cadagua, en este
término municipal./
<UT_3>()ko eremuan/
<UT_2>()ko Sorospidezko Arauen aladarazpena,/
<UT_4>udalerritik igarotean Cadaguako pasabidea barru
sartzeko,
<UT_1>behin betiko onestea./
<UT_5>El Ayuntamiento de ()/
```

```
<UT_6>deberá remitir un ejemplar del documento/
<UT_7>en el plazo de un mes/
<UT_8>para proceder a su autentificación./
<UT_5>()ko udalak,/
<UT_7>hilabeteko epearen barruan/
<UT_6>agiriaren ale bat bidaliko du/
<UT_8>kautotua izan dadin./
```

Es fácil reconocer una UTC en la siguiente expresión:

```
<UT_8>para proceder a su autentificación./
<UT_8>kautotua izan dadin./
```

Esta UTC no es autónoma, ya que forma parte de un párrafo en el que su orden de aparición no se puede predeterminar de manera sencilla. Su tratamiento es complicado ya que el párrafo en cuestión se compone de unidades dispares, es decir, que la mencionada UTC coaparece junto a UTTs y UTLs.

Ejemplo de unidad terminológica (UTT):

```
<UT_2>la Modificación de las Normas Subsidiarias de ()/
<UT_2>()ko Sorospidezko Arauen aladarazpena,/
```

Ejemplo de combinación de UTT (en cursiva) y UTLs:

```
<UT_4>para la inclusión del <UTT>Corredor del Cadagua</UTT>,
en este término municipal./
<UT_4>udalerritik igarotean <UTT>Cadaguako pasabidea</UTT>
barru sartzeko,
```

Este caso ilustra además que la traducción de UTLs, cuando la realiza un traductor humano, suele ser poco literal y está más sujeta a variación, como demuestra la retraducción literal de dicha expresión en euskara: *udalerritik igarotean <UTT> barru sartzeko* ("a [su] paso por el término municipal <UTT> se incluya [en él]"). Esta muestra pretende emular la solución que probablemente produciría un generador de transferencia y da idea de la dificultad que entraña la resolución acertada de las referencias anafóricas "su" y "en él", que son pronombres nulos o vacíos, es decir, no expresados léxicamente en la versión en euskara.

Ejemplo de oración compuesta por unidades de la lengua común (UTL):

```
<UT_9>Se suprimirá la acera que circunda el jardín privado de
las <UTT> edificaciones consolidadas</UTT> que pasará a formar
parte de esta calificación./
<UT_9><UTT>Eraikin sendotu<UTT>etako lorategi pribatuari bira
ematen dion espaloia kenduko da, kalifikazio honetako atal egina
izango dela./
```

Esta oración que hemos identificado mediante la etiqueta <UT_9> se compone mayoritariamente de unidades léxicas que no aparecerán contempladas en el diccionario de UTTs, exceptuando quizá el término complejo "edificación consolidada" ("eraikin sendotua"). El tratamiento de las UTLs escapará a los procedimientos utilizados para el reconocimiento de UTCs y UTTs y no forma parte del proyecto LEGEBIDUN. El estudio de UTLs en euskara entra dentro de los estudios generales de lexicografía, como el que llevan a cabo Ezeiza *et al.* 1996.

8. El alineamiento

Una definición informal de alineamiento es la siguiente: Dados dos conjuntos de información entre los que existe una cierta relación de similitud en función de un determinado criterio, el alineamiento consiste en identificar correspondencias entre subconjuntos de ambos conjuntos. De todos los posibles emparejamientos que se pueden establecer, se seleccionará aquél que satisfaga en mayor medida el criterio en que se basa el alineamiento. Los trabajos realizados a este respecto, toman como unidad de alineamiento bien la oración o bien la palabra. En nuestro caso, se reconocen y emparejan *unidades de traducción variables* (UTV).

Dentro de las principales propuestas de alineamiento de corpus paralelo basadas en métodos estadísticos podemos distinguir diferentes estrategias. La mayoría de los proyectos internacionales que incorporan técnicas de alineamiento de corpus paralelos se basan en las propuestas de Gale y Chuch 1991, por un lado, y de Brown *el al.* 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 *et al.* 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 UTC, 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 UT en ambas lenguas. Aún a pesar de tener que salvar la complejidad añadida, consideramos que las UTVs representan un modelo de unidad de

traducción considerablemente más adecuado que la oración. La memoria de traducción que se está obteniendo mediante este procedimiento es más eficaz y económica que las que se basan en segmentos del tamaño de oraciones. La utilización de SGML como sistema de etiquetado permite utilizar las DTDs que se generan a partir de los textos etiquetados como gramáticas que dan cuenta de la estructura y distribución de las UTVs en los documentos. Con todo, el límite de la aplicación de las DTDs viene marcado por las UTLs. Las divisiones o párrafos de los documentos que se compongan mayoritariamente de UTLs, deberán, en su caso, resolverse mediante procedimientos clásicos de base sintagmática, o bien mediante la intervención del traductor humano.

Conclusión

Se ha mostrado la metodología del proyecto *LEGEBIDUN* como propuesta de optimización de la edición de documentación administrativa bilingüe mediante la aplicación de distintas tecnologías desarrolladas por la lingüística computacional. Concretamente, se han discutido las posibilidades de explotación de un corpus bilingüe de textos administrativos como fuente de datos para la creación de entornos de edición y traducción. El corpus ha sido tratado mediante la incorporación de etiquetas descriptivas que identifican *unidades de traducción variables*, las cuales, una vez alineadas, constituyen listas de equivalencias. De esta manera se disponen de memorias de traducción que permiten automatizar gran parte del proceso de traducción, si bien no lo consiguen en su totalidad. Para completar dicha automatización se deben incorporar en el proceso gramáticas de corte sintagmático sobre textos que contengan etiquetas con información morfosintáctica, a la manera descrita por Ezeiza *et al.* 1996, con cuyo proyecto se complementa LEGEBIDUN.

REFERENCIAS

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

H. Bergenhotz y S. Tarp, *Manual of Specialized Lexicography* (John Benjamins 1995).

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

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

M. T. Cabré, *La terminología. Teoría, metodología, aplicaciones* (Antártida/Empúries, Barcelona 1993).

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

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

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

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

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

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

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

A. K. Melby, *The possibility of language* (John Benjamins 1995).

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

J. C. Sager, *Curso práctico sobre el procesamiento de la terminología*, Fundación Germán Sánchez Ruipérez, (Madrid 1993).

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

H. L. Somers, "Current Research in Machine Translation", *Machine Translation* 7 (1993) 231-246.

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

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

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