Actas del I Congreso Internacional
de la Asociación Ibérica de Estudios de Traducción e Interpretación

Una aproximación a la generación automática de textos jurídicos

Gloria Corpas Pastor
Universidad de Málaga

Como citar este artículo:

Corpas Pastor, Gloria (2003) «Una aproximación a la generación automática de textos jurídicos», en Muñoz Martín, Ricardo [ed.] *I AIETI. Actas del I Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación. Granada 12-14 de Febrero de 2003*. Granada: AIETI. Vol. n.º 2, pp. 709-720. ISBN 84-933360-0-9. Versión electrónica disponible en la web de la AIETI: <http://www.aieti.eu/pubs/actas/I/AIETI_1_GCP_Aproximacion.pdf>.

Una aproximación a la generación automática de textos jurídicos

Gloria Corpas Pastor
Universidad de Málaga
gcorpas@uma.es

Resumen

El presente trabajo describe de forma sucinta los objetivos y resultados conseguidos en el proyecto Diseño de un tipologizador textual para la traducción automática de textos jurídicos (español ↔ inglés/alemán/italiano/árabe) (PB98-1399, DGICYT). Dicho proyecto ha tenido una duración de tres años (1999-2002) y en la actualidad se encuentra en la fase final de su implementación. En él han participado un total de 25 investigadores (profesores y becarios) de tres Universidades: Málaga, Alcalá de Henares y Pablo de Olavide (Sevilla), siendo la primera el centro neurálgico del proyecto. El objetivo central del proyecto es el diseño y posterior implementación de una herramienta informática capaz de generar automáticamente un determinado tipo de texto jurídico en cada una de las lenguas implicadas en el proyecto, a saber, el español, el alemán, el árabe, el inglés y el italiano. Nos hemos centrado en un tipo de contrato: el de compraventa de bienes inmuebles, para el cual se ha elaborado una interlengua a partir de una relación de conceptos, en forma de gramática subyacente y orientada hacia la creación de un generador automático multilingüe. Dicho generador forma parte de un software de traducción basado en cuatro componentes: bases de datos relacionales, interlengua de representación, interfaz de usuario y motor de traducción (aún en fase de diseño y experimentación). Para la alimentación del generador se ha compilado un corpus multilingüe (comparable y paralelo) procedente de textos digitalizados y documentos localizados en Internet. La metodología desarrollada en el proyecto puede ser aplicada a otros tipos de textos jurídicos, e, incluso, a otras parcelas de especialización. En cualquier caso, las investigaciones que venimos realizando supondrán un beneficio considerable para la ingeniería lingüística en general, y sin duda fomentarán la aplicación de las nuevas tecnologías en un campo tan eminentemente multidisciplinar como es, hoy en día, la traducción.

1. Introducción

El presente trabajo describe de forma sucinta los objetivos y resultados conseguidos en el proyecto *Diseño de un tipologizador textual para la traducción automática de textos jurídicos (español↔inglés/alemán/italiano/árabe*) (PB98-1399, DGICYT). Dicho proyecto ha tenido una duración de tres años (1999-2002) y en la actualidad se encuentra en la fase final de su implementación. En él han participado un total de 25 investigadores (profesores y becarios) de tres Universidades, Málaga, Alcalá de Henares y Pablo de Olavide (Sevilla), siendo la primera el centro neurálgico del proyecto.

En Ricardo Muñoz Martín, ed. 2003. *I AIETI. Actas del I Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación. Granada, 12-14 de febrero de 2003.* Granada: AIETI, pp. 709-720. ISBN 84-933360-0-9.

El objetivo central del proyecto es el diseño y posterior implementación de una herramienta informática capaz de generar automáticamente un determinado tipo de texto jurídico en cada una de las lenguas implicadas en el proyecto, a saber, el español, el alemán, el árabe, el inglés y el italiano. Nos hemos centrado en un tipo de contrato: el de compraventa de bienes inmuebles, para el cual se ha elaborado una interlengua a partir de conceptos jurídicos, en forma de gramática subyacente y orientada hacia la creación de un generador automático multilingüe. Dicho generador forma parte de un *software* de traducción basado en cuatro componentes: bases de datos relacionales (SGBDR), interlengua de representación (*tags*/etiquetas), interfaz de usuario y motor de traducción (aún en fase de diseño y experimentación).

Para la alimentación del generador se ha compilado un corpus multilingüe (comparable y paralelo) procedente de textos digitalizados (formularios y contratos reales) y documentos localizados en Internet (legislación civil o mercantil, formularios, etc.) a través de los recursos informáticos y telemáticos que proporcionan la red mundial. Una vez evaluadas las fuentes de documentación concretas, se han extraído los documentos. Éstos han sido limpiados y etiquetados convenientemente, para, posteriormente, ser manipulados mediante un programa de gestión de corpus. Nuestra orientación surge de un convencimiento profundo de la imposibilidad de construir prototipos textuales sin manejar extensos corpus en soporte informático. Los datos obtenidos de este tipo de análisis proporcionan información real y fiable sobre aspectos terminológicos, fraseológicos, macro- y superestructurales que caracterizan de forma bastante precisa el *esqueleto* formal del tipo textual que hemos investigado.

La metodología desarrollada en el proyecto puede ser aplicada a otros tipos de textos jurídicos, e, incluso, a otras parcelas de especialización. En cualquier caso, las investigaciones que venimos realizando aspiran a potenciar la ingeniería lingüística y la aplicación de las nuevas tecnologías en un campo tan eminentemente multidisciplinar como es, hoy en día, la traducción.

2. Compilación de corpus y establecimiento de prototipos textuales

Como se puede deducir de los objetivos propuestos, nuestras bases metodológicas conforman un triángulo equilátero cuyos vértices se asientan sobre la lingüística del texto, la lingüística del corpus y la generación textual automática. Por una parte, estamos convencidos de la necesidad de construir prototipos que respeten las convenciones retóricas, lingüísticas y discursivas de los tipos textuales objeto de estudio. Centrándonos en el género jurídico en nuestro caso, difícilmente se obtendrá un texto traducido que pase por original en la comunidad meta, si no se respeta la idiosincrasia del tipo textual correspondiente, ni se tienen en cuenta su superestructura y sus movimientos retóricos básicos. Por otra parte, somos conscientes de la necesidad de consultar corpus especializados para la construcción de prototipos textuales, esto es, el esqueleto formal de un determinado tipo o variedad textual, e, incluso, su macroestructura semántica básica.

Un primer paso en la ejecución del proyecto ha sido, precisamente, el estudio de las características del discurso jurídico en cada una de las lenguas implicadas, con especial referencia al contrato de compraventa de bienes inmuebles (CCVBI), a partir de monografías y artículos especializados sobre el tema, cuya lista, por extensa, eludiremos enumerar. Llegados a este punto conviene indicar los límites geográficos en el caso de las lenguas transnacionales, habida cuenta de que cada variedad diatópica implicaba una diferencia sustancial en cuanto a sus respectivos ordenamientos jurídicos: así, nos hemos centrado en el español peninsular, excluyendo el español de América; en el caso del alemán, hemos descartado la documentación procedente de Austria o Suiza; Yemen y Arabia Saudita son los únicos países árabes que hemos seleccionado; mientras que para el inglés sólo hemos tenido en cuenta el País de Gales e Inglaterra.

Una vez identificadas las características propias discurso jurídico en las distintas lenguas, se pasa a la construcción de un prototipo textual monolingüe para cada una de ellas. El análisis anterior se valida y enriquece con los análisis de los diversos corpus compilados, siguiendo la metodología propuesta en un principio por Biber (1995) y Biber, Conrad y Reppen (1998). Para la parte textual, hemos seguido a Stubbs (1996) y Garside, Leech y McEnery (1997). De esta forma podemos comprobar si las características reseñadas en los estudios teóricos sobre el tema se corresponden con los datos reales extraídos del discurso real. Además, es posible identificar un texto de origen (TO) en español o en las demás lenguas de trabajo como un contrato de compraventa a partir de los datos que arroja el corpus, así como cuantificar el grado de «prototipicalidad» del TO, señalando las divergencias más notables; y, posteriormente, ofrecer pautas de traducción en cuanto a las características propias del contrato de compraventa en la lengua meta (LM) mediante los algoritmos correspondientes que permitan el establecimiento automático de las correspondencias.

Con esta finalidad, se ha compilado un corpus comparable multilingüe de formularios de CCVBI y de legislación civil (y mercantil, en su caso); así como un pequeño corpus paralelo multilingüe de contratos reales. Nuestro método de trabajo ha seguido de cerca los postulados ya clásicos de Baker (1995), Shreve (1993) y Johansson y Hofland (1994). Nuestro corpus textual multilingüe está compuesto, pues, por un *corpus central* de textos en español y sus traducciones hacia y desde las demás lenguas (inglés, alemán, italiano y árabe; el corpus paralelo); y un *corpus complementario* de textos originales en ambas lenguas (el corpus comparable), en el que los distintos subcomponentes están equiparados según similitudes en cuanto al tipo textual o género, el contenido, la finalidad, el tipo de autor y el tipo de receptor al que va dirigido el texto. Se compondría, en principio, de contratos de compraventa de bienes inmuebles, textos de derecho civil sobre el tema y formularios de contratos en español, inglés, alemán, italiano y árabe.

Aunque en un futuro se puede ampliar el análisis a otros tipos de documentos legales, para el proyecto de investigación que describimos nos hemos ceñido a una forma textual concreta: el contrato de compraventa, dentro del registro jurídico. Nuestra decisión ha obedecido a motivos prácticos. En primer lugar, resultaba conveniente realizar un estudio piloto mediante el cual pudiéramos comprobar la utilidad de nuestra investigación y sus limitaciones antes de embarcarnos

en un estudio a gran escala. En segundo lugar, no todos los tipos de textos constituyen un *input* adecuado para los sistemas de traducción por ordenador. Como hemos indicado más arriba, el registro legal, y especialmente si nos ceñimos a un tipo de contrato particular, constituye un tipo de *input* adecuado para tales sistemas. En definitiva, se trata de crear una base de datos que incluya los siguientes tipos de textos:

a) Contratos de compraventa de bienes inmuebles en las distintas lenguas, traducciones modelo y textos paralelos. Se complementaría con textos de referencia que ayuden al traductor a comprender el tema del que se trata, como artículos del código civil o estudios sobre este tipo de contratos.

b) Textos especiales, que consisten en manuales de estilo y, sobre todo, formularios de contratos.

c) Prototipos textuales, es decir, textos preparados artificialmente en soporte informático, elaborados según la superestructura y otros elementos macro y microestructurales típicos del contrato de compraventa en español, inglés, alemán, italiano y árabe.

La recogida de datos se ha realizado de forma semiautomática, en tanto parte de los documentos que integran el corpus proceden de la red Internet, mientras que otros proceden de CD-ROM en forma electrónica y el resto de documentos, en soporte tradicional de papel, han tenido que ser escaneados. Para el análisis de los datos contamos en un primer momento con dos programas de concordancias y cómputo de frecuencias léxicas como son el MicroOCP y el TACT. Posteriormente nos hemos decantado por la suite *Wordsmith's Tools*, dado que este programa de gestión de corpus es mucho más versátil y flexible que los anteriores, además de incluir una opción de alineamiento automático de corpus paralelos. Si la recogida de datos en soporte electrónico no ha supuesto mayores dificultades, la recogida de documentos legales sí ha presentado problemas, en tanto la recogida de originales y traducciones ha dependido casi exclusivamente de la buena voluntad de los despachos jurídicos, asesorías, agencias de traducción e instituciones contactadas.

Como hemos indicado más arriba, una vez analizadas las características léxico-discursivas, fraseológicas y macroestructurales del CCVBI, una vez que éstas se han contrastado con los datos empíricos que arroja el corpus, pasamos a construir el prototipo para cada una de las lenguas implicadas en el proyecto. Conviene señalar que sólo hemos tenido en cuenta las generalidades de este tipo de contrato, dentro del cual se puede incluir la compraventa de viviendas, solares, parcelas, etc., sin abarcar todas sus especialidades, tanto respecto a la persona (por ejemplo, compraventa de bienes de incapacitados, venta a cooperativas de viviendas, etc.), como respecto al objeto (por ejemplo, compraventa con agrupación de fincas, de finca sin título inscrito, etc.) y al precio (por ejemplo, compraventa con precio aplazado, con retención de parte del precio, etc.). Tampoco hemos incluido la elevación a escritura pública de documento privado de compraventa, ciertos pactos especiales (a satisfacción del compra-

dor, posible cláusula penal, pacto de retroventa, etc.), así como los modelos de resolución del contrato y diversas fórmulas de condición resolutoria.

Con objeto de ilustrar nuestra metodología de trabajo, hemos tomado una cláusula de general inclusión en el contrato de CCVBI que se refiere a las partes contratantes y, dentro de ésta, a la parte vendedora cuando se trata de una persona física. Esta subsección podría presentar diversas realizaciones textuales, como se observa a continuación:

[1] *Reunidos*
De una parte, D. Luis López Asturias, en adelante el vendedor, con NIF 29075476-N, casado, en régimen económico de gananciales, mayor de edad, abogado, con domicilio en Málaga, calle Armengual de la Mota, nº 7, portal 5, 3º B, 29007 ...

[2] *Comparecen,*
De una parte, D. Luis López Asturias, en adelante el vendedor, titular del NIF 29075476-N, casado, en régimen de gananciales, mayor de edad, abogado, domiciliado en Málaga, C./ Armengual de la Mota, nº 7, portal 5, 3º B, 29007 ...

[3] *Reunidos,*
De un lado, D. Luis López Asturias, en adelante el vendedor, titular del NIF 29075476-N, casado, en régimen de gananciales, mayor de edad, abogado, vecino de Málaga, con domicilio en Armengual de la Mota, nº 7, portal 5, 3º B, 29007 ...

A partir de este momento, se procede a la abstracción y generalización de la cláusula en un primer nivel, seguida de las distintas secciones que la conforman, de acuerdo con las diversas nociones (jurídicas y generales) en las que éstas se asientan, las cuales hemos representado mediante códigos:

C/P/A	Calle, plaza o avenida
D	Domicilio
EC	Estado civil
L	Lugar
ME	Mayoría de edad
NA	Nombre y apellidos
NIF	NIF
O	Cláusula de general inclusión en el contrato
O2a	Con la intervención aparte
PRO	Profesión
RE	Régimen económico matrimonial

O2a [P] *Reunidos/Comparecen* ***[En medio y separado]

VENDEDOR	*De una parte, D./Dª* PERSONA FÍSICA NA Luis López Asturias**,** (*en adelante, el vendedor)* NIF *con NIF/titular del NIF* 29075476-N,	
De un lado	EC *casado//soltero//viudo//separado//divorciado*	
		RE ***[si está casado] *en régimen (económico) de gananciales// separación de bienes* ***[se hacen constar las capitulaciones matrimoniales si las hay.]
	ME *mayor de edad,*	
		PRO abogado ***[Este campo no es frecuente en los contratos actuales].
		D *con domicilio en/domiciliado en/ vecino de/vecino de ... con domicilio en* L [[Málaga]], C/P/A Armengual de la Mota, nº 7, portal 5, 3º B, 29007

Tras analizar las distintas variantes y posibilidades, se pasa a la construcción del prototipo. Los corchetes indican microunidades dentro de la cláusula <partes contratantes> y, dentro de ésta, la sección <vendedor::persona física>. Cuando la microunidad presenta algún tipo de fijación posicional, viene indicada entre paréntesis (en este caso, la primera microunidad suele aparecer al principio, centrada y separada del cuerpo de la cláusula). A su vez, las distintas variantes sinónimas aparecen reflejadas mediante el uso de una barra oblicua simple. La barra oblicua doble indica que se trata de opciones alternativas, esto es, no sinónimas. El uso de la negrita tras la llave se reserva para indicar que se trata de un dato variable que ejemplifica una determinada subsección dentro de la cláusula. Esta ejemplificación no pretende ser exhaustiva por cuanto sólo hemos tomado las generalidades del CCVBI, sin tener en cuenta sus especialidades ni otras modalidades.

([Reunidos/Comparecen],)
[de una parte/de un lado,]
[D./Dª {**Luis López Asturias**}, (en adelante el vendedor),]
[con NIF/titular del NIF {**29075476-N**},]
[casado, [en régimen (económico) de gananciales//de separación de bienes]//soltero//viudo//separado//divorciado]
[mayor de edad//menor de edad [representado por ...]//{**mayor de edad**},]
[{[[**abogado**]]}]
[con domicilio en/domiciliado en/vecino de/vecino de {[[**Málaga**]]} [con domicilio en] calle/C. //plaza/ Pl. //Avenida/ Avd.//Ø { [[**C./ Armengual de la Mota, nº 7, portal 5, 3º B, 29007 ...**]] }.]

3. Generación automática: la fase de etiquetado

Una vez construido el prototipo del CCVBI en cada lengua, la siguiente fase consiste en la implementación de un sistema de generación automática basado en el etiquetado de las cláusulas (secciones, subsecciones y variables) que, a su vez, alimentan la base de datos central del programa. La *generación de lenguaje natural* (GLN) (en inglés, *Natural Language Generation* (*NLG*)), como indica su propia denominación, consiste en la producción de lenguaje natural a partir de un *input* de naturaleza no lingüística: esto es, se parte de estructuras semánticas para llegar a construir un texto (cf. Reiter y Dale, 2000). Esta es la característica que lo diferencia del *procesamiento de lenguaje natural* (*PLN*), donde se parte del texto para llegar a una estructura semántica subyacente. En realidad, en una primera fase del proyecto PB98-1399, la distinción entre GLN y PLN aparece borrosa, ya que se observa un movimiento pendular que parte del texto para llegar a la macroestructura y, desde ahí, de vuelta al texto. En otras palabras, el análisis del corpus formado por todos los ejemplares de CCVBI nos ha permitido construir un prototipo genérico, que, a su vez, ha servido de base para el desglose semántico de las distintas cláusulas. El alto grado de abstracción inherente a la estructura semántica subyacente (ESS) ha constituido el punto de partida de la gramática base que sustenta nuestro sistema de GLN.

En la actualidad nos encontramos en la fase de generación textual monolingüe. En estos momentos acabamos de diseñar una interlengua única, independiente de las lenguas individuales implicadas en el proyecto, a modo de gramática que sustente la generación textual monolingüe de este tipo de contrato. Una vez diseñada y alimentada la base de datos correspondiente, esta interlengua permite la realización formal de los contenidos proposicionales de cada una de las cláusulas que componen el prototipo contractual en cada lengua. En un primer nivel, nuestro sistema de GLN presupone la selección de unos contenidos más o menos fijos: en este caso, un acuerdo contractual en el que dos partes proceden a la venta de una propiedad inmobiliaria según unas condiciones específicas que se pueden extraer a partir del clausulado correspondiente. En un segundo nivel, dichos contenidos se expresan a través de la elección léxica más adecuada en cada momento. Dichas secuencias léxicas se ordenan posteriormente para ofrecer una frase, cláusula u oración completa en el eje sintagmático, teniendo en cuenta las propiedades de linealidad y referencialidad de las lenguas naturales. De este tercer nivel se pasa al cuarto, en el cual se dota de coherencia discursiva a las oraciones producidas de forma aislada, de manera que se las conecta y vincula convenientemente a fin de producir párrafos completos.

Los cuatro niveles anteriores requieren previamente el diseño de una arquitectura adecuada para la generación automática. Como decíamos más arriba, el generador constituye parte integrante de un *software* de traducción basado en cuatro componentes: bases de datos relacionales (SGBDR), interlengua de representación (*tags*/etiquetas), interfaz de usuario y motor de generación/traducción. Sería muy extenso explicar cada uno de los módulos que integran el sistema, por lo cual nos vamos a centrar en la fase de etiquetado previa a la programación propiamente dicha.

Para la especificación de secuencias léxico-discursivas se suelen seguir dos enfoques básicos: los basados en la gramática sistémica (*Systemic Grammar*) y aquellos que hacen uso de la gramática funcional (*Functional Unification Grammar*). Ambos enfoques no agotan las posibilidades de diseño de la arquitectura de generación, por cuanto también se utilizan otros mecanismos, como los derivados de la inteligencia artificial (*AI-style planning*) o los del tipo pizarra (*blackboard architectures*). En nuestro caso hemos optado por un diseño ecléctico que aprovecha la rigidez relativa de nuestro *input* lingüístico, esto es, un tipo textual con una superestructura más o menos fija que descansa sobre una macroestructura determinada con cambios puntuales en lo que a la elección conceptual se refiere. La gramática del sistema de GLN (la interlengua) está representada mediante un sistema de etiquetado, desarrollado en lenguaje XML (versión 1.0), que contempla el prototipo del CCVBI como una sucesión de cláusulas ordenadas e interrelacionadas, divididas en secciones obligatorias y opcionales, algunas de ellas con fijación discursiva y situacional, que a su vez constan de subsecciones y, dentro de éstas, hay elementos fijos y variables, susceptibles de ser traducidos o no.

CCVBI
CL1 [S1 (OB=f/v || OP=f/v) + S2 (OB=f/v || OP=f/v) + Sn] + CL2 + CLn

Para ilustrar el sistema de notación empleado tomaremos de nuevo la sección referida a la parte vendedora. Como podemos apreciar, hemos determinado una etiqueta que indique el orden de la cláusula en cuestión dentro del contrato: <CLAUSULA N = 02>, que significa que esta cláusula aparece en segundo lugar, tras la cláusula de la fecha y el lugar.

```
<CLAUSULA N = 02>
<CN n=a>
```

Seguidamente se hace referencia a lo que hemos denominado *identidad conceptual*, la cual hemos representado mediante la notación <CONCEPTO ID=”....”>, donde los puntos suspensivos indican el valor del elemento variable. La etiqueta </CONCEPTO> marca el final de la identidad conceptual concreta. En este caso, se trata de un concepto de naturaleza discursiva, por cuanto marca el inicio de la descripción de las partes contratantes, por lo que presenta fijación discursiva y posicional: ambas formas textuales (FT) alternativas aparecen centradas y separadas del resto del cuerpo textual del contrato:

```
<FB>
<CONCEPTO ID=”reuComp”>
<FT n=1 F=CEN> Reunidos </FT>
<FT n=2 F=CEN> Comparecen </FT>
</CONCEPTO>
</FB>
```

Dentro de una misma cláusula se pueden encontrar varias identidades que funcionan a modo de bloques prefabricados, en tanto hay que unirlos para formar secuencias lineales completas. Las etiquetas <FB> y </FB> indican respectivamente el comienzo y el final de una secuencia lineal, es decir, si entre ambas figura una única identidad conceptual, ello significa que se trata de una realización textual más o menos independiente (como ocurre en el caso anterior); si, por el contrario, entre ambas figuran varias identidades, ello significa que están concatenadas, y que, por tanto, su realización textual implica su vinculación y reorganización en contexto, como se aprecia en el ejemplo que sigue a continuación. La secuencia presenta las partes contractuales, tal como se aprecia en la identidad conceptual, cuyo comienzo viene indicado mediante la etiqueta <FB>. Pero son varias las identidades que están en juego, además de la secuencia de presentación de las partes, a saber, el sexo (<CONCEPTO ID =”sexo”>), que a su vez determina la forma textual adecuada (*D./Dª*).

```
<FB>
<CONCEPTO ID=”partes”>
<FT n=1> De una parte, </FT>
<FT n=2> De un lado </FT>
</CONCEPTO>

<CONCEPTO ID=”sexo”>
<FT n=1> D. </FT>
<FT n=2> Dª </FT>
</CONCEPTO>
```

A partir de ahí se selecciona el bloque de identidad conceptual: la parte vendedora es una persona física (<BLOQUE ID=”perFisicaV”>). Ello, a su vez, repercute en la selección simultánea de identidades conceptuales, que, en conjunto, difieren de las que, por defecto, se seleccionarían en el caso de tratarse de una persona jurídica. Las etiquetas <NB n= ...> y </NB> señalan que se trata de una selección en bloque. Las identidades que se contemplan son la de identificación (nombre y apellidos), calidad de la contratación (vendedor/ comprador), su número de identificación fiscal, estado civil (régimen matrimonial, en su caso), mayoría de edad, domicilio, etc.

<BLOQUE ID=”perFisicaV”> <NB n=1> <CONCEPTO ID=”nombreV”> -- la BD <FT n=1 T=1+msv> %Nom, </FT> </CONCEPTO> <CONCEPTO* ID=vend”> <FT n=1> en adelante el vendedor </FT> </CONCEPTO> <CONCEPTO ID=”nif”> <FT n=1> con NIF %NIF, </FT> <FT n=2> titular del NIF %NIF, </FT> </CONCEPTO> <CONCEPTO ID=”eC”> <FT n=1 T=1+mdu > %EC < /FT> </CONCEPTO>	<FT n=1> en régimen económico de </FT> </CONCEPTO> <CONCEPTO* ID= “ganSep”> <FT n=1 T=1+olu > gananciales %C </FT> <FT n=2 T=1+olu > separación de bienes %C </FT> </CONCEPTO> <CONCEPTO ID=”me”> <FT n=1> mayor de edad </FT> </CONCEPTO> <CONCEPTO ID=”profV”> <FT n=1 T=1+mdu > %PROF </FT> </CONCEPTO>	<CONCEPTO ID= “domic”> <FT n=1> con domicilio en </FT> <FT n=2> vecino de </FT> </CONCEPTO> <CONCEPTO ID= “domic”> <FT n=1> con domicilio en </FT> <FT n=3> domiciliado en </FT> <FT n=2> vecino de </FT> <FT n=3> domiciliado en </FT> </CONCEPTO> <CONCEPTO ID=”ciudVen”> <FT n=1 T=1+mdu > %L </FT> </CONCEPTO>	<CONCEPTO ID=”direcV”> <FT n=1 T=1+msu > CALLE %NC </FT> <FT n=2 T=1+msu > PLAZA %NC </FT> <FT n=3 T=1+msu > AVENIDA %NC </FT> </CONCEPTO> </NB>

Posteriormente a la validación del *software* monolingüe, la siguiente fase consistirá en vincular los distintos campos de las bases de datos multilingües con objeto de alcanzar la generación multilingüe. Es decir, el tipologizador permitirá: *a)* producir un texto meta (TM) a partir del texto de origen (TO), en cualquier dirección (hacia o desde el español); o bien, *b)* producir varios textos originales, paralelos entre sí, automáticamente (generación textual automática). Se trata de ofrecer al usuario una herramienta de traducción que le permita, por un lado, identificar el subtipo de contrato de que se trate; y, por el otro, que le muestre una plantilla con un texto equivalente en la(s) otra(s) lengua(s) en la que sólo tenga que rellenar las casillas vacías referentes a los datos específicos del contrato en cuestión, tales como los nombres de las partes intervinientes, la fecha, el lugar, la identificación del bien o bienes inmuebles y las condiciones de compra (precio, formas de pago, etc.).

El sistema de etiquetado que forma la base del tipologizador, o generador multilingüe, permite la equiparación interlingüística automática, donde cada sección posee una selección de equivalentes alternativos determinados, extraídos de nuestro análisis basado en corpus. Dichos equivalentes forman también el esqueleto formal del prototipo de CCVBI en cada una de las cinco lenguas implicadas en el proyecto (véanse los informes y análisis contenidos en Corpas 2003/ en prensa). Diversas dificultades de índole teórica y técnica han hecho imposible hasta la fecha el desarrollo de un *software* de traducción automática en el sentido estricto del término. En cualquier caso, conviene tener presente que la generación automática multilingüe se presenta hoy día como una de las áreas de investigación más prometedoras, que ya empieza a desbancar a la

traducción automática, y muy especialmente en el ámbito de la traducción técnica especializada.

4. Conclusión

La investigación en traducción especializada pasa por el estudio de las diversas tipologías textuales que conforman un determinado género. Ésta es, por ejemplo, la filosofía que subyace al proyecto PB98-1399 de la DGICYT *Diseño de un tipologizador textual para la traducción automática de textos jurídicos (español ↔ inglés/alemán/italiano/árabe)*. Y lo mismo se puede decir de otros proyectos de reciente concesión, como el proyecto coordinado con las Universidades de Las Palmas, Vigo, Salamanca y Málaga, que dirige Zinaida Lvovskaya (MCYT, ref. BFF2000-0512-C04-03); el dirigido por I. García Izquierdo en la Universidad Jaume I (ref. BFF2002-01932, MCYT); o el coordinado por R. Rabadán, en la Universidad de León, y J. Mª Bravo Gozalo, en la Universidad de Valladolid/ ITBYTE (MCYT, ref. BFF2001-0112). El primero se ocupa de analizar las convenciones de los textos de especialidad de diversas lenguas desde el análisis del discurso y la estilística textual; mientras que el segundo aspira a compilar un corpus multilingüe comparable de géneros textuales de lenguas de especialidad (jurídico, médico y económico), con vistas a etiquetar su microestructura y realizar estudios discursivos, fraseológicos y textuales; y el tercero estudia el discurso especializado y su traducción mediante entornos informáticos.

Otra pieza clave en el desarrollo del proyecto es la compilación y explotación de corpus multilingües, muy en la línea otros proyectos de investigación, como el *OncoTerm: Sistema Bilingüe de Información y Recursos Oncológicos* (PB 98-1342), que dirige P. Faber en la Universidad de Granada. Si bien nos hemos ocupado de este tema en otro lugar (*cfr.* Corpas Pastor, 2002), nos interesa destacar el uso conjunto de corpus paralelos (formado por textos originales y su traducción a otra lengua) y corpus comparables (formados por textos originales equiparados en cuanto al género, la tipología textual, el tema, la función y los límites diasistemáticos), donde se han incluido textos preparados artificialmente (formularios) y textos reales, descargados de la red Internet a modo de gigantesca base de datos.

Los datos procedentes de ambos presupuestos han alimentado las bases de datos que integran nuestro sistema de generación textual multilingüe. A lo largo de estas páginas sólo hemos mostrado brevemente el sistema de etiquetado utilizado y la interlengua de representación. Como decíamos más arriba, nos encontramos en una fase de implementación del *software* de generación monolingüe y, en cuanto éste se encuentre listo, procederemos a la programación del *software* de equiparación interlingüística y generación multilingüe, con objeto de comprobar la validez de este generador piloto y pulir cuantas inconsistencias hubiéramos detectado durante el proceso de evaluación técnica. Es nuestra intención utilizar también los corpus compilados como referente externo y objetivo de la evaluación y comprobación del sistema de generación.

Tras validar y probar el generador piloto estaremos en condiciones de hacerlo extensivo para las distintas especialidades del contrato base (el CCVBI), así como para otras modalidades de contratación civil o mercantil, por citar sólo algunas de las posibilidades que este tipo de estudio, interdisciplinar y multilingüe, ofrece al investigador en la lingüística del corpus, la lingüística del texto, el procesamiento de lenguaje natural y la traductología.

El presente trabajo ha sido realizado en el seno del proyecto PB98-1399 «Diseño de un tipologizador textual para la traducción automática de textos jurídicos (español↔ inglés/alemán/italiano/árabe)» (DGICYT, 1999-2002).

Referencias

BAKER, M. 1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*. 7 (2). 223-243.

BIBER, D. 1995. *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

BIBER, D., S. CONRAD y R. REPEN. 1998. *Corpus Linguistics. Investigating Language Structure and Use*. Series: Cambridge Approaches to Linguistics. Cambridge, Melbourne: Cambridge University Press

CORPAS PASTOR, G. 2002. Utilización de corpus multilingües en traducción: introducción al tipologizador textual automático para textos jurídicos. En S. GAMERO y A. ALCINA, eds. *La traducción científico-técnica y la terminología en la sociedad de la información*. Castellón: Servicio de Publicaciones de la Universidad Jaume I, pp. 155-162

CORPAS PASTOR, G. 2003/En prensa, ed. *Recursos documentales y terminológicos para el estudio del discurso jurídico: el contrato de compraventa de bienes inmuebles (español, alemán, árabe, inglés, italiano)*. Granada: Comares.

GARSIDE, R., G. LEECH y A. MCENERY, eds. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Londres: Longman.

JOHANSON, S., K. HOFLAND. 1994. Towards an English-Norwegian parallel corpus. In FRIES, U., G. TOTTIE y P. SCHNEIDER, eds. *Creating and using English language corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora*. Zürich, Amsterdam: Rodopi, pp. 25-37.

REITER, E., & M. DALE. 2000. *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press.

STUBBS, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.