[Ponencia originalmente leída en el seminario "La ingeniería lingüística en la sociedad de la información", Fundación Duques de Soria. Soria, 17-21 de julio de 2000. Posteriormente publicada en M. A. Martí y J. Llisterri.2002. *Tratamiento del lenguaje natural*. Edicions Universitat de Barcelona: 61-90]

Tratamiento de corpora bilingües

Joseba Abaitua
www.deli.deusto.es
Universidad de Deusto

Resumen

Los corpora bilingües son una fuente inagotable de recursos lingüísticos útiles para el desarrollo de aplicaciones como la lexicografía, la terminografía, la traducción automática, la enseñanza de segundas lenguas, la edición plurilingüe y la búsqueda translingüística de información. El avance de la tecnología de tratamiento de corpora de lengua escrita ha coincidido con el desarrollo de la tecnología *web* en Internet. Esta coincidencia ha propiciado una confluencia de estrategias en los dos campos, lo que se ha traducido en un interés común hacia los mecanismos de codificación y anotación, así como hacia la explotación de los contenidos. El plurilingüismo de Internet y los corpora multilingües son el tema principal de esta ponencia.

0 Introducción

En la edición navideña de 1998 de *Language International* 10-6, en lo que era equivalente a una carta a los Reyes Magos, Eduard Hovy, presidente de la *Association for Machine Translation in the Americas* (AMTA), pedía dos cosas: un *gusano* que recorriera Internet recogiendo los vocablos nuevos que fueran surgiendo, tanto en su versión original como en sus traducciones; y un reconocedor de géneros y de "tipos" de texto. Con ello, estaba señalando dos de los problemas más acuciantes de la tecnología del lenguaje, la diversidad lingüística y la dispersión documental, y además anticipaba cuál iba a ser su principal campo de aplicación, Internet.

Una de las líneas de trabajo más productivas en el contexto actual de la ingeniería lingüística es el tratamiento de corpora multilingües. La importancia de esta línea de trabajo está refrendada precisamente por el papel que ha adquirido Internet como vehículo de comunicación y depósito de información a escala planetaria. La red se ha convertido en el primer destinatario de aplicaciones lingüísticas, a la vez que en la más importante fuente de recursos. Es por ello que muchos estudiosos del lenguaje contemplan Internet como un inmenso corpus de información lingüística.

El primer apartado de la ponencia está dedicado a la lingüística de corpus, que es el área de la lingüística especializada en el aprovechamiento de los corpora. El segundo apartado presentará los corpora bilingües y multilingües, con sus características y variedades, y las razones por las que se cotizan al alza (utilizaremos indistintamente los términos "multilingüe" y "plurilingüe"). El tercer apartado aborda las técnicas de tratamiento. El cuarto expone los tipos de anotaciones. El quinto y sexto se dedican a las tareas de segmentación y de alineación. De las aplicaciones se habla en el séptimo y por último se aportan algunos datos que ilustrarán la importancia del plurilingüismo en Internet.

1 La lingüística de corpus

Como paso previo a la discusión sobre los corpora multilingües, en este apartado se va a reseñar brevemente el enfoque metodológico de la lingüística de corpus. Las obras de McEnery y Wilson

1996, o Pérez Guerra 1998 son dos introducciones muy recomendables para el lector que desee ampliar los datos aquí presentados.

Los primeros estudios basados en corpus se remontan a los años del estructuralismo prechomskyano, que consideraba el acopio de datos una tarea esencial del análisis lingüístico. Suele citarse la gramática de Fries 1952 como ejemplo pionero de aprovechamiento de datos extraídos de textos reales, con un adelanto de más de treinta años sobre la obra *A comprehensive grammar of the English language* (de Quirk y otros, 1985). Pero la revolución del generativismo, a finales de los cincuenta, hizo que objetivos y metodologías del estructuralismo se cuestionaran. Chomsky identificó las limitaciones del corpus para explicar el carácter productivo del lenguaje y propició un cambio radical en la forma de estudiar la gramática. La lingüística abandonó los métodos empiristas basados en la observación de los datos, para promover la prospección introspectiva. Durante las décadas de los sesenta y setenta, la escuela generativa ha protagonizado la escena lingüística, desplazando los trabajos de la lingüística de corpus a un segundo plano.

Sin embargo, a finales de los ochenta algunos teóricos comenzaron a alzar la voz alertando de las limitaciones de los modelos racionalistas. Birdsong 1989, por ejemplo, cuestionó las deficiencias de algunos análisis gramaticales basados únicamente en intuiciones, a menudo imprecisas y mal contrastadas, que no consideraban los datos empíricos observados en textos reales. Pero es sobre todo en áreas aplicadas, como la lingüística computacional, donde, pese al éxito en el desarrollo de modelos formales útiles para tratar computacionalmente muchos aspectos del lenguaje, comienzan a notarse las limitaciones del enfoque racionalista.

Los formalismos gramaticales de los ochenta (FUG, LFG, HPSG, etc.) basan su potencial en información de tipo simbólico, y carecen de capacidad para expresar datos relacionados con la frecuencia o la probabilidad. Pese a algunos intentos para combinar ambos tipos de información, los resultados han tenido poca difusión fuera del laboratorio. Entre tanto la industria necesita aplicaciones para textos reales, y eso requiere el desarrollo de gramáticas muy complejas. La enorme cantidad de variantes oracionales y sintagmáticas del lenguaje escrito es un escollo que complica esta tarea. Muchos sistemas acaban acumulando un número excesivo de reglas, lo que irremediablemente conduce a la redundancia, a la explosión combinatoria (de opciones alternativas - con devastadores efectos sobre la resolución de ambigüedades-), así como a la aparición de inconsistencias y contradicciones. Otra fuente de preocupación ha sido el reducido tamaño de los diccionarios en la mayoría de los sistemas desarrollados y la baja cobertura de texto real que ello implica.

Movidos por estas inquietudes, desde mediados de los ochenta distintos colectivos han decidido recopilar y preparar para su aplicación colecciones de corpora. Con ellos se pretende conocer mejor la realidad de los textos que se van a tratar. Merecen ser destacadas las siguientes iniciativas:

- en EEUU
    - *Association for Computational Linguistics/Data Collection Initiative* (ACL/DCI)
    - *Linguistic Data Consortium* (LDC)
    - *Consortium for Lexical Research* (CLR)
- en Europa
    - *European Language Resources Association* (ELRA)

La utilidad de los corpora depende de los criterios que se aplican en el momento de selección y compilación de los textos. Varios autores (Atkins y otros 1992, Biber 1993, o McEnery y Wilson

1996) subrayan la influencia que estos criterios van a tener sobre el corpus resultante. Los corpora pueden ser de tipos muy variados. Una primera clasificación depende de la propia naturaleza física de los datos. Marcos Marín 1994 distingue tres clases:

- *Corpus oral*: contiene sonidos, material fonético sin transliteración que sirve sobre todo para trabajos específicos de síntesis y reconocimiento de habla.
- *Corpus de lengua hablada*: contiene transliteraciones de textos grabados del registro oral.
- *Corpus de lengua escrita*: contiene textos pertenecientes a todas las modalidades de lengua escrita, incluyendo la comercial, publicitaria, escolar y literaria.

Cada una de estas clases cumple funciones distintas y requiere por ello de tecnologías muy diferentes para su aprovechamiento. Aquí vamos a abordar la problemática de los corpora multilingües en la variedad de lengua escrita, que es la más habitual.

Otro criterio de clasificación tiene que ver con el género y la tipología textual. Este criterio afecta fundamentalmente a las propiedades de representatividad del corpus. Los criterios de selección serán muy diferentes según se pretenda diseñar un corpus especializado, como el *Aarhus Corpus* - sobre derecho contractual europeo -, o se desee crear un corpus de referencia, con una cobertura amplia de estilos y registros. Por definición, los corpora de referencia abarcan, de la manera más exhaustiva posible, todos los aspectos relevantes sobre una lengua. Es el caso notorio del *British National Corpus* (BNC), que sobrepasa los 100 millones de palabras (90% de lengua escrita y 10% de lengua hablada).

Es fácil confundir nociones como "dominio de especialidad", "campo temático", "categoría textual" o "género". Todas tienen que ver con grupos textuales que el compilador de un corpus debe considerar. Biber y Finegan 1986 y también Nakamura 1991 aplican la noción de género para distinguir los textos por su función pragmática: novela, artículo periodístico, ensayo, etc., es decir, atendiendo a factores extralingüísticos. La noción de "tipo" de texto la emplean para distinguir los textos según las propiedades lingüísticas relacionadas con aspectos cuantitativos: longitud de oraciones, utilización de perífrasis verbales, densidad léxica, uso de conectores, etc. Laviosa 1998, por ejemplo, ha encontrado diferencias cuantitativas importantes entre textos traducidos y originales en un corpus de inglés.

La tabla 1 muestra los criterios recomendados para el diseño del corpus de referencia del español, según la información proporcionada por Marcos Marín 1999 (y se comparan con los resultados del subcorpus elaborado en Argentina).

Porcentaje recomendado en %	**Género**	**C. argentino %**
10-15	Científico	16,11
8-12	Comercial	3,25
15-20	Escolar	9,20
5-6	Humanístico	21,66
5-6	Jurídico	6,30
5-10	Literario	9,09
20-25	Periodístico	28,00
5-6	Publicitario	-
10-15	Técnico	6,79
nº de palabras del corpus argentino: 2.008.969		

Tabla 1 Porcentaje de textos según géneros para corpus de referencia

De los corpora de lengua española, destacan los de la Real Academia de la Lengua (RAE), ocupada desde 1995 en compilar y anotar un corpus de carácter histórico y otro de referencia del español actual:

1. El Corpus Diacrónico del Español (CORDE) contiene textos de tres épocas fundamentales, Edad Media, Siglos de Oro y Época Contemporánea, y pretende ser representativo del español a lo largo de su historia.
2. El Corpus de Referencia del Español Actual (CREA) cubre veinticinco años desde 1975 hasta 1999.

Cada corpus contiene 125 millones de palabras, e intenta representar todos los territorios de habla hispana, tanto peninsulares como extrapeninsulares. A estos dos proyectos hay que sumar lógicamente algunos más.

Varias editoriales disponen también de corpora de español: Vox Bibliograf posee uno de 10 millones de palabras, la editorial SGEL otro de 8 millones (denominado CUMBRE, Sánchez y Cantos 1997) y la editorial SM otro de 60.000 palabras (Pérez Hernández y otros 1999). En Cataluña , el *Institut d'Estudis Catalans*, así como TERMCAT y *l'Institut Universitari de Lingüística Aplicada* (IULA) de la Universitat Pompeu Fabra (Bach y otros 1997) han compilado corpora del catalán; en el País Vasco *Euskaltzaindia* y UZEI, así como las universidades del País Vasco y Deusto, disponen de distintas colecciones en euskara; y en Galicia la Academia de la Lengua, ayudada por el *Centro de Investigacións Ramón Piñeiro* (Magán 1996), está estudiando el etiquetado de su propios recursos textuales. El principal objetivo de todos estos esfuerzos es su aplicación en el campo de la lexicografía y terminografía. A partir de ahora vamos a centrarnos en los corpora multilingües.

2 Tipos de corpora multilingües

Con la rara excepción de Eaton 1940, la mayor parte de los corpora que se han recopilado en el mundo hasta fechas recientes han sido monolingües. Sin embargo, la apreciación de los corpora multilingües aumenta cada día, debido sobre todo a la riqueza de información y posibilidades de aprovechamiento que aportan. Es importante distinguir entre dos tipos*:*

- *Corpora de textos en distintos idiomas*. Colecciones de textos en varios idiomas recopiladas con la intención de servir para estudios cuantitativos o estadísticos. Los criterios de selección pueden ser muy diversos, desde la simple disponibilidad de los textos, hasta la selección según géneros y tipos similares (pero sin llegar a ser comparables). Un ejemplo es el *Multilingual Corpus* de la *European Corpus Initiative* (D. McKelvie y H.S. Thompson, 1994).

- *Corpora comparables*: Baker 1995 introdujo este término para corpora monolingües compuestos por textos originales en una lengua y traducciones de otros textos semejantes en la misma lengua. Martínez 1999 amplía el término a corpora multilingües que contienen textos en distintos idiomas, que sin ser traducciones, comparten similar origen, temática, extensión y número: partes meteorológicos, ofertas laborales, artículos

periodísticos, etc. Es decir, que los textos no se reúnen de manera arbitraria, sino que se escogen de acuerdo con unos criterios de selección comunes (Hallebeek 1999). Es el caso del *Corpus Aarhus*, compuesto por textos de derecho contractual en danés, francés e inglés; o también la colección bilingüe de textos chinos e ingleses que Fung 1995 utiliza para generar diccionarios bilingües.

- *Corpora paralelos*: se aplica a corpora que contienen la misma colección de textos en más de una lengua, es decir, cuando a las versiones originales les acompañan sus traducciones. El caso óptimo de paralelismo se produce cuando las traducciones son un reflejo simétrico de la versión original. El caso más conocido es el *Hansard Corpus*, que son actas del parlamento canadiense publicadas en francés e inglés.

Existen otras consideraciones relacionadas con la traducción que se deben considerar. Las traducciones pueden responder a tipologías muy distintas. Así, por ejemplo, los textos del *China News Service*, que Xu y Tau 1999 alinearon como base de un sistema de traducción asistida, o los resúmenes en inglés y japonés del *National Center for Sience Information Systems* (NACSIS), que Kando y Aizawa 1998 utilizaron para probar la recuperación translingüística de información, tienen muy poco que ver con las actas en francés e inglés del *Hansard*. Tampoco es comparable una traducción de una novela de Julio Verne o Ken Follett con las traducciones que los autores hacen de su obra (como las traducciones al castellano que Pere Guinferrer hace de sus poemas en catalán, o la traducción al castellano que Bernardo Atxaga hizo de su colección de relatos *Obabakoak*). Existen más de veinte versiones españolas de *Romeo y Julieta*, cada una con su rasgos y particularidades.

Para valorar adecuadamente los factores que afectan a la categorización de un corpus bilingüe, es pertinente considerar aspectos que han sido estudiados en traductología y que están relacionados con conceptos como *status*, *función*, o las distintas dimensiones de *equivalencia*. Sager 1993 ha introducido la noción de "status" para describir la dependencia del texto traducido respecto al original. Propone tres tipos:

- *Tipo A*: cuando los textos traducidos son autónomos y sustituyen a los originales en la lengua de llegada, pudiendo incluso desempeñar una función distinta. Es el caso más normal: traducciones de novelas de autores como Agatha Chirstie, Tom Clancy o Stephen King.

- *Tipo B*: cuando las traducciones complementan al texto original, coexistiendo en el tiempo y en el espacio con él. El mejor ejemplo son las ediciones bilingües de obras literarias.

- *Tipo C*: cuando las traducciones intentan reflejar de manera simétrica el texto original y mantienen la misma función. Además de los típicos documentos institucionales bilingües o multilingües, como las actas del Hansard, se considerarían de este tipo las llamadas "traducciones canónicas", como la traducción inglesa de la *Biblia* del Rey Jacobo, el *Hamlet* de Moratín, o la versión inglesa del Guzmán de Alfarache realizada por James Mabbe en 1662 (Rabadán 1994).

Otro aspecto importante es la "función" del texto traducido. Según Rabadán 1994, la función la determina la intencionalidad comunicativa del traductor, quien puede actuar movido por alguno de los siguientes objetivos:

1. Presentar un contenido temático, un argumento, una historia, un relato (es la función más común, como en las novelas policíacas de Dashiel Hammet o Raymond Chandler).
2. Presentar el estilo y el punto de vista del autor original (adaptaciones de Borges de la poesía anglosajona, las traducciones de Ezra Pound de los clásicos griegos y latinos).
3. Introducir elementos culturales o tecnológicos nuevos en la sociedad destinataria de la traducción (traducciones técnicas, traducciones de obras de culturas "exóticas").
4. Introducir nuevas formas literarias y textuales en la lengua de llegada (traducciones de sonetos italianos de Boscán y Garcilaso, la traducción de la *Biblia* de Lutero en 1534, que supuso la normalización del *Hochdentch* como forma estándar del alemán).
5. Facilitar la comprensión del texto original por medio de la traducción (obras bilingües de poesía).
6. Recrear la obra original en un texto nativo nuevo (el *Teatro Nuevo Español* de 1800, la versión de Edward Fitzgerald del *Rubáiyat* de Omar Kayyan).
7. Difundir o reforzar una ideología literaria, filosófica, política o religiosa (la versión de *Macbeth* de Michel Garneau en 1978, que sirvió para crear un modelo de teatro nacional *québécois* y legitimar las aspiraciones de independencia de este estado francófono de Canadá).

Otro factor clave en la traducción es el valor de equivalencia. Nord 1994 propone tres dimensiones de equivalencia:

- *Pragmática*: Cuando el original y su traducción comparten la misma función, el mismo efecto comunicativo y van dirigidos al mismo grupo de receptores.
- *Estilística*: Cuando la traducción intenta reflejar la forma y belleza del original.
- *Semántica*: Cuando el texto traducido transmite el mismo mensaje o tiene el mismo significado que el original.

Según el tipo de traducción, tendrá más sentido considerar una dimensión u otra de equivalencia. Así, en la traducción de *Obabakoak* de Bernardo Atxaga, tiene prioridad la equivalencia estilística frente a la semántica. Por contra, en un documento jurídico bilingüe o multilingüe, como es la *Constitución* española, o cualquier otra normativa europea, la correspondencia semántica deberá ser fiel, casi literal, respecto al original.

3 Tratamiento de corpora multilingües

A los niveles típicos de procesamiento monolingüe (análisis morfológico, lematización, desambiguación, análisis sintáctico) se añade en los corpora multilingües un tipo de tratamiento particular mediante el que se establecen las equivalencias entre las unidades de textuales. Esta fase recibe distintos nombres: emparejamiento, correspondencia, alineación (el más utilizado). La alineación es el proceso que mayor valor añadido aporta a un corpus multilingüe.

La forma más común de marcar los resultados del procesamiento es mediante etiquetas (*tags*), códigos (*codes*) o anotaciones (*annotations*) -estos tres términos suelen utilizarse de manera indistinta. Las anotaciones suponen un mecanismo importante en el tratamiento de los corpora y,

salvo para algunos estudios cuantitativos, resultan prácticamente imprescindibles.

3. 1 Estudios cuantitativos

Los corpora monolingües se han utilizado durante años como fuente de datos cuantitativos con aplicación en la lexicografía (generación de lexicones, comprobación de frecuencias, estudio de colocaciones, concordancias, etc.), en filología (verificación de la autoría de una obra, descripción del estilo, etc.), además de en otras disciplinas cercanas: lingüística cuantitativa, lingüística diacrónica, dialectología, psicolingüística, psicología social, sociolingüística, etc. (McEnery y Wilson 1996).

Podemos ilustrar brevemente la utilidad de los datos cuantitativos con algunos resultados obtenidos por Laviosa 1998. Se trata de una comparación de textos originales y traducciones, que pretendía identificar los hábitos de escritura de los traductores. El estudio se realizó sobre un corpus comparable en inglés, compuesto de textos periodísticos y prosa narrativa con un tamaño de unos 2 millones de palabras, de los cuales la mitad eran textos originales y la otra mitad traducciones. Los datos cuantitativos mostraron que en las traducciones había una proporción menor de palabras léxicas frente a funcionales, con independencia de cuál hubiera sido la lengua de la que se había traducido. Asimismo se observaba que las 108 palabras más frecuentes, o lista nuclear (*list head*), se repetían más a menudo, que las palabras menos frecuentes variaban menos y que el tamaño medio de las oraciones era menor en las traducciones. El estudio también indicaba un uso distinto de los auxiliares. En conjunto, las traducciones mostraban un *densidad* léxica (Stubbs 1996) menor, es decir, una proporción de palabras funcionales muy alta en relación al total. En definitiva, los resultados de Laviosa concluían que es posible distinguir entre traducciones y originales a partir únicamente de datos cuantitativos.

Con todo, los estudios cuantitativos son más productivos si se realizan sobre corpora anotados. Algunos datos cuantitativos - como la longitud media de las oraciones- requieren un proceso previo de segmentación/anotación. La oración como unidad ortográfica contenida entre dos signos de punto es un concepto que se apoya en las convenciones de la escritura, que pueden variar de una lengua a otra. En árabe por ejemplo los signos de punto se utilizan para distinguir párrafos, mientras que las oraciones se representan coordinadas mediante conjunciones. Las convenciones de escritura presentan dificultades de reconocimiento también para lenguas occidentales, como el inglés o el español, cuyas normas más simples son también ambiguas. El signo de punto, por ejemplo, no sólo señala el límite de la oración, sino que se utiliza también en acrónimos, cifras, iniciales, nomenclaturas, etc. Por ello las anotaciones resultan en la práctica imprescindibles para el aprovechamiento adecuado de los corpora.

3. 2 Anotaciones

Los datos lingüísticos inherentes a un corpus adquieren mayor valor cuando se explicitan mediante la incorporación de anotaciones metalingüísticas y extralingüísticas. Durante décadas han convivido distintos sistemas de anotación, con propiedades y fisonomías de muy diversa índole. En los noventa la situación ha ido paulatinamente regularizándose. Son conocidas, en este sentido, las recomendaciones de Leech 1993, que se resumen en estas siete máximas:

1. Facilitar la eliminación de las anotaciones, de forma que sea posible recuperar la versión original de los textos.

2. Permitir la extracción de las anotaciones por sí mismas, de manera que puedan constituir

una base de conocimientos autónoma, independiente del texto al que se deben.

3. Distribuir las normas en las que se basan las anotaciones para que los usuarios finales puedan interpretarlas sin dificultad.

4. Indicar el procedimiento por el que se introdujeron las anotaciones en los textos y las personas responsables del proceso.

5. Alertar sobre la posibilidad de que el corpus anotado contenga errores. La anotación de un corpus es un acto de interpretación de estructuras y de contenidos y no es infalible.

6. Permitir la más amplia funcionalidad y reutilización del corpus acudiendo a propuestas con mayor aceptación y neutras en lo posible respecto a formalismos o teorías gramaticales concretas.

7. Admitir la existencia de otras normas y estándares de anotación.

La última máxima de Leech refleja la situación anterior a la publicación de las directrices del *Text Encoding Initiative* (Sperberg-McQueen y Burnard 1994), que una parte mayoritaria de la comunidad lingüística ha adoptado ya como propuesta de anotación estándar. TEI ha sido respaldada por colectivos de gran peso, como son la *Association for Computational Linguistics* (ACL), la *Association for Computers and the Humanities* (ACH), la *Association for Literary and Linguistic Computing* (ALLC), o la *Modern Languages Association* (MLA). Vamos a repasar los logros más importantes de esta y otras iniciativas de normalización:

- TEI fue creado en 1987 con el objetivo principal de proporcionar directrices que facilitasen el flujo de corpora y de herramientas entre la comunidad científica. Para el etiquetado TEI adoptó SGML (*Standard Generalized Markup Language*), que desde 1986 estaba registrado como norma ISO 8879 en el epígrafe para la documentación electrónica. Los esfuerzos de TEI han sido posteriormente ampliados por otras iniciativas.

- El grupo EAGLES (*Expert Advisory Groups on Language Engineering Standards*), que ha sido auspiciado por la Comisión Europea, ha concretado los contenidos de algunas propuestas TEI, añadiendo criterios para la codificación de un amplio abanico de cuestiones lingüísticas (que abarcan desde rasgos fonéticos hasta cuestiones de pragmática y discurso).

- En paralelo con EAGLES, otro proyecto comunitario, PAROLE, se ha centrado sobre todo en la creación de recursos léxicos en varias lenguas europeas. Uno de los mayores logros de PAROLE ha sido la propuesta de un conjunto homologado de etiquetas morfosintácticas para varias lenguas europeas (inglés, danés, neerlandés, francés, italiano, catalán, español, entre otras).

- Otro proyecto europeo destacable es MULTEXT (*Multilingual Text Tools and Corpora*) que ha desarrollado programas modulares para la segmentación y etiquetado de corpus en varias lenguas europeas. Algunas de sus herramientas (como el segmentador MtSeg) han sido reutilizadas en otros proyectos (CRATER y CREA).

- CES (*Corpus Encoding Satandards*) ha ampliado la cobertura de las anotaciones y ha abarcado un numero mayor de lenguas. Entre sus principales logros está haber recopilado

y anotado corpora multilingües con lenguas de Europa oriental.

- CRATER es otro proyecto comunitario que ha permitido la construcción de un corpus trilingüe (inglés, francés y español) especializado en el campo de las telecomunicaciones y que ha sido anotado con etiquetas morfosintácticas, lematizado y alineado.

- Otro logro comunitario ha sido el *Multilingual Corpus I* del *European Corpus Initiative* (ECI/MCI), que en 1994 puso a disposición de la comunidad científica una colección de 98 millones de palabras con textos de 27 lenguas europeas.

4 Tipos de anotaciones

La aportación principal de estas iniciativas, más que la propia compilación de corpora, reside en la metodología que se aplica en el etiquetado. Debemos considerar tres tipos de anotaciones:

- Anotaciones con información extralingüísitca
- Anotaciones tipográficas
- Anotaciones lingüísticas

4. 1 Información extralingüística

En el diseño de un corpus es importante tener registrados adecuadamente todos los datos relacionados con el origen de los textos: la lengua, la datación, el autor, la edición, el trancriptor, fecha de cada una de revisiones, el dominio al que pertenece; y cuanta otra información que pueda ayudar a catalogar el corpus: el género, el tipo, el status, la función, etc. Todos estos datos suelen agruparse en lo que se denomina "cabecera" (*header*) del corpus. TEI ha desarrollado varios modelos de cabeceras.

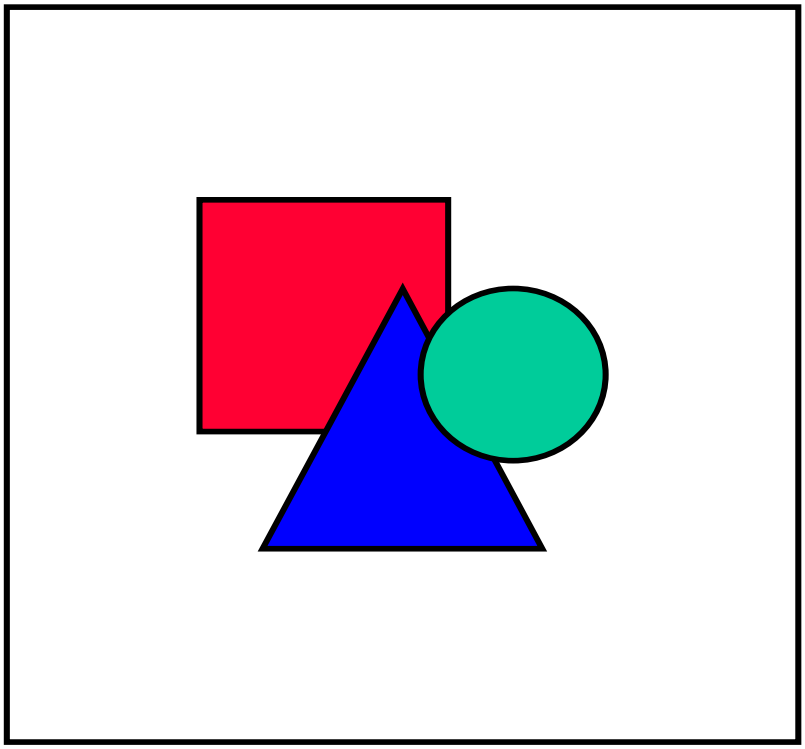
4. 2 Problemas tipográficos

Durante algún tiempo muchos sistemas informáticos sólo han sido capaces de tratar sistemas de escritura cuyos signos pertenecieran a la tabla ASCII (*American Standard Coding for Information Interchange*), es decir los mismos caracteres que el inglés. Si se hacían transcripciones de otros idiomas, había que adoptar tablas de caracteres distintas, que no eran intercambiables entre sí (como JUNET para el japonés, o el ASCII extendido para las lenguas latinas, con versiones incompatibles según pertenecieran a un sistema operativo u otro). Internet ha propiciado una vía de solución:

- El protocolo HTTP (Yergeau y otros 1997) para cuestiones de transferencia de archivos: fue diseñado para permitir la transmisión de metainformación sensible al idioma. La norma RFC 2068 de la versión HTTP 1.1 contempla la codificación de caracteres y la negociación "lingüística" cliente-servidor. De acuerdo con RFC 2068, la codificación de los caracteres se indica mediante un parámetro en el campo de cabecera. Un archivo en japonés codificado en JUNET, por ejemplo, contendrá en la cabecera los atributos Content-type: text/html; charset=iso-2022-JP. El cliente puede indicar preferencia por una determinada codificación (Accept-Charset) o idioma (Accept-Language).

- La norma RFC 2070 para cuestiones relacionadas con los conjuntos de caracteres. La norma RFC 1886, adoptada en las primeras versiones de HTML, restringe el conjunto de caracteres al conjunto del ISO-8859-1 o ISO-Latin-1, que sólo sirve para lenguas con el

alfabeto latino. ISO-Latin-1 es de 8-bits, de forma que permite un máximo de 256 caracteres. La norma posterior RFC 2070 ha añadido propiedades a HTML para soportar documentos en distintos idiomas. Para ello ha reemplazado el ISO-Latin por el ISO 10646 de 1993, conocido como UCS (*Universal Universal Character Set*) y que coincide en todo con otra norma, UNICODE 1.1. UNICODE es un sistema de 16-bits con capacidad para representar cualquier colección de símbolos, con un techo amplio de 65.000 caracteres suficiente para todos los sistemas de escritura en el mundo. XML ha adoptado UNICODE.



Los problemas relacionados con los códigos de caracteres no son exclusivos de los corpora multilingües, también puede suceder con los monolingües. El grupo encargado del tratamiento del corpus histórico CORDE informa de problemas de segmentación, derivados de "las peculiaridades ortotipográficas, que desafían el esquema de anotación, que trata de ser respetuoso con las intervenciones de autores, copistas y editores en el texto y que, en consecuencia, dificulta las unidades de análisis lingüístico" y las diferentes fases del etiquetado (Sánchez León 1999).

4. 3 Información lingüística

El tipo de información más valiosa de un corpus es la información de tipo lingüístico que las anotaciones harán explícitas. Podemos considerar varios niveles:

- *Anotaciones estructurales* donde se identifican los elementos lógicos que componen un texto: epígrafes, párrafos, etc.

- *Anotaciones morfosintácticas* que asignan a cada unidad léxica un código que identifica su categoría morfosintáctica (su *part of speech*, POS), así como otras propiedades morfológicas generalmente asociadas a la flexión (género, número, persona, caso, tiempo, etc.)

- *Lematización*, proceso mediante el cual las formas flexionadas del corpus se emparejan con sus lexemas respectivos, es decir con la forma de citación tal como aparecen en los diccionarios.
- *Análisis sintáctico* de las categorías sintagmáticas intraoracionales: grupos verbales y nominales, cláusulas subordinadas, etc.
- *Anotaciones orientadas a la tarea*: como pueden ser las unidades de traducción (Martínez 1999) o el etiquetado de referencia (numeración, citas, etc.).
- *Códigos de correspondencia*: etiquetas que hacen explícita la correspondencia entre unidades de traducción y que se asignan en el proceso de alineación.

En los siguientes apartados se van a agrupar los procesos que incorporan las anotaciones lingüísticas en dos grupos. A la etapa propia del etiquetado monolingüe la vamos a llamar segmentación; y al proceso distintivo de los corpora multilingües, alineación.

5 Segmentación

La segmentación tiene por objetivo identificar y anotar los elementos distintivos en todos los niveles de análisis. El efecto de la segmentación se suele representar mediante etiquetas o anotaciones que se incorporan al corpus. Existen varias propuestas de segmentación morfológica, de lematización y del análisis sintáctico para el español y para el euskara. Por lo general, estos tres procesos se simultanean con la tarea de desambiguación, es decir, de elección de la categoría más probable entre distintas opciones posibles en cada nivel. En este apartado vamos a repasar los niveles de segmentación, reseñando los trabajos que más nos pueden interesar.

5.1 Etiquetado estructural

En el nivel de análisis más genérico se identifican los elementos lógicos que reflejan la estructura del texto: epígrafes, apartados, párrafos, enumeraciones, oraciones, signos de puntuación, etc. Dentro de la oración se pueden reconocer las palabras del léxico genérico distinguiéndolas de otros elementos como nombres propios, números, fechas, siglas, acrónimos, etc. Martínez 1999 ha desarrollado un conjunto de segmentadores modulares que añaden etiquetas estructurales de tipo SGML/TEI para el corpus bilingüe LEGEBIDUN/BOB. Estos segmentadores pueden adaptarse con facilidad a otros corpora.

5. 2 Etiquetado morfosintáctico

Para el tratamiento morfosintáctico del español se han elaborado varios recursos. La RAE ha aplicado el generador morfológico del proyecto MULTEXT, *mmorph* (Bel y otros 1996). Se trata de un programa que combina morfología de dos niveles (para problemas de morfografía) y gramáticas de unificación y que reutiliza parte de los desarrollos del proyecto EUROTRA (Sánchez León 1999), los resultados se recogen en SGML/TEI (Pino y Santalla 1996). Por otro lado, en la Universidad Politécnica de Catalunya se ha desarrollado MACO+, un etiquetador basado en restricciones contextuales que resuelve mediante árboles de decisión estadísticos (Márquez y Padró 1997). Otros etiquetadores morfosintácticos para el español son SPOST (Farwell y otros 1995), SMORPH (Aït-Mokhtar y Rodrigo Mateos 1995), o el de Gala 1999. En la Universitat Pompeu Fabra se ha desarrollado CATMORF para el catalán (Badía 1997) y en la Universidad del País Vasco se ha desarrollado MORFEUS para el euskara, un sistema que

también está basado en la morfología de dos niveles (Alegria 1995 y Urkia 1997). La asignación de etiquetas morfosintácticas (etiquetas POS) suele simultanearse con la lematización.

5. 3 Lematización

Lematizar es reducir las formas flexivas de un texto a sus lexemas respectivos, es decir, a la forma de citación que se utiliza en los diccionarios (infinitivo para las formas verbales, masculino singular para las forma nominales). Así, tras la lematización, las formas *soy, era, fui o seré* serán reducidas al lexema *ser*. El conjunto de todas las variantes flexivas de un lexema es lo que se conoce como "lema". La lematización es importante porque permite conocer con mayor exactitud la composición léxica de los textos, y tiene especial relevancia para aplicaciones como la categorización textual o la recuperación de información. Para el español se han publicado resultados de varios lematizadores, siendo los más conocidos los de Sánchez-León 1995, Márquez y Padró 1997, o Gala 1999. Todos ellos realizan la lematización dentro del proceso de etiquetado morfosintáctico. En la Universidad del País Vasco se ha desarrollado el lematizador EUSLEM para euskara (Aduriz y otros 1996).

forma flexiva	**lexema**	**etiqueta POS**	**entidad SGML**
que	que	PR3CN00	**que**&pr3cn;
agota	agotar	VMIP3S0	**agota**&vmip3s;
la	la	TFFS0	**la**&tffs;
vía	vía	NCFS00	**vía**&ncfs;
administrativa	administrativo	AQ0FS00	**administrativo**&aq0fs;
podrá	poder	VMIF3S0	**podrá**&vmif3s;
interponerse	interponer	VMN000	**interponerse**&vmn;
recurso	recurso	NCMS00	**recurso**&ncms;
contencioso	contencioso	AQ00000	**contencioso**&aq0;
ante	ante	SPS00	**ante**&sps;

Tabla 2. Resultados del lematizador de Márquez y Padró, con muestras de etiquetas morfosintácticas (según los códigos propuestos por PAROLE)

Para resolver la ambigüedad entre varias categorías candidatas existen tanto métodos de base estadística, como de base simbólica. Por los resultados obtenidos en algunas comparativas (Chanod y Tapanainen 1995) se deduce que los sistemas estadísticos tienen un comportamiento aceptable cuando el corpus es homogéneo, pero son problemáticos si no lo es. Por lo general, resulta más sencillo entrenar un desambiguador estadístico que elaborar uno simbólico. La compilación de reglas gramaticales para desambiguar es un proceso manual, particularmente largo y costoso, pero que da mejores resultados a largo plazo. El grupo de la RAE (Sánchez León 1999) ha desarrollado una herramienta de desambiguación basada en la gramática de restricciones (*Constraint Grammar*, Karlsson y otros 1995). El grupo IXA de la Universidad del País Vasco también ha utilizado este método, aunque lo ha combinado con otros programas de base estadística, para aprovechar las ventajas de ambas estrategias (Aduriz y otros 1999, Gojenola 2000). Márquez y Padró 1997, por su parte, resuelven la ambigüedad mediante árboles de decisión estadísticos. La desambiguación de las categorías léxicas favorece considerablemente la eficacia del análisis sintáctico.

5. 4 Análisis superficial

Después de segmentar el texto en oraciones y anotar las palabras mediante etiquetas POS, el siguiente paso es reconocer las categorías sintagmáticas intraoracionales: sintagmas nominales, verbales, adjetivos y preposicionales, cláusulas subordinadas, etc. Para este proceso hacen falta analizadores sintácticos robustos, capaces de abordar cualquier tipo de construcción. Los modelos de gramáticas formales que se desarrollaron en la década de los ochenta están en su mayoría basados en reglas independientes de contexto. Tienen la ventaja de ofrecer mayor poder expresivo, pero a costa de una mayor complejidad computacional. Desde hace algunos años, para el análisis sintáctico de textos reales se ha optado por el diseño más simple de modelos basados en reglas de estados finitos. El procesamiento suele acometerse por segmentos menores a la oración, a los que se asignan categorías parciales. Es lo que se conoce como análisis superficial (*shallow parsing*).

Gala 1999 clasifica los analizadores superficiales en dos tipos:

- Los que aplican un enfoque construccionista: que funciona mediante la adición progresiva de restricciones durante el proceso de análisis (Joshi 1996)
- Los que aplican un enfoque reduccionista: que funciona mediante restricciones que sirven para eliminar análisis posibles (Karlsson y otros 1995, Chanod y Tapanainen 1996).

Para el español se han elaborado dos analizadores superficiales: TACAT, de Atserias y otros 1998, e IFSP (*Incremental Finite-State Parsing)* de Gala 1999. TACAT está basado en *charts* con una metodología incremental que utiliza gramáticas independientes de contexto en lugar de reglas de estados finitos.

Por otro lado, IFSP aplica un enfoque construccionista por medio de análisis parciales en tres fases (tabla 3):

1. Segmentación primaria: para reconocer secuencias de palabras con relación sintagmática (fase 1).
2. Una vez reconocidas las principales agrupaciones sintagmáticas, se asignan funciones sintácticas (fase 2).
3. El proceso termina resolviendo las relaciones de dependencia (fase 3).

Fase 1: etiquetado morfosintáctico	**Fase 2: etiquetado funcional**	**Fase 3: relaciones de dependencia**
[SC [NP La^el+DETSG posicion^posicion+NOUNSG NP]/N [PP del^de=el+PREPDET Gobierno^gobierno+NOUNSG PP] [AP frances^frances+ADJSG AP] : v ha^haber+HAB sido^ser+PAPUX interpretada^interpretarPAPSG SC] como^como+COMO [NP una^un+DETQUANTSG manera^manera+NOUNSG NP]/N [IV de^de+PREP_DE eludir^eludir+VERBINF IV] [NP el^el+DETSG problema^problema+NOUNSG NP]/N .^.+SENT	[SC [NP El^el+DETSG problema^problema+NOUNSG NP]/SUBJ : v tiene^tener+VERBFIN SC] [NP una^un+DETQUANTSG dimension^dimension+NOUNSG NP]/OBJ [AP mayor^mayor+ADJSG AP].^.+SENT	[SC [NP Las relaciones NP]/SUBJ [AP sociales AP] : v son SC] [AP muy informales AP], [PP en el sentido PP] [SC [PP de que PP] [NP las personas NP]/SUBJ :v se visitan SC] [PP sin previo aviso PP] ; SUBJ(relación,ser) SUBREFLEX(persona,visitar) ATTR(relación informal) VMODOBJ(ser,en,sentido) PADJ(relación,social) ADJ(previo aviso)
La posición del gobierno francés ha sido interpretada como una manera de eludir el problema.	*El problema tiene una dimensión mayor.*	*Las relaciones sociales son muy informales, en el sentido de que las personas se visitan sin previo aviso;*

Tabla 3. Fases del etiquetador IFSP de Gala 1999.

Para euskara se ha utilizado el enfoque reduccionista de la gramática de restricciones de Karlsson (Aduriz y otros 1999, Gojenola 2000, Arriola 2000). No conocemos de momento ninguna evaluación que permita comparar el rendimiento de estas estrategias.

6 Alineación

El proceso que más valor añade a un corpus multilingüe es la alineación. Alinear es hacer explícitas las relaciones de correspondencia entre segmentos de una y otra lengua. La alineación no depende obligatoriamente del resto de los procesos de etiquetado, aunque una segmentación básica previa será siempre necesaria. Según Martínez 1999, existen tres enfoques principales:

1. *Enfoque estadístico*: método de alineación que aprovecha la similitud de algunos rasgos cuantitativos identificados en el corpus, como la longitud de oraciones, el número de palabras o de caracteres, etc. (Brown y otros 1991, Gale y Church 1991).

2. *Enfoque lingüístico*: la alineación se basa en el emparejamiento previo de unidades sintagmáticas o de estructuras dependenciales (Sadler 1991, Kaji y otros 1992, Matsumoto y otros 1993).

3. *Enfoque mixto*: método que aprovecha la identificación de categorías gramaticales como apoyo para la alineación estadística (Chen 1993).

Las técnicas probabilísticas que se basan en anotaciones sintácticas requieren textos etiquetados de antemano (Black y otros 1993). El *Corpus Brown* o el *Penn Treebank* (Marcos y Santorini 1991) pueden servir de modelo para el inglés. Para otras lenguas también se conocen corpora con etiquetas sintácticas: en turco (Skut y otros 1997), en checo (Hajic y Hladká 1998), en alemán (Oflazer y otros 1999). En euskara se han etiquetado sintácticamente 10.000 palabras (Ezeiza y otros 1998). Estos procesos son muy costosos, algunas métricas (Voutilainen 1997) han concluido que es necesario el trabajo de una persona entrenada durante un año para etiquetar sintácticamente un corpus de 200.000 palabras.

Martínez 1998 y 1999 obtiene muy buenos resultados sobre un corpus bilingüe en español y euskara (tabla 4) que no dispone de etiquetas sintácticas, aplicando técnicas que aprovechan las etiquetas estructurales introducidas en el proceso de segmentación monolingüe.

Foru Agindua	**Orden Foral**
Foru Agindua, 767/1994 zk., urriaren 24ko. Aipatutako Foru Aginduaren bidez hurrengo hau xedatu da: Lurzoru batzuk dentsitate txikiko lurzoru urbanizagai gisa birsailkatzeko Zallako Udalerriko Planeamenduari buruzko Sorospidezko Arauen aldarazpena ukatzea. Erabaki honen aurka, haren jakinarazpenetik zenbatu beharreko hilabete biko epearen barruan, administraziozko liskarrauzi-errekurtsoa jarri ahal izango da, Euskal Herriko Justizia Auzitegi Nagusiko Administraziozko Liskarrauzietarako Salan, komeniesten diren beste defentsabideak erabil daitezkeelako kalterik gabe. Adierazi den epearen barruan, BHI-015/94-P05-A espedientea Bilbaoko Gran Vía, 19-21eko 5gn. solairuan egongo da ageriko, azter dadin. Bilbon, 1994.eko urriaren 24an.-Hirigintzako foru diputatua. Pedro Hernández González.	Orden Foral número 767/1994 de 24 octubre. Mediante la Orden Foral de referencia se ha dispuesto lo siguiente: Denegar la Modificación de las Normas Subsidiarias de Planeamiento del municipio de Zalla para la reclasificación de unos terrenos como Suelo Apto para Urbanizar de Baja Densidad. Contra dicha Orden Foral podrá interponerse, en el plazo de dos meses desde su notificación, recurso contencioso-administrativo ante la Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco, sin perjuicio de la utilización de otros medios de defensa que estime conveniente. Durante el referido plazo el expediente BHI-015/94-P05-A, quedará de manifiesto para su examen en las dependencias situadas en Bilbao calle Alameda Rekalde, 30, 5.a y 6.a plantas. Bilbao, 24 de octubre de 1994.-El Diputado Foral de Urbanismo.- Pedro Hernández González.

Tabla 4. Muestra del corpus paralelo LEGEBIDUN/BOB

Como resultado de la segmentación monolingüe Martínez 1998, 1999 identifica y categoriza unidades textuales - nombres propios, fórmulas y términos (tabla 5)- sobre los que posteriormente aplica técnicas de alineación.

<**rs type=organization**>Euskal Herriko Justizia Auzitegi Nagusiko Administraziozko Liskarrauzietarako Salan</**rs**>	<**rs type=organization**>Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco</**rs**>
<**rs type=law**>Zallako Udalerriko Planeamenduari buruzko Sorospidezko Arauen aldarazpena</**rs**>	<**rs type=law**>Modificación de las Normas Subsidiarias de Planeamiento del municipio de Zalla </**rs**>
<**term**>Lurzoru batzuk dentsitate txikiko lurzoru urbanizagai gisa birsailkatzeko</**term**>	<**term**>para la reclasificación de unos terrenos como Suelo Apto para Urbanizar de Baja Densidad</**term**>
<**seg type=9**>Erabaki honen aurka, haren jakinarazpenetik zenbatu beharreko hilabete biko epearen barruan, administraziozko liskarrauzi-errekurtsoa jarri ahal izango da, Euskal Herriko Justizia Auzitegi Nagusiko Administraziozko Liskarrauzietarako Salan, komeniesten diren beste defentsabideak erabil daitezkeelako kalterik gabe. Adierazi den epearen barruan, BHI-015/94-P05-A espedientea Bilbaoko Gran Vía, 19-21eko 5gn. solairuan egongo da ageriko, azter dadin. Bilbon, 1994.eko urriaren 24an.- Hirigintzako foru diputatua. Pedro Hernández González. </**seg**>	<**seg type=9**>Contra dicha Orden Foral podrá interponerse, en el plazo de dos meses desde su notificación, recurso contencioso-administrativo ante la Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco, sin perjuicio de la utilización de otros medios de defensa que estime conveniente. Durante el referido plazo el expediente BHI-015/94-P05-A, quedará de manifiesto para su examen en las dependencias situadas en Bilbao calle Alameda Rekalde, 30, 5.a y 6.a plantas. Bilbao, 24 de octubre de 1994.-El Diputado Foral de Urbanismo.- Pedro Hernández González. </**seg**>

Tabla 5. Segmentación en unidades de traducción.

Como consecuencia de los procesos de segmentación y alineación, el corpus se enriquece con etiquetas que hacen explícita las relaciones de correspondencia entre las dos versiones. Un corpus etiquetado y alineado en todos los niveles de análisis es un recurso lingüístico de extraordinario valor (Abaitua y otros 1998).

... <seg type=9 id=9EU2 corresp=9ES2> <s id=sEU11 corresp=ES11> <rs type=law id=LEU10 corresp=LES12>**Foru agindu** </rs> **horrek amaiera eman dio administrazio bideari; eta beraren aurka** <rs type=organization id=OEU10> **Administrazioarekiko** </rs> **auzibide-errekurtsoa jarri ahal izango zaio** <rs type=organization id=OEU11 corresp=OES9> **Euskal Herriko Justizi Auzitegi Nagusiko Administrazioarekiko Auzibideetarako Salari** </rs>**, bi hilabeteko epean; jakinarazpen hau egiten den egunaren biharamunetik zenbatuko da epe hori; hala eta guztiz ere, egokiesten diren beste defentsabideak ere erabil litezke.** </s> </seg> <seg type=10 id=10EU1 corresp=10ES1> <s id=sEU12 corresp=ES12> **Epe hori amaitu arte BHI-**<num num=10094> **100/94** </num>**-P05-A espedientea agerian egongo da, nahi duenak azter dezan,** <rs type=place id=PEU2 corresp=PES3> **Bilboko Errekalde zumarkaleko** </rs> <num num=30> **30.eko** </num> **bulegoetan,** <num num=5> **5** </num**>** **eta** <num num=6> **6.**</num> **solairuetan.**</s> </seg> <closer id=pEU13> <docAuthor> <s id=sEU13 corresp=ES13> <rs type=title id=TLEU4 corresp=TLES4> **Hirigintzako foru diputatua** </rs>. </s> <s id=sEU14 corresp=ES14> _ <rs type=name id=NEU4 corresp=NES4> **Pedro Hernández González** </rs>.</s> </docAuthor>	... <seg type=9 id=9ES2 corresp=9EU2> <s id=sES11 corresp=EU11> **Contra dicha** <rs type=law id=LES12 corresp=LEU10> **Orden Foral** </rs>**, que agota la vía administrativa podrá interponerse recurso contencioso-administrativo ante la** <rs type=organization id=OES9 corresp=OEU11> **Sala de lo Contencioso-Administrativo del Tribunal Superior de Justicia del País Vasco** </rs>**, en el plazo de dos meses, contado desde el día siguiente a esta notificación sin perjuicio de la utilización de otros medios de defensa que estime oportunos.**</s> </seg> <seg type=10 id=10ES1 corresp=10EU1> <s id=sES12 corresp=EU12> **Durante el referido plazo el expediente BHI-**<num num=10094> **100/94** </num>**- P05-A quedará de manifiesto para su exámen en las dependencias de** <rs type=place id=PES3 corresp=PEU2> **Bilbao calle Alameda Rekalde** </rs>**,** <num num=30> **30** </num>**,** <num num=5> **5.a** </num> **y** <num num=6> **6.a** </num> **plantas.** </s> </seg> <closer=pES13> <docAuthor> <s id=sES13 corresp=EU13> **El** <rs type=title id=TLES4 corresp=TLEU4> **Diputado Foral de Urbanismo** </rs>. </s> <s id=sES14 corresp=EU14> - <rs type=name id=NES4 corresp=NEU4> **Pedro Hernández González** </rs> </s> </docAuthor> </closer>

Tabla 6. Muestra de sección del corpus alineada (Martínez 1999)

Durante la década de los noventa, la alineación de corpora ha sido una de las líneas de actuación más esperanzadoras de la ingeniería lingüística. Gale y Church 1991, Dagan y otros 1993, Macklovitch 1994, McEnery y Oakes 1996, Schmied y Schäffler 1996, o Melamed 1997, son otras referencias destacables.

7 Aplicaciones

A principios de la década de los noventa, los campos de la lexicografía y terminografía, así como la traducción automática protagonizaron la actividad en torno a los corpora multilingües. En la actualidad el interés se ha ampliado hacia otras áreas como la enseñanza de segundas lenguas o

la didáctica de la traducción, herederas de experiencias muy anteriores (McEnery y Wilson 1996). Pero las dos aplicaciones que más atención reciben en estos momentos son la recuperación translingüística de información y la internacionalización de productos.

7.1 Enseñanza de segundas lenguas

Una aplicación tradicional de los corpora multilingües es la enseñanza de segundas lenguas. Como actividad más reciente destacamos el *International Corpus of Learner English* (ICLE) de la Universidad de Lovaina-*la Neuve*. Este corpus contiene una colección de composiciones cortas hechas por estudiantes de niveles altos de inglés de procedencia dispar: chinos, checos, holandeses, finlandeses, franceses, alemanes, japoneses, polacos, rusos, españoles y suecos. El objetivo de esta recopilación es comprobar la hipótesis del "modelo de lengua común" al que se llega en los niveles avanzados. Se ha comprobado, por ejemplo, que los alumnos abusan de ciertas palabras (verbos auxiliares, pronombres personales, algunas conjunciones, como *and* y *but*, los verbos *get* y *think*, algunas palabras de significado vago, como *people*, *things*, del cuantificador *very*) y que infrautilizan otras (como *the, this, these*, o *by*). También se ha estudiado la influencia de las respectivas lenguas maternas sobre el modelo de adquisición de la segunda lengua.

Cuando una palabra se infrautiliza a menudo indica que se produce un alerta inconsciente ante un problema de aprendizaje. Así en el corpus alemán, el verbo *become* se utiliza mucho menos que en los otros corpora, porque tiene un falso amigo en *bekommen*, palabra muy frecuente en alemán que tiene un significado muy distinto del inglés. Los estudiantes tienden a evitar palabras que perciben como potencialmente conflictivas y tienden a utilizar el léxico que les es más familiar. Esta actitud ha sido muy estudiada y se denomina "principio del oso de peluche", *teddy bear principle*.

7. 2 Didáctica de la traducción

Baker 1996 y 1997 ha defendido la utilidad de los corpora comparables para estudiar la traducción, sobre todo en su aplicación a la enseñanza. Han comprobado que factores extralingüísticos, como el sexo del traductor, su edad, la lengua de origen, etc. influyen en las traducciones. Atendiendo a la propuesta de Toury 1995, el grupo de Baker (Laviosa 1997 y 1998, Sardinha 1997) pretende descubrir las "leyes probabilísticas y condicionales del comportamiento traduccional" (*laws of translational behaviour*), a partir de los datos de un extenso corpus comparable de textos originales y traducciones.

Con fines análogos se han realizado estudios contrastivos entre el inglés y el sueco (Johansson. 1996), así como entre el inglés y el noruego (Johansson y Ebeling 1996), o el inglés y el polaco (Piotrowska 1997). Otro corpus diseñado para servir de modelo en la didáctica de la traducción es el *Corpus LSP*, del *Institut Universitari de Lingüística Aplicada* de la Universitat Pompeu Fabra (Vivaldi 1996, Bach y otros 1997).

7. 3 Lexicografía y terminografía

Los lexicógrafos siempre han recurrido a grandes colecciones de textos para desarrollar su trabajo de creación y actualización de diccionarios. Este trabajo ha sido tradicionalmente lento y laborioso. Por ello se ha recibido con júbilo la disponibilidad de corpora en formato electrónico, así como las técnicas y herramientas que hacen posible su procesamiento (Pérez Hernández y otros 1999). La Real Academia de la Lengua ha comenzado a basar sus trabajos lexicográficos

en los corpora CORDE y CREA (Sánchez-León 1999). Pero, además de la RAE, varias editoras de diccionarios de español también disponen de corpora: Vox Bibliograf posee uno de 10 millones de palabras, la editorial SGEL uno de 8 (corpus CUMBRE, Sánchez y Cantos 1997) y SM otro de 60.000 palabras (Pérez Hernández y otros 1999). En Cataluña , el *Institut d'Estudis Catalans*, así como TERMCAT e IULA; en el País Vasco *Euskaltzaindia* y UZEI; en Galicia la Academia de la Lengua y el *Centro de Investigacións Ramón Piñeiro* (Magán 1996), realizan labores de lexicografía y terminografía sobre la base de datos obtenidos de corpora.

Si los corpora monolingües contribuyen sustancialmente en el desarrollo de diccionarios monolingües, no menos útiles son los corpora multilingües (Álvarez 1999). Numerosos autores han contribuido en los últimos años a la creación y enriquecimiento de diccionarios y tesauros bilingües aprovechando los datos disponibles en corpora: Gale y Church 1991; Daille y otros 1994; Catizone y otros 1993; Kumano y Hirakawa 1994; Klavans y Tzoukermann 1995, Langlois 1996. Aplicando similares técnicas a textos de especialidad ha sido posible extraer glosarios terminológicos, o identificar términos compuestos y construcciones colocacionales: Eijk 1993; Kupiec 1993; Dagan y Church 1994; Samajda y otros 1996; Resnik y Melamed 1997.

Por otro lado, los trabajos de Briscoe y Carroll 1997, o Lapta 1999, han abierto nuevas vías de investigación en la lexicografía computacional al haber ensayado la extracción automática de patrones de subcategorización verbal a partir de corpora. Arriola 2000 ha realizado un experimento similar para euskara.

7. 4 Traducción automática

La disponibilidad del *Hansard Corpus* en formato electrónico despertó el interés de los investigadores del Watson Centre de la IBM (Brown y otros 1990) que lo aprovecharon para ensayar métodos alternativos de traducción automática. Los métodos basados en reglas (conocidos por el acrónimo inglés RBMT, *Rule-based Machine Translation*) habían llegado a finales de los ochenta a un punto de estancamiento y la comunidad investigadora comenzaba a buscar nuevos enfoques. Es el retorno de los métodos empíricos que ya habían sido probados en los albores de la disciplina (Weaver 1949) . El cambio de enfoque en los noventa se ve favorecido por el drástico abaratamiento de los microprocesadores y las unidades de almacenamiento. Con ello comienzan a proliferar las colecciones de textos en formato electrónico y su disponibilidad favorecida por Internet es una invitación a probar los métodos probabilísticos y conexionistas, que tan buenos resultados habían dado ya en el tratamiento de corpora orales. El número de sistemas diseñados se multiplica (Catizone y otros 1993, Kay y Röscheisen 1993; Vogel y otros 1996, Wu 1996 y Tillmann y otros 1997) de forma que puede decirse que el paradigma de la traducción por reglas ha perdido numerosos adeptos en beneficio de la traducción por analogías, ABMT, *Analogy-based Machine Translation* (Jones 1992).

Nagao 1984 ya había anticipado este cambio con su propuesta de traducción basada en ejemplos (EBMT, *Example-based Machine Translation*), técnica que ha tenido gran eco en la comunidad científica. Sadler 1989 se sirvió de un corpus alineado para crear una base de ejemplos bilingües, utilizada luego como recurso de traducción automática. Tsuji y otros 1991 y Sumita e Iida 1991 aplican enfoques híbridos. La traducción basada en ejemplos ha tenido su mayor aplicación en una tecnología conocida como "memoria de traducción" (MBMT, *Memory-based Machine Translation*). Consiste en el almacenamiento de traducciones realizadas manualmente y validadas por el traductor, de forma que puedan ser reutilizadas posteriormente para textos similares que se reconocen mediante umbrales de similitud basados generalmente en lógica difusa. Esta tecnología ha sido llevada al mercado con un considerable éxito: Déjà-Vu (ATRIL),

Translator's Workbench (TRADOS), Transit (STAR), SDLX, son algunas de las herramientas de mayor difusión.

7. 5 Edición plurilingüe

Un enfoque alternativo a la traducción automática es la generación de textos multilingües (Kittredge 1989, Hartley y Paris 1997). Movidos por el éxito de los sistemas de escritura asistida por ordenador (Gómez Guinovart 1999, Gojenola y Oronoz 2000), varios proyectos han probado a simultanear los procesos de composición y traducción. Esta técnica ha resultado muy adecuada para textos cuyo esquema se ciñe a un guión preestablecido y recurrente. A partir de una muestra representativa de partes meteorológicos en varios idiomas, Chevreau y otros 1999 han desarrollado MultiMétéo, un sistema que permite generar automáticamente partes multilingües. Otro sistema de este tipo es TREE, que genera ofertas de empleo multilingües (Somers 1992). En nuestro entorno más cercano se debe reseñar GIST, proyecto europeo para la generación plurilingüe de textos instructivos (Lavid 1995 y 1996), así como LEGEBIDUN (Abaitua y otros 1997, Casillas y otros 1999, 2000). A partir de un corpus paralelo alineado de documentos administrativos en euskara y castellano, Casillas 2000 ha diseñado un generador de nuevos documentos bilingües.

7. 6 Internacionalización de productos

Pero el sector de mayor crecimiento es la internacionalización de productos. Desde hace poco más de un lustro, la edición en CD-ROM, la producción de páginas *web* y, sobre todo, la industria de software han hecho que el enfoque de la traducción cambie radicalmente. La necesidad de actualizar cíclicamente los productos y su distribución en los mercados internacionales obliga a las empresas a realizar importantes esfuerzos para satisfacer las exigencias lingüísticas y culturales de esos mercados. Microsoft, por ejemplo, adapta a más de 25 idiomas sus principales productos, y tiene muy buenos motivos para hacerlo así. Según datos publicados por la propia empresa (Brooks 2000), más de la mitad de sus ingresos proceden del comercio exterior (unos 5.000 millones de dólares).

La adaptación de un programa de software no es un mero problema de traducción. Hay muchos factores que deben estar previstos en la propia fase de diseño. Además de traducir los títulos, los mensajes, las ayudas *on-line*, o los menús que interactúan con el usuario, para adaptar un programa de sofware hay que resolver otros aspectos que son difíciles de notar a simple vista. El cambio de alfabeto es una fuente obvia de dificultades, pero los problemas más delicados son los que tienen que ver con el código del programa. La traducción de las llamadas a función de los menús, por ejemplo, puede afectar al tamaño de los registros, o al código de las instrucciones que los ejecutan. Los enlaces de Internet y el tratamiento de cifras y fechas también deberán ser adaptados. Si el programa además dispone de herramientas de tratamiento propiamente lingüístico - relacionadas con los formatos, la corrección ortográfica, gramatical o de estilo, los diccionarios y glosarios- el proceso se complicará considerablemente.

La adaptación lingüística y cultural de software se conoce como "localización" (calco directo del término anglosajón *localisation*) y forma parte de lo que en el contexto empresarial se relaciona con la "internacionalización" de un producto. Es una actividad que está en expansión y que ha generado un importante nicho de mercado. Si se menciona aquí es porque la tecnología que se utiliza en la localización está relacionada con la lingüística de corpus. En el sector aeronáutico, por ejemplo, la documentación de cualquier producto es muy voluminosa y se distribuye en formato electrónico, generalmente en CD-ROM. Estos manuales y libros de referencia, cuando

se traducen, se convierten en genuinos corpora bilingües. Para su traducción y actualización se utilizan los gestores de memorias de traducción.

7. 7 Búsquedas translingüísticas

Las redes internacionales y la difusión de bases documentales hace cada vez más relevante la cuestión del acceso y recuperación multilingüe de la información. Los proyectos de digitalización de bibliotecas que se iniciaron primero en países de habla inglesa, se han extendido ya a todas las regiones del mundo. Tanto Asia como Europa están experimentando un rápido desarrollo de infraestructuras para la distribución de sus fondos documentales, lo que significa que materiales en multitud de idiomas han comenzado a ser accesibles por red.

Unido al crecimiento de bibliotecas digitales monolingües en distintos idiomas, existen también bibliotecas multilingües en países con más de una lengua nacional, o en países donde el inglés es la lengua usada para la documentación técnica o científica, en instituciones paneuropeas o en empresas transnacionales. Todo esto ha despertado la conciencia de que deben diseñarse herramientas con capacidad translingüística para la recuperación y extracción de información. Es una de las áreas de investigación de la ingeniería lingüística que ha experimentado mayor crecimiento y en apenas unos años han proliferado los foros, reuniones internacionales y publicaciones sobre este tema. Algunas referencias obligadas son: Oard y Dorr 1996, Sheridan y Ballerini 1996, Picchi y Peters 1996, Gilarranza, Gonzalo y Verdejo 1997, Kando y Aizawa 1998, Grefenstette 1998, o Pazienza 1999; sobre bibliotecas digitales, Peters y Picchi 1997; sobre técnicas de categorización de documentos, Yang 1998.

8 Internet multilingüe

Internet se ha convertido en el principal campo de aplicación de las técnicas de tratamiento de corpora multilingües. Pese a un acusado desequilibrio inicial favorable al inglés, la tendencia actual es hacia la corrección de la balanza lingüística. En la medida en que la situación se vaya normalizando, aumentará la información en idiomas distintos del inglés, como cabe prever de la comparación del porcentaje de páginas *web* con el de publicaciones o traducciones en formato tradicional (Lockwood 1998). En la actualidad la proporción de páginas *web* en inglés casi triplica al de publicaciones en papel. Para otras lenguas con presencia internacional, sólo el japonés mantiene una relación equitativa entre lo que se publica en *web* y en papel. La desproporción es acusada para el resto, y particularmente grave en el caso del chino.

Unión Latina (UL) publicó en 1998 una encuesta basada en consultas para 57 palabras (del tipo "ambigüedad" o "rodilla") entre los principales buscadores de Internet. En 1999 se publicó una segunda encuesta realizada por Pedro Maestre para el Instituto Cervantes (IC), sobre los 180 millones de páginas indizadas en Altavista Magallanes. Los resultados de ambas encuestas se parecen y dan unas tasas de presencia similares. El español ocupa en ambos casos el quinto lugar, con 2,6 millones de páginas según el cálculo de Maestre, de las que más de la tercera parte corresponden a España. Se sitúa muy cerca del francés, pero bastante por debajo del japonés y el alemán. El inglés exhibe una supremacía absoluta, con índices entre el 70 y el 75%. La ubicuidad de la red y la condición del inglés como *lingua franca* son los dos factores a los que se atribuye esta desproporción. Un dato ilustrativo de la encuesta de Masetre es la significativa presencia de lenguas pequeñas como el neerlandés. Su presencia cuantitativa, en número de páginas, es equiparable a la del chino, con una tasa de páginas por habitante que resulta ser 657,18 veces más alta que la tasa de páginas en chino. (Los datos de la última columna de la tabla 7

corresponden a la proporción de páginas por cada 1.000 habitantes en los países de origen - España, Francia etc.- y se han extraído de la encuesta de Maestre, salvo las cifras entre paréntises, que se han calculado sobre el total de hablantes, y no sobre el de habitantes).

lenguas	**en las que se publica**	**a las que se traduce**	**páginas web (UL)**	**páginas web (IC)**	**pág./hab.**
inglés	28%	5%	75%	70,05%	(81,51)
chino	13%	0,5%	-	0,71%	0,86
alemán	12%	17%	4,02%	3,34%	51,77
francés	8%	6%	2,81%	1,96%	23,26
español	7%	16%	2,53%	1,51%	22,94
japonés	5%	5%	-	5,01%	(69,96)
ruso	5%	3%	-	-	
portugués	5%	0%	0,82%	0,73%	(7,51)
neerlandés	2%	7%	-	0,71%	56,85
otras	15%	41%	14,82%	15.98%	
	100%	100%	100%	100%	

Tabla 7. Comparativa de presencia internacional de las principales lenguas

No disponemos de datos sobre la presencia de páginas bilingües o multilingües en la red, aunque hay claros indicios de su rápido crecimiento. Una parte importante de las publicaciones en Internet proviene de los medios de comunicación, de las empresas transnacionales y de las instituciones internacionales. Todos ellos se afanan para que sus presencia en la red supere las barreras lingüísticas. Por ello, pese a la supremacía de los textos monolingües, la red se ha convertido también en un vasto corpus multilingüe que crece exponencialmente. Esto ha disparado la demanda de tecnologías con capacidad de procesamiento multilingüe: buscadores inteligentes, sistemas de indización y catalogación, extractores de información, gestores de conocimientos, generadores de textos, generadores de resúmenes, etc. La lingüística de corpus y las técnicas de alineación tienen el campo abonado en Internet.

9 Referencias

Joseba Abaitua, Arantza Casillas, Raquel Martínez. 1997. Segmentación de corpus paralelos para memorias de traducción. *Procesamiento del Lenguaje Natural* 21:17-30.

Joseba Abaitua, Arantza Casillas, Raquel Martínez. 1997. Tratamiento de textos administrativos bilingües: el proyecto LEGEBIDUN. *Philologia Hispalensis* 11-2:115-130.

Joseba Abaitua, Arantza Casillas, Raquel Martínez. 1998. Value added tagging for multilingual resource management. *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada: 1003-1007.

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegría, Nerea Ezeiza, Ruben Urizar. 1996. Del analizador morfológico al etiquetador/lematizador: unidades léxicas complejas y

desambiguación. *Procesamiento del Lenguaje Natural*.

Itziar Aduriz, Jose Maria Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Koldo Gojenola, A. Maritxalar. 1997. Morphosyntactic disambiguation for Basque based on the Constraint Grammar formalism. *Recent Advances in Natural Language Processing (RANLP'97)*.

Itziar Aduriz, Eneko Agirre, Izaskun Aldezabal, Iñaki Alegría, Jose Maria Arriola, Xabier Artola, Koldo Gojenola, A. Maritxalar, Kepa Sarasola, Miriam Urkia. 2000. A word-grammar based morphological analyzer for agglutinative languages. *Proceedings of the 18th International Conference on Computational Linguistics* (*COLING'2000*). Saarbrücken.

Salah Aït-Mokhtar y José Lázaro Rodrigo Mateos. 1995. Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH. *Procesamiento del Lenguaje Natural* 17: 29-41.

Iñaki Alegria. 1995. *Euskal morfologiaren tratamendu automatirako tresnak*. Tesis doctoral. Universidad del País Vasco.

Alberto Álvarez Lugrís. 1999. Técnicas de representación en la lexicografía plurilingüe. *Revista española de lingüística aplicada*. Volumen monográfico: 215-245.

Jose Maria Arriola. 2000. *Euskal Hiztegia-ren azterketa eta egituratzea ezagutza lexikalaren eskurateze automatikoari begira.*Tesis doctoral. Universidad del País Vasco.

S. Atkins, J. Clear, N. Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7-1: 1-16.

Jordi Atserias, Irene Castellón, M. Civit. 1998. Syntactic parsing of unrestricted Spanish text. *Proceedings of the 1st International Conference on Language Resources and Evaluation*. Granada.

Carme Bach, Roser Saurí, Jordi Vivaldi, M. Teresa Cabré. 1997. El Corpus de l'IULA. *IULA/INF017/97* Universitat Pompeu Fabra.

Toni Badía. 1997. CATMORF: multi two-level steps for Catalan morphology. *Applied Natural Language Processing (ANLP'97)*. Washington.

Mona Baker. 1996. Corpus-based translation studies: the challenges that lie ahead. Harold Somers (comp.) *Terminology, LSP and translation: studies in language engineering in honour of Juan C. Sager*. John Benjamins.

Mona Baker. 1999. The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4-2: 1-18.

Núria Bel, J. Marimón y J. Porta. 1996. Etiquetado morfosintáctico de corpus en el proyecto MULTEXT. *Actas del XXVI Simposio de la Sociedad Española de Lingüística*. Madrid.

Douglas Biber y Edward Finegan. 1986. An initial typology of English text types. Jan Aarts y Willen Meijs (comp.) *Corpus Linguistics II: New Studies in the Analysis and Exploitation of Computer Corpora*. Rodopi: 19-46.

D. Birdsong. 1989. *Metalinguistic performance and interlinguistic competence*. Sringer-Verlag.

P.F. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. L. Mercer y P. Roosin. 1990. A statistical approach to machine translation. *Computational Linguistics* 16-2.

P.F. Brown, J. Lai y R.L. Mercer. 1991. Aligning sentences in Parallel Corpora. *Proceedings of the Association for Computational Linguistics (ACL'91)*. Berkeley: 169-176.

Ralf D. Brown. 1996. Example-Based Machine Translation in the Pangloss System. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*: 169-174.

T. Briscoe y J. Carroll. 1997. Automatic extraction of subcategorization from corpora. *Proceedings of Applied Natural Language Processing (ANLP'97)*, Washington.

David Brooks. 2000. What prize globalization? *Language International* 12-1: 17-20.

Arantza Casillas. 2000. *Explotación de corpus alineados para el desarrollo de entornos de composición de documentos estructurados bilingües.* Tesis doctoral. Universidad de Deusto.

Arantza Casillas, Joseba Abaitua, Raquel Martínez. 1999. Extracción y aprovechamiento de DTDs emparejadas en corpus paralelos. *Procesamiento del Lenguaje Natural* 25: 33-41.

Arantza Casillas, Joseba Abaitua y Raquel Martínez . 2000. DTD-driven bilingual document generation. *International Natural Language Generation Conference (INLG'2000)*. Mitzpe Ramon, Israel.

Arantza Casillas, Joseba Abaitua y Raquel Martínez. 2000. Toward a Document-grammar definition: Experiences with an aided document composition and translation tool. *Extreme Markup Languages*. Montreal

Arantza Casillas, Joseba Abaitua y Raquel Martínez. 2000. Recycling annotated paralell corpora for bilingual document composition. *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas (AMTA'2000).* Cuernavaca.

R. Catizione, G. Russel y S. Warwick. 1993. Deriving translation data from bilingual texts. *Proceedings of the 1st International Lexical Acquisition Workshop*. Detroit.

J.P. Chanod y P. Tapanainen. Tagging French. 1995. Comparing a statistical and a contraint-based method. *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'95)*, Dublín.

Karine Chevreau, José Coch, José A. García Moya, Margarita Alonso. 1999. Generación multilingüe de boletines meteorológicos. *Procesamiento del Lenguaje Natural* 25: 51-58.

M. Collins. 1997. Three new probabilistic models for statistical parsing. *Proceedings of the Association for Computational Linguistics (ACL'97)*, Madrid.

M. Collins, J.Hajic, L. Ramshaw, C. Tillman. 1999. Statistical parsing for Czech. *Proceedings of the Association for Computational Linguistics (ACL'99)*, Maryland.

B. Daille, E. Gaussier, J.M. Lange. 1994. Towards automatic extraction of monolingual and bilingual terminology. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*: 515-521.

I. Dagan, K. Church y W. A. Gale. 1993. Robust word alignment for machine aided translation.

Proceedings of the 1st Workshop on Very Large Corpora. Columbus.

I. Dagan y K. Church. 1994. Termight: Identifying and translating technical terminology. *Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP'94)*: 34-40.

T. Erjavec, N. Ide,V. Petkevic, J. Véronis. 1996. Multext-East: Multilingual text tools and corpora for Central and Eastern European languages. *Proceedings of the European TELRI Seminar: Language Resources for Language Technology*, 87-98.

Rosa Estopà. 1999. Eficiencia en la extracción automática de terminología. *Perspectives: Studies in Translatology* 7-2: 277-286.

Nerea Ezeiza. 2000. *Corpusak ustiatzeko tresna linguistikoak/ Herramientas lingüísticas para la exploración de corpus*. Tesis doctoral. Universidad del País Vasco.

David Farwell, Stephen Helmreich, Mark Casper. 1995. SPOST: a Spanish part of speech tagger. *Procesamiento del Lenguaje Natural* 17: 42-53

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-paralell English-Chinese corpus. *Proceedings of the 3rd Workshop on Very Large Corpora*. Boston.

Núria Gala. 1999. Using the incremental finite-state architecture to create a Spanish shallow parser. *Procesamiento del Lenguaje Natural* 25: 75-82.

W. A. Gale y K. W. Church. 1991. Identifying word correspondences in parallel texts. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*: 152-157.

W. A. Gale y K. W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19-1: 75-102.

J. Gilarranz, Julio Gonzalo y Felisa Verdejo. 1997. Language-independent text retrieval with the EuroWordNet multilingual semantic database. Workshop on multilinguality in software industry: the AI contribution. *Proceedings fo the International Joint Conference on Artificial Intelligence (IJCAI'97)*.

Gregory Grefenstette. 1998. *Cross-language information retrieval*. Kluwer Academic Press.

Koldo Gojenola. 2000. *Euskararen sintaxi konputazionalerantz*. Tesis doctoral. Universidad del País Vasco.

Koldo Gojenola y M. Oronoz. 2000. Corpus-based syntactic error detection using syntactic patterns. *Proceedings of the Student Research Workshop at Applied Natural Language Processing (ANLP'2000)*. Seattle.

Javier Gómez Guinovart. 1999. *La escritura asistida por ordenador*. Universidad de Vigo.

Javier Gómez Guinovart y Anxo M. Lorenzo Suárez. 1996. *Lingüística e informática*. Ediciones Tórculo.

J. Hajic y B. Hladká. 1998. Tagging inflective languages: prediction of morphological categories for a rich, structured tagset. *Proceedings of the Association for Computational Linguistics (COLING-ACL'98).* Montreal.

Jos Hallebeek. 1999. El corpus paralelo. *Procesamiento del Lenguaje Natural* 24: 49-55.

A. Hartley C. Paris. 1997. Multilingual document production from support for translating to support for authoring. *Machine Translation*, 12:109-128.

Amparo Hurtado (comp.) 1994. *Estudis sobre la traducció*. Publicacions de la Universitat Jaume I.

Nancy Ide y J. Véronis. 1994. MULTEXT (Multilingual Text Tools and Corpora). *Proceedings of the International Wordshop on Shareable Natural Language Resources*: 90-96.

Mats Johansson. 1996. Fronting in English and Swedish: a text based contrastive analysis. Percy y otros (comp.): 29-39.

Stig Johansson y Jarle Ebeling. 1994. Exploring the English-Norwegian parallel corpus. Udo Fries, Totti Gunnel y Peter Schneider. *Creating and using English language corpora.* Rodopi.

Daniel Jones. 1996. *Analogical Natural Language Processing*. UCL Press.

Noziko Kando y A. Aizawa. 1998. Cross-lingual information retrieval using automatic generated multilingual keyword clusters. *Proceedings of the 3rd International Workshop on Information Retrieval with Assian Languages.*

F. Karlsson, A. Voutilainen, J. Heikkilä, A. Anttila. 1995. Constraint Grammar: a language independent system for parsing unrestricted text. Mouton de Gruyter.

Philip King. 1997. Parallel Corpora for translator training. Barbara Lewandowska-Tomaszczyk and Patrick James Melia (comp.) *PALC '97 Practical Applications in Language Corpora*. Lodz University Press. 393-402.

R. Kittredge. 1989. Multilingual Text Generation as an Alternative to Machine Translation. *Proceedings of the 30th Annual Conference of the American Translators Association*: 465-469.

J. Klavans y E. Tzoukermann. 1995. Combining corpus and machine translation dictionary data for building bilingual lexicons. *Machine Translation* 10: 185-218.

A. Kumano y H. Hirakawa. 1994. Building an MT dictionary from paralell texts based on linguistics and statistical information. *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94*): 76-81.

J. Kupiec. 1993. An algorithm for finding noun phrase correspondance in bilinbual corpora. *Proceedings of the Association for Computational Linguistics (ACL'93): 17-22.*

Stig Johansson y Jarle Ebeling. 1996. Exploring the English-Norwegian parallel corpus. Percy y otros (comp.): 3-15.

L. Langlois. 1996. Bilingual concordances: A new tool for bilingual lexicographers. *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA'96).*

M. Lapta. 1999. Acquiring lexical generalizations from corpora: a case study of diathesis alternations. *Proceedings of the Association for Computational Linguistics (ACL'99).* Maryland.

Julia Lavid. 1995. From interpersonal options to thematic realization in multilingual

administrative forms. *Working notes of the IJCA'95 workshop on multilingual text generation*: 75-83.

Julia Lavid. 1996. Generating thematic choices for multilingual text generation. *Book of the ALLC-ACH'96*. A. Lindebjerg, E. Ore y O. Reigem (comp.) Norwegian Computing Centre for the Humanities: 183-188.

Sara Laviosa. 1997. How comparable can 'comparable corpora' be? *Target*, 9-2: 289-319.

Sara Laviosa. 1998. The English Comparable Corpus: A resource and a methodology. Lynne Bowker, Michael Cronin, Dorothy Kenny y Jennifer Pearson (comp.). *Unity in Diversity? Current Trends in Translation Studies*. St. Jerome Publishing.

Geofrey Leech. 1993. Corpus annotation schemes. *Literary and Linguistic Computing* 8-4: 257-281.

Geofrey Leech, R. Garside, E. Atwell. 1983. The automatic grammatical tagging of teh LOB corpus. *ICAME news* 7:13-33-

Rose Lockwood. 1998. Global English and language market trends. *Language International* 10-4: 16-18.

Pedro Maestre Yenes. 1999. La utilización de las diferentes lenguas en Internet. *Centro Virtual Cervantes*. http://cvc.cervantes.es/obref/anuario_99/pmaestre/

Fernando Magán Muñoz. 1996. Estándares de representación da información textual e multimedia. Gómez y Lorenzo (comp.): 153-186.

Francisco A. Marcos Marín. 1991. ADMYTE (Archivo Digital de Manuscritos y Textos Españoles); the Digital Archive of Spanish Manuscripts and Texts. *Literary and Linguistic Computing* 6-3: 221-224.

Francisco A. Marcos Marín. 1994. *Informática y Humanidades*. Gredos.

Lluis Márquez y Lluis Padró. 1997. A flexible POS tagger using an automatically acquired language model. *Proceedings of the Association for Computational Linguistics (ACL'97)*: 238-245.

Mitch Marcus, B. Santorini. 1991. Building very large natural language corpora: the Penn Treebank. CIS Report. University of Pennsylvania.

Raquel Martínez. 1999. *Alineación automática de corpus paralelos: una propuesta metodológica y su aplicación a un dominio de especialidad*. Tesis doctoral. Universidad de Deusto.

Raquel Martínez, Joseba Abaitua, Arantza Casillas. 1998. Bitext correspondences through rich markup. *Proceedings of the Association for Computational Linguistics* (COLING-ACL'98), Montreal: 812-818.

Raquel Martínez, Joseba Abaitua, Arantza Casillas. 1998. Aligning tagged bitexts. *Proceedings of the 6th Workshop on Very Large Corpora*, Montreal: 102-109.

Tony McEnery y Michael Oakes. 1996. Sentence and word alignment in the CRATER proyect.

Thomas y otros (comp.): 211-231.

Tony McEnery y Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh University Press.

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.

César Montoliu. 1998. La traducción automática en El Periódico de Catalunya. *Puntoycoma* 51: 6-8.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. Elithorn and Nanerji.

Junsaku Nakamura. 1991. The relationships among genres in the LOB corpus based upon the distribution of grammatical tags. *Jacet Bulletin* 22: 55-74.

G.T. Nicol. 1995. The Multilingual World Wide Web. Electronic Book Technologies. http://mirage.irdu.nus.sg/multilingual/unicode/misc/multilingual-www.html

Christiane Nord. 1994. Traduciendo funciones. Hurtado (comp.) 97-112.

D. W. Oard y B.J. Dorr. 1996. A survey of multilingual text retrieval. *Technical Report UMIACS-TR-96-19*. University of Maryland.

D. W. Oard. 1997. Alignment of Spanish and English TREC topic descriptions. *5th TREC Conference (TREC-5)*.

K. Oflazer, D. Zeynep, H. Tür, G. Tür. 1999. Design for a Turkish treebank. *Proceedings of Workshop on Linguistically Interpreted Corpora (EACL'99)*, Bergen.

M.T. Pazienza. 1998. *Information Extraction: A multi-disciplinary approach to an emerging information technology*. Springer-Verlag.

Javier Pérez Guerra. 1998. *Introducción a la lingüística de corpus: un ejercicio con herramientas informáticas aplicadas al análisis textual*. Ediciones Tórculo.

Chantal Pérez Hernández, Antonio Moreno Ortiz y Pamela Faber. 1999. Lexicografía computacional y lexicografía de corpus. *Revista española de lingüística aplicada*. Volumen monográfico: 175-213.

Carol Peters y Eugenio Picchi. 1997. Across languages, across cultures: issues in multilinguality and digital libraries. *D-Lib Magazine*.

Eugenio Picchi y Carol Peters. 1996. Cross-language information retrieval: a system for comparable corpus querying. *Working Notes of the Workshop on Cross-Linguistic Information Retrieval, ACM SIGIR'96*, 24-33.

M. Pino y M. O. Santalla. 1996. Codificación morfosintáctica de corpus en lenguaje SGML. *Procesamiento del Lenguaje Natural* 20: 101-117.

Maria Piotrowska. 1997. Criteria for selecting parallel texts in teaching a translation course. Barbara Lewandowska-Tomaszczyk and Patrick James Melia (comp.) *PALC '97 Practical*

Applications in Language Corpora. Lodz University Press. 411-420.

Rosa Rabadán. 1994. Traducción, intertextualidad, manipulación. Hurtado (comp.): 129-139.

C. Samuelsson y A. Voutilainen. 1997. Comparing a linguistic and a stochastic tagger. *Proceedings of Association for Computational Linguistics (ACL'97)*. Madrid: 246-251.

A Sánchez y P. Cantos. 1997. Predictability of word forms (types) and lemma in linguistic corpora. A case study based on the analysis of the CUMBRE corpus: an 8 million word corpus of contemporary Spanish. International Journal of Corpus Linguistics 2-2: 259-280.

Fernando Sánchez-León. 1995. Desarrollo de un etiquetador morfosintáctico para el español. *Procesamieto del Lenguaje Natural* 17:14-28.

Fernando Sánchez-León, Jordi Porta, José Luis Sancho, Amalio Nieto, Almudena Ballester, Adelaida Fernández, Javier Gómez, Laura Gómez, Encarnación Raigal, Rafael Ruiz. 1999. La anotación de los corpus CREA y CORDE. *Procesamiento del Lenguaje Natural* 25: 175-182.

Berber A. P. Sardinha. 1997. Patterns of lexis in original and translated business reports: textual differences and similarities. Karl Simms (comp.) *Translating sensitive texts: linguistic aspects*. Rodopi. 147-153.

Harold Somers. 1992. Interactive multilingual text generation for a monolingual user. *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TIM'92)*:151-161.

S. Sato y M. Nagao. 1990. Toward Memory-based Translation. *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*. Helsinki: 247-252.

Josef Schmied y Hildegard Schäffler. 1996. Approaching translationese through parallel and translation corpora. Percy y otros (comp). 41-56.

P. Sheridan y J.P. Ballerini. 1996. Experiments in multilingual information retrieval using the SPIDER system. *Proceedings of the 19th ACM SIGIR Conference*, 58- 65.

W. Skut, T. Brants, B. Krenn, H. Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. *Proceedings of 1st International Conference on Language Resources and Evaluation (LREC'98)*, Granada.

Frank Smadja, Kathleen MaKeown y Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* 22-1:1-38.

C.M. Sperberg-McQueen y Lou Burnard (comp.) 1994. *Guidelines for Electronic Text Encoding and Interchange*.

Robert Sprung. 1999. Crossing cultures: *Time* magazine opens new markets with Spanish and Portuguese editions. *Language International* 11-6: 22-25.

Michael Stubbs. 1996. *Text and corpus analysis: computer-assisted studies of language and culture*. Blackwell.

E. Sumita y H.Iida. 1991. Experiments and Prospects of Example-Based Machine Translation. *Proceedings of the Association for Computational Linguistics (ACL'91)*. Berkeley: 185-192.

C. Tillmann, S. Vogel, H. Ney, A. Zuriaga. 1997. A DP based search using monotone alignments in statistical translation. *Proceedings of the Association for Computational Linguistics (ACL'97)*: 289-296.

Jenny Thomas y Mick Short (comp.) 1996. *Using corpora for language research. Studies in honour of Geoffrey Leech*. Longman.

Gideon Toury. 1995. *Descriptive translation studies and beyond*. John Benjamins.

Margherita Ulrych. 1997. The impact of multilingual parallel concordancing on translation. Barbara Lewandowska-Tomaszczyk and Patrick James Melia (comp.) *PALC '97 Practical Applications in Language Corpora*. Lodz University Press. 421-436.

Miriam Urkia. 1997. *Euskal morfologiaren analisi automatikorantz*. Tesis doctoral. Universidad del País Vasco.

Jorge Vivaldi. 1996. Proyectos del IULA: Corpus técnico. V.M. Forcada, A.G. Carrasco y J.C. Sager. *Estudios computacionales del español y el inglés*. Instituto Cervantes.

Jorge Vivaldi. 1996. A LSP multilingual corpus. *TermNet*.

S. Vogel, H. Ney, C. Tillmann. 1996. HMM-based word alignment in statistical translation. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*: 836-841.

A. Voutilainen. 1997. Designing a parsing grammar. R. Roche y Y. Schabes (comp.) *Finite-state processing*. MIT Press.

A. Winarske, S. Warwick-Armstrong, J. Hajic. 1992. Tagging and alignment of parallel texts: current status of BCP. *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP'92)*. Terento: 227-228.

D. Wu. 1996. A polynomial-time algorithm for statistical machine translation. *Proceedings of the Association for Computational Linguistics (ACL'96*): 152-158.

Donghua Xu y Chew LIm Tau. 1999. Alignment and matching of bilingual English-Chinese news texts. *Machine Translation* 14: 1-33.

Y. Yang.1998. An evaluation of statistical approaches to text categorization. *Information Retrieval Journal*.

F. Yergeau, G. Adams y M. Duerst. 1997. Internationalization of the Hypertext Markup Language. RFC 2070. Network Working Group. http://babel.alis.com:8080/web_ml/html/rfc-i18n/rfc-i18n-0.en.html