PONENCIAS

Los bancos de conocimiento especializado multilingüe: un nuevo recurso para la traducción

M. TERESA CABRÉ CASTELLVÍ
Universitat Pompeu Fabra

1 Introducción

No hay duda de que las tecnologías lingüísticas y los recursos digitalizados prestan un gran servicio a la traducción, tanto para la búsqueda de información como utilizadas como herramientas. Los traductores utilizan las tecnologías para efectuar consultas lingüísticas y temáticas y para localizar nuevos recursos que puedan ayudarles en la traducción. Y asimismo se sirven de las tecnologías aplicadas a la traducción para acelerar su productividad traductora y asegurar un mayor grado de sistematicidad en los textos traducidos. Tal sería el caso de la utilización de programas de memorias de traducción asociadas a gestores de terminología, que, a la vez que facilitan la traducción automática de fragmentos traducidos con anterioridad y aseguran el uso de las mismas unidades terminológicas para una unidad del original, permiten acumular una gran cantidad de términos en formato de base de datos.

2 Evolución de los recursos lingüísticos

Sin embargo, la concepción de los recursos de búsqueda, almacenamiento y selección de información adecuados a las necesidades de los lingüistas aplicados (término genérico que utilizaremos para referirnos a todo profesional de la lengua o las lenguas para fines aplicados) así como también la concepción de las vías de búsqueda de datos ha variado con el tiempo.

En un inicio los recursos más representativos fueron los bancos de datos terminológicos y léxicos, que se constituían para extraer de ellos la nomenclatura de glosarios temáticos y diccionarios, o también para que fueran fuente de resolución de dudas lingüísticas sobre unidades terminológicas o léxicas en general tanto en contextos de traducción

como en contextos de normalización lingüística. Los datos contenidos en estos recursos tenían el formato de registros de base de datos en los que a cada entrada léxica o terminológica se le asociaban tipos de información estructurada en campos. Los campos más representativos en los bancos terminológicos, por ejemplo, eran la categoría gramatical, el ámbito temático, la definición, y en el caso de bancos plurilingües, los equivalentes en otras lenguas. Era preceptivo que para todos los datos figurara la fuente de la que procedían. Cada registro correspondía a una unidad. Esta unidad se presentaba en forma descontextualizada y solo a veces se acompañaba de muestras, y solo muestras, de contextos en los que se habían encontrado. La funcionalidad más prominente de estos recursos era la de ser repositorios de datos léxicos o terminológicos y su valor aumentaba en consonancia con la cantidad de entradas que la base contenía.

Una segunda etapa en la evolución de los recursos al servicio de los traductores la representan los bancos de textos. Los bancos textuales aportaron la novedad de presentar los datos en contexto real no fragmentado y proporcionar tantos contextos como ocurrencias de una misma unidad en los textos. Mediante motores de búsqueda los traductores acceden a ocurrencias de una unidad en los textos, con lo que disponen de usos documentados de dicha unidad. Los bancos textuales se convierten además en laboratorios adecuados para la descripción de las unidades en discurso. Es relevante en este período la constitución de grandes corpus textuales monolingües denominados *corpus de referencia,* cuya información se selecciona a partir de criterios de representatividad y equilibrio. El ejemplo históricamente más representativo es el corpus COBUILD desarrollado en la Universidad de Birmingham para la construcción de diccionarios de la empresa *Collins*. Un ejemplo de corpus de referencia del español es el CREA (*Corpus de Referencia del Español Actual*) de la Real Academia Española, que contiene unos ciento cuarenta millones de registros en textos en español desde 1975 hasta la actualidad.

Muestra de corpus de referencia son también todos los corpus constituidos en el Proyecto LE-PAROLE (LE2-4017) (http://www.ub.es/gilcub/SIMPLE/simple.html), financiado por la Unión Europea en el marco del IV Programa Marco de I+D de la UE. Se trata del desarrollo de recursos lingüísticos, corpus y léxicos electrónicos para todas las lenguas de la UE, para que sean directamente utilizadas en aplicaciones de ingeniería lingüística y destinado a dotar a todas las lenguas oficiales (y a dos lenguas no oficiales: el catalán y el danés) de un corpus de referencia con el que poder construir recursos informáticos comparables destinados al tratamiento automático de estas lenguas. La continuación del Proyecto

PAROLE fue la constitución de un diccionario de 10.000 entradas sobre la base de la extracción de información de los corpus PAROLE. Los corpus textuales, por lo tanto, aunque representan un avance en la creación de recursos lingüísticos en soporte digital, son recursos utilizados fundamentalmente hasta hace poco tiempo para la creación de diccionarios. Así lo muestran las palabras del académico Manuel SÁNCHEZ RON referidas al CREA de la Real Academia, pronunciadas en el Seminario sobre «El español como lengua científica» organizado por la FECYT en diciembre de 2003: «gracias al CREA es posible extraer información para estudiar las palabras, sus significados y contextos».

Pero la evolución de los recursos lingüísticos no ha cesado desde sus inicios en los años sesenta hasta hoy, y de la fase de construcción de grandes corpus de textos cuyo valor radicaba en la cantidad de millones de ocurrencias hemos pasado a corpus más pequeños, pero de contenido más afinadamente seleccionado. Este paso nos hace entrar en la tercera etapa de la evolución de los recursos lingüísticos.

3 Vías de innovación en los recursos lingüísticos

A nuestro entender tres han sido las vías de refinamiento de los bancos de textos como recursos al servicio de las necesidades de consulta y extracción de información por parte de los lingüistas aplicados:

A por una parte, la constitución de bancos textuales específicos temáticamente, que ofrecen información más selectiva y permiten una recuperación más afinada de información;
B en segundo lugar, la creación de bancos de textos clasificados por criterios pragmático-comunicativos;
C en tercer lugar, la constitución de bancos etiquetados gramaticalmente, que permiten recuperar información seleccionada por criterios lingüísticos estrictos.

3.1 Constitución de bancos textuales de tema específico

Los bancos textuales temáticamente específicos se han convertido hoy en día en uno de los recursos más valorados para cubrir necesidades expresivas y comunicativas en contexto profesional. Las razones que explican este interés son conocidas. La información de tema específico tiene actualmente un gran valor por la importancia que el conocimiento especializado ha adquirido en las sociedades avanzadas. Gracias a las tecnologías de la información es posible almacenarla, actualizarla y

acceder a ella de forma selectiva, amigable y eficiente. El crecimiento exponencial de la ciencia y de la técnica, así como el desarrollo y la transferencia de tecnología y servicios requieren gran cantidad de terminología disponible y estandarizada para dar respuesta denominativa a estas necesidades expresivas y comunicativas tanto en el ámbito nacional como internacional. Es relevante también la importancia que ha adquirido la terminología aplicada en sectores de intensa actividad y necesidad, fundamentalmente aplicada a la resolución de situaciones plurilingües a través de la traducción y la enseñanza de lenguas para propósitos específicos, y aplicada al tratamiento automático de las lenguas a través del desarrollo de tecnologías lingüísticas. Una muestra de banco textual especializado plurilingüe es el que hemos constituido en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra de Barcelona, que comprende textos en castellano, catalán, inglés, francés y alemán (en diferente proporción cada una) y en cinco materias: informática, medicina (incluyendo genómica humana), medio ambiente, derecho y economía (http://www.iula.upf.edu). El estado actual del corpus (Informe de 24 de marzo de 2004) expresado en número de ocurrencias por tema y lengua se muestra en el cuadro siguiente:

Área	Catalán	Castellano	Inglés	Francés	Alemán
Derecho	1463	2085	431	44	16
Economía	1776	1091	274	78	27
Medio ambiente	1506	1062	599	230	429
Informática	655	1227	338	194	83
Medicina	2619	4001	1555	27	198
Total	8019	9466	3197	573	753

Una parte de este corpus está integrado por textos paralelos. Las parejas de lenguas más representadas en el corpus paralelo son catalán-castellano, catalán-inglés y castellano-inglés. La situación actual en relación con el número de documentos del corpus paralelo es la siguiente:

Área	Catalán-Castellano	Catalán-Inglés	Castellano-Inglés
Derecho	460	12	57
Economía	600	250	283
Medio ambiente	214	213	144
Medicina	118	40	640
Informática	28	-	300
Total	1.420	515	1424

Los bancos textuales especializados se han convertido hoy en la base para el desarrollo de sistemas de identificación y extracción automática de candidatos a términos, así como en el banco de base para el diseño y desarrollo de sistemas generadores de resúmenes, y de sistemas de extracción de información para la construcción de bancos de datos terminológicos y de ontologías.

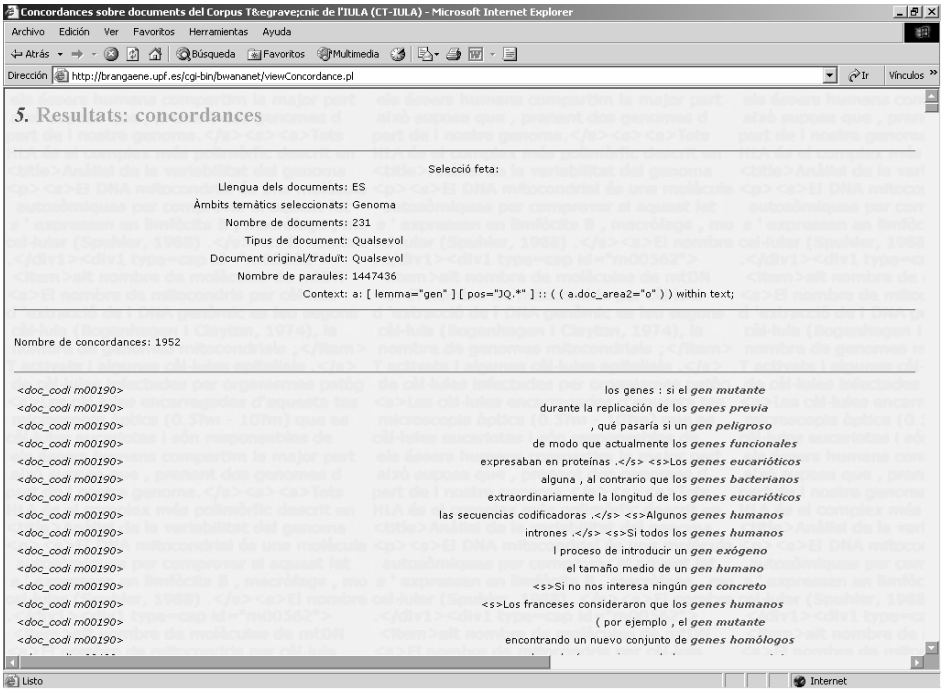
3.2 Creación de bancos de textos clasificados por criterios pragmático-comunicativos

La segunda vía de refinamiento de los bancos textuales la constituye el haber creado bancos textuales seleccionados por géneros, o bien el hecho de disponer para cada texto de información explícita sobre el género y el tipo textual al que pertenece, lo que facilita el análisis lingüístico descriptivo de los textos especializados y el contraste entre ellos sobre la base de la selección y la frecuencia de uso de recursos gramaticales distintos. Buena muestra de ello es el banco de géneros textuales constituido en la Facultad de Traducción e Interpretación de la Universitat Jaume I de Castellón dentro de un proyecto dirigido por la profesora Isabel GARCÍA IZQUIERDO. Véase Isabel GARCÍA IZQUIERDO / Esther MONZÓ NEBOT (2002) *La traducción científico-técnica y la terminología en la sociedad de la información*, Universitat Jaume I, Castelló de la Plana [ISBN: 84-8021-409-0], *Enciclopèdia electrònica de gèneres d'especialitat per a la traducció*.

3.3 Constitución de bancos etiquetados

La tercera vía de innovación de los bancos textuales la constituyen los bancos etiquetados o, dicho de otro modo, los bancos enriquecidos con etiquetas gramaticales atribuidas a las unidades que constituyen los textos. Así se habla de bancos marcados o etiquetados con información

morfológica, sintáctica, semántica y pragmática. Los bancos etiquetados más habituales son los que contienen información morfológica flexiva sobre las terminaciones de cada unidad en tanto que perteneciente a una clase gramatical que permite agrupar bajo un mismo lema todas sus formas gramaticales, o bien desplegar todas las formas gramaticales del mismo lema. El marcaje morfológico, como todos los tipos de marcaje, lo realizan herramientas que se conocen en general como *etiquetadores*. El marcaje morfológico flexivo es de obligada presencia en un tratamiento automático de la información basado en criterios lingüísticos y no simplemente a través del reconocimiento de cadenas de caracteres como realizan, por ejemplo, herramientas tan habituales como WordSmith o TACT. Podemos ver su utilidad en las búsquedas de patrones gramaticales previos a la detección de unidades que pueden ser candidatos a términos, descartando paralelamente todas aquellas combinaciones que nunca podrían constituir una unidad terminológica sintagmática. Es también este marcaje el que permite obtener fraseología sobre la base de la combinación de verbos y nombres deverbales con sus complementos. Como muestra podemos aportar la herramienta de exploración *BwanaNet* (http://www.iula.upf.edu) desarrollada en nuestro Instituto. BwanaNet permite buscar combinaciones de unidades definidas como cadenas de caracteres o bien como categorías, como muestra el siguiente ejemplo:



Menos frecuente es la marcación morfológica léxica que descompone cada unidad construida en los formantes que constituyen su estructura interna. Este etiquetaje permite recuperar grupos de unidades que poseen la misma configuración morfológica o que poseen el mismo número de formantes. Permite asimismo generar unidades morfológicamente construidas si dispone de una gramática de formación de palabras que contenga las reglas léxicas y sus restricciones.

En cambio es cada vez más usual disponer de un analizador sintáctico, de los denominados de primer nivel, que atribuyan funciones sintácticas básicas a los grupos de palabras estructurados en forma de sintagmas que aparecen a la derecha y a la izquierda de determinadas categorías simples (verbo, adjetivo, nombre) o complejas (sintagma nominal, sintagma verbal, sintagma adjetivo, sintagma preposicional). La base de este marcaje está en la explicitación de las condiciones argumentales de las unidades léxicas predicativas, ya sean verbos, ya sean nombres o adjetivos con valor predicativo. Así, por ejemplo, la presencia en el corpus de un verbo como *poner* activará en el analizador la búsqueda, normalmente a su derecha, de un nombre o un sintagma nominal que actúe de complemento directo o complemento interno, y a su izquierda de otro SN o nombre que seguramente actuará como complemento externo o sujeto. Y si aparece un sintagma preposicional se espera que la preposición que encabece este sintagma sea en castellano *en.*

De hecho, estos sistemas que atribuyen funciones sintácticas básicas a los complementos de unidades predicativas no son analizadores sintácticos o *parsers* en el sentido literal del término, tan necesarios para la traducción automática, sino solo *chunkers* o segmentadores que reconocen cuándo una determinada combinación constituye una estructura y por lo tanto puede cumplir una función. Si se da el caso de que el diccionario de tratamiento automático de la información contiene una unidad de carácter predicativo con la descripción de cuantos complementos suele llevar, complementos obligatorios y opcionales básicos, y a qué tipo de estructura pertenece cada complemento, el sistema reconoce la pieza léxica predicativa y busca en su entorno los complementos descritos, atribuyéndoles una etiqueta sintáctica. Si, además de la información sintáctica, cada complemento está asociado a un papel argumental del tipo *Tema, Objeto, Agente* o *Locativo*, y cada argumento además está especificado semánticamente como perteneciente a una clase semántica, el sistema intentará marcar cada unidad o estructura léxica con una etiqueta semántica, la etiqueta exigida en la descripción de cada unidad del diccionario. Este marcaje es muy

importante para un tratamiento automático refinado de la información y para la extracción selectiva de datos. Sin embargo son muy escasos los corpus etiquetados semánticamente. Como paliativo existen herramientas desarrolladas para el español y el catalán con financiación pública que facilitarían este marcaje, como es por ejemplo el sistema WordNet (http://www.cogsci.princeton.edu/~wn/index.shtml) o el uso del diccionario del Proyecto SIMPLE (http://www.ub.es/gilcub/SIMPLE/simple.html), continuación del Proyecto PAROLE.

Más escasos son todavía los corpus generales o especializados etiquetados pragmáticamente. Existen, esto sí, corpus muy pequeños con marcas pragmáticas, elaborados para estudios muy específicos. Una muestra puede ser el Corpus Oral de Conversa Col·loquial del Corpus de Català Contemporani de la Universitat de Barcelona, de lenguaje oral espontáneo en catalán desarrollado por el profesor Lluís PAYRATÓ de la Universitat de Barcelona.

Por las dificultades y el coste que comporta la creación de corpus etiquetados es absolutamente necesario que la comunidad científica actúe coordinadamente en el desarrollo de recursos digitales para las lenguas: cuanto más ilustrados gramaticalmente, más provechosos para la investigación descriptiva y aplicada.

4 Un recurso digital integrador: los bancos de conocimiento especializado

Los bancos de conocimiento (*knowledge databases*) son una propuesta que intenta aunar en un mismo recurso información gramatical, textual, terminológica, documental y semántica. A pesar de que su desarrollo es costoso en tiempo y recursos, la información que permiten obtener es de gran valor para la enseñanza de lenguas y la traducción.

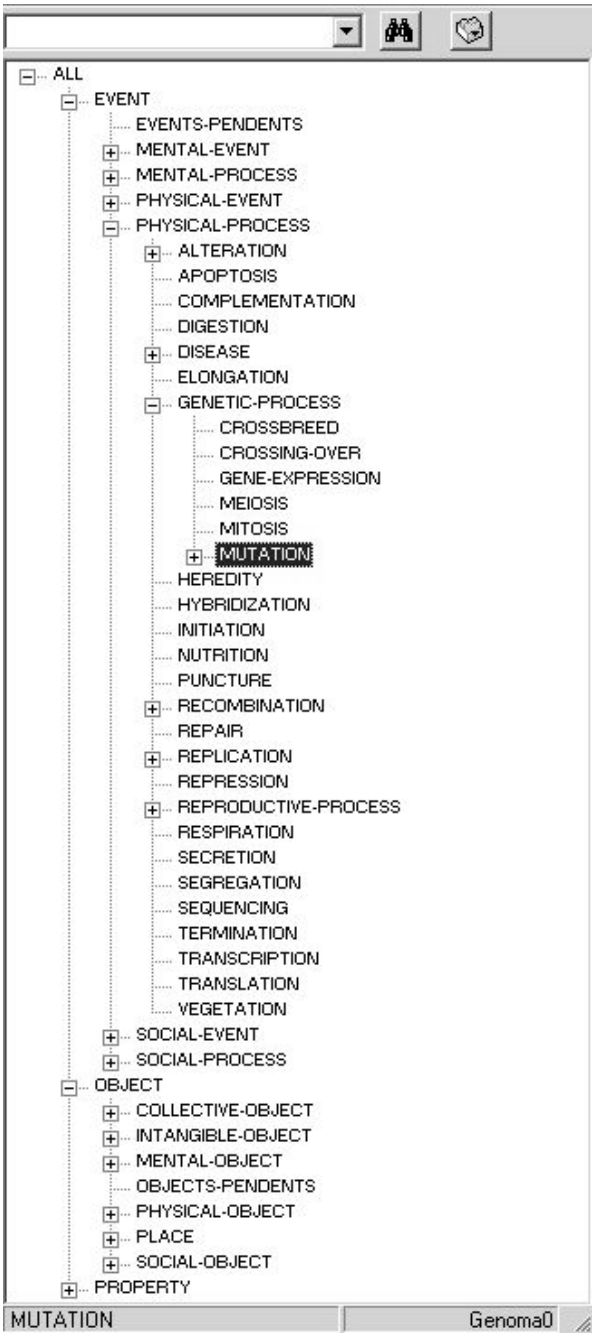
4.1 Concepto de banco de conocimiento

Se han definido los bancos de conocimiento como «repositorios de conocimiento representados en un lenguaje formal a los que se accede mediante un sistema experto basado en unidades terminológicas relacionadas conceptualmente entre sí por diferentes tipos de conexión». Así lo definen los diccionarios

> ***Knowledge database***: A collection of knowledge expressed using some formal knowledge representation language. A knowledge base forms part of a knowledge-based system (KBS).

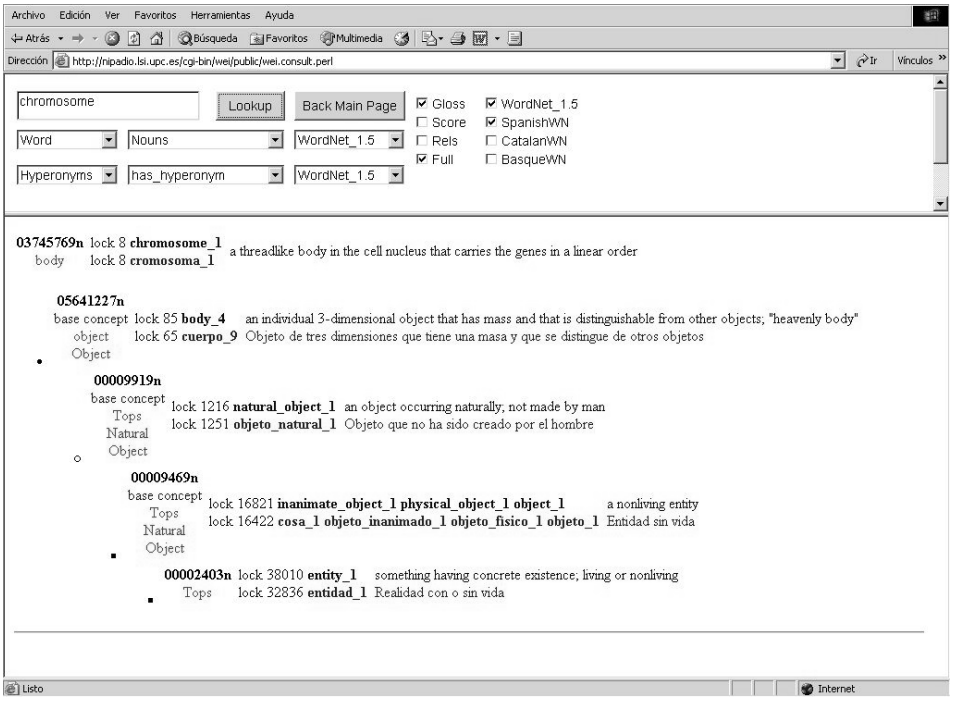
> **knowledge-based system** *<artificial intelligence> (KBS) A program for extending and/or querying a knowledge base. The related term expert system is normally used to refer to a highly domain-specific type of KBS used for a specialised purpose such as medical diagnosis.* [http://dict.die.net/]

Los bancos de conocimiento representan una clara superación de los bancos terminológicos, por varias razones: en primer lugar porque ofrecen la posibilidad de recuperar todos los contextos en los que aparece una unidad, pero por encima de todo incorporan conocimiento semántico a la terminología, tanto a través de los textos (permiten recuperar los contextos definitorios de cada unidad) como a través de las ontologías, por cuanto cada término está asociado a un concepto ubicado en una estructura de conocimiento que lo vincula a otros conceptos del mismo campo a través de relaciones diversas (hiperonimia, hiponimia, meronimia, holonimia, causalidad, localización, etc.). De esta forma el contenido de un término representa el conjunto de las relaciones que el concepto asociado a este término establece con el resto de conceptos del ámbito especializado. Véase a continuación una muestra de la ontología sobre genoma humano:



A diferencia de las clasificaciones léxicas tipo WordNet, las ontologías parten de la ficción de que lo que se representa es el

conocimiento y no las unidades léxicas. Por ello, la representación de un concepto comprende el conjunto de relaciones que establece este concepto en la ontología, y se representa mediante una cadena de caracteres que, aunque puedan coincidir formalmente con una unidad léxica, se presupone que no es la unidad léxica sino solo una etiqueta funcional. Las clasificaciones léxicas, en contraste, son estructuras de unidades léxicas por su contenido, como muestra la figura siguiente:



Los bancos de conocimiento están al servicio de múltiples y variadas profesiones implicadas en la expresión y transferencia de conocimiento especializado:

Sirve a los traductores e intérpretes para resolver consultas lingüísticas y cognitivas. Sirve a los terminólogos y lexicógrafos para facilitar la elaboración de diccionarios generales y especializados. Sirve a los documentalistas para elaborar tesauros y clasificaciones, para indexar documentos y facilitar la recuperación de información. Sirve también a los redactores técnicos como recurso de búsqueda de información. Y a los especialistas y profesores para enseñar una materia. Y a los profesores de lenguas para propósitos específicos como fuente de selección de materiales y recurso didáctico en sí mismo. Sirve también a los periodistas

científicos en su preparación para divulgar el conocimiento especializado de manera ortodoxa, sin saltar los límites de la especialidad.

4.2 El Proyecto GENOMA[1]

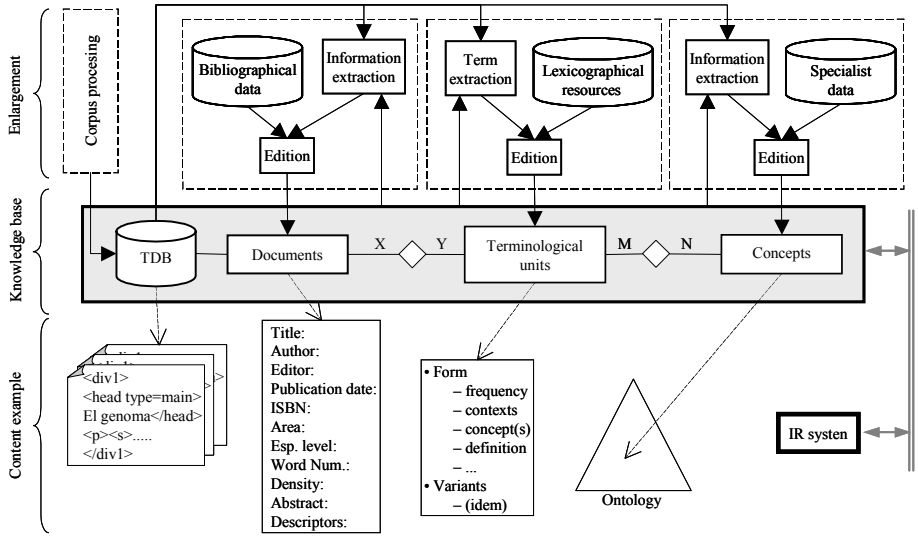
En nuestro grupo de investigación IULATERM, en el marco del Proyecto TEXTERM (*Textos especializados y terminología: selección y recuperación automática de la información*) BFF2000-0841, hemos desarrollado un banco de conocimiento sobre genómica humana.

En este proyecto hemos definido un banco de conocimiento especializado sobre un ámbito como una integración en una misma plataforma de cuatro recursos distintos:

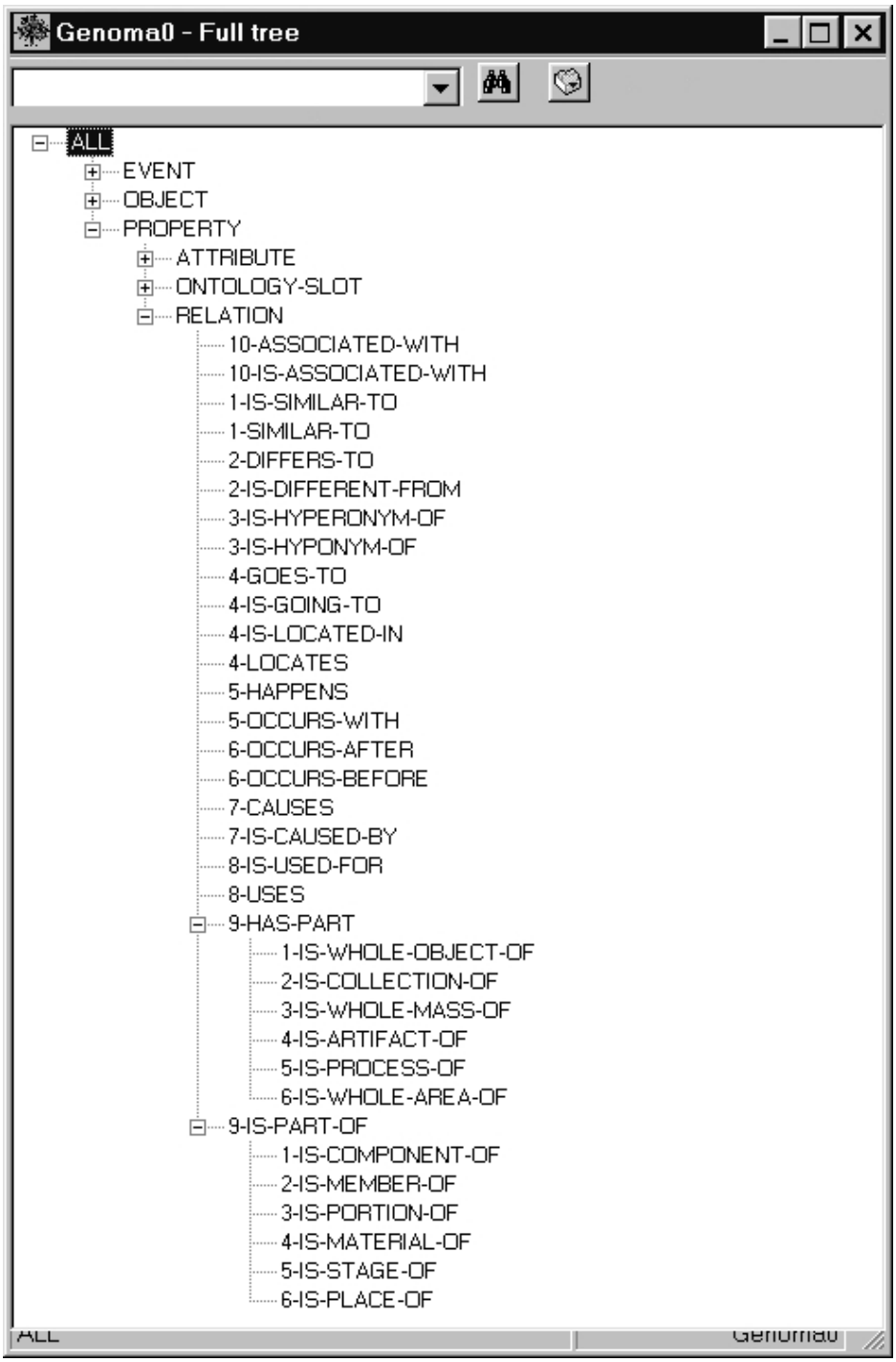
A un banco textual que comprende textos sobre el tema del ámbito
B un banco documental sobre los textos y factográfico con relación al ámbito
C un banco terminológico que comprende las unidades pertinentes sobre el tema
D una ontología que representa la estructura de contenido del ámbito.

La estructura del banco de conocimiento, que hemos desarrollado sobre el tema del Genoma Humano, puede representarse con la figura siguiente:

[1] En el desarrollo del proyecto GENOMA han colaborado además C. BACH, J. FELIU, G. MARTÍNEZ y J. VIVALDI. Han cooperado como becarios del módulo terminológico J. J. GIRALDO y V. VIDAL.



En esta figura podemos ver la conjunción de tres módulos, cada uno correspondiente a una base de información: la textual, la documental y factográfica, y la terminológica. Y podemos observar además cómo la base terminológica aparece ligada a una ontología en la que el contenido de cada término se asocia a un concepto representado por una etiqueta funcional que reúne el conjunto de relaciones de este concepto con el resto de conceptos de la estructura global del ámbito. Las relaciones pertenecen a una lista cerrada, previamente declarada (FELIU 2004):

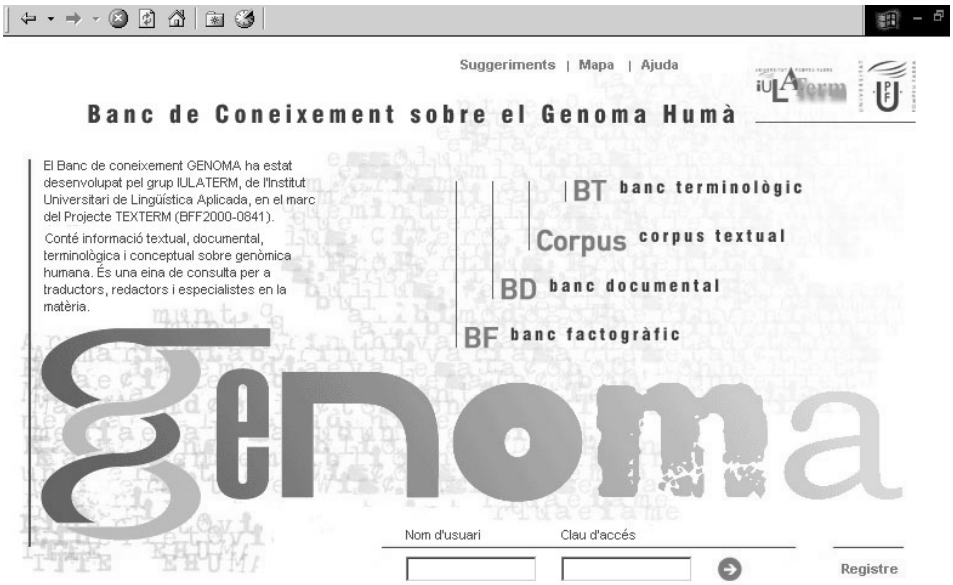

Genoma0 - Full tree
ALL
EVENT
OBJECT
PROPERTY
ATTRIBUTE
ONTOLOGY-SLOT
RELATION
10-ASSOCIATED-WITH
10-IS-ASSOCIATED-WITH
1-IS-SIMILAR-TO
1-SIMILAR-TO
2-DIFFERS-TO
2-IS-DIFFERENT-FROM
3-IS-HYPERONYM-OF
3-IS-HYPONYM-OF
4-GOES-TO
4-IS-GOING-TO
4-IS-LOCATED-IN
4-LOCATES
5-HAPPENS
5-OCCURS-WITH
6-OCCURS-AFTER
6-OCCURS-BEFORE
7-CAUSES
7-IS-CAUSED-BY
8-IS-USED-FOR
8-USES
9-HAS-PART
1-IS-WHOLE-OBJECT-OF
2-IS-COLLECTION-OF
3-IS-WHOLE-MASS-OF
4-IS-ARTIFACT-OF
5-IS-PROCESS-OF
6-IS-WHOLE-AREA-OF
9-IS-PART-OF
1-IS-COMPONENT-OF
2-IS-MEMBER-OF
3-IS-PORTION-OF
4-IS-MATERIAL-OF
5-IS-STAGE-OF
6-IS-PLACE-OF
ALL
Genoma0

El módulo documental, que en nuestro proyecto corresponde a la base de datos GENDOFAC, se ha desarrollado en una plataforma Access 2000 y contiene todas las referencias bibliográficas de los textos del módulo corpus textual y de las fuentes de las que proceden los contextos de las unidades de la base terminológica. Comprende monografías, revistas, artículos de revista, tesis y entidades, clasificadas siguiendo la estructura elaborada por un experto para el corpus textual, que presenta los siguientes apartados o subámbitos:

- Estructura interna
- Ingeniería genética
- Enfermedades
- Investigación genética
- Diferenciación
- Inmunología
- Biotecnología
- Filogenia
- Neurociencia
- Farmacogenómica
- Eugenesia

El módulo textual incluye el conjunto de textos sobre genoma humano ordenados temáticamente a partir de la estructuración anterior. Los textos, en catalán, castellano e inglés, poseen distintos niveles de especialización, aunque todos ellos han sido producidos por especialistas.

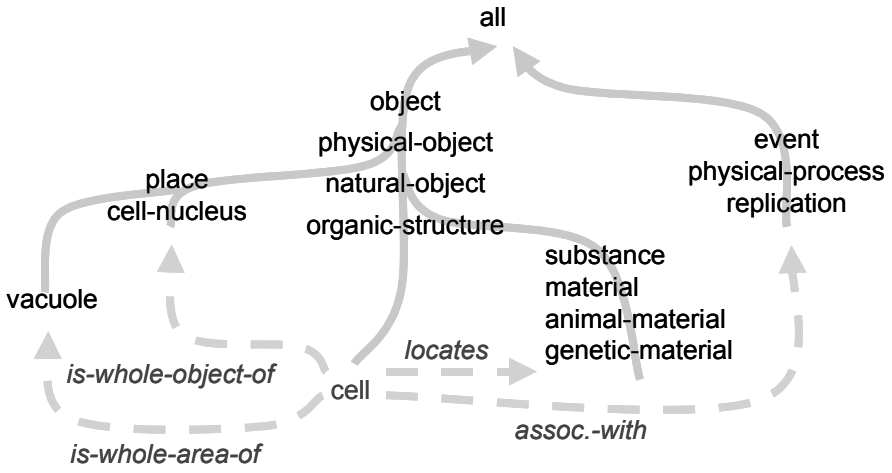
Mediante la aplicación de la cadena de procesamiento automático disponible en nuestro Instituto, este corpus está etiquetado morfológicamente, con lo que es factible la búsqueda en línea de información por cadenas de caracteres, por formas, por lemas, por categorías, o por la combinación de cualquiera de estas posibilidades, a través de la herramienta BwanaNet, de la que ya hemos hablado anteriormente, que, aunque se encuentra disponible en la Web del IULA, se ha incorporado en este caso directamente a la plataforma GENOMA, como veremos a continuación.



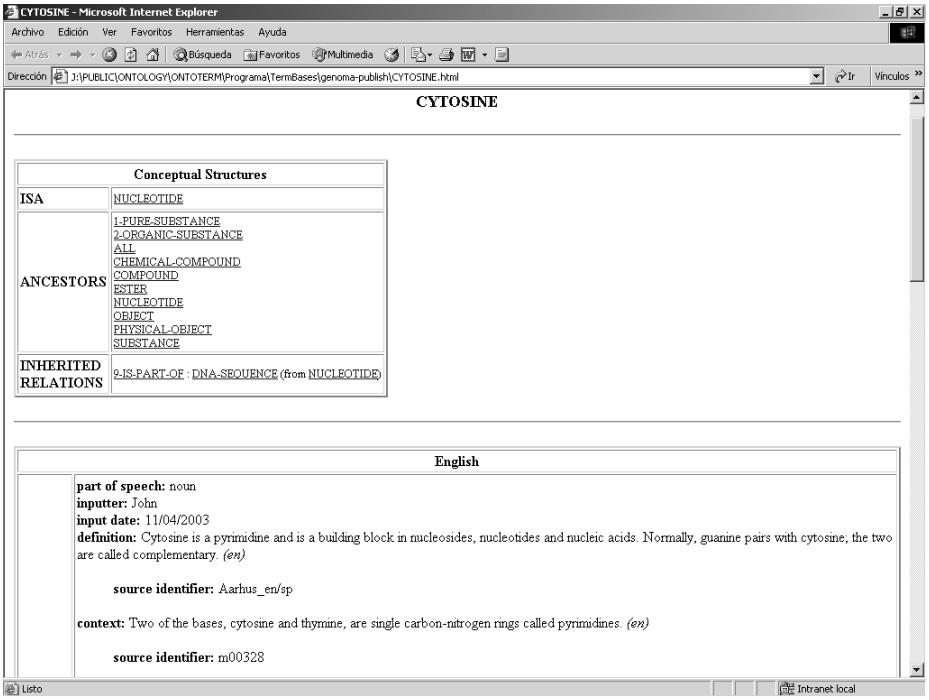
El desarrollo del módulo terminológico ha sido paralelo al del módulo ontológico. Ello obedece a uno de los principios que, sobre la terminología, asume nuestro marco teórico denominado Teoría comunicativa de la Terminología o TCT (CABRÉ 1999, 2000 y 2003): desde una aproximación lingüística a la terminología no es posible acceder directamente a los conceptos, sino solo cabe hacerlo a través de las unidades terminológicas que los representan. El respeto a este principio requiere que para cada término existirá en la ontología un concepto asociado, al cual solo se accederá partiendo del término que le corresponde.

La creación de un registro del módulo terminológico ha supuesto declarar previamente un concepto en la ontología, y solo posteriormente se ha llevado a cabo la creación del registro terminológico asociado.

Para el desarrollo del módulo ontológico hemos utilizado el gestor OntoTerm desarrollado por el profesor Antonio MORENO de la Universidad de Málaga a partir de la ontología de Mikrokosmos. OntoTerm está compuesto por un editor de ontologías, un editor de terminología, un navegador y un generador de información en formato HTML. Véase a continuación la información conceptual que para el concepto CELL contiene la ontología:



Y véase además la información conceptual que, procedente de la ontología, contiene un registro terminológico, en este caso del término inglés *cytosine*:



CYTOSINE - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Dirección J:\PUBLIC\ONTOLOGY\ONTOTERM\Programa\TermBases\genoma-publish\CYTOSINE.html

CYTOSINE

Conceptual Structures	
ISA	NUCLEOTIDE
ANCESTORS	1-PURE-SUBSTANCE 2-ORGANIC-SUBSTANCE ALL CHEMICAL-COMPOUND COMPOUND ESTER NUCLEOTIDE OBJECT PHYSICAL-OBJECT SUBSTANCE
INHERITED RELATIONS	9-IS-PART-OF : DNA-SEQUENCE (from NUCLEOTIDE)

English

part of speech: noun
inputter: John
input date: 11/04/2003
definition: Cytosine is a pyrimidine and is a building block in nucleosides, nucleotides and nucleic acids. Normally, guanine pairs with cytosine; the two are called complementary. *(en)*

source identifier: Aarhus_en/sp

context: Two of the bases, cytosine and thymine, are single carbon-nitrogen rings called pyrimidines. *(en)*

source identifier: m00328

Listo Intranet local

4.3 Situación actual

El corpus genómica humana incluido en el de medicina comprende el siguiente número de ocurrencias:

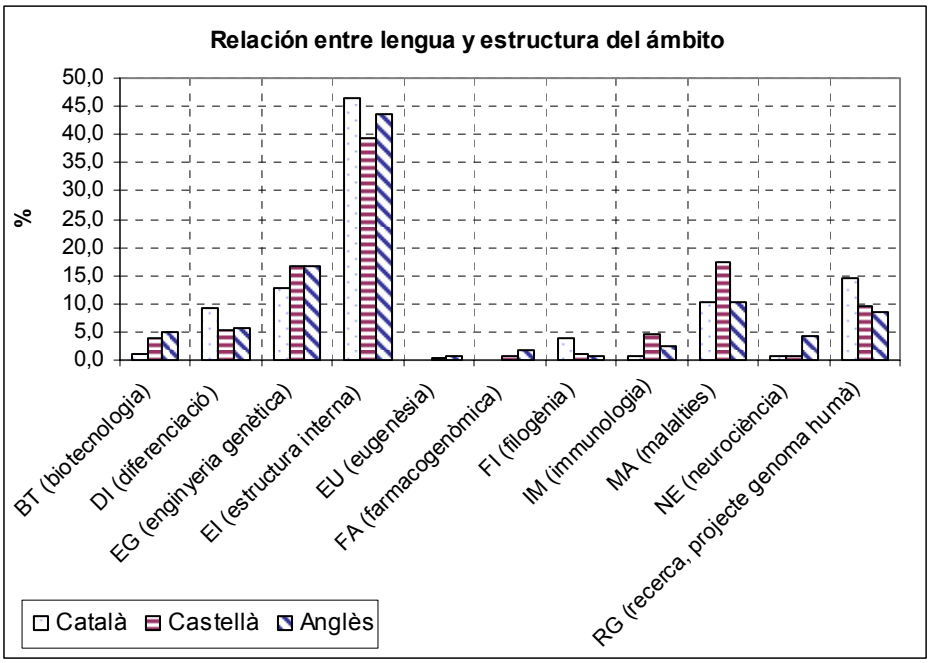
	Catalán	Castellano	Inglés	Total
Número de ocurrencias	945.164	1.371.104	1.119.347	3.435.615
Número de documentos	133	218	223	574

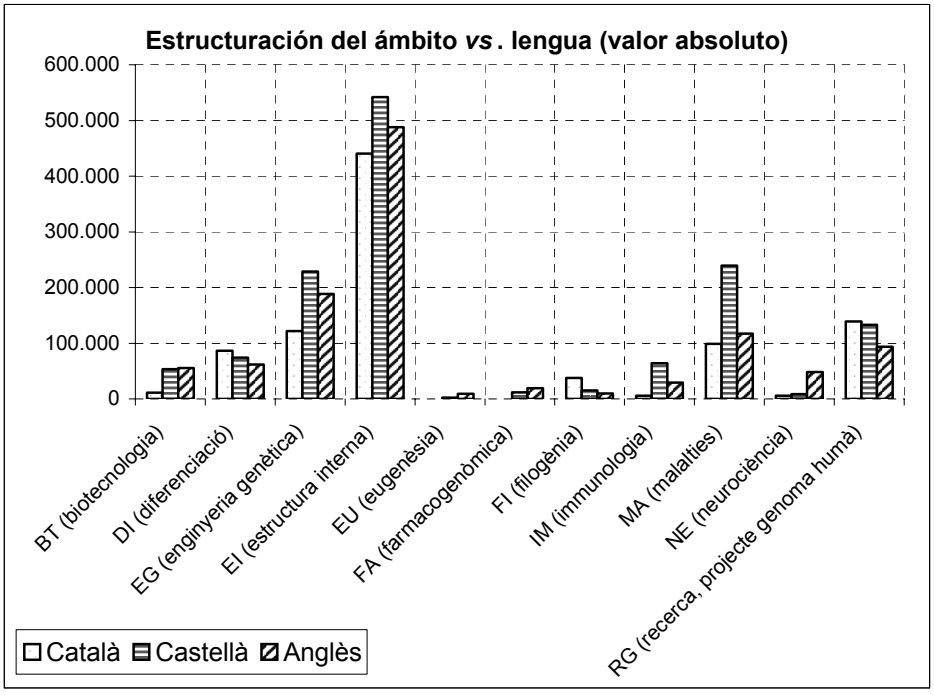
Las ocurrencias por pares de lenguas son en este momento las siguientes:

Castellano-Inglés	Catalán-Castellano
515.613	10.310

El número de ocurrencias por subámbitos es el siguiente:

Estructuración del ámbito BIOM	Catalán	Castellano	Inglés	Total
BT (biotecnología)	11.425	53.128	5.5557	120.110
DI (diferenciación)	86.180	74.211	61.746	222.137
EG (ingeniería genética)	121.598	228.585	188.022	538.205
EI (estructura interna)	440.189	541.687	487.900	1.469.776
EU (eugenesia)	-	2.473	8.959	11.432
FA (farmacogenómica)	-	11.877	18.940	30.817
FI (filogenia)	37.302	15.056	9.634	61.992
IM (inmunología)	5.394	63.952	29.402	98.748
MA (enfermedades)	98.838	239.161	117.015	455.014
NE (neurociencia)	5.454	8.435	48.440	62.299
RG (investigación)	138.814	132.539	93.732	365.085
Total	945.164	1.371.104	1.119.347	3.435.615


Relación entre lengua y estructura del ámbito
%
50,0
45,0
40,0
35,0
30,0
25,0
20,0
15,0
10,0
5,0
0,0
BT (biotecnologia)
DI (diferenciació)
EG (enginyeria genètica)
EI (estructura interna)
EU (eugenèsia)
FA (farmacogenòmica)
FI (filogènia)
IM (immunologia)
MA (malalties)
NE (neurociència)
RG (recerca, projecte genoma humà)
Català
Castellà
Anglès

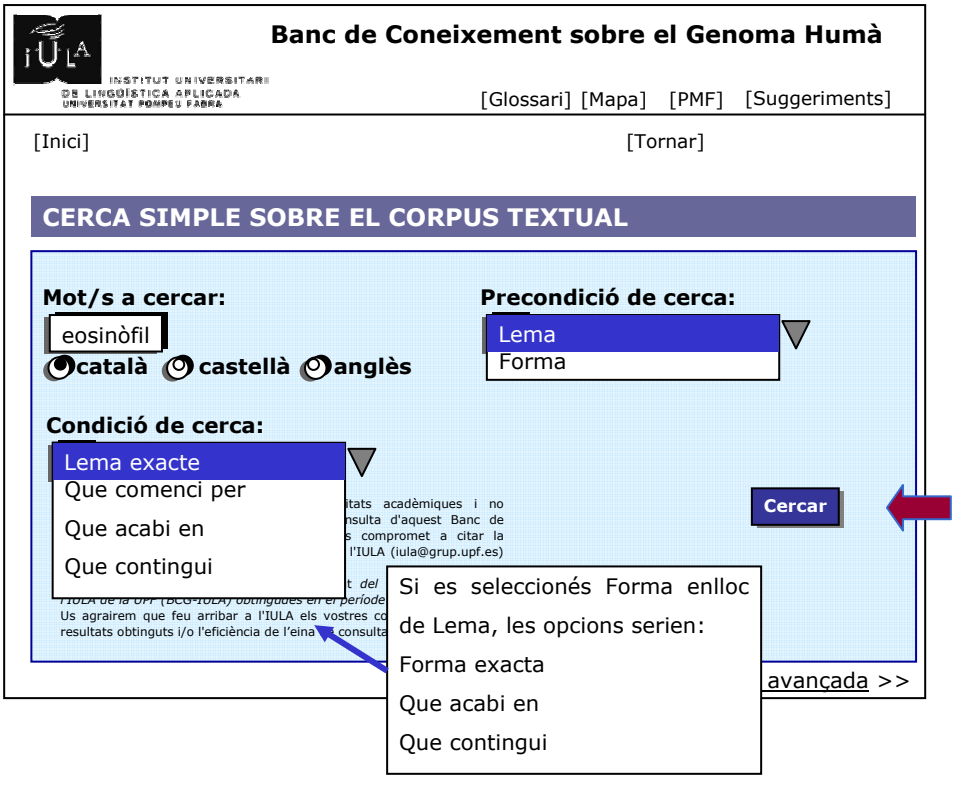

Estructuración del ámbito *vs*. lengua (valor absoluto)
600.000
500.000
400.000
300.000
200.000
100.000
0
BT (biotecnologia)
DI (diferenciació)
EG (enginyeria genètica)
EI (estructura interna)
EU (eugenèsia)
FA (farmacogenòmica)
FI (filogènia)
IM (immunologia)
MA (malalties)
NE (neurociència)
RG (recerca, projecte genoma humà)
Català
Castellà
Anglès

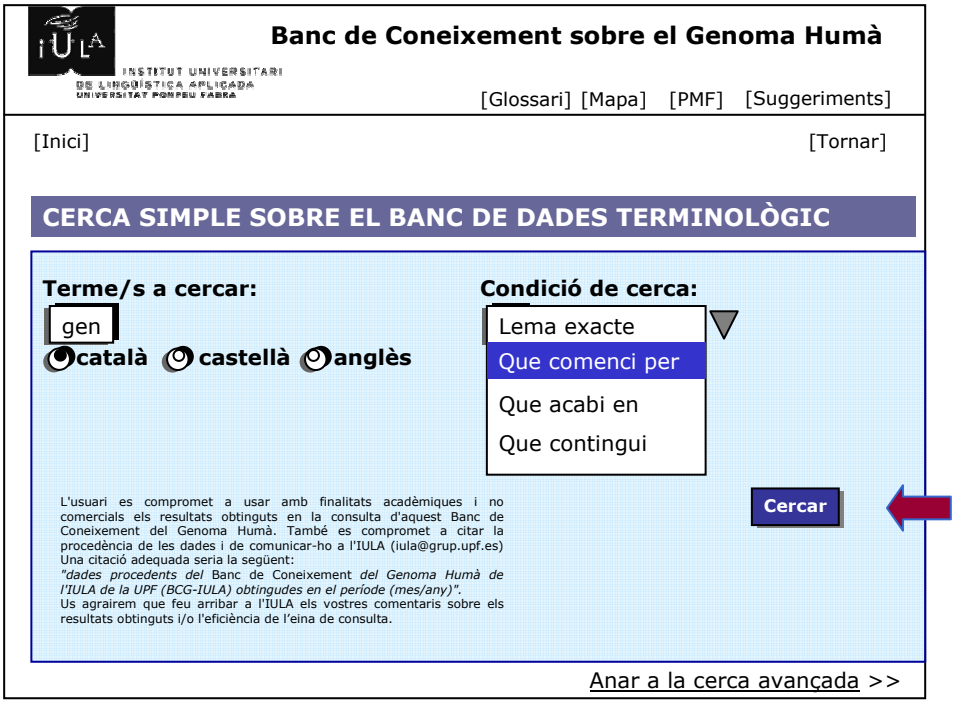
4.4 Acceso a la información del banco GENOMA

La consulta de GENOMA puede hacerse a través de tres tipos de búsqueda:

A la búsqueda simple,
B la búsqueda compleja,
C la búsqueda combinada (todavía no desarrollada).

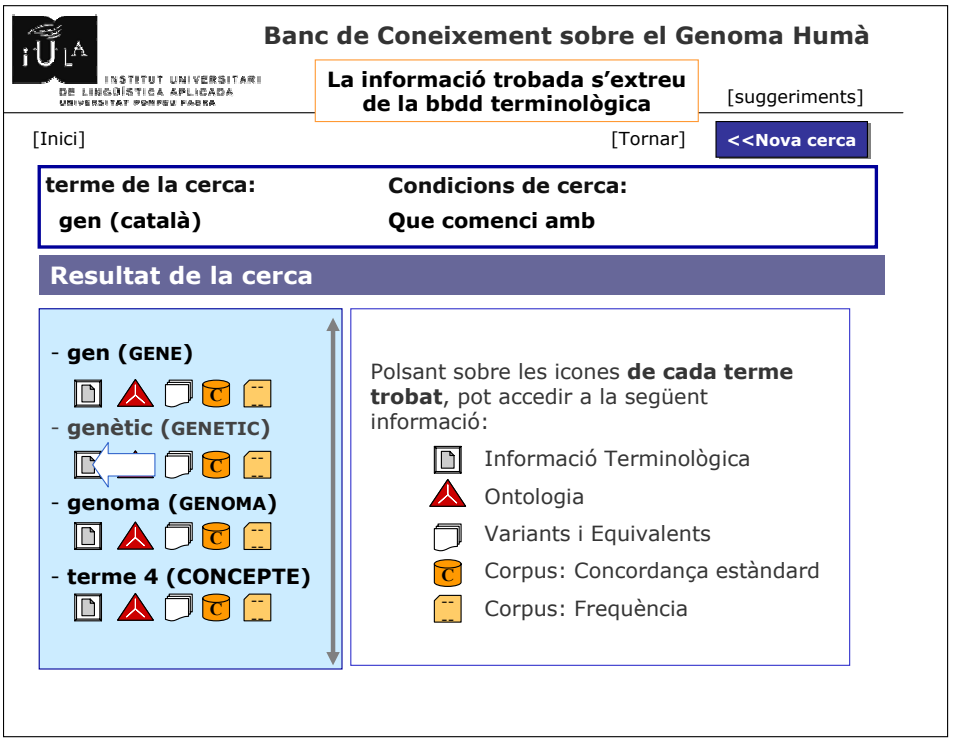
Mediante una búsqueda simple puede consultarse información sobre una unidad (forma o lema) o una cadena de caracteres en el corpus textual o en el banco terminológico, como muestran las figuras siguientes:



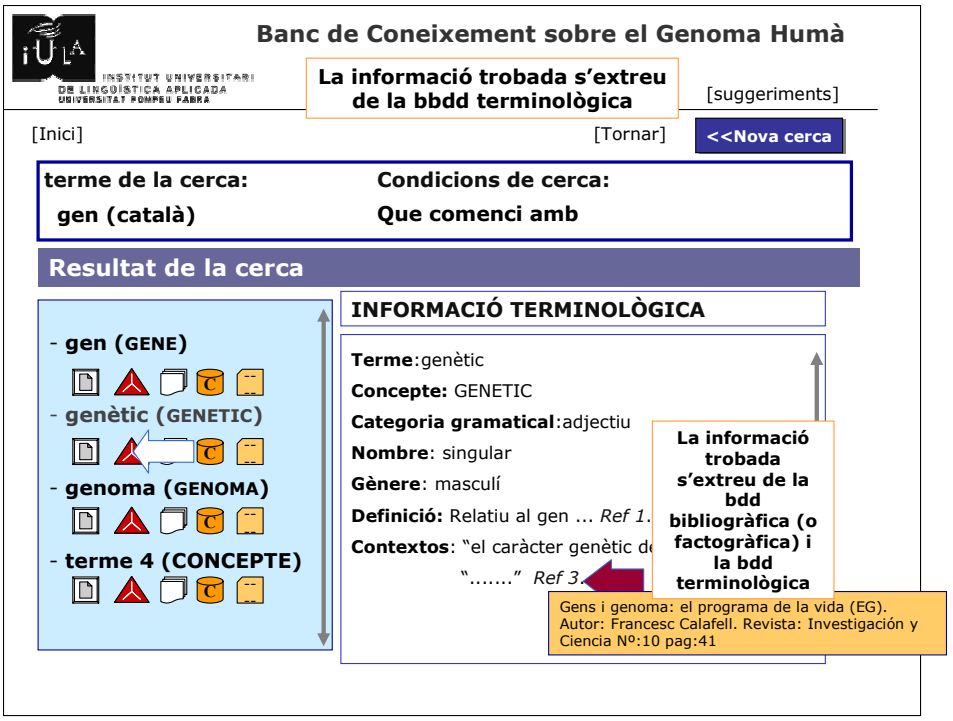


Una vez definida la búsqueda y reconocida la unidad, podemos obtener de ella información asociada, ya sea procedente del corpus textual, del banco documental, del banco terminológico o de la ontología:

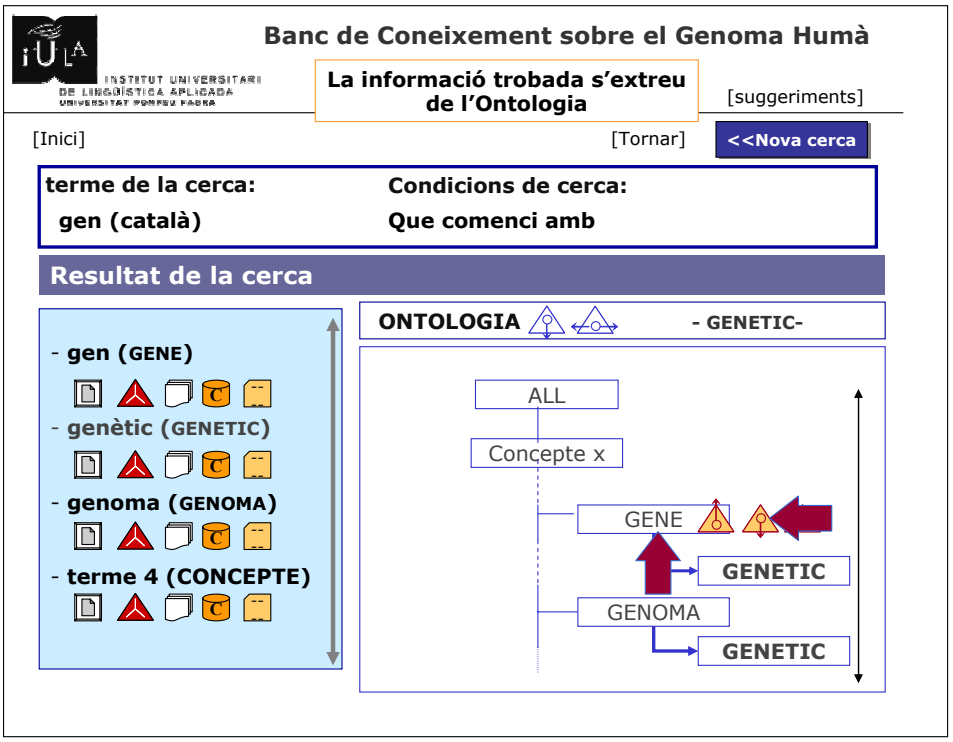
- su frecuencia en un corpus de textos o subcorpus previamente seleccionado,
- sus ocurrencias en el corpus o subcorpus presentadas en forma de concordancias,
- sus ocurrencias en un entorno gramatical o léxico predefinido,
- las fuentes documentales en que tal unidad aparece,
- su registro en la base de datos terminológica, si esta es su condición léxica,
- su valor conceptual en la ontología.



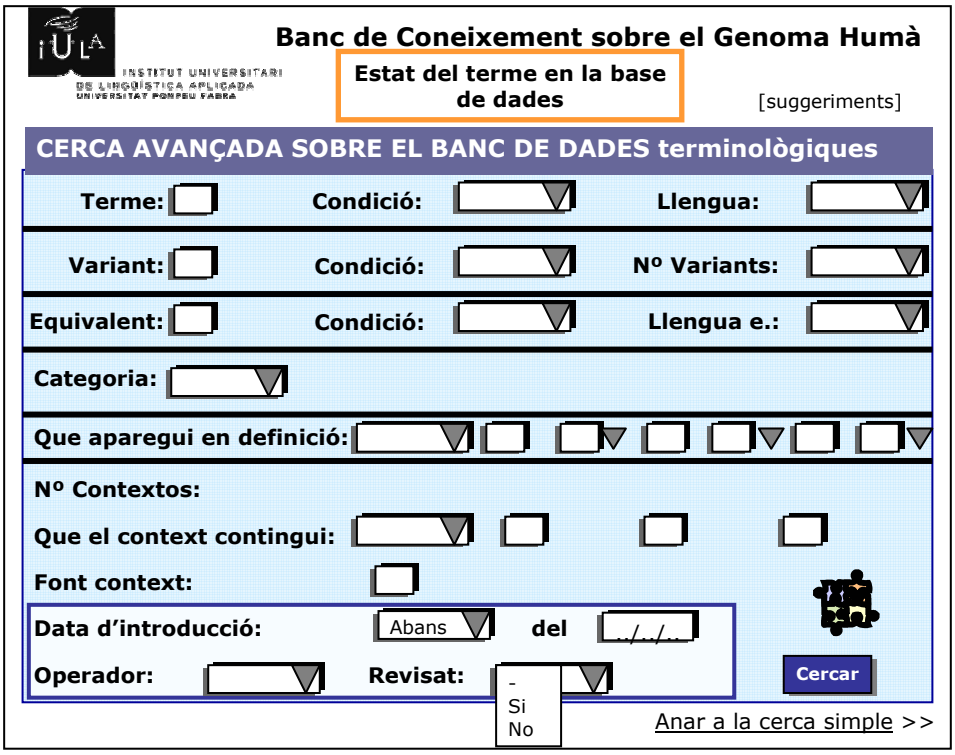
A través de la búsqueda de una unidad en la base terminológica podemos obtener para cada unidad incluida en ella información sobre su categoría y subcategorías, su definición, sus variantes, si las posee, en la misma lengua y sus equivalentes en otras lenguas:



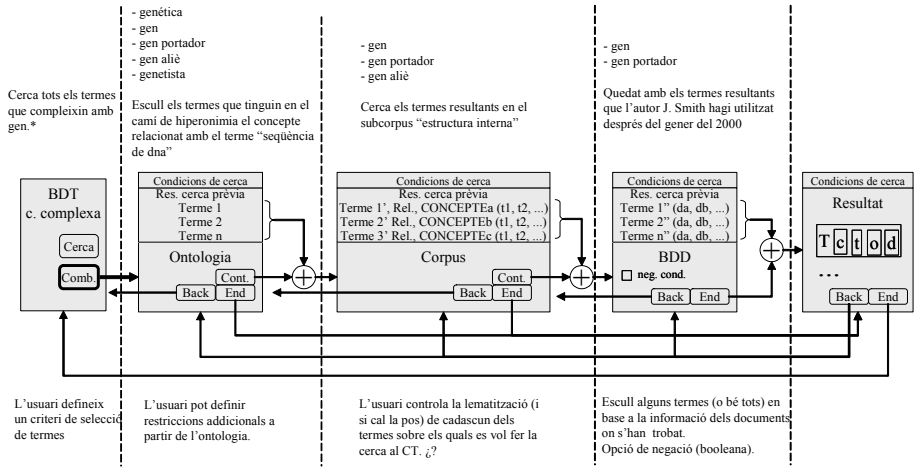
Y además todas las relaciones conceptuales que la unidad establece con otros conceptos del mismo ámbito, esta información, como hemos visto en la figura incluida en 4.2, procedente de la ontología.



Una búsqueda compleja permite obtener listas de unidades que cumplen determinadas condiciones en uno o más campos, combinadas mediante los operadores booleanos.



Finalmente, la búsqueda combinada, actualmente solo diseñada, pero no desarrollada, permitirá, a partir de cualquiera de los módulos, definir una búsqueda de todas las unidades que cumplan determinadas condiciones en el primer módulo, y restringir los resultados de esta primera búsqueda con búsquedas progresivas con condiciones en el resto de los módulos.



5 Como conclusión

No hay duda a nuestro entender de que las tecnologías de acceso a la información han modificado muy significativamente el proceso de trabajo de un traductor. Las consultas habituales a los diccionarios impresos se ven hoy acompañadas, por no decir superadas, por las consultas en línea a diccionarios electrónicos en red y por la búsqueda de información sobre las unidades de traducción o directamente la obtención de unidades equivalentes en la lengua hacia la que se traduce.

Pero paralelamente al desarrollo de motores de búsqueda más potentes y de filtros más refinados, la ingeniería lingüística ha ido explorando en el diseño de aplicaciones más adecuadas a la resolución de las necesidades de los colectivos profesionales dedicados a las lenguas. Uno de estos recursos son los bancos de conocimiento especializado, que, sobre la base de la terminología estructurada conceptualmente, es decir, sobre la base de un banco de términos asociado a una ontología, se han aplicado a la gestión del conocimiento de las organizaciones complejas.

En nuestro caso, hemos concebido los bancos de conocimiento especializado como un recurso lingüístico que integra en una misma plataforma las unidades terminológicas, su estructuración conceptual, los textos reales de los que proceden y las referencias de dichas fuentes. Y

toda esta información estructurada en bases de datos relacionados entre sí y accesibles transversalmente. Esperamos que los traductores puedan reconocer en este tipo de recurso una vía de resolución de algunas de sus necesidades.