Anna Aguilar-Amat (anna.aguilar-amat@uab.es)
Pilar Sánchez-Gijón (pilar.sanchez.gijon@uab.es)
Grupo Tradumática, Departamento de Traducción e Interpretación de la Universidad Autónoma de Barcelona (España)

BACUS (Base de Conocimiento Universitario): una ontología para el lenguaje especializado.

1. En torno a la Terminología

En la actualidad existen muchas herramientas que pueden facilitar el trabajo del mediador lingüístico; en algunas ocasiones incluso obviarlo. Nos referimos a programas de Traducción Asistida y/o Automática, y también a algunos sistemas que permiten obtener información léxica de los textos (o córpora), mediante la combinación de recursos estadísticos y de análisis gramatical[1]. Sin embargo, aunque éstas herramientas pudieran ponerse al servicio de un programa de Traducción Automática, existen razones sociolingüísticas que afectan al proceso de Traducción de tal manera que sólo podemos concebir un proceso de Traducción Automática que trabaje con la información anteriormente procesada por un traductor humano. Si tomamos como ejemplo el procesamiento de un término acuñado en los años 90 como "ordenador de quinta generación" (cf. Aguilar-Amat, 2002), que, como término complejo, supuso en su momento un reto interesante, y lo situamos en el contexto actual, vemos que, en España, ha sido substituido por el término simple 'Pentium'. Éste es un proceso muy común en todos los lenguajes de especialidad, los únicos lenguajes que, en principio, pueden ser traducidos automáticamente. Por lo general los términos inician su decurso conceptual con una *anatomía* fuertemente descriptiva, para luego metamorfosear en formas más analógicas o reducidas. Esto es mucho más acusado en el caso de las lenguas que importan (traducen) tecnología, y más todavía cuando la lengua receptora es de origen latino. Las lenguas románicas tienen menos agilidad que, por ejemplo, las anglogermánicas, para crear términos por composición de modificadores en número superior a uno, y a menudo prefieren la analogía:

carretilla hidráulica de elevación de horquilla (traducción literal del alemán) -> toro

También sucede a menudo que los términos son substituidos por sus siglas o, simplemente, elididos en parte o substituidos.

pequeña y mediana empresa -> PYME
procesador esclavo -> servidor

Existe, pues, un conflicto que se resuelve sólo con el tiempo entre la necesidad de traducir rápidamente para obtener la información que precisa la ciencia y el mercado y la consolidación de los términos, la cual viene determinada solamente por el uso de los mismos.
Este último tipo de cuestiones supone, tal como sugiere el tipo de inversión pública en el sector educativo, un campo de estudio tal vez mucho más apasionante que el de la Traducción Automática. Dónde, cuándo, qué, cómo y a qué se traduce es, al fin y al cabo, mucho más importante que el instrumento o la velocidad, si lo que estamos traduciendo no es una declaración de amor o de guerra.
Para estos otros casos, y cuando desconocemos por completo la lengua a la que nos enfrentamos, es suficiente con el hecho de que la Traducción Automática nos ofrezca un mensaje 'correcto' desde el punto de vista binomial (sí cuando es sí, no cuando es no), algo que se puede lograr con un buen diccionario y algunas reglas. En lo que se refiere al lenguaje natural, el control de la producción y la síntesis de significado puede efectuarse mediante la utilización de ontologías como la que se desarrolla en el Computing Research Laboratory de la Universidad de Nuevo Méjico, a cargo de Sergei Nirenburg (http://crl.nmsu.edu/). Una ontología es una representación del mundo donde los conceptos y sus denominaciones se definen en razón de sus características como procesos, entes, o propiedades. La ontología es una jerarquía taxonómica que permite desambiguar la polisemia por medio de la selección de los significados próximos en el árbol. Podemos ver un ejemplo en las frases siguientes:

a) La multiplicación de grupos de immigrantes preocupa a las autoridades.

---

[1] De entre los pioneros podemos destacar OCP (Oxford Concordance Program) y Tact (Universitat de Toronto), pero existen muchas otras herramientas que trabajan sobre textos monolíngües o bilíngües (ver Sánchez-Gijón, 2002). Estos productos permiten obtener listas de palabras ordenadas de diversas maneras, extraer contextos y efectuar comparaciones de textos en distintas lenguas. Actualmente estos programas (Word Smith Tools, Kwick, System Quirck) son de gran utilidad para el traductor humano. Existen también programas que ayudan en la recopilación y clasificación de los córpora.

b) La multiplicación de grupos posee la propiedad conmutativa.

El conocimiento del mundo permite, en el caso de un lector humano, distinguir entre los diferentes significados de 'multiplicación' y de 'grupo'. Un programa de TA basado en una ontología como la del CRL reconocerá la polisemia de estas unidades para luego decidir una lectura en función de la coincidencia y proximidad que poseen algunas de las características en a) geografía humana, y en b) matemáticas.
Pero el magnífico trabajo ontológico subvencionado por la NASA que puede traducir titulares de periódico de lenguas como el persa, el coreano o el ruso, no es lo bastante sutil como para acercarse a los lenguajes especializados. Un concepto como 'ojo anterior lateral' que pertenece a la anatomía de la araña no puede ser procesado por la ontología de Nirenburg. En estos casos la traducción palabra a palabra es la única solución, aunque el resultado puede que no sea muy convincente. Una expresión como 'agua física metálica' sería el resultado de traducir del griego, palabra por palabra, el concepto de 'agua mineral natural', por poner un ejemplo.
Lo cierto es que la Traducción Automática que será tendrá que pasar necesariamente por una combinación de estrategias: lógicas, estadísticas, probabilísticas, pero en cualquier caso no podrá eludir la Ingeniería del Conocimiento.

2. El proyecto BACUS (Base de Conocimiento UniverSitario)

BACUS (BAse de Conocimiento UniverSitario, 1996-2001) es una base de datos que resulta del trabajo de investigación terminológica multilingüe que realizan los alumnos de cuarto curso de la Facultad de Traducción e Interpretación de la Universidad Autónoma de Barcelona. Se trata de un trabajo descriptivo que parte de textos originales en diferentes lenguas sobre un mismo tema, con el requisito de homogeneidad de los textos en cuanto a nivel de lengua y actualidad. Se exige a los alumnos la colaboración de un experto en la especialidad que van a tratar. Aunque los datos no se someten, de momento, a un criterio de normalización, las referencias contextuales, lexicográficas y terminográficas colaboran a documentar al máximo cada caso. Se calcula un incremento aproximado de BACUS en unos 1500 nuevos términos por año académico.
Las áreas de trabajo son, básicamente, técnicas y científicas, pero no se excluye la posibilidad de que otro tipo de áreas temáticas puedan ser representadas. Los alumnos escogen el área temática libremente, acogiéndose, si lo desean, a las propuestas de los profesores de nuestra universidad que se prestan como asesores. Cada término está orientado temáticamente a través de tres ejes: el área temática, el título del trabajo (el cual indica el nivel máximo de especifidad) y el código de la UNESCO que más se ciña a ese nivel de especifidad.
La definición de esta base de conocimiento se ha realizado en el entorno MultiTerm para Windows, y pretende ser una plataforma para representar la información léxica de modo jerarquizado y relacionado. En su día se escogió este programa por haber sido acordado como estándar por la Coordinadora Interuniversitaria para la Normalización Lingüística e Informática (CINLI) de los países de habla catalana.
Para establecer una base de conocimiento multilingüe hemos partido de cuatro criterios: relatividad conceptual, atomización del conocimiento, exclusión de la gramática y navegación conceptual.
Cuando se comparan distintas lenguas se están comparando sistemas conceptuales diferentes, lo cual presenta una dificultad en la búsqueda de equivalencias léxicas. Debemos aceptar, pues, la relatividad de nuestro propio sistema y sustituir el ideal de biunivocidad y exactitud por un sistema plural, para que, a ser posible, ningún sistema conceptual ejerza de *director de orquesta*. Cada tema tiene como lengua de referencia aquella en la que cuenta con mayor refinamiento conceptual.
La información debe representarse en la forma más atomizada posible. Esto puede provocar que la información contenida en la definición y/o el contexto sea repetida en un campo particularizado (por ejemplo *color*, *forma*, etc.). Pero, precisamente, la finalidad de la atomización del conocimiento es evitar caer en la trampa lingüística de la paráfrasis definitoria. Esto puede ser importante cuando queremos representar y comparar terminologías de culturas distantes. Además, aunque la estructura de campos semánticos (unos 50) no es un conjunto completo ni perfecto, una información atomizada puede facilitar el tratamiento y la utilización de la información si se pretende integrar la base de datos en un sistema de procesamiento más amplio, como pueda ser un sistema de TA.
Las cuestiones gramaticales no atañen a la base de datos, la cual se centra en la estructura conceptual. Esto significa, por ejemplo, que los nombres predicativos y los verbos son considerados sinónimos conceptuales. Así, *ejecución* y *ejecutar* son dos índices para el mismo concepto. Se adjunta información sobre la categoría gramatical de cada denominación; así se indica género femenino en los sustantivos de lenguas flexivas como *silla* o *mesa*, pero se dejará abierto en los demás casos para que pueda flexionar según la gramática de la lengua en cuestión.
La flexibilidad en la representación de la información de la que BACUS hace gala también afecta al nivel monolingüe. Por ejemplo: podemos recoger un concepto constituido por la expresión *ejecutar un programa*, pero este término complejo también puede ser representado si creamos una entrada para *programa* y la relacionamos con otra entrada correspondiente al concepto *ejecutar* a través de un determinado campo que

indique el tipo de relación verbo-objeto existente (el campo 'argument of'). Esta última representación es preferible, porque todos los conceptos que tengan a *programa* como hiperónimo pueden quedar asimismo relacionados con *ejecutar* a través de las relaciones cruzadas.
El léxico no se compone sólo de denominaciones que vinculan la realidad con los conceptos, sino también de relaciones interléxicas que configuran una red de familias. Estas familias también se relacionan unas con otras a través de relaciones lógicas, ontológicas, analógicas, infralógicas o argumentales. Una base de datos terminológica que reflejase solamente una información glosariada, documentada y ponderada, no sería una base de conocimiento. Son precisamente las remisiones las que permiten al usuario navegar a través de un sistema conceptual.
Una muestra de la base de datos BACUS:

(gráfico)

2.1. Hipótesis conceptual de BACUS
(Aguilar-Amat, 1996)

Relaciones conceptuales

- **lógicas - verticales:**
  - **superconcepto/subconcepto**
    - **. genérico/específico**
      - **(pala/spade-shovel)**
    - **. hiperónimo/hipónimo**
      - **(mineral/antracita)**
- **- horizontales:**
  - **. coordinación (pragmáticas)**
    - **(mesa/silla)**
  - **. cohiponimia**
    - **(antracita/pirita)**
- **analógicas - transferéncia:**
  - **. metáfora**
  - **. metonimia**
- **ontológicas - parte/todo:**
  - **. parte/parte (radio/llanta)**
  - **. todo/parte (rueda/radio)**
  - **- causa/efecto (VIH/sida)**
- **infralógicas**
  - **. localizaciones (cebra/África)**
  - **. cronológicas (Edad Media/Edad Moderna)**
  - **. sustancia (chupete/caucho,silicona)**
  - **. genéticas (Microsoft/MS-DOS)**
  - **. transmisión (televisión/telespectadores)**
  - **. instrumentales (agua oxigenada/desinfección)**
  - **. pragmáticas (divorcio/juicio, pensión, tutela)**
- **argumentales (biomasa/extracción)**

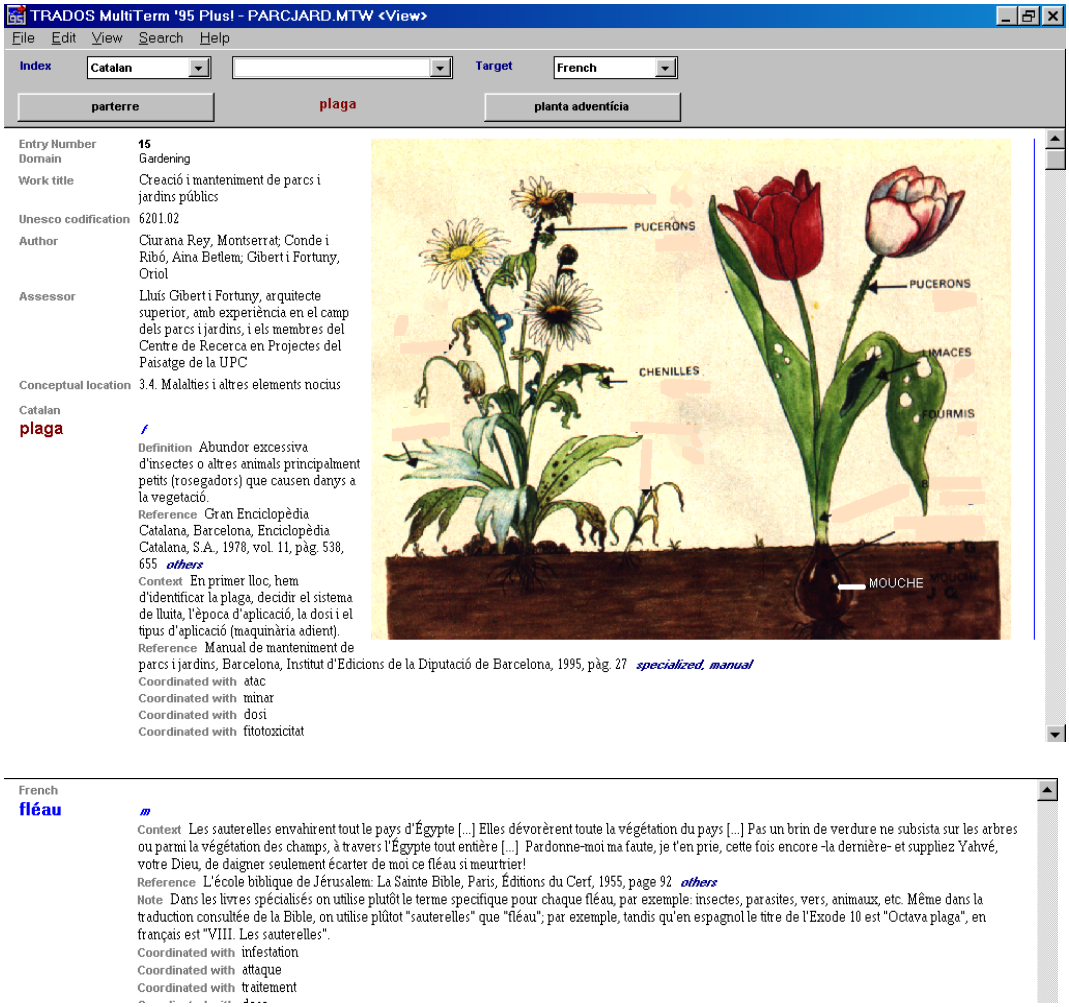
Referencias

AGUILAR-AMAT, A. (1996): *Logical Organization of Information at BACO: A Knowledge Multilingual Database for Translation Purposes*. Proceedings of the TKE'96 4th International Congress on Terminology and Knowledge Engineering, p. 238-244.

AGUILAR-AMAT, A. (2002): *Traducción, computación, utopia*. Conferencia impartida en la sede de la Comisión Europea en Bruselas. http://europa.eu.int/comm/translation/bulletins/puntoycoma/73/pyc734.htm

SÁNCHEZ GIJÓN, P. (1999): *Documentación y construcción de un corpus digital monolingüe sobre Astronomía: criterios de clasificación para su aplicación a la investigación y a la docencia de la traducción especializada.* Trabajo de investigación, Universidad Autónoma de Barcelona.

SÁNCHEZ GIJÓN, P. (2002): *Explotación de un corpus electrónico con textos de Internet sobre Astronomía.* Conferencia impartida en la sede de la Comisión Europea de Bruselas.



French

plaie *f*

Context Les chancres sont provoqués par des champignons se développant sur les plaies (de taille, de gel, etc.)

Reference Guide Clause, Brétigny-sur-Orge, édité par L. Clause, 1979, page 584 *specialized*

Spanish

plaga *f*

Context Las enfermedades producidas por organismos animales, especialmente insectos, se vienen a denominar plagas, mientras que las enfermedades propiamente dichas consisten en alteraciones que sufren los árboles debidas al ataque de microorganismos parásitos.

Reference F. Navés: El árbol en jardinería y paisajismo, Barcelona, Ediciones Omega, S.A., 1992, pág. 200 *specialized*

Coordinated with roer

Coordinated with insecticida

Coordinated with plaguicida

Coordinated with acaricida

Coordinated with caldo

Coordinated with caldo nicotinado

Subsequent tratamiento fitosanitario

Subsequent fumigar

Agent procesionaria del pino

Agent heteróptero

Agent pulgón