EUROTRA OPERATIVO (1991-1992)

T.Badia, N.Bel, J.Vidal

1. Introducción

El principal objetivo político de la Comunidad Económica Europea es la creación de un mercado unificado en 1992 para competir y cooperar con las configuraciones geopolíticas de dimensiones comparables (EEUU, Japón, China, el mundo árabe, la URSS, etc.). Sin embargo, la gran diferencia entre la CEE y los países mencionados reside en el hecho de que la Comunidad mantiene doce lenguas oficiales mientras que los demás sólo una. La lengua puede convertirse, pues, en el mayor obstáculo: cualquier producto que haya de ser distribuido por el territorio europeo ha de ir acompañado de documentación en el idioma del usuario y, cuánto más sofisticado sea el producto mayor será la documentación y mayor, por tanto, el coste de su traducción. Como cualquier intento de reducir el número de lenguas ha fracasado estrepitosamente, parece que la única solución es reducir el coste de la traducción gracias a la automatización.

Como respuesta a este planteamiento, el Consejo de la CEE aprobó el 4 de noviembre de 1982 la adopción de un programa de investigación y desarrollo de un sistema de traducción automática multilingüe de diseño avanzado: EUROTRA (82/752/EEC).

El objetivo operativo del programa era la creación de un prototipo capaz de trabajar con todas las lenguas oficiales de la Comunidad en un campo limitado y para textos de categorías limitadas, que proporcionara la base para un ulterior desarrollo a escala industrial. Asimismo el programa tenía como segundo objetivo el impulsar los estudios y investigaciones en Lingüística Computacional en general y en Traducción Automática en particular en los países de la Comunidad, dado que en los años 70 y 80 se había observado que los Estados Unidos de América y el Japón conseguían un nivel de investigación y desarrollo en este área mucho más elevado que el de los países de la Comunidad en su conjunto.

2. Planificación y organización de EUROTRA-I

Como consecuencia del segundo de los objetivos mencionados no se podía concebir que el programa EUROTRA tuviera una estructura altamente centralizada, con poca participación de equipos de investigadores de los países miembros de la Comunidad. Un equipo de trabajo, altamente especializado, radicado en un único centro, podía haber alcanzado, quizás, el primero de los objetivos propuestos y haber conseguido un prototipo preindustrial de traducción automática entre las distintas lenguas oficiales de la Comunidad. Ciertamente, ello habría acarreado numerosas dificultades, puesto que difícilmente se habrían podido tener en cuenta las tradiciones lingüísticas de los diferentes países, pero con toda seguridad se puede afirmar que se habría podido elaborar un prototipo de traducción automática similar al que finalmente se ha conseguido. No obstante, lo que no se habría logrado es el creciente interés por los temas de la Lingüística Computacional en las comunidades de investigación y educación superior de los distintos países miembros de la CE, así como el alto grado de interacción entre ellas.

2.1 Organización general

De ahí, pues, que desde un principio se estructurara EUROTRA como un proyecto descentralizado con núcleos de investigación en cada uno de los países de la CE. En cada país, pues, se ha creado uno o más grupos de investigación que han llevado a cabo las tareas básicas

de investigación y desarrollo del prototipo. La distribución de las tareas entre todos ellos ha sido la siguiente:

- Los grupos de Alemania, Dinamarca, España, Grecia, Italia, Portugal y el Reino Unido se han ocupado, cada uno de ellos, de su propia lengua (alemán, danés, español, griego, italiano y portugués, respectivamente).

- El grupo belga fue dividido en dos subgrupos que han tratado el neerlandés y el francés, en colaboración con los grupos holandés y francés, respectivamente.

- Finalmente, los grupos de Luxemburgo e Irlanda se han ocupado de tareas específicas, al margen de la investigación sobre una lengua concreta. El grupo de Luxemburgo ha cumplido con las funciones de distribuidor de material lógico, y el irlandés ha trabajado en terminología, produciendo y manteniendo un programa para el tratamiento de entradas consideradas 'términos' en EUROTRA y sirviendo de monitores para todo lo relacionado con este campo.

Los grupos dedicados a una lengua han tenido a su cargo los módulos de análisis y síntesis de esta lengua y todos los módulos de traducción de las demás lenguas a ella.

El trabajo de estos grupos ha contado con la ayuda de un equipo de la Comisión que ha aportado el soporte administrativo y técnico necesario en cada momento.

Naturalmente, el trabajo de todos estos equipos ha tenido que ser coordinado y dirigido de tal manera que el conjunto de los esfuerzos haya servido para la consecución del objetivo primario del proyecto. Por ello se han creado una serie de equipos centrales encargados de esta tarea. Entre ellos destacan:

- el *Grupo de Enlace*, al que han asistido los jefes de los grupos de investigación de cada país y, por parte de la Comisión, el jefe del proyecto. Las funciones del Grupo de Enlace han sido el seguimiento del trabajo día a día, la planificación general de las tareas lingüísticas y la decisión sobre las opciones científicas y técnicas que afectaban a todo el proyecto.

- el *Grupo de Investigación Lingüística*, que ha asumido y coordinado toda la planificación de la investigación. Este grupo a su vez se hallaba dividido en otros dos con responsabilidades específicas:

a) el *Grupo de Especificaciones Lingüísticas*, encargado de coordinar la investigación dentro del proyecto y autor del llamado 'Manual de Referencia'. Este Manual, una vez aprobado por el Grupo de Enlace, pasaba a ser vinculante para todos los grupos, que habían de producir gramáticas conforme a lo establecido en éste.

b) el *Grupo de Pragmática*, que tenía las tareas siguientes:

* suministrar ejemplos específicos en una lengua ilustrando así el uso óptimo de los mecanismos propuestos en el Manual de referencia.

* responder a las demandas y problemas expresados por los grupos nacionales.

* revisar los informes de los grupos nacionales para extraer ideas útiles.

* validar la legislación suministrando material para la formalización o sugiriendo cambios en la legislación.

- El *Grupo de Diccionarios*, que elaboró un manual específico para los diccionarios generales del sistema. Además este grupo se encargaba de estudiar, en colaboración con el grupo de terminología, el diseño de un procedimiento para incorporar la terminología en el

diccionario general de EUROTRA.

- El *Grupo de Software*, encargado de desarrollar el sistema en todo lo referente al prototipo, interficies con bases de datos y entorno de usuario.

- El *Grupo de Planificación*, encargado de dar directrices sobre las actividades y objetivos a desarrollar en cada período de tiempo tanto por los grupos centrales como por los nacionales.

Toda la investigación lingüística (excepto la específica de problemas monolingües) se ha planificado, controlado y coordinado de forma centralizada y ha sido responsabilidad del Grupo de Investigación Lingüística y del 'Problem Office' (oficina encargada de la gestión directa de la investigación). La selección de prioridades y la clasificación de los temas de investigación, tanto a corto como a largo plazo, dependían de las necesidades generales de la traducción.

El procedimiento de funcionamiento del 'Problem Office' ha sido el siguiente:

1.El Grupo de Especificación Lingüística preparaba descripciones de los problemas mencionados en el plan de investigación, se encargaba de sugerir propuestas de tratamiento y de remitirlas al 'Problem Office'. Éste, a su vez, se encargaba de difundirlas invitando a todos los grupos nacionales a enviar sus propias propuestas de tratamiento.

2.Cada grupo nacional podía entonces preparar las propuestas y hacerlas públicas.

3.El 'Problem Office' recogía las propuestas, las evaluaba -mediante consulta con el Grupo de Especificaciones Lingüísticas- y asignaba la investigación a las partes interesadas.

Este procedimiento no excluía la posibilidad de incoar iniciativas por parte de los grupos nacionales.

Gracias a este procedimiento del 'Problem Office', los resultados obtenidos para cualquier investigación se convertían en legislación -es decir en criterios de seguimiento obligatorio- que el Grupo de Especificaciones Lingüísticas publicaba en el Manual de Referencia. En él se incluían:

. descripciones semiformales de los fenómenos lingüísticos.
. especificaciones formalizadas.
. ejemplos y ayudas pragmáticas.

Aparecía un Manual de Referencia cada seis meses, coincidiendo así con el principio de cada período de trabajo.

2.2 Participación del grupo español

El grupo español ha estado formado desde su incorporación al proyecto por dos grupos, uno ubicado en Barcelona y el otro en Madrid.

El equipo de Barcelona ha producido el módulo de análisis del español a partir del nivel de constituyentes y el de síntesis hasta este mismo nivel (tratando, por lo tanto, la estructura de constituyentes, la estructura relacional y la estructura de interficie en ambos sentidos), y los módulos y diccionarios de traducción de todas las lenguas tratadas en EUROTRA al español, y se ha llevado a cabo la investigación básica y aplicada relacionada con ellos.

En Madrid se han producido los módulos morfológicos de análisis y síntesis y se confecciona el diccionario monolingüe junto con el trabajo lexicográfico y terminológico. También se lleva a cabo la investigación básica y aplicada para su desarrollo.

Durante los dos últimos años uno de los miembros del equipo español ha estado formando parte del grupo central de Investigación Lingüística, primero en el Grupo de Pragmática y luego en el de las Especificaciones Lingüísticas.

Distintos miembros del equipo español han participado en varias áreas de investigación en el marco de los trabajos gestionados por el 'Problem Office', contribuyendo, por tanto, a la definición de la legislación establecida en el 'Manual de Referencia'. Los temas en los que esta participación ha sido más importante son los siguientes:

- tiempo y aspecto
- modalidad
- estructura argumental de los verbos
- estructura argumental de los nombres
- relaciones semánticas para los modificadores
- rasgos semánticos léxicos
- morfología: composición y derivación
- determinación y semántica formal
- anáfora pronominal
- coordinación
- dependencias no ligadas

3.Los fenómenos gramaticales que han sido tratados en la gramática del español producida en el proyecto EUROTRA.

3.1. Introducción

La gramática contiene el tratamiento de la mayoría de los fenómenos que pueden encontrarse en el ámbito oracional y produce representaciones de cualquier tipo de frase aislada. Sin embargo, no proporciona un tratamiento de los fenómenos que se producen en el ámbito del discurso.

El sistema comprende, como ya hemos explicado en otros artículos[1], un analizador morfológico, una representación configuracional de las relaciones sintagmáticas (ECS), una representación relacional utilizando las funciones sintácticas (ERS) y una representación de interficie que además de las relaciones de dependencia contiene importantes parcelas de información semántica. Es obvio que no podemos presentar aquí todas las soluciones que hemos dado a todos y cada uno de los problemas gramaticales que hemos abordado, por limitaciones de espacio. Para ilustrar el alcance y sobre todo la línea de las orientaciones gramaticales que hemos seguido, añadiremos a los comentarios generales acerca de la cobertura de la gramática, algunas explicaciones adicionales sobre el tratamiento del nudo frase (O).

[1] Ver, por ejemplo, una somera presentación de EUROTRA en Cerdá, R. et al. "El tratamiento computacional de la lengua española en el proyecto de traducción automática EUROTRA", en *Actas del Congreso de la Sociedad Española de Lingüística XX Aniversario* (Tenerife, abril 1990); por otra parte puede encontrarse una descripción más técnica en el artículo de Arnold, D. y L. des Tombes "Basic theory and methodology in EUROTRA", publicado en el libro editado por S. Nirenburg, *Machine translation. Theoretical and methodological issues* , Cambridge University Press, 1987.

3.2. Tablas de fenómenos que cubre la gramática

tc: tratamiento completo
tp: tratamiento parcial
nt: no tratado

Fenómenos	análisis	generación
1. Oraciones principales	tc	tc
2. Categorías sintagmáticas Sintagmas nominales Sintagmas adjetivales Sintagmas adverbiales Sintagmas preposicionales	tc	tc
3. Oraciones subordinadas finitas (nominales y adverbiales)	tc	tc
4. Oraciones de infinitivo	tc	tc
5. Oraciones de participio y gerundio	tc	tc
6. Complementos oracionales de nomb.	tc	tc
7. Complementos oracionales de adj.	tc	tc
8. Coordinación	tp	tp
9. Tiempo y aspecto en or. principales	tc	tc
10. Tiempo y aspecto en subordinadas	tc	tc
11. Modalidad	tp	tp
12. Aposiciones	tp	tp
13. Movimiento de constituyentes - coindexación	tc	tc
15. Referencia pronominal	tp	tp
16. Construcciones independientes no oracionales (títulos, paréntesis)	tp	tp
17. Comparación	tp	tp
18. Elipsis	nt	nt
19. Negación	tc	tc
20. Determinación	tc	tc
21. Compuestos y lexías complejas	tp	tp

22. Coordinación con elipsis	nt	nt
23. Construcciones de verbo soporte	tp	tp
24. Modificadores temporales	tc	tc
25. Transconstruccionales	tp	tp

La gramática es capaz de analizar todo tipo de frases (principales y subordinadas, finitas e infinitas). También está preparada para analizar todos los otros nudos sintagmáticos que pueden aparecer autónomamente, como los sintagmas nominales de los títulos de los párrafos. En lo que se refiere al resto de los nudos sintagmáticos (SN, SP, SADJ, SADV, SDET), se recogen la mayoría de las configuraciones posibles. En el ámbito de los SSNN quedan fuera del análisis, por el momento, las aposiciones del tipo N + N (el presidente González), las oraciones de relativo que van entre comas y los sintagmas sin núcleo aparente (algunos, todos, etc.). Todos los predeterminantes se recogen en un único nodo SDET que se define a través del estudio de las restricciones de coaparición entre determinantes y sus posibilidades de combinación para formar estructuras partitivas. Los adverbios se clasifican sintáctica y semánticamente para calcular correctamente las distintas maneras de combinarse entre sí y con el resto de los constituyentes.

Las reglas de coordinación admiten cualquier combinación de SSNN, SADJ, SSPP y frases, pero no de sintagmas adverbiales, de constituyentes intermedios como N' o SV o constituyentes de distinta categoría.

Esto hace que por ejemplo no se puedan analizar frases como:

(1) La chica está ((preocupada) y (de mal humor))
SADJ SP

La coordinación que implica elipsis verbal, aunque ha sido estudiada y experimentada parcialmente en la gramática, no ha sido incorporada finalmente por el alto grado de sobregeneración que produce la necesidad de recuperar la estructura elidida. Esto sucede en frases del tipo:

(2) a. La comisión ha aprobado la medida pero el consejo no
b. La comisión ha aprobado^{1} la medida^{2} pero el consejo no (e^{1}) (e^{2})

El tiempo y el aspecto están tratados de forma completa tanto para las formas del indicativo como las del subjuntivo. También las formas no finitas están sujetas al cálculo temporal con el fin de facilitar la traducción, cuando las condiciones de infinitivización varían de una lengua a otra. Los modificadores temporales (adverbios, SSPP, SSNN y subordinadas adverbiales) reciben un tratamiento especial que permite clasificarlos como locativos, aspectuales o cuantificadores temporales. También reciben información temporal y aspectual que es utilizada para la desambiguación.

Algunas perífrasis aspectuales ("estar a punto de", "acabar de", "soler", "volver a", "seguir -ndo", etc.) son analizadas (en estructuras de control o de 'raising') y etiquetadas semánticamente para facilitar su traducción, que a veces requiere ciertos cambios estructurales. También las perífrasis modales reciben un tratamiento que se basa en la utilización de un conjunto de rasgos semánticos que las clasifican en torno a dos parámetros: el carácter epistémico o deóntico y el origen de la actitud proposicional (si la obligación o la información emana de un sujeto o no). Para ello se tienen en cuenta, además de los rasgos estructurales,

algunos valores semánticos como la agentividad.

Las oraciones de relativo y las interrogativas han sido objeto de una exhaustiva investigación, por lo que se contempla la posibilidad de que los constituyentes desplazados procedan de cualquier nivel de profundidad. Esto permite, por ejemplo, el análisis de frases como:

(3) El programa que dijo el presidente que quería aprobar el consejo es muy interesante.

Quedan por resolver algunos casos marginales como la extracción múltiple, y los pocos casos en que se puede relativizar un posesivo (p.e. "Aristóteles, de quien se conservan pocas obras, ...").

La gramática contiene también un tratamiento propio de las categorías vacías que son generadas en el análisis, tanto en los infinitivos (PRO), como en los desplazamientos de constituyentes (trazas) o los casos de sujeto de oración finita no especificado (pro). Para facilitar la recuperación de los pronombres en el caso de que los infinitivos hayan de ser transformados en formas finitas, el análisis proporciona la coindización de los constituyentes.

Los verbos que pueden llevar oraciones de infinitivo como complementos contienen información sobre el tipo de control que realizan. Los sujetos vacíos de las oraciones adverbiales están también coindizados con los constituyentes de los que se infiere que tienen la misma referencia. Los pronombres reflexivos y recíprocos no han recibido todavía un tratamiento adecuado que permita calcular su referencia.

En cuanto a la comparación, se han tratado las formas de comparación de todos los tipos de complejidad siempre y cuando se de en el ámbito del sintagma adjetival. Así pues, todas las frases de (4), en las que aparecen algunas elipsis, son cubiertas por nuestra gramática.

(4) a. La hija de Juan es más guapa que la de Pedro
b. Juan es tan inteligente que no necesita estudiar
c. Juan es demasiado joven para leer estas cosas

Las lexías complejas se tratan a través de unas reglas de reconocimiento de tipo general que se basan en la información que sobre las posibilidades de constitución de unidades de palabra contienen las unidades léxicas en el diccionario. Una vez reconocidas se convierten las secuencias en constituyentes únicos utilizando líneas de unión. Las que son invariantes ("encima_de", "a_pesar_de", etc.) se identifican en el nivel morfológico. Otras lexías más complejas que admiten inflexión, se reconocen, sin embargo, en los niveles superiores del análisis (p.e "poner_en_marcha", "dar_la_lata", etc.)

Otro tipo de estructuras a las que se ha concedido especial atención son las llamadas *estructuras de verbo soporte*. Llamamos estructuras de verbo soporte a aquellas construcciones del tipo V(transitivo)+ SN (O.Directo) en las que la fuerza de la predicación reside más en el nombre que en el verbo (p.e. "dar un paseo", "tomar la decisión", etc.). El problema que plantean estas estructuras es el de la traducción del verbo porque, al ser casi vacío de significado, su traducción en las diferentes lenguas responde a criterios heterogéneos y arbitrarios. Por ejemplo encontramos las siguientes equivalencias:

(5) a. dar un paseo (español)
b. take a walk (inglés)
c. faire un promenade (francés)

La representación que se propone consiste en eliminar el verbo soporte y sustituirlo por una unidad 'abstracta' que en principio no requiere traducción. De este modo el verbo se genera posteriormente con criterios exclusivamente monolingües. Los nombres que pueden formar parte de

este tipo de predicaciones llevan en el léxico información que da cuenta del verbo soporte con el que se combinan. La gramática no contiene la implementación de todas las estructuras de verbo soporte que encontramos en nuestra lengua pero sí un método para tratar todas ellas. Incluso se contempla la posibilidad de que un mismo nombre predicativo admita varios verbos soportes que corresponden a diferentes variantes aspectuales (p.e. "tomar la iniciativa", "perder la iniciativa", etc.).

Otro de los fenómenos que ha sido tratado con cierta extensión es el de los trasconstruccionales, que son aquellos modificadores que introducen información sobre la predicación que no es de tipo situacional (lugar, tiempo, modo, instrumento, etc.). Estos adjuntos se caracterizan porque introducen información de varios tipos:

- sobre actos de habla (p.e. "francamente, no quiero decirlo")
- información sobre la estructura del discurso (p.e. "por otro lado, ...")
- actitud frente a la proposición o evaluación de su contenido (p.e. "Evidentemente, la proposición es buena").

Su tratamiento ha requerido un estudio exhaustivo de los modificadores (adverbios, oraciones subordinadas, sintagmas preposicionales, construcciones absolutas, etc.) que nos ha permitido extraer de todo el conjunto los que deben considerarse como transconstruccionales y no como modificadores situacionales.

3.3.El tratamiento de la oración principal

Así como en el resto de los nudos sintagmáticos hemos seguido un patrón de representación que se ajusta más o menos al esquema general NUCLEO + COMPLEMENTOS nuestro planteamiento del nudo O se basa en criterios más pragmáticos que lingüísticos. En principio, disponemos de dos únicos patrones para analizar la estructura de la frase:

- oraciones finitas
- oraciones infinitas (infinitivo, gerundio y participio)

La diferencia entre ambas estriba en que la complejidad de los modificadores es mucho menor en el caso de las oraciones infinitas y por lo mismo las posibilidades de sobregeneración son menores.

Para facilitar la proyección de la estructura configuracional en la relacional hemos procurado que la estructura de frase sea lo más plana posible lo que nos ha conducido a una representación por campos que se definen por la función sintáctica que adquieren sus elementos. El esquema de la frase puede representarse más o menos como se indica en (5).

(5)

```
o |mod1      (grupo de modificadores)
  |sn/o      (posible sujeto u objeto directo extrapuesto)
  |mod2      (grupo de modificadores)
  |par   |coma
  |      |mod3    (modificadores entre comas)
  |      |coma
  |negación
  |sv    |sn      (clíticos)
  |      |gv      (grupo verbal)
  |      |sn      (posible sujeto invertido)
  |      |par2  |mod4 (modificadores que aparecen entre argumentos)
  |      |sn/sp/o/sadv/sadj  (posibles argumentos)
  |      |mod5    (modificadores en posición final)
```

Esta concepción de la frase presenta las siguientes características:

- Está basada en el estudio estadístico de las secuencias más frecuentes en la lengua.

- Es excesivamente poco restrictiva en cuanto a los órdenes posibles de palabras (sobre todo en relación a la coaparición de constituyentes), pero aumenta su poder restrictivo gracias a la acción de filtros que amortiguan sus efectos.

- Para aumentar su poder restrictivo en el nivel configuracional hemos hecho un uso parcial de valores semánticos. Por ejemplo, los SSNN pueden ser temporales y no temporales y esto nos ayuda a la hora de expresar que solamente los temporales pueden ir en la posición reservada a los modificadores.

- Está basada en el hecho de que los modificadores en español poseen una gran libertad de posicionamiento en la frase, especialmente si van entre comas.

- Para evitar que se produzcan enlaces inadecuados de los sintagmas preposicionales disponemos de una serie de filtros que son el reflejo de las siguientes regularidades:

. Si un verbo subcategoriza un sintagma preposicional con una preposición determinada P, siempre que aparezca en su entorno un sintagma preposicional con esa preposición, se toma como bueno el resultado en que éste constituyente aparece ligado al verbo y no a otro componente. Nótese que ello exige la codificación en el léxico del primer nivel, información sobre la forma de los SSPP que subcategorizan los verbos, los nombres y los adjetivos, que en teoría correspondería al segundo nivel de análisis. El proceso de selección del constituyente al que debe ir ligado un SP se completa en el último nivel donde las unidades léxicas contienen información sobre los rasgos semánticos que deben poseer sus complementos.

. Los constituyentes de más peso suelen aparecer al final de los nudos sintagmáticos.

. Los sintagmas preposicionales de la forma DE + SN reciben un tratamiento especial dada su abundancia y sus especiales caraterísticas. Sólo se admite que vayan ligados directamente al nudo O cuando el SN posee ciertos rasgos semánticos.

La representación de la frase en el nivel funcional se consigue gracias a una reglas de proyección que tienen en cuenta dos tipos de información:

. El orden en que aparecen los constituyentes
. La codificación completa de la valencia verbal que no sólo informa sobre las funciones sintácticas de los constituyentes que saturan al verbo, sino sobre todas sus peculiaridades sintácticas (preposiciones regidas fuerte o débilmente, categoría de los constituyentes, si admiten complementos oracionales o no, si admiten construcciones pasivas o no, si admiten la reducción de argumentos (voz media) o no, etc.

En el último nivel de representación la frase adquiere una fisonomía a medio camino entre la representación sintáctica profunda y la representación semántica. En una palabra, los argumentos reciben un conjunto de etiquetas que tienen un cierto parecido con los llamados roles semánticos. Algunos roles que se utilizan son: segundo participante, argumento de medida, argumento de lugar, atributo, etc. Los adjuntos oracionales aparecen también etiquetados respondiendo a otro sistema más complejo que el destinado a la catalogación de los constituyentes regidos. El cálculo del valor semántico de los sintagmas preposicionales se realiza mediante un sistema de reglas bastante complicado que tiene en cuenta los valores atribuidos en el léxico a las preposiciones y los valores semánticos del núcleo del sintagma nominal. Además, las frases contienen información acerca de la diátesis, entendida este

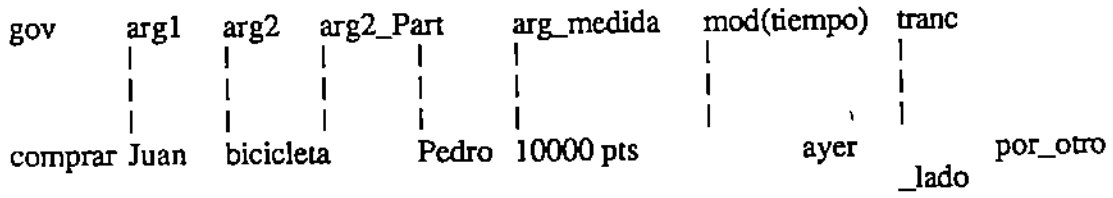
concepto no en un sentido morfológico sino semántico. Las predicaciones son clasificadas como construcciones ergativas, impersonales, pasivas, decausativas, etc. Con ello se consigue representar convenientemente oraciones como las de (6) y se facilita la traducción de los diferentes tipos de "se" que pueden aparecer en español.

(6) a. Juan ha roto el vaso
b. El vaso se ha roto

Consideremos, por ejemplo, la frase de 7a, que se representa en el nivel IS como queda indicado en 7b.

(7) a. Por otro lado, ayer le compró Juan a Pedro su bicicleta por 10.000 pts
b.

O



Como se puede apreciar, solamente los argumentos primero y segundo no reciben una etiqueta semántica. La razón de ello es que estos argumentos son los que presentan mayor regularidad en cuanto a la forma y al contenido en las diferentes lenguas europeas y no se precisa conocer el papel semántico que comportan para poder traducirlo correctamente. Otro aspecto de la representación que conviene hacer notar es que la diferencia entre modificadores y transconstruccionales nos permite representar parcialmente ciertas relaciones de alcance.

4. EUROTRA OPERATIVO

El proyecto EUROTRA finalizó formalmente en diciembre pasado. En este momento, como continuación de los estudios generados por él, nos encontramos con tres líneas de trabajo paralelas. Por una parte, se están elaborando las bases para el diseño de un nuevo formalismo para EUROTRA, que incorpore los últimos adelantos en lingüística computacional. Al mismo tiempo, prosigue el trabajo en el marco de los Contratos de asociación entre la Comisión y los países miembros de la CE, con la finalidad de mejorar el producto resultante del proyecto. Finalmente, se ofrecen a concurso las principales líneas de investigación que deben ofrecer como resultado las bases científicas sobre las que se debe asentar el nuevo sistema operativo de traducción automática.

4.1. Estudios para el diseño de un nuevo formalismo para EUROTRA

Durante 1989, la Comisión sacó a concurso la realización de estudios sobre la viabilidad de una nueva implementación, que pudiera ser usada a escala industrial, y sobre la definición de un diseño del mismo.

Dentro de estos estudios se contempla la realización de un sistema de análisis de la arquitectura general de EUROTRA que habrá de constituir las bases para desarrollos tecnológicos ulteriores tanto de sistemas de traducción automática como de otras aplicaciones y programas.

El desarrollo de este proyecto está bastante avanzado en la actualidad. Las líneas maestras que van a definir el nuevo formalismo pueden quedar reflejadas en las siguientes características:

- se distingue entre un núcleo central del formalismo y elementos periféricos.

- el núcleo central va a estar basado en una gramática libre de contexto con unificación de estructuras de rasgos.

- el formalismo es neutro en cuanto a la distinción entre niveles de descripción lingüística y/o análisis.

- las operaciones semánticas van a poder ser implementadas en el formalismo.

- un formalismo de transferencia, basado también en la unifica ción, va a ser implementado para operar entre las representa ciones de una lengua a las de otra.

Como puede verse, se pretende que el formalismo incorpore las mejoras que han aportado los sistemas que actualmente se están desarrollando (LFG, CUG, HPSG, CLG, PATR-II, CLE, FUG y GPSG). Por otra parte se ha tenido presente en todo momento que se trata de un formalismo que va a ser usado con léxicos y gramáticas grandes y no sólo para fines experimentales, por lo que la eficiencia en la computación es uno de los requerimientos básicos del sistema; de ahí, por ejemplo, que se siga pensando en una base libre de contexto.

3.2. Desarrollo de un sistema EUROTRA operativo (1991-1992)

Una vez obtenido un producto pre-industrial en el marco de EUROTRA, y habiendo sido elaborado ya un estudio sobre la viabilidad y la optimización del sistema, se ha previsto una nueva fase del proyecto de dos años de duración durante la cual llegue a implementarse un sistema operativo desarrollable industrialmente. Esta nueva fase tiene los siguientes objetivos:

. Implementación de un entorno estable de desarrollo de sistemas eficientes y de evaluación.

. Trabajo de desarrollo y de investigación monolingüe, destinado a la extensión del sistema y a la mejora del tratamiento lingüístico. Este punto debe incluir:

- revisión de los módulos lingüísticos ya existentes.
- extensión de la cobertura lingüística
- experimentación con nuevos modelos de traducción (lengua pivot)
- incorporación de los resultados de nueva investigación y la comprobación exhaustiva de los módulos

. Investigación lingüística (básica y aplicada)

. Investigación en arquitectura de sistemas

. Desarrollo de métodos e instrumentos para la reutilización de recursos léxicos en aplicaciones informáticas

. Creación de estándares para datos lexicográficos y terminológicos

La organización de esta nueva fase se basa en la participación de todos los países miembros renovando los contratos de asociación ya establecidos. La tarea de cada grupo nacional de investigación consiste, como en el pasado, en la elaboración de los módulos de análisis y generación de su propia lengua y de los de transferencia de otras lenguas hacia la suya

propia. Todos los grupos nacionales participan además en la investigación que es coordinada directamente por la Comisión, ya que los grupos centrales han dejado de existir.

En el caso español, el plan de trabajo para estos dos años contempla las siguientes tareas:

- cobertura gramatical: los fenómenos gramaticales descritos en la última edición del Manual de Referencia, que incorpora los resultados de la investigación llevada a cabo durante 1990.

- cobertura léxica: el léxico contemplado en el Manual de Referencia, para las palabras de clases cerradas, y el léxico de los textos de referencia (en el campo de las telecomunicaciones) para el vocabulario general y el terminológico.

- elaboración de los módulos de análisis y síntesis, según la cobertura gramatical y léxica.

- elaboración de los módulos de transferencia del inglés, alemán y holandés al español.

- investigación contrastiva: temas de investigación en los que los estudios contrastivos entre distintas lenguas tienen un papel primordial. Prevemos estudios contrastivos sobre los siguientes aspectos:

. tiempo y aspecto (en colaboración con los grupos inglés y holandés/belga)

. compuestos (en colaboración con los grupos inglés, alemán, danés y holandés/belga)

. estructura argumental (en colaboración con el grupo alemán)

. alcance/determinación/negación/cuantificación (en colaboración con los grupos inglés, alemán, portugués y holandés/belga)

- investigación monolingüe: temas de investigación que tienen una relevancia especial para el tratamiento del español. Prevemos estudios monolingües sobre los siguientes aspectos:

. oraciones sin sujeto explícito

. oraciones existenciales

. el orden de palabras en castellano

- investigación sobre metodología de la transferencia: estudio sobre la reversibilidad de las reglas de transferencia (en colaboración con los grupos alemán, portugués y holandés).

3.3. Temas presentados a concurso para la investigación lingüística básica

Esta investigación tiene como objetivo "mejorar gradualmente las prestaciones y los resultados lingüísticos del sistema y la calidad de la traducción". Está previsto que se organice siguiendo tres direcciones básicas:

(i) investigación lingüística general para aumentar la interlingualidad de la estructura de interficie y reducir la sobregeneración;

(ii) utilización de conocimientos específicos de un determinado campo temático;

(iii) utilización de restricciones de contexto y tipo de texto para reducir la sobregeneración del sistema.

Así, los temas de investigación se han agrupado a su vez en tres áreas: conocimiento lingüístico, terminología y conocimiento específico de un determinado dominio, teoría de la traducción automática y reutilización de recursos gramaticales, y subsistemas y aplicaciones. A continuación presentamos un pequeño resumen del contenido de cada una de ellas.

(a) Conocimiento lingüístico

El sistema de traducción EUROTRA está basado en la transferencia de la información de estructuras de interficie dentro de un sistema estratificacional. La 'estructura de interficie' es por lo tanto el área básica de investigación ya que se han de conseguir estructuras que sean al mismo tiempo interlinguales y lingüísticamente motivadas. La investigación ha de cubrir todos los elementos de la estructura de interficie y proveer un modelo lingüístico que sea mejorable y extensible incluso para incorporar conocimiento extralingüístico.

Subáreas

(1) Lexicón, morfología, Estructura y
Semántica formal en la estructura de interficie
(2) Rasgos semánticos léxicos
(3) Discurso

(b) Terminología y conocimiento específico de un determinado dominio

El sistema EUROTRA se ha basado principalmente en un análisis sintáctico profundo de oraciones. La investigación en este área debería estudiar la manera de incorporar información extralingüística que pueda ayudar a reducir la sobregeneración en el sistema de análisis. Sin embargo y como el campo de la representación del conocimiento del mundo es muy amplio, el estudio se ha de restringir al tipo de información relevante e incorporable a un sistema de traducción automática. Dentro de este área, la terminología tiene un lugar privilegiado ya que puede simplificar la tarea de la traducción. Sin embargo en este apartado el estudio se ha de fijar en la información útil que puede extraerse de un tratamiento no tanto lingüístico como de 'conocimiento del mundo' inherente a la definición terminológica.

Subáreas

(1) Representación del conocimiento
(2) Terminología

(c) Teoría de la traducción automática y reutilización de recursos gramaticales

La justificación del estudio sobre la reutilización de recursos gramaticales es la necesidad de explorar métodos y posibilidades de compartir conocimiento gramatical sobre determinadas lenguas con independencia del formalismo utilizado para expresarlas.

El estudio tendría que fijarse en las posibles interpretaciones de 'reutilización':

- organizar el conocimiento de forma sistemática, de tal forma que los usuarios puedan preguntar sobre las propiedades gramaticales de una lengua.

- representar el conocimiento como una gramática completa, codificada en una representación genérica que permita transformarla fácilmente para diferentes formalismos y diferentes aplicaciones.

Así, los temas de investigación se han agrupado a su vez en tres áreas: conocimiento lingüístico, terminología y conocimiento específico de un determinado dominio, teoría de la traducción automática y reutilización de recursos gramaticales, y subsistemas y aplicaciones. A continuación presentamos un pequeño resumen del contenido de cada una de ellas.

(a) Conocimiento lingüístico

El sistema de traducción EUROTRA está basado en la transferencia de la información de estructuras de interficie dentro de un sistema estratificacional. La 'estructura de interficie' es por lo tanto el área básica de investigación ya que se han de conseguir estructuras que sean al mismo tiempo interlinguales y lingüísticamente motivadas. La investigación ha de cubrir todos los elementos de la estructura de interficie y proveer un modelo lingüístico que sea mejorable y extensible incluso para incorporar conocimiento extralingüístico.

Subáreas

(1) Lexicón, morfología, Estructura y
Semántica formal en la estructura de interficie
(2) Rasgos semánticos léxicos
(3) Discurso

(b) Terminología y conocimiento específico de un determinado dominio

El sistema EUROTRA se ha basado principalmente en un análisis sintáctico profundo de oraciones. La investigación en este área debería estudiar la manera de incorporar información extralingüística que pueda ayudar a reducir la sobregeneración en el sistema de análisis. Sin embargo y como el campo de la representación del conocimiento del mundo es muy amplio, el estudio se ha de restringir al tipo de información relevante e incorporable a un sistema de traducción automática. Dentro de este área, la terminología tiene un lugar privilegiado ya que puede simplificar la tarea de la traducción. Sin embargo en este apartado el estudio se ha de fijar en la información útil que puede extraerse de un tratamiento no tanto lingüístico como de 'conocimiento del mundo' inherente a la definición terminológica.

Subáreas

(1) Representación del conocimiento
(2) Terminología

(c) Teoría de la traducción automática y reutilización de recursos gramaticales

La justificación del estudio sobre la reutilización de recursos gramaticales es la necesidad de explorar métodos y posibilidades de compartir conocimiento gramatical sobre determinadas lenguas con independencia del formalismo utilizado para expresarlas.

El estudio tendría que fijarse en las posibles interpretaciones de 'reutilización':

- organizar el conocimiento de forma sistemática, de tal forma que los usuarios puedan preguntar sobre las propiedades gramaticales de una lengua.

- representar el conocimiento como una gramática completa, codificada en una representación genérica que permita transformarla fácilmente para diferentes formalismos y diferentes aplicaciones.

- investigar la posibilidad de convertir el material lingüístico en soporte lógico en un componente de trabajo, por ejemplo, en un CD-ROM.

Por su parte, la investigación sobre la teoría de la traducción automática debe centrarse en cuestiones teóricas que puedan contribuir al desarrollo de la T.A., más que a la traducción en general. Algunas de las preguntas que el estudio ha de tener en cuenta son:

(i) sobre la naturaleza de la traducción: ¿qué propiedades de un texto son las que tienen que preservarse al traducirse éste de una lengua a otra?

(ii) ¿qué puede aprovecharse de la teoría de la traducción?

(iii) teóricamente, ¿cuáles son las ventajas y los problemas de reducir el alcance de los sistemas a lenguajes restringidos?

(iv) criterios de evaluación de los sistemas

Subáreas

(1) Reutilización de recursos gramaticales
(2) Terminología

(d) Subsistemas y aplicaciones

Se espera que este área se concentre en las especificaciones (e incluso la realización) de un prototipo preoperativo para investigación, capaz de ejecutar tareas bien definidas y que estén relacionadas o derivadas del trabajo realizado en EUROTRA. En ambos casos -subsistemas y aplicaciones- está prevista la participación de empresas privadas, que pueden participar tanto en calidad de futuro proveedor de productos que incorporen los resultados de los estudios de la investigación, o como usuarios potenciales que quieran invertir en actividades relacionadas con la solución de sus problemas específicos.

Subáreas

(1) Subsistemas
(2) Aplicaciones

El objetivo del estudio sobre los subsistemas es la realización de un prototipo para investigación de un sistema de traducción automática de alcance más limitado que el actual sistema Eurotra. El dominio del estudio estará entonces relacionado con la definición de las restricciones que habrá de tener el mencionado sistema para que resulte viable: definición de sublenguajes, interacción con el usuario, arquitectura del sistema, etc.

Por aplicaciones se entiende los productos que pueden derivarse del actual sistema Eurotra que no incorporen traducción. En relación con esto, hay que tener en cuenta que un sistema de traducción automática incorpora, por su naturaleza, todos los componentes que se relacionan con el Procesamiento del lenguaje natural. En este sentido, los diferentes componentes pueden emplearse por separado para la definición de productos con aplicaciones varias (correctores de estilo, programas educativos, indexación de textos, interficies con bases de datos, etc.).

5. Hacia la elaboración de un nuevo programa

Dentro del tercer Programa Marco de acciones comunitarias de I+D (1990-1994), aprobado por el Consejo de Ministros de la CE el pasado 23 de abril, la lingüística computacional y áreas relacionadas están contempladas dentro de la línea "Sistemas telemáticos de interés general" en el apartado 6, "Investigación e Ingeniería lingüística". El objetivo de este área es desarrollar una tecnología lingüística básica que pueda incorporarse a las aplicaciones computacionales donde el lenguaje natural es un ingrediente esencial y tiene en cuenta la necesidad de superar las dificultades que suponen la diversidad lingüística de la CE. Está directamente dirigido a la creación de recursos lingüísticos (gramáticas, diccionarios, colecciones terminológicas, y corpora de textos para las nueve lenguas oficiales) y a la definición de normas estándar para la codificación y almacenamiento de estos datos. Además prevé el desarrollo de aplicaciones piloto y proyectos de demostración que muestren cómo se puede utilizar esta tecnología y demostrar la viabilidad técnica y económica de las soluciones adoptadas.

El área está dividida en tres partes: investigación, desarrollo de recursos y aplicaciones piloto. Cada una de ellas tomará como base los resultados y la experiencia ganada en programas específicos de la Comunidad (EUROTRA, proyectos ESPRIT, ...) y proyectos de investigación de los diferentes países.