4. Lexicografía computacional

TGE: Un entorno para la generación de enlaces léxicos multilingües

Alicia Ageno, Francesc Ribas, German Rigau, Horacio Rodríguez.
Departament de Llenguatges i Sistemes Informàtics.
Universitat Politécnica de Catalunya.
horacio@lsi.upc.es

1. Introducción

El objetivo último de los proyectos Acquilex y Acquilex II[1] , en el marco de los cuales se inscribe este trabajo, es la exploración de la viabilidad del uso de fuentes de conocimiento ya existentes para la construcción de sistemas de procesamiento del Lenguaje Natural. Acquilex se centró en el uso de diccionarios de uso humano en soporte magnético (**MRD**, machine readable dictionaries) mientras que Acquilex II presta más atención a la utilización de córpora.

En el marco de Acquilex se logró la construcción de Bases de Datos Léxicas (**LDB**, Lexical Data Bases) a partir de los correspondientes **MRDs** y la posterior integración en una Base de Conocimientos Léxica Multilingüe (**LKB**, Lexical Knowledge Base) de la información semántica extraída de la **LDB**. Este proceso de integración se llevó a cabo mediante la utilización de un entorno (**SEISD**) que actuaba como soporte de una metodología concreta de extracción de información (ver [Ageno et al.,92a]).

El diseño y construcción de la **LKB** (ver [Copestake,91b]) se basó en la expresión de la información sintáctica y semántica extraída en forma semiautomática de las diferentes **LDBs** en un lenguaje de expresión léxica (**LRL**, Lexical Représentation Language) basado en mecanismos de unificación por defecto y herencia por defecto y capaz de soportar reglas léxicas (ver [Copestake,91a] para más detalles). Todo ello se apoyaba en una estructura de tipos subyacente que permitía la definición, en forma declarativa, de la información presumiblemente común a las diferentes lenguas presentes en el proyecto. Aunque Acquilex no se planteaba en sus

[1] *Acquilex: Esprit BRA 3030, Acquilex II: Esprit BRA 7315*

objetivos el estudio concreto de la posible aplicación de la **LKB** a la traducción automática (T.A.), parece obvio que una Base de Conocimientos Multilingüe puede constituir un elemento importante para realizar T.A.

En los últimos tiempos han aparecido varias aproximaciones a la T.A. que proponen el uso de formalismos basados en la Unificación para modelizar el proceso de Transferencia. Los motivos, por otra parte igualmente aplicables a otras áreas de tratamiento del L.N., se refieren a la expresividad del modelo, a su bidireccionalidad y al uso de una forma de expresión declarativa. Aproximaciones en las que la transferencia se realiza a nivel léxico, como el "Shake and Bake" de Whitelock (ver [Whitelock,92]) o a nivel lógico, como el **BCI** de Alshawi (ver [Alshawi, 92]) son ejemplos conocidos de esta tendencia.

Podemos considerar que en cualquiera de los modelos anteriores el proceso de traducción (análisis, transferencia, generación) se lleva a cabo a través de la unificación de estructuras sujetas a determinadas restricciones monolingües o bilingües. Si las reglas de análisis o generación pueden considerarse (en los sistemas basados en unificación) como restricciones monolingües que se aplican sobre estructuras de rasgos bien formadas, podemos considerar las reglas de transferencia como restricciones bilingües, igualmente expresadas en forma declarativa, que se aplican entre las estructuras de rasgos que corresponden a los textos fuente y objetivo.

La **LKB** supone un marco excelente para expresar tanto las restricciones monolingües correspondientes a las unidades léxicas de las lenguas fuente y objetivo como las restricciones bilingües que ligan entradas léxicas de las dos lenguas. Más aún, la capacidad expresiva del **LRL** nos permite no limitarnos a los signos léxicos sino establecer restricciones a nivel de signos frasales.

El trabajo que aquí presentamos se centra en la construcción de un entorno que permite la extracción en forma semiautomática de estas relaciones entre entradas léxicas de lenguas distintas. La organización del documento es la siguiente: Tras esta introducción, la sección 2 nospresenta

el tema de los tlinks o relaciones de transferencia entre entradas léxicas. En la sección 3 se introduce el entorno de desarrollo **TGE** (Tlinks Generation Environment). En la sección 4 describimos el proceso de trabajo dentro de **TGE**. La sección 5, finalmente, presenta unas conclusiones y líneas futuras de actuación.

2. Tlinks

Para realizar la integración de lexicones monolingües, que resulte en una sola **LKB** multilingüe (partiendo de la teoría de que el desarrollo de lexicones es un proceso incremental) , es necesaria una metodología, así como la correspondiente herramienta que la implemente. Esta se encargará de llevar a cabo (en paralelo con la definición de las reglas de transferencia correspondientes a las diferentes lenguas) la proyección sentido a sentido, de forma semiautomática. Para ello se usarán reglas de transferencia puramente léxicas, y como fuentes de conocimiento, tanto las representaciones de las entradas léxicas de la **LKB** como la información procedente de **LDBs** bilingües. Con este sistema se tratará de conseguir lexicones multilingües integrados no triviales, sino esencialmente ilustrativos de la casuística que se descubre al efectuar la mencionada proyección.

Las relaciones entre acepciones de palabras las representaremos en términos de lo que en el entorno **LRL/LKB** se denominan **tlinks** (translation links) [Copestake et al., 92]. Un tlink es una estructura por medio de la cual se expresa que dos estructuras de rasgos (las denominadas "estructuras de salida") se van a considerar equivalentes (es decir, traducción una de la otra). El tlink codifica la relación existente entre las acepciones de palabras de entrada y las correspondientes estructuras de salida, de manera que esta relación se puede considerar como una forma de transformar entradas léxicas en pares de equivalentes en cuanto a traducción.

Aunque la implementación más inmediata del concepto tlink habría sido mediante un enlace entre las estructuras de rasgos de la entradas léxicas

fuente y destino, para facilitar la utilización del mecanismo formal de las reglas léxicas, se ha optado por la definición de un tlink en términos de las relaciones entre una regla léxica en el nivel del lenguaje fuente y otra en el nivel del lenguaje destino (aunque frecuentemente ocurra que estas reglas se reduzcan a la identidad). Así, *fs1* representa la primera regla mencionada y *fs0* la segunda, y en ambas, los rasgos *1* y *0* indican la entrada y salida de la regla respectivamente. En términos de la **LKB**, el tlink se define dentro de la jerarquía de tipos de la siguiente forma:

tlink (top)
<fs0> = rule
<fs1> = rule
<fs0 : 0 : sem : id> = <fs1 : 0 : sem : id>.

(donde *(top)* indica que el tipo tlink es subtipo del tipo principal; *<fs0 : 0>* y *<fs1 : 0>* serán equivalentes en cuanto a traducción, lo cual se expresa mediante la tercera línea, que establece la coincidencia entre los valores de la interpretación semántica en ambas estructuras de rasgos ; además, *<fs0 : 1>* y *<fs1 : 1>* habrán de estar instanciadas a las correspondientes entradas léxicas para producir un tlink completo).

El más común y a la vez más simple de los casos es obviamente el de de equivalencia directa entre acepciones de palabras; este caso se representa por medio de un subtipo de tlinks, lo que se denominará un **simple-tlink**:

simple-tlink (tlink)
<fs0 : 0> = <fs0 : 1>
<fs1 : 0> = <fs1 : 1>

En el cual las dos entradas léxicas son una traducción directa de la otra, sin precisar ninguna clase de transformación. Un posible ejemplo es la equivalencia entre la primera acepción de la entrada *absinth* de la **LKB** y la también primera de la entrada *absenta*[2]:

[2] *En este y el resto de los ejemplos del artículo, los MRD's a partir de los cuales se han derivado las correspondientes LDB's española e inglesa (y a partir de éstas, los correspondientes lexicones de la LKB) han sido respectivamente el diccionario VOX (Vox, 87) y el LDOCE (Ldoce, 78). Asimismo, el diccionario bilingüe a partir del cual se ha originado la LDB usada en la traducción ha sido el VOX Harrap's (Vox 92).*

simple_tlink
<fs0 : 1> = absinth_L_0_1
<fs1 : 1> = absenta_X_1_1.

O bien, en notación abreviada:

absinth_L_0_1 / absenta_X_1_1 :
simple-tlink.

Aunque un amplio número de las entradas léxicas presentan este tipo de equivalencia (Alshawi, para el sistema BCI, cita unos resultados de un 69.9% de los casos [Alshawi, 92]), se comprueba la existencia de situaciones en las cuales no hay una equivalencia directa acepción-acepción, sino que aparecen lo que se denomina "desajustes" (mismatches) léxicos. Estos desajustes pueden deberse a procesos morfológicos en la formación de las palabras en una de las dos lenguas (derivaciones y composiciones), a procesos semánticos (extensiones de sentidos metafóricas o metonímicas), etc. (véase en [Soler, 93] un estudio más detallado de un subconjunto de los *mismatches* entre español e inglés). Por esta causa, aparecen otros tipos más complejos de tlinks, algunos de los cuales se definen a continuación (no se presentará la definición de estos tlinks en términos del sistema de tipos; para más detalles acerca de ésta, ver [Ageno et al. 93]).

En algunos casos no existe una traducción equivalente simple para una entrada, pero sí para su hiperónimo. Cuando esto ocurre se dice que hay un **partial-tlink** entre la acepción original y la traducción directa de su hiperónimo. Es el caso por ejemplo de la tercera acepción de la entrada *agasajo* ; para ella no se encuentra traducción directa, pero sí para su hiperónimo (la primera acepción de *refresco* , que presenta un simple-tlink con la primera acepción del segundo homónimo de la entrada *drink*). Por tanto, tendríamos el siguiente tlink:

agasajo_X_1_3 / drink_L_2_1 :
partial-tlink.

El **LRL** permite varias formas de expresar generalizaciones (definición de nuevos tipos de tlinks que incorporen restricciones sobre los ya existentes,

expresión de restricciones sobre las circunstancias en que se puede llevar a cabo una traducción,....).

Aún así, existen casos más complejos, como aquellos en los que una de las estructuras de rasgos ha de sufrir una cierta transformación (vía regla léxica) para que se pueda establecer la equivalencia. Las reglas a aplicar pueden variar desde una pluralización (en casos como la equivalencia entre el plural de *mueble* (acepción 1) y la 1ª acepción de *furniture*) a casos más complejos en que una palabra ha de ser traducida por medio de una frase (compuesta por más de una palabra). De esta última forma se definen los **phrasal-verb-tlinks** y los **phrasal-tlinks**.

El **phrasal-verb-tlink** es un tipo de tlinks bastante usual en el caso del español-inglés, necesario cuando un solo verbo español equivale a un *phrasal verb* inglés, como en el siguiente ejemplo:

aupar_X_1_1 / lift_L_1_1 + up_1 :
phrasal-verb-tlink.

(en este caso la estructura de rasgos destino resulta de la composición , mediante una regla léxica, de las dos entradas léxicas verbo y partícula).

El caso general de traducción de una sola palabra por una frase es el de los **phrasal-tlinks**. Normalmente las traducciones se compondrán de un término genérico y algun(os) modificador(es). Por ejemplo, en el caso de *amontillado* (acep.1), que resulta ser una clase de *sherry* que es pálido y seco, la conexión definida sería:

amontillado_X_1_1 / sherry_L_0_0 :
phrasal-tlink.

Se observa pues una casuística que, aún siendo compleja, no es en absoluto aleatoria, sino que presenta una serie de pautas bien claras, lo cual sugiere, a la hora de diseñar un sistema automático que lleve a cabo la proyección entre lexicones monolingües, la idea de un entorno interactivo que permita extraer las singularidades, pero que a su vez sea capaz de, automáticamente, detectar las estructuras sistemáticas existentes. Este

entorno no precisará de otro aparato formal que el ya usado para el lenguaje de representación de la **LKB** monolingüe, y en él se aprovechará el mecanismo de tipos para definir generalizaciones interesantes sobre las clases de equivalencias en traducciones. En la próxima sección se describe este Entorno semiautomático para la Creación de Tlinks (**TGE**), integrado dentro del entorno **SEISD** [Ageno et al., 92a] con el cual pretendemos ser capaces de generar una **LKB** multilingüe substancial.

3. Entorno de generación de enlaces de traducción, TGE

Dadas por un lado, la variedad de fuentes de conocimiento (Bases de Datos Léxicas bilingües, Bases de Conocimiento Léxicas monolingües) con las que debe trabajar este entorno, y por otro, la multiplicidad de casos que pueden presentarse para la formación de tlinks entre distintos pares de lenguas, hemos decidido implementar el entorno **TGE** usando una aproximación de reglas de producción. Esta aproximación ya ha sido usada anteriormente en el entorno **SEISD** [Ageno et al. 92a] para implementar el proceso de conversión de las entradas léxicas desde la **LDB** a la **LKB** [Ageno et al. 92b]. Nuestra elección ha sido motivada principalmente por la necesidad de proveer al sistema **SEISD** de una forma abierta y flexible de definir mecanismos de formación de tlinks.

El núcleo de **TGE** lo constituye el lenguaje **PRE** (Production Rules Environment). **PRE** ha sido diseñado como un lenguaje de descripción de reglas de producción de uso general, de forma que incluye las capacidades ofrecidas por **SEISD** y las necesarias para **TGE**. El lenguaje **PRE** se descrive en la próxima sección. La sintaxis de **PRE** proporciona una forma amigable, versátil y potente de expresar las reglas de formación de tlinks por los distintos usuarios del sistema. Las reglas en **PRE** están estructuradas en paquetes de reglas o rulesets. Cada ruleset agrupa las reglas que realizan una determinada funcionalidad. Los rulesets estan a su vez estructurados en forma taxonómica.

3.1 El lenguaje de reglas de producción PRE

PRE es un lenguaje de reglas de producción de propósito general. **TGE** incorpora un intérprete de dicho lenguaje. **PRE** sigue la filosofía de muchos Sistemas de Reglas de Producción (como **OPS5** [Brownston et al. 86]): hay un conjunto de objetos con sus características en un dispositivo de almacenamiento de datos o memoria de trabajo (**WM**, working memory) y un conjunto de reglas que se activan si se cumplen ciertas condiciones de la **WM**. Las capacidades de estos lenguajes incluyen consultar, crear y modificar objetos en la **WM**. Los sistemas concretos difieren en la potencia (dependiendo de la expresividad) de sus reglas, la posibilidad de definir mecanismos de control de alto nivel, como conjuntos de reglas o metareglas, y la posibilidad de escoger entre distintas estrategias de control, ya sean definidas por el usuario o proporcionadas por el sistema. **PRE** incorpora buena parte de estas características. Otras características importantes de **PRE** son:

- Tratamiento modular de las reglas.
- Mecanismos de control explícitos a nivel de regla y de paquete de reglas.
- Carecer de conjuntos de conflictos.
- Un *pattern matching* versátil y potente.
- Integración en un entorno de programación **Lisp**.

La **WM** en **PRE** es un conjunto de elementos de la forma:

(objecto ^atributo 1 valor 1 ... ^atributo n valor n)

Donde *objeto* denota el tipo de objeto dentro de la **WM**, seguido por una lista de pares atributo-valor. Los objetos de un mismo tipo se ordenan en la **WM** en orden inverso a su antigüedad.

En **PRE**, las reglas están agrupadas en conjuntos de reglas (rulesets). El paquete de reglas más alto en la jerarquía es el ruleset *top*. Este ruleset debe estar presente en toda aplicación **PRE**, mientras que los demás son opcionales. La estructura más simple sería aquella con un único paquete de

reglas, el ruleset *top*, en el que estuvieran todas las reglas. El resto de los rulesets están enlazados en la jerarquía mediante la relación clase-subclase (isa).

Un ruleset es un conjunto de reglas identificable con unas características específicas de control. Los rulesets proporcionan una forma versátil de controlar la activación de las reglas de producción. Las funciones principales para la aplicación de rulesets son , en **PRE**:

(apply-ruleset-top)
(apply-ruleset <nombre de ruleset>)

Estas funciones provocan la activación de las reglas incluidas en los paquetes de reglas *top* o *<nombre de ruleset>*, siguiendo el mecanismo de control definido en el propio ruleset. El proceso de aplicación de un paquete de reglas viene regido por los valores de los descriptores definidos en el ruleset:

- *Ruleset*: identificador del ruleset.
- *Control*: permite definir el modo de activación de las reglas del ruleset. Valores posibles son: *one-cycle* y *forever*.
- *Sort-proc*: indica al ruleset la función de ordenación que debe aplicar al paquete de reglas.
- *Sort-type*: permite definir el momento en que se evalúa la función de ordenación.
- *Final-cond*: define, si procede, la condición de finalización del ruleset.

Veamos, por ejemplo, cómo se definen los rulesets *top* y *simple-tlink* como subclase de *top*.

```
(ruleset top
  control one-cycle
  sort-proc standard-sort-proc
  sort-type static
  final-cond nil)

(ruleset simple-tlink
  isa top)
```

Así, al definir el control del ruleset *top* como "one-cycle" el intérprete **PRE** ejecuta las reglas del ruleset *top* una sola vez. Declarando *standard-sort-proc* como función de ordenación de las reglas del ruleset *top* el intérprete las ordena según la prioridad con la que están declaradas. La función de ordenación es definida por el usuario. Indicando que el tipo de ordenación es estático el intérprete ejecuta el proceso de ordenación sólo durante la carga de los rulesets. Además, el comportamiento del ruleset *simple-tlink*, al heredar sus descriptores, es idéntico al de *top*.

Las reglas en **PRE** son del tipo:

```
(<lista de descriptores>
 <lista de condiciones sobre la WM>
 :
 <lista de condiciones genéricas>
 -->
 <lista de acciones>)
```

Así, en una regla en **PRE** como:

```
(rule rule-2-simple-tlink
  ruleset  simple-tlink
  control  one
  priority 2
  (tlink-type ^type ?type)
  (translation-bil  ^trans-record  ?translation)
  :
  (not (null ?type))
  ->
  (delete 1)
  (?translation-psorts :=
    (filter-psorts  ?translation
          (get-lex-entry (translation-record-target-orth  ?translation))))
  (create translation
    ^trans-psorts ?translation-psorts
    ^trans-record ?translation
    ^tlink-type ?type
    ^checked nil))
```

la lista de descriptores indica que la regla *rule-2-simple-tlink* pertenece al ruleset *simple-tlink*. El descriptor control, permite definir el modo de activación de la regla. Valores posibles de este descriptor son: *one, forever* y

top. El descriptor *priority* puede ser consultado por el procedimiento de ordenación del ruleset. En realidad, la función de ordenación del ruleset puede consultar toda la estructura de las reglas.

La lista de condiciones sobre la **WM** son descripciones parciales de los objetos de ésta. En nuestro ejemplo, compueba en primer lugar, la existencia de un objeto de tipo *tlink-type* del cual, si lo encuentra, se capturará el valor del atributo ^*type* en la variable *?type*. En segundo lugar, comprueba la existencia de un objeto de tipo *translation-bil* del cual, si lo encuentra, se capturará el valor del atributo ^*trans-record* en la variable *?translation*. La lista de condiciones genéricas sólo comprueba que el valor de la variable *?type* no sea nulo.

Si se cumplen las anteriores condiciones, se ejecutará la lista de acciones de la regla, que incluyen: la eliminación de la **WM** del objeto *tlink-type* instanciado, la asignación a la variable *?translation-psorts* del resultado de la aplicación de la evaluación de la función *filter-psort* y la creación de un nuevo objeto de tipo *translation* con los valores de sus atributos instanciados en la ejecución de la regla.

Como puede verse, las condiciones son consultas a objetos de la **WM** o a sus características, mientras que las acciones permiten la actualización de los mismos (creación, eliminación o modificación). Otras posibles acciones permiten aplicar otros rulesets y la asignación a variables de valores calculados por funciones **Lisp**.

Una descripción completa y detallada de la sintaxis y funcionalidad de **PRE** puede encontrarse en [Ageno et al. 93].

3.2 Funcionamiento general del Sistema

El entorno **TGE** facilita el conjunto de tareas que deben realizarse para la creación de tlinks. **TGE** engloba **PRE** [Ageno et al. 93] y utiliza los módulos **LDB** [Carroll 90] y **LKB** [Copestake, 91a] incluídos en **SEISD**. La **LDB** facilita el acceso y extracción de información de los diccionarios presentes en el

sistema. La **LKB** permite la representación, acceso y manipulación de las unidades léxicas activas en él. **PRE** facilita la creación de los tlinks entre las entradas léxicas monolíngües representadas en la **LKB**.

El propósito del sistema **TGE** es producir tlinks para un conjunto de entradas léxicas de la lengua origen, conectándolas con las correspondientes entradas léxicas de la lengua destino, mediante el acceso a diversas fuentes de información y siguiendo distintas estrategias dependiendo de la información disponible para cada entrada léxica. Las fuentes de conocimiento utilizadas en nuestra implementación son las siguientes:

1) Un léxico multilingüe representado como un conjunto de estructuras de rasgos en el formato de la **LKB**, en el que deben estar incluídas las entradas léxicas de los lexicones origen y destino, además de las relaciones taxonómicas existentes en el lexicón origen.

2) Uno o más diccionarios bilingües en formato **LDB**, y por lo tanto accesibles automáticamente por el propio entorno.

3) Y finalmente, las interacciones con el lexicógrafo que es el que en última instancia decide cuales de los tlinks propuestos por el sistema son válidos y deben ser generados.

La operativa de **TGE** es básicamente la siguiente:

a) Seleccionar un conjunto de entradas léxicas de la **LKB** del lenguaje fuente (por ejemplo, español). Se admiten varias posibilidades de selección, por ejemplo, se pueden seleccionar todas las entradas léxicas hipónimas de una entrada léxica específica.

b) Seleccionar un grupo de entradas léxicas de la **LKB** del lenguaje destino (por ejemplo, inglés).

Estos dos primeros pasos, nos permiten construir la primera fuente de conocimiento.

c) Seleccionar un diccionario bilingüe mediante la **LDB** que vincule los lenguajes fuente y destino (en nuestro caso, español-inglés). Esta constituye la segunda fuente de conocimiento.

d) Aplicar los distintos mecanismos de enlace entre entradas léxicas mediante las reglas definidas en **PRE**. Actualmente, para el caso español-inglés disponemos de seis tipos de tratamiento (o paquetes de reglas) para la creación de tlinks, comentados con detalle en la sección 4. La operativa seguida en **TGE** para producir y validar tlinks es altamente interactiva, pudiendo definir el léxicografo distintas opciones de funcionamiento. Es aquí dónde se aplica la tercera fuente de conocimiento.

Las estrategias de control disponiles son cuatro:

- **all**: se aplican todos los rulesets acumulando los tlinks generados y presentándolos al usuario para una selección final.
- **collect:** se aplican todos los rulesets uno por uno. Si un ruleset tiene éxito, los tlinks propuestos son mostrados al usuario, que puede seleccionar los que considere correctos. Al final todos los tlinks acumulados son mostrados al usuario para una selección final.
- **one-by-one**: se van aplicando los rulesets uno por uno hasta que un paquete de reglas tenga éxito y algunos de los tlinks propuestos por él sean seleccionados por el usuario, en cuyo caso se finaliza el proceso para la entrada léxica origen.
- **select**: el usuario aplica selectivamente los rulesets disponibles, acumulando los resultados hasta una selección final.

e) El Sistema facilita al usuario la elección definitiva de los tlinks correctos mediante una utilidad de comparación entre estructuras de entradas léxicas llamada **Lucifer** [Copestake et al. 92]. El Sistema ordena las entradas léxicas destino encontradas como posibles candidatas a formar tlinks según su "parecido" con la entrada léxica fuente. A medida que se realizan sesiones de formación de tlinks, el sistema aprende a ponderar cuales son los rasgos más importantes para la comparación.

3.3 El diccionario bilingüe

El diccionario bilingüe es la fuente de conocimiento básica para la creación de tlinks. Dada una entrada léxica en el idioma fuente correspondiente a un concepto o sentido, mediante su ortografía podemos averiguar todas sus posibles traducciones en el idioma destino. El diccionario empleado es el Vox Harrap's esencial español-inglés [Vox 92] con unas dieciséis mil entradas por cada idioma. La descripción del proceso de carga del **MRD** en el entorno **LDB** y de las caracerísticas del diccionario están ampliamente descritas en [Hastings et al. 93]. Por ejemplo, la entrada de la palabra coñac en formato **LDB** tiene el siguiente aspecto:

((coñac)
(NS 1 > 1)
(CG m.)
(TR cognac, brandy)
(PL coñacs))

4 . Generando tlinks con TGE

Para realizar la generación de tlinks, **TGE** usa distintas estrategias. Cada una de ellas hace uso de distintas fuentes de conocimiento. Hasta ahora se han desarrollado seis módulos diferentes que cubren una parte significativa de los casos detectados. En todos ellos la fuente básica de conocimiento es la **LDB** bilingüe. Cada uno de los módulos genera uno de los cuatro tipos de tlink (*simple-tlink, compound-tlink, phrasal-verb-tlink,* y *general-phrasal-tlink* , introducidos en la sección 2), dependiendo de la correspondencia conceptual entre las entradas léxicas enlazadas. Cada módulo ha sido implementado en un paquete de reglas distinto. A continuación se introducen cada uno de los seis módulos y se detallan sus condiciones de éxito:

1- **Simple**: se aplica cuando hay una traducción directa en el diccionario bilingüe, que además existe como entrada léxica en el lexicón destino. El tipo de tlink producido es *simple-tlink*. Así, por ejemplo, al generar el tlink

correspondiente a la entrada léxica *absenta_X_1_1*, este módulo busca la traducción de *absenta* en el bilingüe donde aparece traducido como *absinth*. En el lexicón destino hay una entrada léxica con esta ortografía, *absinth_L_0_1*, y por lo tanto se produce un *simple-tlink* entre *absenta_X_1_1* y *absinth_L_0_1*.

2- **Compound**: en este caso la traducción en el bilingüe está compuesta de dos palabras, que existen en el lexicón destino como una sola entrada léxica, obtenida mediante la concatenación de las dos palabras con un guión. Se produce un *simple-tlink* entre la entrada léxica origen y la de destino. Así, por ejemplo, la traducción de la ortografía correspondiente a la acepción española *pelel_X_1_1* (*pelel*) es *pale ale*. Estas dos palabras no aparecen en el lexicón destino, pero su concatenación *pale_ale* si, correspondiéndole la entrada léxica *pale_ale_L_0_0*. Por lo tanto se genera un *simple-tlink* enlazando ambas extradas léxicas.

3- **Phrasal Verb**: se aplica en caso de que la traducción sea un verbo frasal (compuesta de dos palabras una de las cuales es una partícula), y que cada uno de sus componentes exista en el lexicón destino como entrada léxica. El tipo de tlink generado es *phrasal-verb-tlink*, en el cual la estructura de rasgos destino es la composición (mediante una regla léxica) de las dos entradas léxicas (verbo y partícula). Así por ejemplo la traducción correspondiente a *aupar_X_1_1* es *lift up*. La primera palabra (*lift*) es un verbo que aparece en el lexicón inglés como *lift_L_1_1*, y la segunda aparece como partícula (*up_1*). La aplicación de este módulo generaría un *phrasal-verb-tlink* conectando *aupar_X_1_1* con el resultado de componer *lift_L_1_1 +up_1*.

4- **General Phrasal**: se aplica cuando la traducción es una frase corta en forma de un genus y algunos modificadores. Para la identificación del genus se utiliza un algoritmo simple que hace uso de heurísticos. Se produce un *phrasal-tlink* entre la entrada léxica de origen y el genus. Por ejemplo en el caso de *amontillado_X_1_1*, su traducción es *pale dry sherry*. El genus de esta traducción es identificado como *sherry*, cuya correspondiente entrada léxica en el lexicón inglés es *sherry_L_0_0*. Por lo tanto se genera un

phrasal-tlink entre *amontillado_X_1_1* y *sherry_L_0_0*.

5- **Parent**: se aplica en caso de que la entrada léxica de origen sea hipónima de otras entradas del lexicón origen (padres), y que éstas tengan una traducción simple en el diccionario bilingüe. Se produce un *partial-tlink* entre la entrada léxica de origen y la correspondiente a la traducción del padre en el lexicón destino. Así, por ejemplo, al generar el tlink correspondiente a la entrada léxica *agasajo_X_1_3*, este módulo busca un hiperónimo suyo en la taxonomía, encontrando la acepción *refresco_X_1_2*, la ortografía de la cual (*refresco*) es traducida en el bilingüe como *drink*. En el lexicón destino hay una entrada léxica con esta ortografía, *drink_L_2_1*, y por lo tanto se produce un *partial-tlink* entre *agasajo_X_1_3* y *drink_L_2_1*.

6- **Grandparent**: es muy similar al anterior, pero en este caso son los hiperónimos de los hiperónimos de la entrada léxica origen los utilizados para producir el tlink. Es decir, que el antecesor es buscado en el diccionario bilingüe, y si su traducción corresponde a alguna entrada léxica destino, se propone un *partial tlink* entre la entrada léxica original y la correspondiente a la traducción del antecesor. Así, por ejemplo, al generar el tlink correspondiente a la entrada léxica *caridad_X_1_4*, este módulo busca un hiperónimo suyo en la taxonomía, encontrando la acepción *agasajo_X_1_3*, la cual a su vez tiene como hiperónimo a *refresco_X_1_2*, la ortografía de la cual (*refresco*) es traducida en el **LDB** bilingüe como *drink*. En el lexicón destino hay una entrada léxica con esta ortografía, *drink_L_2_1*, y por lo tanto se produce un *partial-tlink* entre *caridad_X_1_4* y *drink_L_2_1*.

Para que el lector se haga una idea más precisa del funcionamiento de los paquetes de reglas en la generación de tlinks, a continuación, nos proponemos repasar la aplicación de **TGE** en estrategia *all*, para generar los tlinks entre español e inglés correspondientes a la entrada *coñac_X_1_1*. Como ya hemos comentado en la sección 3.2, la estrategia *all* aplica todos los módulos para generar tlinks, efectuandose posteriormente el proceso de selección por parte del lexicógrafo, de los tlinks que considera correctos. Así pues analizaremos las condiciones de éxito/fracaso módulo por módulo

para el caso de *coñac_X_1_1*, y posteriormente explicaremos con la ayuda de la pantalla 1 qué tipo de interacciones están permitidas en el proceso de selección posterior.

El primer módulo aplicado, **simple**, tiene éxito gracias a la segunda regla, *rule-2-simple-tlink* (ya mostrada en la sección 3.1.), y por lo tanto, genera un *simple-tlink*. Revisemos con más detalle el funcionamiento de dicha regla. *Rule-2-simple-tlink* se aplicará siempre que se cumplan tanto las condiciones que impone sobre la WM como las condiciones genéricas. La condiciones sobre la WM se cumplirán siempre que existan dos objetos de tipo *tlink-type* y *translation-bil*. El primero habrá sido creado por la primera regla del paquete de reglas (**simple**) con el atributo ^*type* con valor *simple-tlink*. El segundo objeto habrá sido creado por el módulo de control de la estrategia *all* el cual habrá guardado en el atributo ^*translation-bil* la traducción obtenida de la consulta al bilingüe de la entrada *coñac*, que en este caso consistirá en un registro conteniendo entre otras informaciones la ortografía de la traducción, *cognac*.

Debido a que las dos condiciones sobre la WM tienen éxito, se procede a comprobar la condición genérica, que se cumple gracias a que la variable *?type* ha tomado por valor *simple-tlink* en la primera condición sobre la WM, y por lo tanto, su valor no es nulo. Gracias a que los dos tipos de condiciones tienen éxito se puede aplicar la regla, lo que implica ejecutar las acciones que aparecen en su parte derecha.

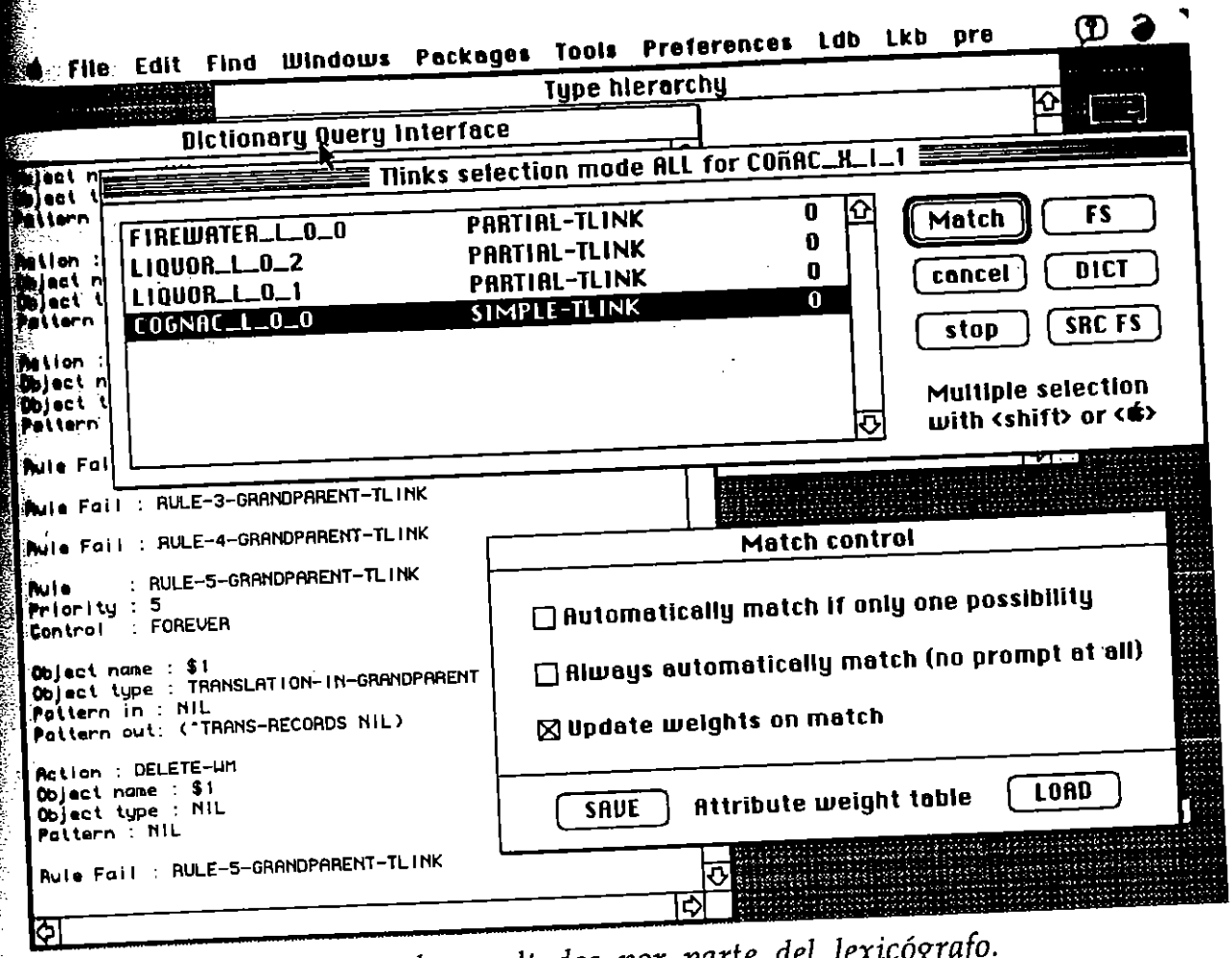
La primera acción borra el objeto identificado por la primera condición sobre la WM. La segunda, obtiene en primer lugar, mediante la función *translation-record-target-orth*, la ortografía de la traducción, *cognac*, que se encuentra en el registro asignado a la variable *?translation* (capturado por la segunda condición sobre la WM). Posteriormente se obtienen las entradas léxicas (mediante *get-lex-entry*) correspondientes a *cognac*, que en este caso será tan solo una, *cognac_L_0_0*. Estas entradas léxicas son filtradas por la función *filter-psorts*, y el resultado final en este caso (*cognac_L_0_0*) es asignado a la variable *?translation-psorts*, la cual guardará, por lo tanto, las entradas léxicas de destino con las que se establece un *simple-tlink*.

La segunda acción crea un objeto de tipo *translation* con la información obtenida en las condiciones y acciones anteriores. Este objeto sirve para guardar información importante para el proceso de selección posterior: los tlinks generados (en el atributo^*trans-psorts*), información del bilingüe (en^*trans-record*), el tipo de tlink que ha generado estos resultados (en^*tlink-type*), y finalmente, la inicialización de información de contabilidad para procesos posteriores (en^*checked*).

Los módulos **compound**, **phrasal verb**, y **general phrasal** fracasan al generar ttlinks para *coñac_X_1_1* porque necesitan que la traducción presente en el bilingüe esté compuesta de más de una palabra, y la correspondiente a *coñac, cognac*, consiste en una sola palabra.

El módulo **parent** también tiene éxito puesto que el hiperónimo de *coñac_X_1_1, aguardiente_X_1_1*, posee un par de traducciones *liquor*, y *firewater*, de las cuáles existe entradas léxicas en el lexicón destino: *liquor_L_0_2* y *liquor_L_0_1* para la primera, y *firewater_L_0_0* para la segunda. La estrategia implementada en este paquete de reglas consiste en: (1) buscar la traducción del hiperónimo en el bilingüe, (2) aplicar el paquete de reglas **simple** para que intente resolver el enlace entre el hiperónimo y sus traducciones, y (3) si éste tiene éxito, recoger los resultados, cambiando los datos necesarios para que el tlink generado sea establecido entre la entrada léxica original y la traducción del hiperónimo, y para que el tlink sea del tipo *partial-tlink*, en vez de *simple-tlink*. Así en este caso se establecerían tres *partial-tlink* entre *coñac_X_1_1* por un lado, y *liquor_L_0_2*, *liquor_L_0_1*, y *firewater_L_0_0* por el otro.

Finalmente, el módulo **grandparent** fracasa porque el hiperónimo de *coñac_X_1_1, aguardiente_X_1_1*, es el elemento cumbre en la taxonomía del lexicón del español sobre el que se desarrolló este ejemplo, y por lo tanto, no hay ningún hiperónimo de *aguardiente_X_1_1* en el lexicón.



Pantalla 1 Selección de resultados por parte del lexicógrafo.

Una vez aplicados los distintos paquetes de reglas sobre la entrada léxica *coñac_X_1_1*, se puede producir un proceso de selección en el que el lexicógrafo, accediendo a fuentes de información auxiliares, decide qué tlinks, de los varios generados, considera correctos. En la pantalla 1 se puede observar un momento en la selección de los tlinks producidos para la entrada léxica *coñac_X_1_1*. En ella hay una ventanas principal, la propia de selección (*Tlinks selection mode ALL..*), y una serie de ventanas auxiliares.

En la ventana principal aparece un menú de selección múltiple en el que TGE muestra los posibles tlinks que se pueden construir a partir de *coñac_X_1_1* aplicando la estrategia *all* (generación de todas las posibilidades). En este caso aparecen tres posibles *partial-tlink* y un *simple-tlink*. Los tres primeros se obtienen a partir del hiperónimo de *coñac_X_1_1*, *aguardiente_X_1_1*. Mientras, que el *simple-tlink* se obtiene mediante la traducción de *coñac* en el bilingüe, *cognac*. En el instante en

que está tomada la imagen de la pantalla, el lexicógrafo solamente había seleccionado en el menú (destacado en imagen inversa) el tlink que enlaza con *cognac_L_0_0*. En esta misma ventana aparecen unos botones a la derecha que sirven para: *Match* finalizar el proceso de selección y por lo tanto dar por buenos los tlinks ya selecionados, *FS* pedirle a **TGE** una ventana en la que muestre la estructura de rasgos correspondiente a las entradas léxicas destino de los tlinks seleccionados, *cancel* no aplicar más paquetes de reglas para la entrada léxica en consideración, pero generar los tlinks que hasta ahora han sido seleccionados (útil en los modos de operación *collect* y *one-by-one*), *DICT* obtener la entrada en el bilingüe correspondiente a la entrada léxica de origen, *stop* abandonar totalmente la generación de tlinks para la entrada léxica en consideración sin guardar ninguno de los tlinks ya seleccionados, y finalmente, *SRC FS* pedirle a **TGE** una ventana en la que muestre la estructura de rasgos correspondiente a la entrada léxica origen.

En las otras ventanas aparece otra información auxiliar que puede ser útil al lexicógrafo: mensajes de aviso o error, la traza de la ejecución de los paquetes de reglas que aplicados hasta el momento, la jerarquía de tipos de la **LKB**, etc.

5. Conclusiones

Hemos presentado un entorno de generación semiautomática de enlaces léxicos multilingües (tlinks) desarrollado en el marco del proyecto Acquilex II. El sistema sirve de soporte a una metodología de extracción de correspondencias entre entradas léxicas de diferentes lenguas. El núcleo de esta metodología es la identificación de pautas sistemáticas en las correspondencias léxicas. La fuente principal para esta identificación es la utilización de Bases de Datos Léxicas Bilingües construídas a partir de diccionarios bilingües. El entorno presentado permite expresar cada una de estas pautas en forma de paquetes de reglas de producción. Cada uno de estos paquetes agrupa un conjunto de reglas capaces de detectar el caso, extraer la información pertinente y generar el tipo correspondiente de tlink.

Una versión anterior del Sistema fue experimentada, en el marco de Acquilex, sobre dominios limitados (alimento, bebida) dando resultados esperanzadores (ver [Acquilex,92]).

Actualmente, ya en el marco de Acquilex II, se está utilizando el sistema para un estudio más exhaustivo de la tipología de los desajustes que se producen entre las entradas léxicas nominales del español y el inglés. Unos primeros resultados se presentan en [Soler,93].

Está previsto un estudio similar para las entradas verbales. El entorno está siendo asimismo utilizado por otros grupos dentro de Acquilex II para realizar tareas similares para otros pares de lenguas.

Bibliografía

[Acquilex, 92] Amsterdam,Barelona,Cambridge,Dublin,Pisa *Final Evaluation of LDB/LKB System* Esprit BRA-3030 Acquilex deliverable nº 11. Barcelona, june 1992.

[[Ageno et al. 92a] Ageno A., Castellón I., Martí M.A., Ribas F., Rigau G., Rodríguez H., Taulé M., Verdejo F. *SEISD: An environment for extraction of Semantic Information from on-line dictionaries.* Proceedings of 3th Conference on Applied Natural Language Processing. Trento. Italia. Abril 1992.

[Ageno et al. 92b] Ageno A., Castellón I., Martí M.A., Ribas F., Rigau G., Rodríguez H., Taulé M., Verdejo F. (1992). *A semiautomatic process to create LKB entries.*
Esprit BRA-3030 Acquilex Working Paper nº38.

[Ageno et al., 93] Ageno A., Ribas F., Rigau G., Rodríguez H., Verdejo F. (1993). *TGE: Tlinks Generation Environment* . Esprit BRA-7315 Acquilex II Working Paper.

[Alshawi, 92] Alshawi, H. (ed) (1992). *The Core Language Engine*. MIT Press. Cambridge & London.

[Briscoe, 91] Briscoe,T., Copestake, A. y de Paiva, V.,1991, *Functionality of*

LKB en Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon, University of Cambridge Computer Laboratory, Technical Report n° 238. Esprit BRA-3030 Acquilex working paper n°040.

[Brownstom et al. 86] Brownston L., Farrell R., Kant, E., Martin N. *Programming Expert Systems in OPS5.* Addison-Wesley. 1986.

[Carroll 90] Carroll J. *Lexical Data Base System. User Manual.* Computer Laboratory, University of Cambridge. Octubre 1990.

[Copestake, 91a] Copestake A (1991). *The LKB: a system for representing lexical information extracted from machine-readable dictionaries, Proceedings of the ACQUILEX Workshop on Default Inheritance in the Lexicon,* Cambridge.

[Copestake, 91b] Copestake A (1991). *LKB implementation outline- version 3* , Cambridge Univ, Computer Laboratoty, ms

[Copestake et al. 92] Copestake, A., Jones B., Sanfilippo A., Rodriguez H., Vossen P. (1992). *Multilingual Lexical Representation*. Esprit BRA-3030 Acquilex Working Paper n°38.

[Hastings et al. 93] Hastings A., Rigau G., Soler C., Tuells A. *Loading a bilingual dictionary into the LDB. Characteristics of Vox Harrap's dictionary.*
Esprit BRA-7315 Acquilex II Working Paper.

[Ldoce 78] Procter, P. et al. (eds) (1987). *Longman Dictionary of Contemporary English.* Longman, Harlow and London.

[Soler, 93] Soler, C. (1993). *Dealing with Spanish-English / English-Spanish mismatches.* Esprit BRA-7315 Acquilex II Working Paper.

[Vox 87] *Diccionario General Ilustrado de la Lengua Española VOX.* Ed. Biblograf S.A. Barcelona, 1987.

[Vox 92] *Vox Harrap's Diccionario esencial Inglés-Español, Español-Inglés.* Segunda Edición. Ed. Biblograf S.A. Barcelona, Febrero 1992.

[Whitelock,92] Whitelock, P. (1992). *Shake-and-Bake Translation* en Proceedings of International Conference on Computational Linguistics, Coling 92. Nantes. págs. 784-790.