Guía de anotación para vincular Ancora-Verb y PropBank

Working paper 2: TEXT-MESS 2.0 (Text-Knowledge 2.0)

Patricia Fernández, Oriol Borrega, Mariona Taulé, M.A. Martí

2010

CLiC
Centre de Llenguatge i Computació

UNIVERSITAT DE BARCELONA

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
TextMESS2.0
Plan Nacional de I+D+I 2008 - 2011

FFI2009-06497-E/FILO TIN2006-15265-C06-06

Índice

Guía de anotación relación AnCora-Verb – PropBank

Descripción

La tarea consiste en vincular cada sentido de cada entrada del léxico verbal AnCora-Verb (CA/ES) con uno o varios de los sentidos especificados en las entradas léxicas de PropBank (http://verbs.colorado.edu/verb-index/), es decir, que el tipo de relación que estableceremos será de 1 a *n*.

Para ello, se ha realizado un *mapping* o relación automática inicial a través de los sentidos de WordNet 3.0. De este proceso se han generado tres tipos de fichero:

1. Ficheros con los verbos cuya relación con PropBank se ha establecido a nivel de palabra (sin descender hasta el sentido): **proposals_wd**
2. Ficheros con los verbos cuya relación con PropBank se ha establecido a nivel de sentido: **proposals_wn**
3. Ficheros con los verbos con los que no se ha podido establecer ninguna relación con PropBank: **empties**

Todos ellos numerados a partir de 0.

Los verbos de los ficheros _wd estarán enlazados a todos los sentidos de un verbo en inglés, mientras que los verbos de los ficheros _wn tendrán como equivalencia únicamente un sentido concreto de los verbos de PropBank. Es decir, que la equivalencia establecida en los ficheros_wd es mucho más generalista.

Ejemplo: Fichero **proposals_wd.1**:
verb.acortar.2.default - …
verb.acortar.2.default - cut.01
verb.acortar.2.default - cut.02 → Equivalencia correcta
verb.acortar.2.default - cut.03
verb.acortar.2.default - cut.04
verb.acortar.2.default - cut.05
verb.acortar.2.default - cut.06
verb.acortar.2.default - cut.07
verb.acortar.2.default - …

En un fichero "**proposals_wn**" el enlace se habría realizado únicamente con el sentido 02 del verbo *cut* (cut.02).

Proceso

Para llevar a cabo el proceso de anotación se utilizará la herramienta AnCoraPipe, para la utilización de la cual se debe abrir la aplicación Eclipse y tener cargado el proyecto Ancora-Net con la carpeta Proposals. Esta carpeta contiene los **ficheros** para el español y los ficheros para el catalán. En cada uno de estos ficheros se ubican los verbos separados por sentidos y sus diátesis, ordenados alfabéticamente y con una o varias opciones de relación con PropBank para cada sentido y su diátesis.

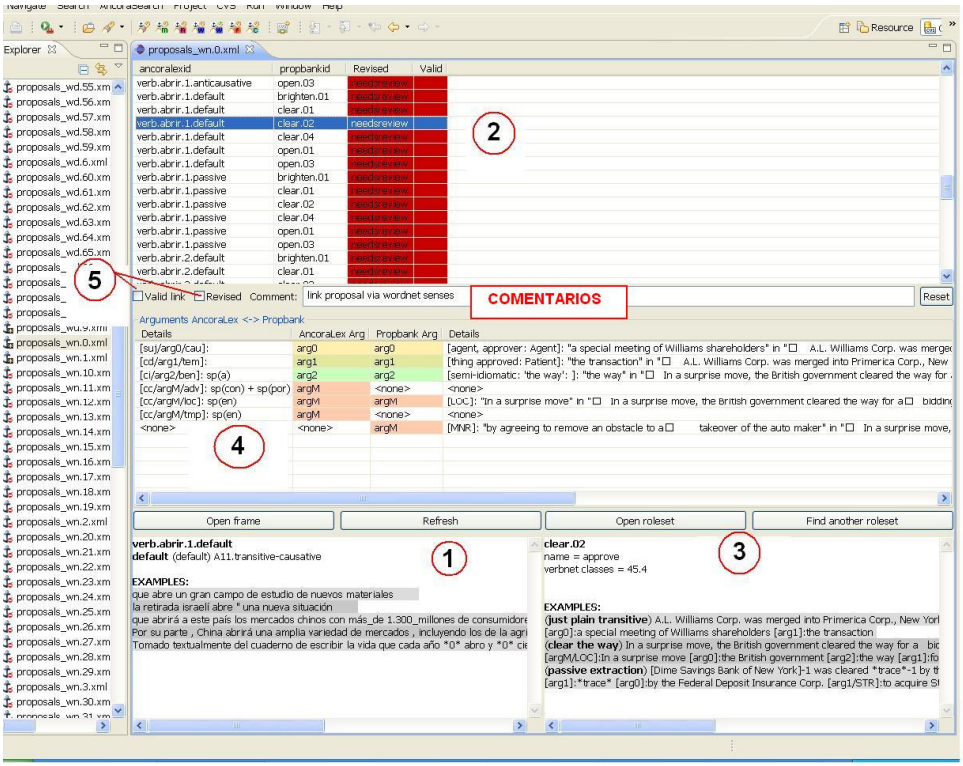
Ejemplo del verbo *acabar* (**proposals.wd0.xml**):

- verb.acabar.1.default
- verb.acabar.2.benefactive
- verb.acabar.2.default
- verb.acabar.3.default
- verb.acabar.4.default
- verb.acabar.4.passive
- verb.acabar.5.default
- verb.acabar.6.default

El proceso que se debe seguir para realizar la anotación es abrir cada fichero con la herramienta "Ancora Resource Link Editor" (botón derecho sobre el nombre del fichero → Open with → Ancora Resource Link Editor). Una vez abierto, se debe hacer clic en cada verbo y seguir los siguientes pasos:

1- Leer el ejemplo del verbo en castellano para detectar el sentido del verbo (puede tener más pero que no estén recogidos en Ancora-Verb por no haber aparecido en el corpus de referencia).
   a. Si fuera necesario, consultar dicho sentido en la RAE para entenderlo bien (sólo en caso de duda).
   b. Consultar en recursos *on-line* las posibilidades de traducción de ese verbo (al final de la guía se incluye una relación de los recursos *on-line* de referencia para el proyecto).
2- Con la propuesta de los diccionarios/corpus de referencia, analizar cada verbo considerado en la equivalencia automática con PropBank.
3- Para analizar cada verbo:
   a. Leer la glosa que lleva asociada y sus ejemplos (en inglés).
   b. Considerar también los argumentos verbales.
4- Cuando se considere que un verbo de PropBank es una equivalencia correcta, se deben establecer equivalencias también entre sus argumentos verbales utilizando los listados desplegables de la ventana intermedia de la aplicación, y escogiendo el argumento de PropBank que se corresponda con el argumento de AnCora-Verb.
5- Tras todo este proceso, se deben marcar las casillas de validación que se ubican bajo el listado de verbos. Si la equivalencia es válida, se debe marcar "Valid lindk" y "Revised"; si no lo es, se debe marcar únicamente "Revised" (para confirmar que el verbo ha sido revisado).

Si fuera necesario, se puede añadir algún comentario en el cuadro de texto de "Comments". Para ver los tipos de comentarios posibles (no incluiremos observaciones aleatorias sino las que definamos en función de las necesidades de la anotación), véase la hoja "Comentarios original" del fichero excel **comentarios.xls.**



Cabe mencionar que es posible no encontrar ejemplos del verbo en castellano (ventana 1) debido a que la aparición de este verbo en el corpus sea únicamente en un sustantivo deverbal. Así, si se quiere buscar la ocurrencia real en el corpus para confirmar su sentido basta con ejecutar la opción de búsqueda y utilizar la siguiente sentencia: //n[ub:matches(@originlexicalid,".*__verbo__.*")]

Criterios de anotación

Los criterios que se deben tener en cuenta a la hora de establecer correspondencias entre verbos son:

-Las **correspondencias** podrán ser de 1 a 1 o de 1 a *n*, considerando como equivalentes tantos verbos del inglés como se crea necesario, es decir, se pueden tener en cuenta sinónimos. Esta equivalencia se establecerá a nivel de sentido. Los sentidos se diferencian por la numeración que acompaña a cada verbo (abrir.1, abrir.2... // get.01, get.02, get.03...).

Por ejemplo: *augurar* → augur, forecast, foretell, foreshadow // augurar, predecir, pronosticar, vaticinar, presagiar… Este verbo tiene muchos sinónimos tanto en castellano como en inglés, por lo que los consideraremos todos si el sentido del verbo recogido en PropBank se corresponde con el sentido del verbo en AnCora-Verb.

-A la hora de localizar todas las equivalencias semánticas (sinónimos) de un verbo nos ceñiremos a las opciones consideradas en el *mapping* automático con WordNet, a las que ofrecen los recursos *on-line* de referencia del proyecto y a nuestro conocimiento lingüístico de la lengua de destino. No se realizará una búsqueda abierta de equivalencias para no eternizar la tarea.

-Las equivalencias consideradas en nuestro repositorio como "propuestas" y que proceden del *mapping* automático con WordNet no se deben considerar válidas *a priori*. Se deben evaluar una por una todas las opciones propuestas y marcar como correctas sólo aquellas que realmente lo sean. Se ha de poner especial atención en los ficheros _wd_ , pues las equivalencias que contienen no son tan específicas como las de los ficheros _wn_. Así, nos encontraremos ejemplos como el siguiente: en WordNet se ha considerado como equivalente o sinónimo "barrer" – "sweep" y "cepillar" – "brush", y aunque tienen un mismo matiz semántico de "limpiar o eliminar suciedad y restos" no los marcaremos como equivalentes porque no significan lo mismo.

-Como ya se ha mencionado, además de las equivalencias propuestas a partir del *mapping* con WordNet, se pueden añadir otras nuevas que no han sido seleccionadas automáticamente (*vid.* sección "Nuevas equivalencias"). Hay cuantiosos casos en los que el verbo en inglés que comparte **raíz latina** con el verbo de AnCora-Verb no se ha considerado desde WordNet, por lo que conviene estar atentos a este tipo de verbos e incluirlos en el proceso. Por ejemplo: *agravar* sólo tenía como propuesta de equivalencia *worsen*, no obstante, en inglés también existe el verbo *aggravate* que comparte el mismo sentido que los dos anteriores. Este verbo se debe considerar como una equivalencia nueva para *agravar*.

-En el análisis de cada equivalencia, es conveniente considerar también la **estructura argumental de los verbos**. Así, un verbo que tanto en AnCora-Verb como en PropBank tiene una estructura sintáctico-semántica benefactiva tendrá más posibilidades de considerarse como equivalencia que un verbo que en AnCora-Verb sí es benefactivo pero que en PropBank no se considera como tal. No obstante, la diferencia en la estructura argumental de los verbos no será NUNCA un criterio excluyente para considerar como válida una equivalencia verbal. Por ejemplo: *agrupar.1* (tem/atr) se ha vinculado a *group*.01 porque su estructura en inglés es *theme/group*; y *agrupar.2* (agt/pat) se ha asociado a *cluster.01* cuya estructura argumental es también *agent/patient*.

-**AnCora-Verb** está creado con un carácter semántico "generalista", es decir, un mismo *frame* puede incluir más de un sentido cuando éste tiene, por ejemplo, su vertiente más física y a la vez metafórica. Si en inglés se puede utilizar una misma traducción para todos los casos, entonces no se separará el sentido del original. En caso de que la diferencia que existe entre estos sentidos recogidos en un mismo *frame* sea muy acusada se incluyen los distintos verbos del inglés correspondientes. En algunos casos puntuales, se ha reconsiderado los distintos sentidos propuestos en AnCora-Verb.

-En el listado de verbos de AnCora-Verb se presentan de manera diferenciada **sentidos** y **diátesis**. Por lo general, para las diátesis de un mismo sentido utilizaremos las mismas equivalencias. De hecho, PropBank no realiza una **distinción de diátesis**, por lo que en sus ejemplos nos encontraremos mezcladas alternancias pasivas, benefactivas, etc., por lo que en el proceso de anotación podremos establecer equivalencias entre verbos que se

definen por medio de ejemplos con diferentes diátesis. Por ejemplo: adornar.1.default, adornar.1.oblique_subject, adornar.1.passive = adorn.01 (simple transitive), decorate.01 (transitive, passive), etc.

-Mientras que las alternancias diatéticas no afectan a las equivalencias verbales, las **diferencias de sentido** sí lo hacen. Así, por lo general, cada sentido de un verbo de AnCora-Verb se traducirá por verbos distintos al inglés. No obstante, puede ocurrir que un mismo sentido verbal de PropBank albergue sentidos distintos de AnCora-Verb. Por ejemplo: adoptar.1 (*adoptar una resolución, adoptar sanciones…*) y adoptar.2 (*adoptar un niño*) se corresponden ambos al mismo sentido verbal de PropBank (adopt.01). En ese caso, los dos sentidos de AnCora-Verb se asocian al mismo sentido de PropBank. Puede darse también el caso contrario, es decir, que AnCora-Verb incluya sentidos distintos de PropBank. En este caso, si existe en inglés un verbo que tenga una misma traducción al castellano para todos sus sentidos entonces se vincula a éste. Si la diferencia que xiste entre los sentidos del inglés y el del español es muy acusada y su traducción sea completamente distinta, se le asocian todas las equivalencias posibles. Por ejemplo: adquirir.1 (adquirir un coche, adquirir renombre…) se vincula a los verbos adquire.01, buy.01 y asume.01 de PropBank.

-En caso de que un mismo verbo se traduzca de manera diferente en función del sustantivo que le acompaña (***collocations***), se considerarán todas las posibles traducciones como equivalencias posibles para este verbo. Ejemplo: *cancelar* se traduce al inglés de forma diferente en función del elemento cancelado (*settle a debt…*). Así, se considerarán como válidas todas las posibles traducciones del verbo *cancelar.*

-**Expresiones idiomáticas**. En caso de que nos encontremos con un verbo con un ArgL (expresión idiomática) buscaremos su equivalencia para toda la expresión, y no sólo para el verbo de forma independiente. Ejemplo: *capear el temporal* → su equivalencia debería ser la adecuada para traducir esta expresión. En caso de que detectemos una expresión idiomática no etiquetada como tal en el lexicón original, se procederá tal y como se describe en el punto 3 del apartado "Problemáticas detectadas en verbos de AnCora-Verb".

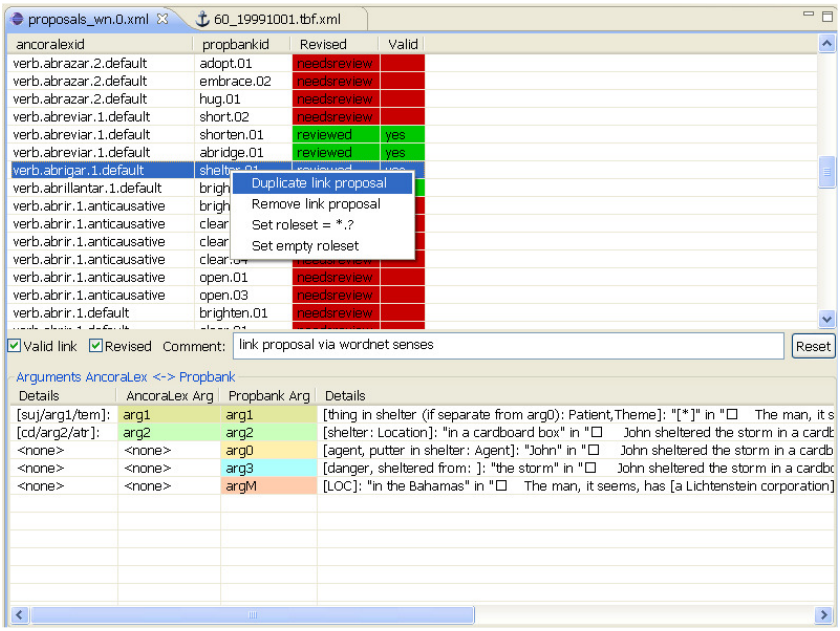
-**Términos de especialidad**. En ámbitos como el jurídico, la terminología latina y sajona difiere en gran medida. En ese caso, buscaremos el verbo más equivalente que logremos encontrar. Ejemplo: *avalar – underwrite* (lit. financiar). Si no se encuentra ninguna equivalencia clara, se deberá dejar en blanco.

-**Verba dicendi**. Cuando un verbo tenga un uso de verbo declarativo y sea equiparable a los declarativos del inglés (*say*, *state*, etc.) no consideraremos todas sus posibles equivalencias como *verbum dicendi* para no generalizar en exceso. Ejemplo: *agregar* tiene un sentido declarativo específico, pero su equivalente será sólo *add.01*, no incluiremos ni *say*, ni *state*, ni *declare*, etc.

-Por último, es posible **no encontrar equivalencias** entre los verbos de AnCora-Verb y los de PropBank. Un ejemplo de este caso sería *azuzar*, que no se corresponde exactamente con ningún verbo de los recogidos en PropBank. En ese caso se validan las opciones ofrecidas por la aplicación con la casilla *Revised* y no se marca ninguna entrada como válida.

Nuevas equivalencias

Si se ha de añadir una equivalencia nueva al verbo porque se considera necesaria pero no ha sido considerada en el *mapping* automático con WordNet, se puede crear una entrada verbal nueva haciendo clic en una entrada del verbo con el botón derecho y eligiendo la opción “Duplicate link proposal”.



Recursos on-line

Para utilizaremos los siguientes recursos *on-line*

- Wordreference: http://wordreference.com/
- Google Translate: http://translate.google.com/
- Corpus UNESCO (Cluvi): http://sli.uvigo.es/CLUVI/index.html#correo
- Merriam-Webster: http://www.merriam-webster.com/
- Diccionario de la RAE: http://www.rae.es/rae.html
- Diccionario de sinónimos: http://sinonimos.org