Estudio preliminar para la creación de Euskal PropBank

Izaskun Aldezabal Roteta
Departamento LSI, Universidad del Pais Vasco (UPV/EHU)
izaskun.aldezabal@ehu.es

Resumen

En este artículo presento el estudio preliminar que hemos realizado en el grupo Ixa[1] para comprobar la validez del modelo PropBank de etiquetado semántico basado en roles (Palmer et al., 2005), para el corpus EPEC del euskera[2]. Hemos realizado el estudio con tres verbos, mediante lo cual hemos establecido una metodología de trabajo, que ejemplificamos en el presente artículo con el verbo *esan*. De hecho, la metodología ha sido otro resultado relevante del estudio.

1. Introducción

El estudio que presento se sitúa en el marco de trabajo establecido en el Grupo Ixa en el que se pretende desarrollar recursos léxicos tales como base de datos y corpus anotados a diferentes niveles lingüísticos (Aduriz et al. 06). Se trata de añadir información semántica en términos de roles semánticos a un corpus previamente etiquetado morfosintácticamente, y crear paralelamente un léxico de verbos.

Uno de los mayores problemas a la hora de abordar esta tarea es que no encontramos un repertorio de roles semánticos unánime. Hasta ahora no ha habido ningún estudio exhaustivo interlingual que haya presentado un modelo como válido para todas las lenguas. Y por otro lado, muchos de los trabajos que optan por un modelo en concreto, inevitablemente acaban teniendo inconsistencias, al no estar todos los fenómenos lingüísticos tratados e identificados a priori.

Así, en el grupo Ixa hemos querido tomar como punto de partida un modelo que ha sido manejado en más de una lengua y ha desarrollado recursos a gran escala. Este es precisamente el caso de PropBank (Palmer et al. 05), basándose en el cual se ha trabajado en lenguas tan diferentes como el inglés y el chino, entre otros. Además, coincidiendo con nuestra línea de trabajo, PropBank se ha desarrollado sobre un corpus previamente etiquetado sintácticamente. Sin pretender encontrar un modelo acabado, hemos querido evitar por un lado emprender un camino desde cero y por otro llegar a un repertorio de roles que nos limite a la hora de hacer comparaciones entre lenguas. De hecho, siguiendo la idea de la comparación interlingual, estamos implicados en un proyecto Mec (CESS-ECE[3]) cuya finalidad es etiquetar corpora de tres lenguas

[1] http://ixa.si.ehu.es
[2] EPEC es el nombre del corpus de referencia que utilizamos para el procesamiento del euskera en el grupo Ixa
[3] CESS-ECE: Corpus etiquetados sintáctico-semánticamente del español, catalán y euskera (HUM2004-21127-E)

(euskera, catalán y castellano) mediante el modelo PropBank, teniendo partes comparables.

También se han utilizado trabajos sobre la subcategorización verbal que se han realizado previamente en el grupo Ixa (Aldezabal 04). Los resultados obtenidos en estos trabajos son parte y base de las fases propuestas en la metodología.

El esquema que seguiré será el siguiente. En la sección 2 describiré brevemente los recursos de los que partimos en el grupo. En la sección 3 entraré a exponer detalladamente los pasos de la metodología, por lo que esta sección se dividirá en 3 subsecciones: preparación de la entrada léxica (3.1), etiquetado del corpus con los roles semánticos (3.2), y realización de la tabla de comentarios de dudas (3.3). Finalmente, en la sección 4, expondré las conclusiones.

2. Recursos

En Ixa contamos con varios recursos que sirven como base del estudio preliminar.

Por un lado, tenemos el corpus EPEC etiquetado morfosintácticamente. En concreto hemos elegido una muestra de 50.000 palabras, que está manualmente etiquetada basándose en dependencias (Aranzabe et al. 04). Este etiquetado sintáctico se realizó dentro del proyecto 3LB, donde al corpus se le dio el nombre de Eus3LB (Palomar et al. 04). Es un corpus equilibrado que consta de textos en euskera batua (o estandar) perteneciente al siglo XX[4].

Por otro lado, tenemos una base de datos creada con la información obtenida de las bases de datos de pago PropBank y Verbnet. Estas bases de datos no venían integradas físicamente. Lo que hemos hecho en el grupo ha sido unir la información de ambas y crear una única base de datos.

Hemos utilizado diccionarios monolingües y bilingües euskera-castellano-inglés: Euskal Hiztegia (Sarasola 97) e Hiztegi Modernoa (Elhuyar 00) entre los monolingües; Morris Hiztegia (Morris 98) para euskera-inglés / inglés-euskera; Elhuyar hiztegia (Elhuyar 00) para euskera-castellano / castellano-euskera; y Word Referente para inglés-castellano / castellano-inglés.

Por último, nos hemos basado en los datos de la tesis de Aldezabal (04), donde se hace un estudio descriptivo-formal de 100 verbos vascos basándose en sus argumentos, roles temáticos y alternancias. Concretamente, se dan para cada verbo unos frames sintáctico-semánticos (fss), donde se codifican sentidos amplios y alternancias aceptadas.

3. Metodología

En la metodología distinguimos tres fases: primero, preparamos la entrada léxica del verbo; posteriormente etiquetamos el corpus con los roles semánticos; y por último, agrupamos las dudas y completamos una tabla de comentarios.

[4] http://www.euskaracorpusa.com

A continuación expongo las tres fases más detalladamente

3.1 Preparación de la entrada léxica

En esta fase planteamos 4 pasos:
- Elegir el verbo en euskera
- Obtener la visión general de los sentidos del verbo en euskera y buscar sus equivalentes léxicos en inglés
- Analizar los equivalentes sintácticos en inglés
- Proponer las entradas para los verbos en euskera al estilo VerbNet/PropBank

Los expongo a continuación.

3.1.1 Elegir el verbo en euskera

Para elegir el verbo, nos basamos en el corpus Eus3lb, donde encontramos 622 verbos diferentes. 40 verbos tienen más de 20 apariciones y 482 verbos menos de 5. Nuestros criterios para el estudio preliminar fueron, por un lado, la frecuencia, y por otro, la facilidad. Normalmente los verbos más frecuentes suelen ser, a su vez, los más complejos, ya que son generalmente los más ambiguos, y los que, a menudo, forma unidades complejas con otras palabras. Tal es el caso de los verbos *egin* ('hacer'[5]) e *izan* ('ser', 'tener'), que aparecen como los más frecuentes en el corpus. En la tabla 1 se pueden ver los porcentajes de los 10 primeros verbos más frecuentes.

Frecuencia	Verbo
%38.0	egin 'hacer'
%37.0	izan 'ser, tener'
%29.0	esan 'decir, llamar'
%16.0	adierazi 'expresar'
%15.0	eskatu 'pedir
%12.0	eman 'dar'
%10.0	azaldu 'exponer'
%9.0	hartu 'coger'
%9.0	jo 'considerar, tocar...'
%9.0	salatu 'denunciar

Tabla 1. Frecuencia de los 10 primeros verbos en Eus3lb.

Ante estos datos, y dejando a un lado los verbos *egin* e *izan*, elegimos los siguientes 3 verbos más frecuentes: *esan* ('decir'), *adierazi* ('expresar'), *eskatu* ('pedir')

3.1.2 Obtención de los sentidos del verbo en euskera y búsqueda de equivalentes léxicos en inglés

Para explicar los siguientes pasos, tomaremos como ejemplo el verbo *esan*. Una vez elegido el verbo, debemos obtener la visión general de sus sentidos en euskera. Para

[5] Las traducciones son orientativas.

ello, comenzamos consultando los diccionarios monolingües: Euskal Hiztegia (Sarasola 97) e Hiztegi Modernoa (Elhuyar 00). En los diccionarios aparecen tres acepciones: 'expresar algo mediante palabras', 'prometer'y 'llamarse'.

A continuación, consultamos los sentidos que se le han asignado a *esan* en la tesis de Aldezabal (04):
- esan-DU-1 y esan-DU-2: actividad (de expresión) de una entidad
- esan-DIO-3: asignación de un atributo o característica a una entidad

Dado que en la tesis de Aldezabal los sentidos y los frames se diferencian por el número de argumentos, por la(s) realización(es) sintáctica(s), y por los roles semánticos, no se hace la distinción entre 'expresar algo mediante palabras' y 'prometer'. Más adelante veremos cómo se codifican los frames en Aldezabal (04).

Siguiendo las lineas de Aldezabal, finalmente definimos el verbo *esan* con dos acepciones:
- alguien dice (a alguien) algo
- alguien llama a algo de una manera

Una vez delimitados los sentidos, analizamos los equivalentes léxicos en inglés de esos sentidos, para lo cual utilizamos el diccionario Morris (Morris 98). Así, elegimos los siguientes equivalentes:
- alguien dice (a alguien) algo : *say, tell*
- alguien llama a algo de una madera: *call*

Para cercionarnos de que los equivalentes son adecuados, consultamos los equivalentes castellanos de los sentidos de *esan* en el diccionario Elhuyar (Elhuyar 00) (*decir* y *llamar* son los más cercanos a los sentidos), así como los equivalentes de *decir* y *llamar* en el Diccionario on-line WordReference, y comprobamos que efectivamente los que hemos elegido (*say, tell, call*) aparecen como equivalentes de los sentidos de *decir* y *llamar*.

3.1.3 Analizar los equivalentes sintácticos en inglés

Una vez elegidos los sentidos en euskera y los verbos en inglés, el siguiente paso importante es elegir los equivalentes sintácticos en PropBank y Verbnet. Este paso y el anterior (búsqueda de equivalentes léxicos) muchas veces se dan conjuntamente, ya que al buscar el equivalente léxico, consciente o inconscientemente se piensa también en la conducta sintáctica. En este paso (3.1.3) precisamente se hace hincapié en localizar esas equivalencias sintácticas.

Para ello, debemos conocer mínimamente la filosofía general de PropBank y la de Verbnet, que es lo que vamos a hacer brevemente a continuación.

Respecto a PropBank, distinguen dos niveles independientes:
- Un nivel de argumentos y adjuntos
- Otro nivel de roles semánticos específicos (*buyer, thing bought, speaker…*)

Los argumentos son numerados del 0 al 4 (Arg0, Arg1, …, Arg4). Los adjuntos no llevan número; simplemente la M de modificador(ArgM)

Cada verbo tiene sus rolsets (sentidos) y cada rolset tiene sus frames (realizaciones sintácticas: frameset). Por ejemplo: el rolset para *tell.01* y los frames asociados al roleset *tell.01*son los siguientes:

rolset *tell.01*:
arg 0 speaker
arg 1 utterance
arg 2 hearer

Los frames relacionados con el rolset *tell.01*:

tell.01
ditransitive (-)
The score tell you what the characters are thinking and feeling.
Arg0: The score
REL: tell
Arg2: you
Arg1: what the characters are thinking and feeling
odd ditransitive (-)[6]
prepositional arg2 (-)
fronted (-)

Respecto a Verbnet[7], la clasificación de verbos está basada en Levin (1993), por lo que a cada verbo se le asigna el número de la clase semántica de Levin (9.1, 9.2, 10.1…) y unos roles más generales que en PropBank (*agent*, *theme*, *topic*, *beneficiary*…). También, si es el caso, se definen las propiedades semánticas: (±animate, ±organization, ±communication, ±concrete, ±location, ±region, ±animal…).

Una vez entendida una entrada verbal al estilo PropBank y Verbnet, consultamos las entradas y las estructuras sintácticas en la base de datos Verbnet/PropBank de los equivalentes léxicos de esan 1 (*say*, *tell.*…) y esan 2 (*call*…). Por ejemplo (1):

(1) say.01 (LEVIN say 37.7)
ARG0-null Sayer (VN Agent 37.7) = **10449** **(by = 1)**
ARG1-null Utterance (VN Topic 37.7) = **10491** **(by = 1)**
10503 ARG2-null Hearer (VN Recipient 37.7) = **12** **(to = 10)**
ARG3-null Attributive (VN) = **37** (null = 2, about = 10, ADV = 1, for = 2, For = 3, in = 1, of = 16, Of = 2)

tell.01 (LEVIN tell 37.1-1 37.2)
ARG0-null Speaker (VN Agent 37.1-1:Agent 37.2) = 323 (by = 4)
356 ARG1-null Utterance (VN Topic 37.1-1:Topic 37.2) = 340 (about = 8, of =5)
ARG2-null Hearer (VN Recipient 37.1-1:Recipient 37.)=310 (to=2, REC=1)

Como vemos, en la base de datos se nos muestra cómo y cuántas veces se han realizado los diferentes argumentos en las oraciones etiquetadas. En la izquierda

[6] Por delimitar el espacio, en el ejemplo sólo hemos puesto completo un frame (ditransitive (-)).
[7] http://www.cis.upenn.edu/group/verbnet/

podemos ver también cuántas apariciones han tenido say.01 y tell.01 (10503 y 356 respectivamente).

Por último, tenemos en cuenta lo propuesto en la tesis de Aldezabal (04), esta vez teniendo en cuenta también las realizaciones sintácticas. En Aldezabal (04) se proponen 3 frames sintáctico-semánticos (fss), correspondientes a dos sentidos:

1. alguien dice algo: 2 argumentos con 2 variantes sintácticas:
   esan-DU-1: experimentador (erg[8]); tema (abs)
   esan-DU-2: experimentador (erg); tema (compl.)

2. alguien a algo/alguien dice de un forma: 3 argumentos, sin variantes sintácticas:
   esan-DU-3: origen (erg); destino (dat); característica (abs)

De esta forma, podemos pasar al siguiente paso.

3.1.4 Proponer las entradas para los verbos en euskera al estilo VerbNet/PropBank

Comparamos las entradas de PropBank/Verbnet con las de Aldezabal (04), y vemos que en PropBank ha y una mayor tendencia a considerar argumentos. Elementos que en Aldezabal aparecen como adjuntos, en PropBank se consideran argumentos. Dado que en este caso la consideración de unos elementos como argumentos o adjuntos no influye en el conjunto de los sentidos del verbo *esan*, y dado que nuestra finalidad es seguir el modelo PropBank, los consideramos como argumentos y mantenemos la información de los casos. Así las entrada léxicas finales del verbo *esan* al estilo PropBank son:

esan.01	(alguien dice algo a alguien sobre algo)			
Arg0	*el que dice*	*sayer*	*agent*	ERG
Arg1	*lo dicho*	*utterance*	*topic*	ABS/COMPL
Arg2	*oyente*	*hearer*	*recipient*	DAT
Arg3[9]	*atributo*	*attributive*		INS, -i buruz

esan.02	(alguien dice a algo/alguien de una manera)			
Arg0	*el que dice*	*caller*	*agent*	ERG
Arg1	*el calificado*	*item being labelled*	*theme*	DAT
Arg2	*la calificación*	*attribute of arg1*	*predicate*	ABS

Una vez definida la entrada, damos paso a etiquetar el corpus.

3.2 Etiquetado del corpus

En esta fase pasamos a la tarea de etiquetado del corpus EPEC. En el ejemplo (2) se puede ver cómo se añaden la etiqueta del argumento numerado (o adjunto) y el del rol

[8] erg, abs, compl. y dat. son abreviaturas de los sufijos flexivos ergativo, absolutivo, completivo y dativo, respectivamente

[9] Como se puede apreciar, en las entradas de say.01 y tell.01 hay una diferencia: say.01 admite un arg3 mientras que tell.01 no. Dado que en euskera el verbo *esan* puede admitir dicho arg3, hemos decidido ponerlo como argumento.

semántico al corpus previamente etiquetado con dependencias sintácticas. Nótese que únicamente se etiquetan los argumentos superficiales.

(2) *Euskal gatazka ez dela armen bidez konponduko erran digu, politikoki baizik, borroka armatua eta errepresio itsua gaitzetsirik* ('Nos ha dicho que el conflicto vasco no se arreglará mendiante las armas…')

ccomp_obj[10] (konp, esan, aditz_aurk, dela) Arg1 *utterance/topic*
ncsubj (erg, erran, pro2.1, pro2.1, subj)
nczobj (dat, erran, pro3.1, pro3.1, zobj)
auxmod (-, esan, digu)

Los criterios generales que se han aplicado son los siguientes:

- Se trata la semántica superficial: sólo se etiquetan los argumentos/adjuntos que aparecen en el corpus.
- La unidad es la oración de cada verbo a tratar
- Las oraciones dudosas se dejan a un lado.

3.3 Completar la tabla de comentarios

Al acabar de etiquetar un verbo, se agrupan las dudas y se completa una tabla con comentarios, tales como los problemas que se han tenido con los argumentos (por ejemplo, ciertos casos que no está claro si etiquetarlos como Arg3 attributive), ejemplos incompletos o dudosos, ambigüedad de los modificadores (pertenecen a un verbo u otro) etc.

4. Conclusiones (14 ptos)

Con este estudio preliminar hemos comprobado que el modelo PropBank es adecuado o se puede adaptar fácilmente a las entradas y el etiquetado del verbo vasco. Con los tres verbos analizados no ha habido mayores problemas. Quizás se debería probar con verbos más difíciles (ambiguos). Por lo tanto, llegamos a la conclusión de que el modelo PropBank es válido para el euskera.

Por otro lado, hemos visto la posibilidad de realizar el etiquetado de forma semiautomática. Es decir, se pueden utilizar heurísticos que garantizan un etiquetado automático correcto. Estos heurísticos pueden ser generales. Por ejemplo, el caso de declinación erg (ergativo) siempre es Arg0. Y pueden ser específicos para cada verbo. Por ejemplo, con el verbo *esan*, las ambigüedades de los casos de declinación según los sentidos son las siguientes:

[10] ccomp_obj, ncsubj, nczobj, auxmod son etiquetas utilizadas en las dependencias que corresponden a objeto completivo, sujeto, objeto indirecto y auxiliar, respectivamente.

Casos de declinación	**Roles**	**Sentidos de *esan***
ERG	Arg0: Agent	01/02
ABS	Arg1: Topic / Arg2: Predicate	01/02
COMP	Arg1: Topic	01
DAT	Arg2: Recipient / Arg1: Theme	01/02
INS / *-i buruz*[11]…	Arg3: Attributive	01

Así, los casos COMPL, INS e *-i buruz*, desambiguarían los sentidos y los roles sin ningún error. En este corpus en concreto, el COMPL ha tenido una presencia del 82 %, y el INS la de un 3 %. Eso significa que sólo un 18 % quedaría ambiguo, para tratarlo manualmente. De todas formas, siempre se deberá hacer una revisión manual final.

En cuanto a los adjuntos, hay que etiquetarlos manualmente.

Por tanto, concluimos diciendo que la tarea principal consiste en definir bien las entradas verbales. A partir de ahí, se pueden proponer métodos semiautomáticos que faciliten y agilicen la tarea de etiquetado.

Referencias

Aduriz I., M. Aranzabe, J.M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, R. Urizar (2006). "Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing". *Corpus Linguistics Around the World. Book series: Language and Computers*. Vol 56 (pag 1- 15). ISBN 90-420-1836-4 Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi. Netherlands.

Aldezabal, I.(2004) *Estudio de la subcategorización verbal. Análisis detallado de 100 verbos en euskera, basándose en Levin (1993) y utilizando métodos automáicos*. Tésis doctoral. UPV-EHU.

Aranzabe M., J.M. Arriola, A. Díaz de Ilarraza (2004)."Towards a Dependency Parser of Basque". *Proceedings of the Coling 2004 Workshop on Recent Advances in Dependency Grammar. Geneva, Switzerland.*

Elhuyar (2000). Elhuyar hiztegia. Elhuyar fundazioa.

Elhuyar (2000). Hiztegi Modernoa. Elhuyar fundazioa.

Levin, B. (1993). *English verbs classes and alternations. A Preliminary Investigation.* The University of Chicago Press

[11] INS es la abreviatura correspondiente a instrumental. Por otro lado, el equivalente de la preposición compleja *-i buruz* sería 'a cerca de'.

Morris (1998). Morris Hiztegia.

Palmer, M., Gildea, D., Kingsbury, P. (2005). “The Proposition Bank: A Corpus Annotated with Semantic Roles”. In Computational Linguistics Journal. 31:1.

Palomar, M., M. Civit, A. Díaz de Ilarraza, L. Moreno, E. Bisbal, M. Aranzabe, A. Ageno, M.A. Martí, B. Navarro (2004). “3LB: Construcción de una base de árboles sintáctico-semánticos para el catalán, euskera y castellano”. *XX Congreso de la SEPLN*

Sarasola, I.(1997). Euskal Hiztegia. Kutxa fundazioa.

Word Referente: http://www.wordreference.com/