ANUARI DE FILOLOGIA. ESTUDIS DE LINGÜÍSTICA (Anu.Filol. Estud. Lingüíst.)
1/2011, pp.167-193, ISSN: 2014-1408

PARALELIZACIÓN DEL CORPUS SENSEM: ESPAÑOL-CATALÁN

GLORIA VÁZQUEZ[1]
Universitat de Lleida
gvazquez@dal.udl.cat
ANA FERNÁNDEZ MONTRAVETA
Universitat Autònoma de Barcelona
ana.fernandez@uab.es

RESUMEN

En ese trabajo presentamos un corpus paralelo para las lenguas española y catalana, SenSem. Se trata de un corpus paralelo anotado a distintos niveles lingüísticos (morfológico, sintáctico y semántico) y que abarca distintas unidades de información (palabra, sintagma y oración). Uno de los valores principales de este recurso es que ha supuesto la creación del primer corpus del catalán con información relativa a aspectos de la semántica oracional, como la aspectualidad y la modalidad. El recurso se ha creado traduciendo al catalán el corpus ya creado para el español en un proyecto anterior y manteniendo los vínculos entre ambas lenguas entre las distintas unidades de información mencionadas. Además, se ha heredado la anotación que se llevó a cabo para el español, corrigiendo cuando ha sido necesario la información para el catalán.

PALABRAS CLAVE: anotación de corpus, corpus del español, corpus del catalán, paralelización de corpus, sintaxis, semántica

PARALLELIZATION OF THE SENSEM CORPUS: SPANISH AND CATALAN

ABSTRACT

This paper presents a parallel corpus for Spanish and Catalan, SenSem. The parallel corpus has been annotated at several linguistic levels (morphological, syntactical and semantic). The information covered at the different levels ranges from words to phrases, and sentences. One of the main values of the resource presented in this work is that it is the first corpus for Catalan that has been annotated with information regarding sentence semantics such as aspectuality and modality. The methodology followed in this project was to translate into Catalan the corpus we already had available from a previous project in Spanish. The link between both corpora has been established at the different informative units mentioned above. All the linguistic annotation has been inherited, correcting necessary aspects when it proved necessary.

KEY WORDS: corpus annotations, corpus for Spanish language; corpus for Catalan language, parallel corpus, syntax, semantics

[1] Trabajo financiado por el Ministerio de Ciencia e Innovación mediante el proyecto "Adaptacion Multilingue del Banco de Datos SenSem (catalán)" - MICINN - FFI2009-06939-E.

1. INTRODUCCIÓN

El corpus SenSem-Cat se ha creado a partir de un corpus ya existente para la lengua española, el corpus SenSem. [2] Este último es el resultado del trabajo desarrollado por los miembros del grupo GRIAL[3] durante los últimos ocho años (Castellón *et al.* 2006; Vázquez y Fernández 2009). Este corpus del español consta de un total de 30.000 frases, 25.000 pertenecientes al registro periodístico y 5.000 al literario, y está anotado a nivel léxico (sentido verbal), morfológico (categoría morfo-sintáctica), sintáctico (funciones y categorías sintácticas) y semántico. En este último nivel se distingue entre la semántica a nivel de palabra (*Aktionsart* y sentido verbal), a nivel de sintagma (roles semánticos de los argumentos y aspecto del sintagma verbal) y a nivel de oración (construcciones, modalidad y aspectualidad).

El corpus SenSem-Cat contiene el mismo tipo de información que el corpus SenSem del español aunque es de tamaño algo más reducido, ya que se compone de la parte correspondiente al registro periodístico, y no de la parte del registro literario. No se ha considerado oportuno la traducción del corpus literario del español en tanto que este registro incorpora nuevas dificultades, como por ejemplo el uso creativo de palabras, y creemos que como primera aproximación es más apropiado excluirlo. Como el registro periodístico constituye aproximadamente el 80% de todo el corpus SenSem-Esp, la diferencia no es muy importante. Así, el corpus confeccionado en lengua catalana contiene 25.000 frases y unas 700.000 palabras.

El objetivo final de este proyecto que aquí presentamos es usar el catalán como banco de pruebas con el fin de comprobar la viabilidad de crear corpus multilingües de lenguas próximas en los que se pueda "reusar" una parte importante del trabajo realizado en una de las lenguas de partida.

Las motivaciones que nos llevaron a trasladar el corpus SenSem-Esp a la lengua catalana y crear la réplica del corpus en esta lengua son diversas. En primer lugar, se consideró pertinente contribuir en el aumento del número de recursos lingüísticos de dicha lengua, el catalán, con una aportación novedosa en el ámbito de la anotación de corpus. Cabe decir que la gran mayoría de corpus del catalán existentes, en el caso de estar anotados, lo están tan solo con información de tipo morfosintáctico, que normalmente incluye la categoría de la palabra y el lema, como sucede con el Corpus Textual Informatitzat de la Llengua Catalana (IEC, 50 millones de palabras), el CUCWeb (Dpto. de Traducció i Filologia; Departamento de Tecnología y Cátedra Telefónica de Producción Multimedia, Universitat Pompeu Fabra, 208 millones de paraules) y los corpus técnicos del

[2] "SenSem: Banco de datos sintáctico y semántico del español" - Ministerio de Ciencia y Tecnología (BFF2003-06456) y "Ampliación de la BD léxica y el corpus sintáctico-semántico de semántica oracional del español SenSem" - Ministerio de Educación y Ciencia (HUM2007-65267).

[3] http://grial.uab.es

IULA del catalán (8 millones de palabras).[4] A nuestro conocimiento, no existe ningún corpus en catalán con anotación semántica oracional que incluya la descripción del aspecto y la modalidad.[5]

En segundo lugar, otra de las motivaciones fue que se disponía ya de herramientas informáticas desarrolladas en el seno del proyecto SenSem-Esp que nos podrían servir como punto de partida para la confección del corpus catalán teniendo en cuenta que habría que realizar las remodelaciones y ampliaciones que fueran necesarias para adaptar dichos recursos a un formato multilingüe.

En tercer lugar, y aún más importante, además de la reutilización y adaptación de las herramientas, se partió de la idea de que se podría también aprovechar todo el trabajo de anotación manual del corpus del español. Dado el grado de similitud entre ambas lenguas y lo costoso que resulta realizar la anotación manual de corpus, se consideró oportuno confeccionar el recurso catalán traduciendo las oraciones del español y traspasando la anotación, junto con toda la estructura de las bases de datos, de una lengua a la otra, evidentemente con las consiguientes modificaciones que fueran necesarias sabiendo, por otro lado, que no iban a ser excesivas (microvariación, Hernanz y Rigau 2003[6]).

Por último, la consecución del proyecto permitía aportar un recurso que puede tener gran utilidad en diferentes ámbitos. Por un lado, el hecho de que se trate de un corpus anotado va a permitir llevar a cabo estudios empíricos sobre diversos aspectos lingüísticos relacionados con la sintaxis y la semántica del catalán y también desde el punto de vista contrastivo con el español, por lo que hay que destacar su interés en los campos de la traducción (Baker 1993) y la enseñanza de lenguas. Asimismo, en el campo del procesamiento del lenguaje la explotación de este corpus del catalán podría mejorar los sistemas automáticos que trabajan con esta lengua. Más concretamente, el reflejo de la información relativa a la aspectualidad y la modalidad que subyace en las oraciones es muy útil tanto para los sistemas que requieran un procesamiento profundo de las lenguas, como los sistemas de extracción de información, las aplicaciones de

[4] Cabe mencionar en este sentido HISPACAT (Villalba 2008), una base de datos de construcciones en catalán y español. Para la creación de dicha base de datos se creó una ontología de conceptos gramaticales, entre los que se incluye la aspectualidad y la modalidad, entre muchos otros, que se usan para definir un conjunto de cerca de 500 construcciones. Para cada construcción se incluyen ejemplos, que han sido extraídos de obras gramaticales que constituyen un marco de referencia en cada lengua. Aunque no hemos podido consultar dicho recurso, por la bibliografía consultada parece que no se trata de un corpus anotado.

[5] Cabe decir que en el resto de lenguas tampoco es común el disponer de corpus anotados con información semántica oracional. Destaca en este ámbito un corpus japonés-inglés (Murata *et al.* 2005 y 2006), así como el TimeBank para el inglés (Pustejovsky *et al.* 2003).

[6] Proyecto MCyT/FEDER BFF2003-08364-C02: "(A)simetrías sintácticas en catalán y español". Universitat Autònoma de Barcelona.

pregunta y respuesta y de traducción automática (Saurí *et al.* 2006; Murata *et al.* 2006).

Aunque la tarea a la que nos hemos enfrentado se concibió como factible desde un inicio por los motivos mencionados, nos hemos encontrado con distintos escollos, aunque consideramos que no son de una gran envergadura. En primer lugar, la adaptación informática de las herramientas ha sido más compleja y costosa de lo esperado en cuanto a tiempo y recursos humanos. Se han tenido que reestructurar las distintas bases de datos y reprogramar las interfaces de anotación de las oraciones y visualización de las mismas. No obstante, la estructura básica de la base de datos se ha mantenido, ampliándola, eso sí, para dar cabida a las equivalencias en catalán. Es decir, la estructura de anotación es compartida y está ligada a las formas que la expresan en cada una de las lenguas tratadas. Como veremos más adelante, esta decisión metodológica nos ha comportado algún problema.

En segundo lugar, la idea inicial de la que partimos fue la de traducir los verbos del español al catalán explotando las relaciones sinonímicas de esta lengua con el fin de aportar un valor añadido al léxico creado. Esta idea ha tenido que ser reelaborada. Hay que tener en cuenta que el léxico de SenSem contiene información sobre número y tipo de argumentos, por lo que el concepto de sinonimia debe ceñirse a este criterio, además de los propios de la semántica léxica. Ello provocó un retraso en el inicio del proceso de traducción y (re)anotación de las oraciones debido a la complejidad del tema.

A continuación, indicamos cómo se estructura el presente artículo. El apartado 1 está dedicado a la descripción de la metodología usada en el proyecto SenSem-Esp, así como de los recursos creados en torno a éste, con el fin de comprender mejor la construcción y el contenido del corpus catalán y su paralelización con el español, que son presentados en el apartado 3. En el apartado 4, describimos los reajustes llevados a cabo para la adaptación multilingüe de las herramientas informáticas creadas en el proyecto SenSem-Esp. Los desajustes más significativos desde el punto de vista lingüístico entre las dos lenguas y sus repercusiones en la estructura de la base de datos son presentados en el apartado 5. Finalmente, dedicamos un apartado final a las conclusiones y las líneas futuras.

2. EL PROYECTO SENSEM DEL ESPAÑOL

En una primera fase, para la recopilación de los textos que forman SenSem, se partió de un corpus inicial (sin anotar) de unos 13 millones de palabras pertenecientes al registro periodístico. A partir de este conjunto de textos se extrajo el listado de los 250 verbos más frecuentes en dicho corpus y se procedió a la selección aleatoria de 100 frases de cada uno de ellos. Estas 25.000 frases constituyeron inicialmente el corpus SenSem.

Uno de los motivos que nos llevó a la elección de textos periodísticos para la ejecución del proyecto fue la existencia de diarios en línea, ya que ello facilita enormemente la recopilación de grandes cantidades de texto en un tiempo razonable. Otro motivo importante por el cual se decidió usar este género fue que los textos escritos en un periódico presentan cierta variedad de registros: por un lado, suelen corresponderse en un alto porcentaje con el registro de la lengua estándar y, por otro, entre los textos que incluye un periódico se encuentran también artículos escritos con una formalidad elevada, rozando el estilo literario. Además, la diversidad temática que caracteriza los textos periodísticos fue también considerada como un aspecto positivo para la consecución de nuestro objetivo, ya que permitiría un uso variado de los ítems léxicos y, por tanto, aportaba mayor riqueza a los textos. Todo ello nos decidió a considerar el uso de textos periodísticos como un buen referente a la hora de pretender describir los usos lingüísticos a partir de textos reales. Luego, en una segunda fase del proyecto se consideró que podría ser interesante comparar los usos periodísticos con los literarios y se procedió a ampliar el corpus introduciendo 5.000 oraciones de registro literario (20 para cada uno de los verbos), extraídas de obras narrativas de autores españoles de los s. XX y XXI.

Previa anotación del corpus, se creó un léxico a partir del listado inicial de los 250 verbos mencionados. Para ello, se procedió a definir para cada uno de estos verbos sus diferentes sentidos partiendo de diferentes fuentes lexicográficas ya existentes. Además, para cada sentido verbal, se estableció su correspondencia con el sentido de la ontología WordNet y también se incluyó la información sobre el tipo aspectual léxico (*Aktionsart*). Por último, se codificaron los roles semánticos de los participantes argumentales requeridos por el verbo. Como resultado de este proceso se obtuvo una base de datos léxica con un total de 1.100 sentidos verbales descritos en estos términos.

Una vez confeccionado el léxico, se procedió a la anotación del corpus. El primer paso en esta tarea implicaba la elección del sentido verbal reflejado en la frase usando el listado de sentidos que se había confeccionado para cada verbo. Seguidamente, se procedía a la localización de los diferentes argumentos que aparecían en dicha oración partiendo del listado de roles semánticos asociados al sentido escogido, sabiendo que algunos de dichos participantes podían no estar presentes sintácticamente puesto que ello dependía de la construcción elegida por el emisor del mensaje.

Por lo que se refiere al *Aktionsart*, el tipo aspectual léxico de cada verbo, éste se usaba como punto de partida, sabiendo que no necesariamente coincide con la aspectualidad reflejada en la oración en que el verbo participa, ya que en ocasiones esta información se modifica en base a otros elementos que puedan aparecer en el contexto de la frase (Smith 1997; Xiao y McEnery 2004). Además, para cada oración también se codificó información sobre la aspectualidad de otro tipo, como la perfectividad y la habitualidad.

Otro tipo de información a nivel oracional que se describió para cada oración fue el tipo de construcción (anticausativa, pasiva, impersonal, construcción de dativo, reflexiva, recíproca), la polaridad (positiva o negativa) y la modalidad (asertividad y algunos tipos de no asertividad).

Por último, en la anotación de las oraciones, se incorporó también información morfo-sintáctica de los constituyentes, como los tipos de sintagmas, sus funciones sintácticas y su valor argumental.

Una vez realizada la anotación completa del corpus, se procedió a la ampliación de la información contenida en las entradas léxicas a partir de los datos extraídos del corpus. Es decir, el léxico inicial, en el que se habían codificado datos básicos sobre los verbos, se amplió con algunos otros recuperados a través de la anotación de las oraciones que aparecen asociadas a cada sentido, como los patrones de subcategorización y su frecuencia, así como la semántica de la construcción con los que dichos patrones han sido relacionados.[7] El tipo de aspectualidad oracional, la polaridad y la modalidad no se incluyó en el léxico a la espera de realizar un estudio pormenorizado de estas cuestiones para cada sentido verbal que permita determinar en qué casos este tipo de información puede ser relevante para la descripción adecuada de la pieza léxica.

3. CONSTRUCCIÓN DEL CORPUS PARALELO ESPAÑOL-CATALÁN

Tal y como hemos avanzado, para la creación de SenSem-Cat se ha partido del corpus SenSem-Esp, tanto por lo que se refiere al contenido textual, ya que se han traducido automáticamente los textos del español al catalán, como por lo que se refiere a la anotación sintáctico-semántica, que inicialmente se ha heredado también del corpus citado, aunque se han realizado las modificaciones convenientes en la anotación cuando ha sido necesario. Se ha elegido esta metodología puesto que se ha partido de la premisa de que las diferencias entre ambas lenguas a nivel estructural son escasas.

Una tarea imprescindible en dicho proyecto ha sido la constitución del léxico verbal catalán correspondiente al léxico verbal SenSem-Esp. Para la consecución de esta tarea no se han podido reutilizar herramientas ya existentes, es decir, diccionarios bilingües en formato electrónico, porque, en primer lugar, la descripción de sentidos del léxico verbal SenSem es original, es decir, aunque se base en otras fuentes lexicográficas no necesariamente coincide la distinción de sentidos realizada en SenSem-Esp con dichas fuentes. Por otro lado, para poder aprovechar diccionarios bilingües existentes habría que previamente establecer correspondencias entre los sentidos de cada verbo de SenSem-Esp y los sentidos

[7] El corpus anotado, el léxico verbal y otros recursos desarrollados como parte del proyecto SenSem-Esp se pueden consultar en línea a través de: http://grial.uab.es/sensem/corpus.

de estas otras herramientas, tarea que no era factible, puesto que era muy difícil la total coincidencia de sentidos.

Por todo ello ha sido necesario traducir manualmente todos los sentidos del español al catalán. Como ya se ha comentado, se ha optado por introducir listados de sinónimos. Esto nos ha supuesto problemas adicionales, ya que a la hora de volcar los lemas del catalán y crear automáticamente el léxico de esta lengua se han sobregenerado sentidos que después se han tenido que revisar manualmente. Esta sobregeneración viene provocada porque, en primer lugar, no siempre es sencillo establecer sinónimos que requieren equivalencia argumental y, en segundo lugar, porque los sinónimos del español dentro del propio léxico de esta lengua no estaban establecidos, lo cual no ha facilitado el establecimiento de las correspondencias con el catalán. Por todo ello, ésta ha resultado una de las tareas más pesadas y costosas, en cuanto a tiempo, que se ha presentado en la primera parte del proyecto.

El siguiente paso ha consistido en traducir las frases del corpus. Esta tarea se ha hecho automáticamente con el traductor en línea Google, al cual accede la interfaz de traducción y anotación del proyecto. Cada palabra se envía al traductor de forma individual, que la devuelve, a su vez, traducida al catalán. No siempre la traducción es correcta, pero constituye un primer paso. En tanto que las diferencias estructurales entre ambas lenguas no son importantes, la traducción que realiza Google palabra a palabra es muy útil como punto de partida. La necesidad de alinear los textos y de conservar para el catalán la anotación asociada al texto en español no nos ha permitido optar por otro tipo de herramientas de traducción existentes, que, sin duda, aportarían mejores resultados desde el punto de vista de la calidad de la traducción. A nuestro conocimiento, no era factible usar estas herramientas y, a la vez, mantener las correspondencias léxicas entre las dos lenguas, ya que estos sistemas devuelven textos completos aislados, sin mostrar las correspondencias con la lengua original. Tampoco permiten reasignar la información anotada en el texto original al texto traducido, ya que suelen aceptar sólo texto plano y, por tanto, no enriquecido.

En cambio, con el motor de Google se han podido mantener las correspondencias de los ítems léxicos entre español y catalán, creando un corpus alineado palabra a palabra entre ambas lenguas, y también reasignar a las frases del catalán la información con la que se anotaron las frases del castellano en los distintos niveles lingüísticos. Respecto a la información asociada al conjunto de la oración (construcción, aspectualidad, modalidad y polaridad), ésta se reasigna a la nueva frase creada, formada por el conjunto de traducciones palabra por palabra que realiza Google y que la herramienta del proyecto recolecta formando una oración del catalán con un identificador nuevo. Respecto al nivel léxico, la palabra que devuelve Google queda relacionada con la palabra del texto origen y mantiene la información asociada (sentido). En cuanto al nivel sintagmática, el alcance de los distintos sintagmas viene delimitado en cada oración por la palabra de inicio y la

final, información que también se hereda en catalán. Como consecuencia, la información asociada a cada sintagma en las distintas oraciones del español (argumentalidad, categoría sintagmática, rol semántico y función sintáctica) también se reasigna a los correspondientes sintagmas del catalán.

Una vez realizado todo este proceso automático, en el que se crean las oraciones del catalán con información lingüística asociada, se ha procedido a ejecutar la fase manual, que consiste en verificar y/o corregir la traducción y la anotación de las frases.

En nuestro caso, tal como se ha avanzado, ha sido necesaria la edición de las frases catalanas en relación a determinadas cuestiones básicas a nivel léxico, como el uso de posesivos compuestos y no simples, que otros sistemas resuelven automáticamente con éxito durante el proceso de traducción y que Google no realiza. Así mismo, se han tenido que corregir también otros aspectos de la gramática catalana que son distintos de la española y que se resuelven a partir del análisis del contexto. Algunos de estos reajustes también serían fácilmente resueltos de forma automática ya que, aunque dependen del contexto, formalmente son predecibles, como la apostrofación. En otras ocasiones, aunque no siempre, la resolución correcta de determinadas cuestiones requeriría más conocimiento, como en el caso del fenómeno de la ausencia o el cambio de determinadas preposiciones y las formas compuestas de determinados relativos, entre otros.[8] Este tipo de reajustes no suelen repercutir en el tipo de información asociada a los sintagmas que contienen las palabras que se modifican o, si lo hacen, las modificaciones que deben realizarse son escasos, como en el caso del tipo sintagmático de los objetos directos de persona, que en español se codifican con un SP y en catalán con un SN.

Sí que tiene mayor relevancia, sin embargo, otros casos en los que no se ha podido establecer una correspondencia de la estructura a nivel formal entre ambas lenguas. Como veremos más adelante (apartado 5), por ejemplo, el paradigma de los pronombres clíticos es más complejo en catalán que en castellano, así como también es distinto el uso de los posesivos en ambas lenguas. Estas diferencias pueden tener consecuencias más graves en la paralelización.

En cualquier caso, la casuística con la que nos venimos encontrando confirma la hipótesis sobre la facilidad de la tarea de la traducción y transposición de la anotación. Cabe tener en cuenta que sólo en aquellas porciones de textos en que se realiza algún reajuste en la traducción aportada por el sistema automático que vaya más allá del cambio de una palabra por otra se va a tener que realizar algún

[8] Aunque la traducción del sistema de Google mejora siendo usada a través de la propia web (http://translate.google.es), no llega a tener la precisión de Internostrum (http://www.internostrum.com), que realiza ajustes teniendo en cuenta el entorno de la palabra. Así, el primer traductor continúa cometiendo errores con los pronombres relativos y también con algunos pronombres. Internostrum, en cambio, es capaz de solucionar problemas de traducción que no resuelve Google, siempre que no dependan de una interpretación semántica.

cambio en la anotación sintáctico-semántica. Además, dichos cambios sólo afectan el nivel de complementos y no el de la semántica de la oración. Así, tanto la aspectualidad como la modalidad y la polaridad son necesariamente las mismas en ambas lenguas. Lo mismo ocurre con la semántica de la construcción.[9] Teniendo en cuenta que la anotación de oraciones es una tarea especialmente ardua cuando hablamos de sintaxis, y sobre todo de semántica, consideramos que el poder reaprovechar en gran medida el trabajo realizado en un proyecto para construir otros recursos en otras lenguas es un gran avance.

Cabe decir, por último, que el proceso de anotación morfosintáctica a nivel de palabra se ha realizado de forma independiente en los dos corpus. Teniendo en cuenta que existen herramientas de anotación automática con cobertura y precisión muy aceptables en ambas lenguas, se consideró más oportuno en este caso no heredar para el catalán la anotación del español, ya que previsiblemente se producirían más errores que realizando la anotación directamente de los textos catalanes. En la fase última del proyecto, en la que estamos en el momento de redactar este artículo, se prevé corregir los errores que se hayan producido en este nivel. La herramienta utilizada ha sido Freeling (Atserias *et al.* 2006).

4. INTERFAZ DE ANOTACIÓN Y TRADUCCIÓN

Para crear el corpus y el léxico del catalán, se han tenido que adaptar una serie de herramientas informáticas provenientes del proyecto SenSem del español y crear otras nuevas. En primer lugar, se ha ampliado el editor del léxico para dar cabida al catalán (apartado 4.1). En segundo lugar, se ha ampliado la interfaz de anotación de las oraciones (apartado 4.2): se ha incluido un nuevo módulo de traducción al catalán de las frases del español y se ha incorporado un módulo propio para el catalán con el fin de poder modificar la información heredada de la etiquetación del español si fuera necesario.

4.1. Editor de sentidos

Como ya se ha indicado, en el proyecto español primero se creó el léxico verbal. Este paso era primordial ya que a la hora de anotar las oraciones era necesario asignar a cada una de ellas el sentido correspondiente. Por este motivo, el primer paso también fue la modificación de dicha herramienta para incorporar los lemas

[9] Lo que puede suceder es que el mecanismo morfológico que adopta el verbo para expresar la semántica de la construcción cambie en algunos casos, eso sí, excepcionalmente, entre ambas lenguas. Por ejemplo, algún verbo pronominal en español no lo es en catalán, como es el caso de *dormir*. Así, cuando este verbo es transitivo o intransitivo se traduce como *dormir* en catalán, pero cuando en español adopta la forma pronominal (*dormirse*) en catalán se usa un verbo derivado (*adormir-se*).

verbales del catalán y sus sentidos. El resultado es un editor a través del cual se establecen las conexiones entre los sentidos de las dos lenguas y en el que se visualiza la descripción lingüística para los sentidos correspondientes. Dicha descripción es compartida, ya que tanto la definición como la lista de roles semánticos y el *Aktionsart* han de ser, desde nuestro punto de vista, los mismos. En algún caso, puede darse alguna variante, pero entre dos lenguas tan cercanas culturalmente y estructuralmente como el castellano y el catalán es poco habitual. Por ejemplo, el verbo *poner* en español se traduce en catalán como *ficar* cuando el lugar de colocación es el interior de algún objeto y *posar* cuando es una superficie. Como puede observarse, en este caso, la divergencia no trasciende en la codificación realizada en el proyecto.

Las modificaciones que se han realizado en la herramienta del léxico se presentan a continuación. En primer lugar, se ha ampliado el editor del léxico español, ya que se ha añadido un campo en cada entrada léxica para la traducción del catalán. Una vez finalizada la tarea de traducción del léxico, se han volcado los nuevos lemas del catalán en un léxico específico para esta lengua, a la vez conectado con el español, ya que las entradas del catalán también contienen un campo para la traducción al español y, por tanto, las correspondencias entre ambas lenguas se pueden realizar y/o modificar desde ambas lenguas. Una vez creado el léxico catalán, la información de los campos de las correspondientes entradas léxicas han sido traducidos del español.

En la figura 1 se visualiza el aspecto de una entrada léxica de un verbo del español. Debajo del lema, en este caso el verbo *acordar,* se indica el idioma al que pertenece (en este caso, el español) y un número identificador en el léxico de esta lengua. En la siguiente sección se presenta el listado de todos los sentidos definidos en el proyecto para dicho verbo y se visualizan algunos de los campos de información para cada uno de ellos: la definición, los roles asociados y las traducciones al catalán. De este modo, aunque esta información es parcial, puesto que la entrada completa incorpora más datos, el usuario se puede hacer una idea general del número y tipología de sentidos del lema verbal en cuestión así como de las correspondencias establecidas entre las dos lenguas. A partir de esta pantalla, se puede editar cada sentido ya creado, así como borrar o añadir sentidos del español para el lema en cuestión. La traducción al catalán está en formato de hipervínculo. Ello permite acceder directamente desde esta pantalla del español a la entrada equivalente en catalán.

Figura 1: Entrada léxica del verbo *acordar*

En la pantalla de la figura 2 se observa la entrada léxica completa para un sentido concreto del verbo español mencionado (*acordar1*), que incluye, además de la definición, los roles y la traducción al catalán, ya mencionados anteriormente, otro tipo de información complementaria: los *synsets* y las *variants* de WordNet, los sinónimos en español dentro del propio léxico SenSem[10] y el tipo aspectual. En el apartado "Traducciones entrantes" se visualizan todos los sentidos del catalán que apuntan a este sentido del español como equivalente conjuntamente con la definición traducida al catalán y los roles asociados, ya que estos campos coinciden (puesto que son heredados) entre las dos lenguas. En el caso de que se quisieran añadir más equivalencias al catalán o bien eliminar las establecidas, el usuario podría hacerlo desde esta pantalla editando la información del apartado "Traducciones".

[10] Cabe decir que en el proyecto del español no se llevó a cabo la codificación de dichos sinónimos.

Figura 2: Entrada léxica del sentido *acordar 1*

4.2. Editor de traducciones y anotaciones

Como avanzábamos, también se ha tenido que rediseñar y reprogramar la interfaz de anotación (Fernández *et al.* 2006a; Fernández *et al.* 2006b) para incorporar las frases del catalán. En la figura 3, se presenta el aspecto de la pantalla inicial de dicha herramienta. Como puede observarse, en el apartado "Corpus" el usuario tiene que escoger la lengua con la que quiere trabajar y, si lo desea, el tipo de registro (periodístico o literario). A partir de ahí, el usuario decide en el apartado "Búsqueda" si quiere acceder a todas las oraciones de la lengua escogida, a todas las oraciones de un verbo concreto para dicha lengua o a una determinada oración. Usando los "filtros", se pueden escoger criterios varios para visualizar un subconjunto de oraciones, como por ejemplo, aquellas que sólo han sido anotadas por lo que se refiere a los complementos y, que, por tanto, requieren todavía ser anotadas en cuanto a la semántica oracional, o bien aquellas cuya traducción

automática no ha sido todavía revisada. Con el uso de estas opciones es más fácil controlar los procesos y hacer un seguimiento de los mismos.

Evidentemente, antes de poder acceder a las frases del catalán hay que crearlas, lo cual se lleva a cabo desde la pestaña "Traducción" del módulo del programa que gestiona las oraciones de la lengua española. Así pues, una vez se accede a una oración en concreto del español, se visualiza un menú en forma de pestañas que permite escoger el tipo de tarea a realizar entre 5 acciones (v. figura 4):[11]

Anotación de la semántica oracional (pestaña "Oración")
Anotación de los complementos (pestaña "Complementos")
Edición (pestaña "Editar")
Traducción de la oración en español al catalán (pestaña "Traducción")
Grabación de datos (pestaña "Guardar")

Figura 3: Pantalla de inicio de la herramienta de anotación de las oraciones de los corpus

[11] Una vez se acceda a una frase del catalán se dispone de la posibilidad de realizar el mismo tipo de tareas que en español excepto la de traducción.

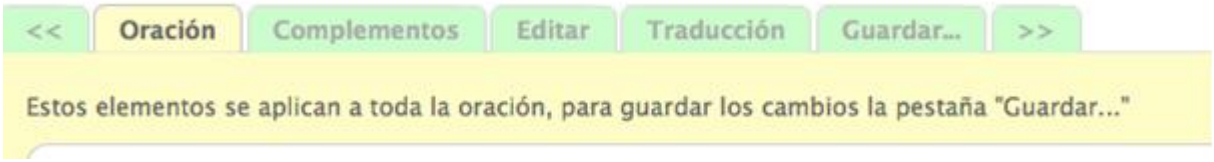

Figura 4: Menú de inicio de la interfaz de traducción

Las tareas de edición, anotación de los complementos y de la semántica oracional se desarrollaron ya como parte del proyecto SenSem del español. La anotación de la semántica oracional (pestaña "Oración") y de los "Complementos" fueron las tareas por excelencia a lo largo del proyecto español. La tarea de "Editar" nos permite modificar el texto original de la oración. Aunque no es un proceso habitual, en su momento fue necesario en algunas ocasiones recomponer una frase, bien porque estuviera cortada o bien porque presentara problemas con respecto a la visualización de caracteres especiales, ya que el volcado de frases se hizo de manera automática. En el proyecto actual, se ha trabajado en la traducción al catalán.[12]

En la figura 5 se presenta, a modo de ejemplo, una parte de la oración obtenida una vez ejecutado el programa de traducción automática para traducir una oración del verbo *citar*, cuyo alcance se indica en el margen izquierdo puesto que dicha oración está subordinada a otra. También se señalan las correcciones que deben realizarse. La traducción se realiza de todo el contexto, es decir, de la oración completa en que se localiza la oración de *citar*, pero la anotación del español sólo se ha llevado a cabo para los elementos que están en la frase específica de dicho verbo, y esto mismo se mantiene para el catalán.

[12] Durante el desarrollo de este proyecto también se ha aprovechado para corregir errores en la edición y/o anotación del español, en el caso de que se detecten al traducir y revisar la anotación del catalán.

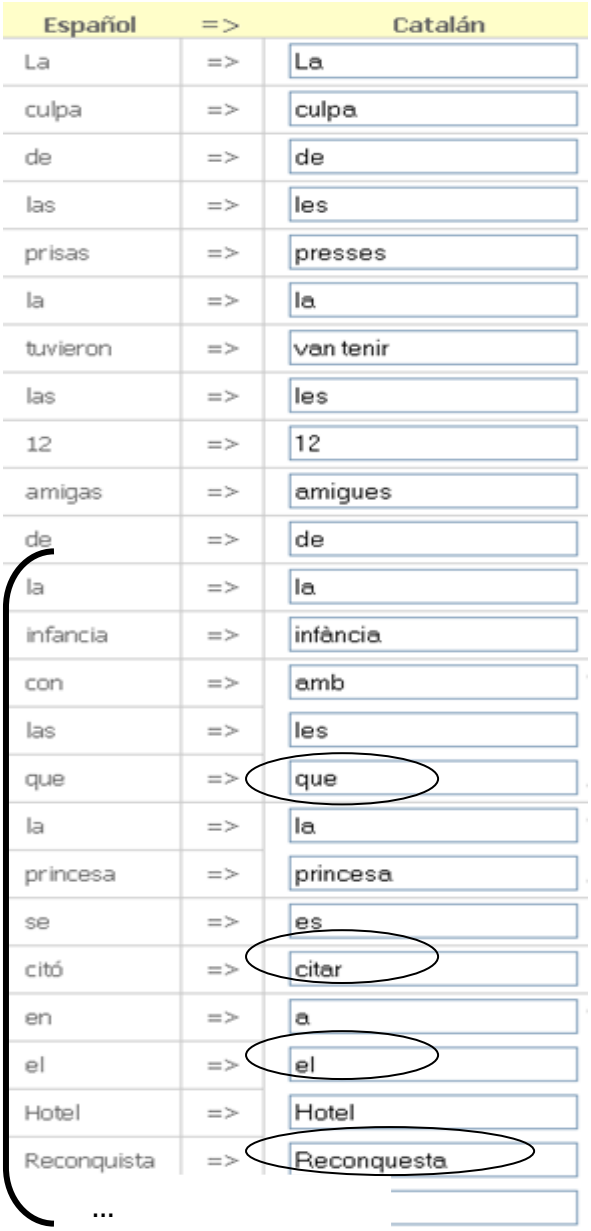

Figura 5: Resultado de la traducción automática y detección de errores

Una vez realizados los cambios oportunos en la frase catalana el usuario graba la traducción y el sistema crea automáticamente un identificador para la nueva frase catalana, que hereda automáticamente la anotación íntegra de la española. A partir de este momento se puede acceder desde esta misma pantalla a la frase catalana para revisar la anotación que se ha heredado, en el caso de que fuera necesario cambiar alguna de las etiquetas.

El sistema presenta desde cualquier pantalla una ficha de control de la frase en la que se está trabajando donde el usuario marca manualmente los procesos finalizados (v. figura 6). En la primera parte de esta ficha se indican los datos básicos de la oración, como su número identificador (12286), el lema con el que se

corresponde (*citar*) y el tipo de registro al que pertenece (periodístico). En segundo lugar, se indica si dicha frase es principal o secundaria, quedando relegadas estas últimas a un bloque de frases complementarias que se utilizan en el caso de que por algún motivo se tenga que prescindir de alguna principal, por ejemplo, porque ha habido algún error en el proceso de volcado y la frase ha quedado cortada o con demasiados problemas de codificación. En la tercera parte de la ficha se indica si la frase en cuestión ha sido anotada a nivel de oración y complementos. Además, en la última sección de la ficha, se indica si la traducción al catalán de dicha frase ha sido revisada manualmente, tanto por lo que se refiere a la traducción ("Tr. revisada") como por lo que se refiere a la anotación ("Tr. anotada"). Además, al final de todo se indica el identificador de la frase catalana con un *hiperlink* que permite el acceso directo a ésta.

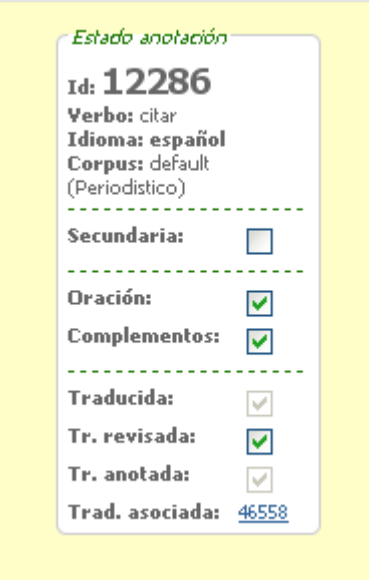

Figura 6: Cuadro de control de la oración del español

La frase resultante en catalán es la siguiente, una vez corregida,: "…amb les quals (amigas de la infancia) la princesa es va citar a l'hotel Reconquista, on s'allotajaran aquests dies"). Esta frase no presenta cambios estructurales respecto al español, por lo que se ha mantenido la anotación que se había efectuado para esta lengua. Por un lado, como puede observarse en la figura 7, donde se visualiza la caracterización de la semántica de la oración que estamos analizando: expresa un evento perfectivo, se trata de una oración recíproca (la princesa y sus amigas quedaron conjuntamente en verse en dicho hotel), con modalidad asertiva y polaridad positiva.

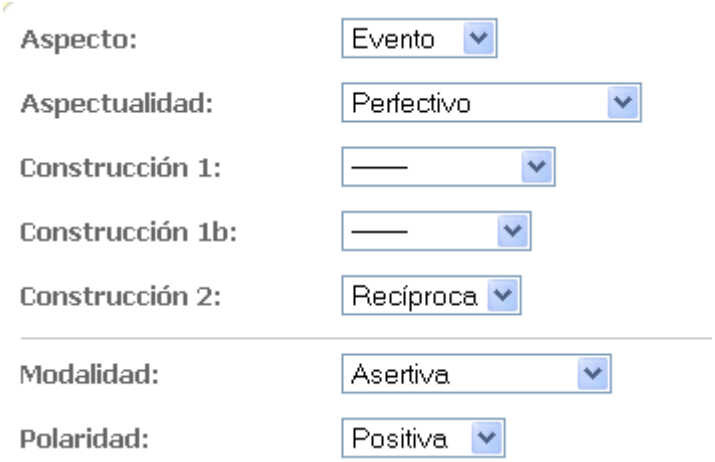

Figura 7: Anotación semántica de la oración

Por otro lado, en la figura 8, se visualiza la anotación de los complementos. El sentido al cual se ha asociado la oración es el número 4 del lema *citar* del catalán, que se trata de un evento con 3 roles: un agente plural (ag(pl)), un localizador temporal (temp-loc) y un lugar (loc). Los roles del tipo ("pl"), como el primero de los mencionados, han sido creado en el proyecto para el caso de verbos recíprocos léxicos, como el sentido de *citar* que estamos tratando, donde el evento siempre tiene lugar entre mínimo dos personas. En la sintaxis estos participantes pueden aparecer en el mismo constituyente (con un SN plural o coordinado) o bien en dos, como en la frase que nos ocupa, donde hay un SN sujeto ("la princesa") y un SP relativo ("amb les quals": con las amigas de la infancia). Respecto al lugar del encuentro es un SP que incluye una oración de relativo "a l'Hotel Reconquista, on s'allotjaran aquests dies". Por último, en esta oración el momento de la cita se obvia, por lo que el rol "temp-loc" queda sin adjudicar. Además de la información sobre el rol semántico, para cada complemento, se da información sobre la categoría sintagmática, la función sintáctica y la argumentalidad, que, como hemos dicho, coincide con el español.

En esta figura, como puede observarse, se visualizan también casillas al lado de cada palabra, que, en algún caso, aparecen marcadas. Así, en la primera columna a la derecha de las palabras se marcan las palabras de la oración que forman parte del verbo. En la siguiente columna se marcan, por un lado, las palabras que constituyen el núcleo de los sintagmas ("bloques") para facilitar su recuperación en la búsqueda de restricciones de selección y también se marcan aquellos núcleos que se han usado de forma creativa en la oración (por ejemplo, de forma metafórica o metonímica) de manera que se puedan separar del resto de núcleos a la hora de realizar estudios o agrupaciones de palabras seleccionadas por el verbo.

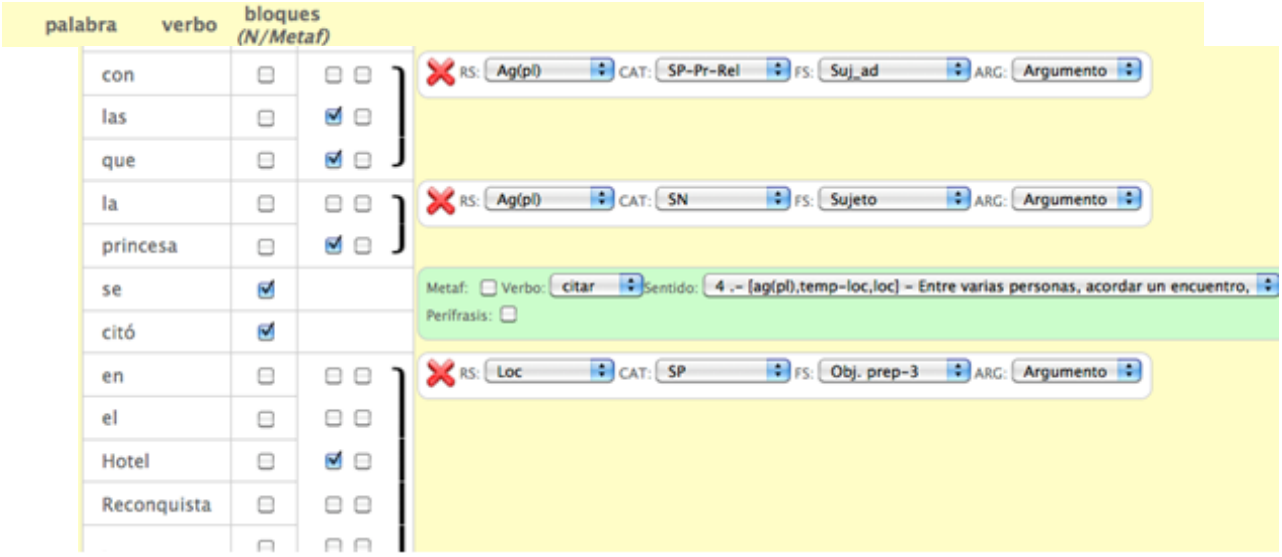

Figura 8: Anotación sintáctico-semántica de los complementos

Aunque no se hayan efectuado cambios en la anotación el paso final es siempre la grabación de los datos y la compleción de la ficha de control. En el caso del catalán, esta ficha contiene los mismos datos que hemos visto para el español, excepto la sección de la traducción, que está ausente. En dicha ficha se indica también el identificador de la frase en español ("Frase orig.") en formato hipertexto. Ello permite un rápido acceso a la oración equivalente en español muy útil en el caso de que durante el proceso de revisión de la frase catalana se detectaran errores en la española.

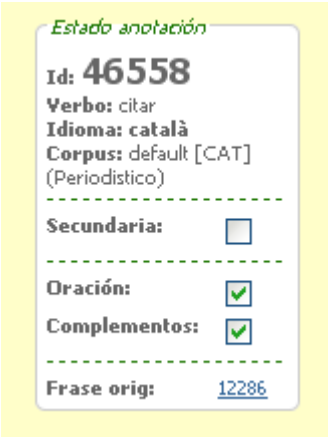

Figura 9: Cuadro de control de la oración del catalán

5. DESAJUSTES ENTRE EL ESPAÑOL Y EL CATALÁN

En esta sección presentamos algunos desajustes que se dan entre las lenguas con las que se ha trabajado y algunos problemas con los que nos hemos encontrado durante el proceso. Cabe señalar que, tal y como ya hemos señalado, dada la

proximidad estructural de ambas lenguas, la gran mayoría de dichos desajustes no presentan una dificultad especial. También es necesario indicar que, teniendo en cuenta el objetivo principal del proyecto y el tiempo del que se ha dispuesto para llevarlo a cabo, se ha optado por la solución más sencilla en cada caso y la menos costosa en términos de recursos. El escollo principal ha sido la paralelización de la información y la estructura de la base de datos en algunos casos.

Por cuestiones de eficiencia, la metodología de partida ha sido la duplicación de toda la base de datos. Ahora bien, dicha metodología, consistente en copiar la estructura del español exactamente igual para el catalán, ha conllevado algunos problemas. Por un lado, puede haber desajustes en el número de las palabras que se requieren en cada una de las lenguas (apartado 5.1), ya sea porque se usan menos palabras en español que en catalán o bien porque se da el caso inverso, es decir, un mayor número de palabras en catalán que en español. Por otro lado, también es necesario en algún caso cambiar el orden de palabras (apartado 5.2). Por último (apartado 5.3) se han detectado otros casos de interés desde el punto de vista contrastivo que también han ocasionado alguna dificultad derivada de la copia de la estructura de campos efectuada.

No obstante, la metodología adoptada ha sido mayormente beneficiosa para el proyecto, porque ha permitido crear de forma rápida y eficaz la estructura de datos sobre la cual trabajar, teniendo en cuenta que los problemas mencionados pueden ser resueltos con postedición automática casi siempre y que las diferencias sustanciales de estructura son muy escasas.

5.1. Desajustes en el número de palabras

Vamos a tratar en primer lugar los casos de creación de campos nuevos en catalán. Así, si una oración catalana contiene más palabras que la correspondiente del español, no se ha podido crear un espacio independiente para dicho elemento léxico, es decir, se ha tenido que incluir dicho elemento en el mismo campo que otro de los elementos léxicos de dicha frase. Este tipo de desajuste puede darse con dos casuísticas distintas: puede ser que una palabra del español requiera dos o más elementos léxicos en catalán, por lo que no ha supuesto ningún problema que todos los elementos del catalán aparecieran en el mismo campo, o puede ser que un concepto no explicitado en español requiera realización sintáctica en catalán, lo cual sí ha supuesto un problema importante en la estructura.

Un ejemplo en que un elemento del español se corresponde con 2 en catalán es el de los posesivos. A continuación, en la figura 10 puede verse gráficamente como queda implementado este desajuste ejemplificado con un ejemplo de posesivo. Como puede verse, se ha establecido la equivalencia incorporando ambos términos en un mismo campo, ya que conceptualmente equivalen al mismo elemento de la lengua origen. La categoría morfosintáctica continúa siendo la misma en ambas lenguas, por lo que la resolución de las correspondencias en estos

casos no es problemática. Esto mismo ocurre al traducir el pretérito indefinido del español al catalán[13], también ejemplificado en la misma oración.

Ahora bien, también puede ser que se tenga que crear uno o más campos en catalán sin una correspondencia en español. Ello supone un problema, ya que la palabra o las palabras del catalán tienen que incluirse en el campo correspondiente al equivalente de otra palabra del español, lo cual supone un error en la alineación que tendrá que corregirse posteriormente. Ello ocurre típicamente con el uso de algunos pronombres débiles en catalán. En la figura 11 puede verse cómo se ha implementado este tipo de desajuste al tener que introducir en la oración del catalán el pronombre en, sin ninguna correspondencia en la sintaxis del español. Como puede observarse se ha incluido dicho pronombre en el campo correspondiente al verbo. Sabemos que la solución no es la adecuada y que queda pendiente de postedición. Otra cuestión pendiente es la anotación de este elemento, que es además un complemento argumental en la oración.

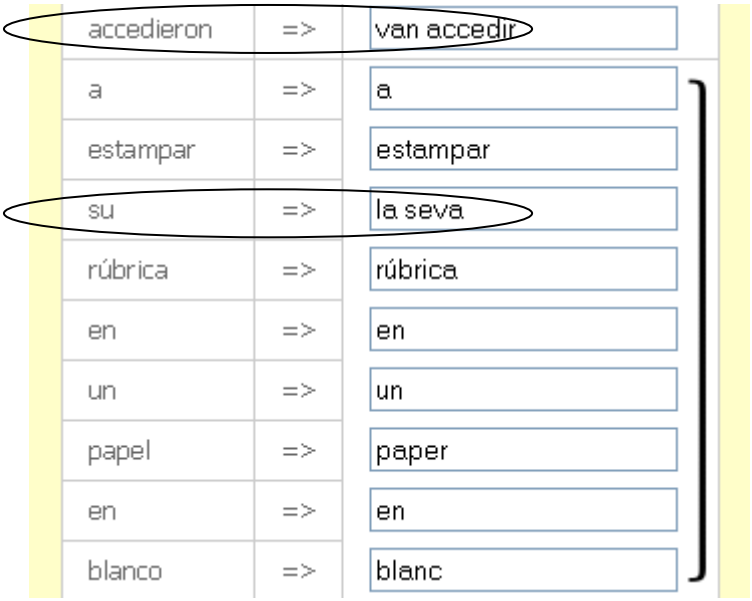

Figura 10: Ejemplos de desajustes entre español y catalán: 1→2

Por otro lado, como ya hemos avanzado, se da el caso de que se tengan que borrar campos en catalán (marcados con el uso de líneas discontinuas en las figuras). De nuevo se presentan dos casuísticas. En primer lugar, puede ser que dos palabras del español se correspondan con una en catalán. En la figura 11 vemos también un ejemplo de este fenómeno, ya que el conector *para que* del español se corresponde con un solo elemento en catalán: *perquè*. Otro ejemplo similar lo constituye el que se presenta en la figura 12, donde se observa como en

[13] En esta lengua este tiempo verbal se denomina pretérito perfecto.

catalán se contrae la preposición con el artículo plural, resultando, de este modo, una sola palabra (*als*) cuando en español había dos (*a los*).

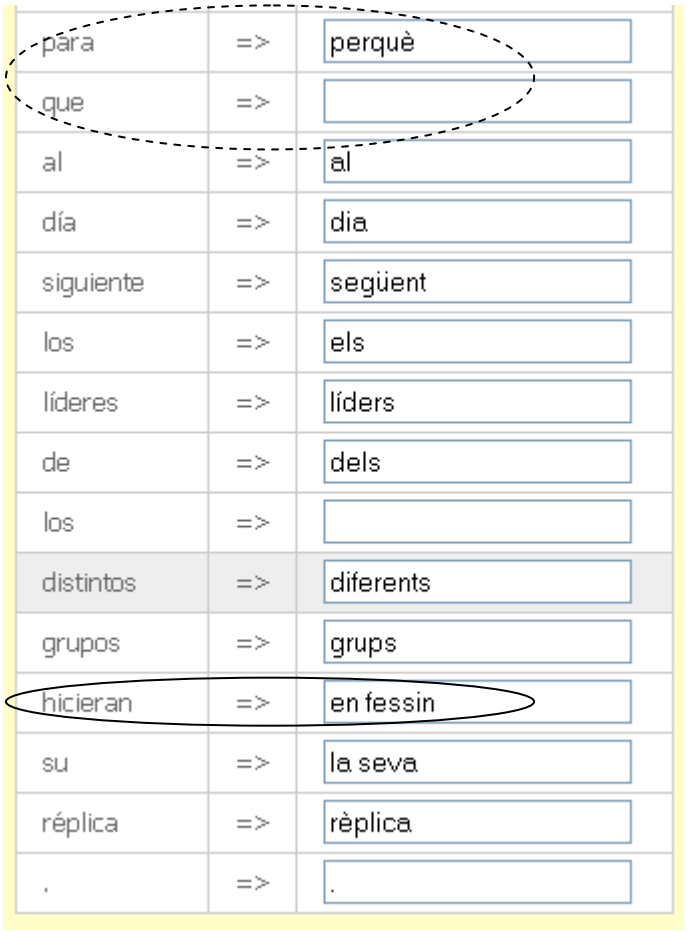

Figura 11: Otros ejemplos de desajuste entre español y catalán respecto al número de palabras.

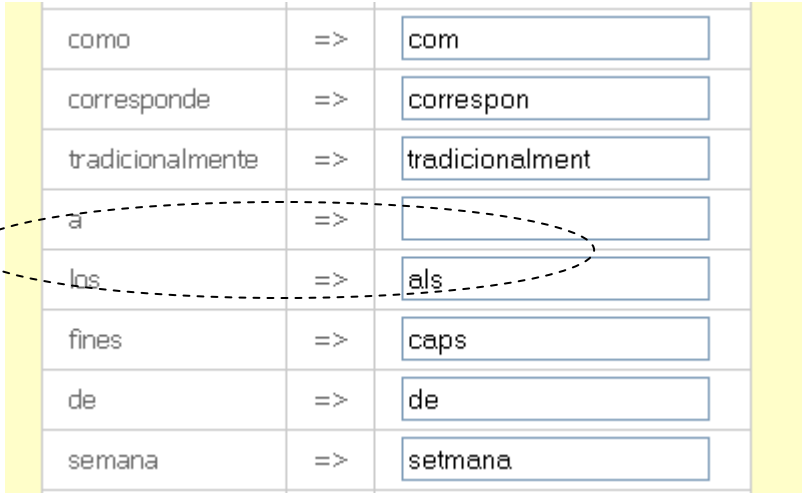

Figura 12: Ejemplo de desajuste entre español y catalán: 2→1.

Dado que la alineación es palabra a palabra en ambos casos queda un campo vacío en la estructura catalana, por lo que la alineación de palabras no queda del todo bien resuelta y también habrá que realizar postedición en estos casos. Desde el punto de vista estructural, el problema no afecta a la alineación, ya que las dos palabras pertenecen al mismo sintagma y, por tanto, se mantiene la equivalencia a este nivel de la alineación a través de la información sintáctica, categoría y función.[14]

En segundo lugar, puede ser que una palabra del español no tenga correspondencia en catalán; en este caso el problema es menos importante porque la alineación de elementos entre las dos lenguas se mantiene, ya que simplemente una palabra del español se corresponde con un espacio vacío. Evidentemente, tampoco supone ningún problema en la anotación morfosintáctica. Un ejemplo de ello es el caso del objeto directo de persona que en español se codifica con la preposición *a*, mientras que en catalán no, como se observa en la figura 13.

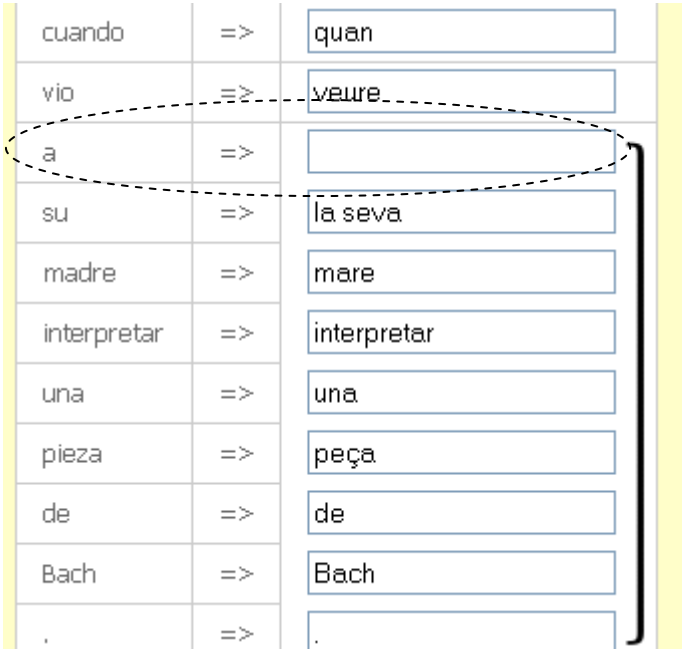

Figura 13: Ejemplo de desajuste entre español y catalán: 1→0.

5.2. Desajustes en el orden de palabras

El orden de palabras en la oración en español y catalán es muy coincidente. Ahora bien, por lo que se refiere al SN el catalán muestra mayor tendencia que el español

[14] Como ya se ha avanzado, no se anotan todas las palabras de una oración, sino sólo aquellas que quedan en el ámbito del verbo que se están analizando. En el caso de las palabras no tengan etiquetas asociadas la alineación a nivel de sintagmas no se lleva a cabo.

a colocar el adjetivo detrás del sustantivo y uno de los casos más recurrentes en este desajuste es con el adjetivo *siguiente* (figura 14).

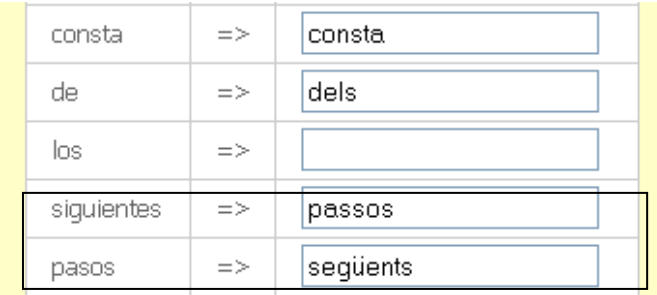

Figura 14: Desajustes del orden en el SN entre español y catalán.

En otras ocasiones, el cambio de orden es una consecuencia del paradigma pronominal del catalán, más rico que el del español. Ello provoca que determinados significados que se expresan en la lengua española con un adverbio se codifiquen en la catlana con un pronombre, lo cual implica un desajuste entre ambas lenguas en el interior del SV (figura 15), ya que los pronombres se sitúan normalmente delante del verbo y los adverbios detrás.

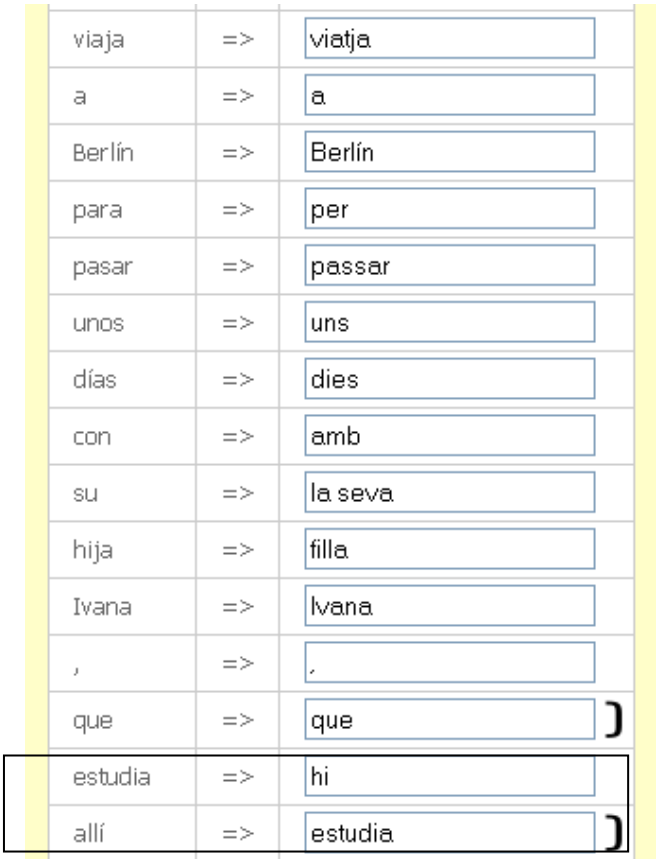

Figura 15: Desajustes de orden en el SV entre español y catalán.

5.3. Otros desajustes

A continuación vamos a presentar algunos ejemplos de cambios estructurales más profundos que los anteriores. Un ejemplo de ello es la traducción de algunas oraciones con posesivo en español, ya que su uso es más restringido en catalán que en español. En aquella lengua, estos elementos prácticamente quedan relegados a la posesión inalienable, o metafóricamente inalienable en algún caso, por lo que en una frase como la de la figura 16 el recurso que se usa en catalán es otro. Una de las posibles traducciones de esta frase implica el cambio del sustantivo del español (*participación*) por un verbo en catalán (*van participar* 'participaron'), lo cual implica a la vez un cambio en la anotación morfosintáctica y la pérdida del artículo en la frase catalana. No obstante, para la alineación ambas cuestiones se resuelven sin problemas.

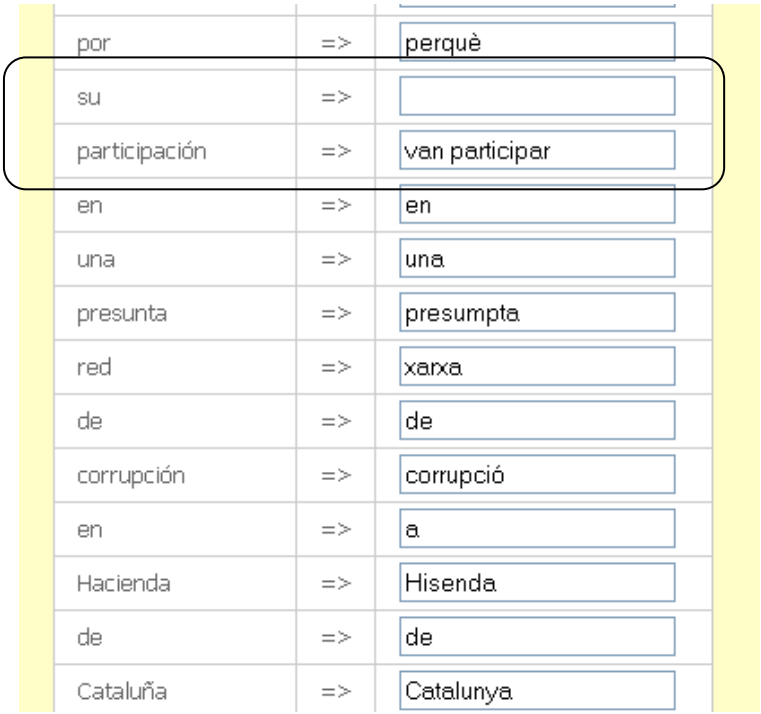

Figura 16: Cambios estructurales derivados del distinto uso del posesivo.

Otra de las diferencias entre el español y el catalán es el uso también restringido de pronombres fuertes en esta última. Así, en la oración de la figura 17 se utiliza un SN (*aquestes persones* 'estas personas') para traducir el pronombre *ellos*.

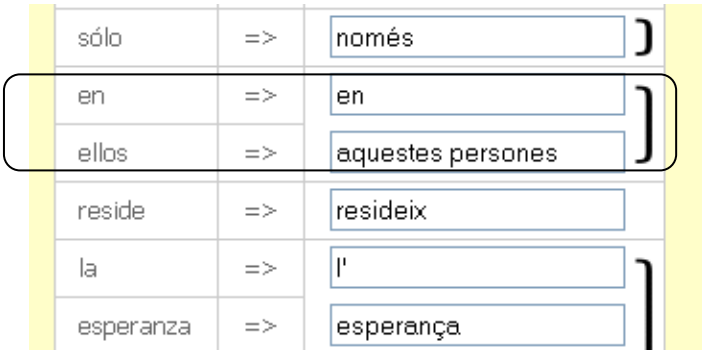

Figura 17. Cambios estructurales derivados del distinto uso del pronombre fuerte.

6. CONCLUSIONES Y LÍNEAS FUTURAS

En este artículo hemos presentado la paralelización del corpus SenSem del español con el correspondiente catalán. Cada una de las lenguas presenta una colección de textos de aproximadamente unas 700.000 palabras anotado a nivel de palabra (morfológico), de sintagma (con información sintáctico-semántica sobre los argumentos) y a nivel de oración (con información semántica referente a la semántica de la construcción, la aspectualidad, la modalidad y la polaridad).

Lo primero que nos gustaría destacar es la originalidad del recurso, por lo que se refiere al tipo de anotación consignada y sus posibles utilidades, ya que no existe ningún otro recurso para estas lenguas con las características aquí descritas. En consecuencia, su creación aporta interesantes ventajas tanto en el campo de la lingüística teórica como aplicada y contrastiva, por ejemplo, en el campo de la traducción, ya sea humana o automática, y también en el de la enseñanza de segundas lenguas.

A lo largo del artículo se ha descrito el proceso utilizado para la composición de dicho corpus y el tipo de herramientas informáticas utilizadas. Como parte del proyecto también se ha creado una base de datos léxica con los verbos tratados en el corpus El corpus se ha creado de forma semiautomática a partir de un corpus del español anotado manualmente.

La automatización ha consistido en la traducción de las frases del español al catalán con el traductor Google y en la conservación de la anotación del idioma original. Por lo que se refiere a las traducciones de las oraciones, se han corregido manualmente a través de una interfaz que permite mantener la paralelización de los textos. Dicha interfaz se ha rediseñado a partir de la ya utilizada en el proyecto del español. En cuanto a la anotación, en el caso de que se den diferencias estructurales entre ambas lenguas, se procede a modificar la información asociada al nivel de sintagma. Por lo que se refiere a la semántica de la oración, se preserva en todos los casos. Por último, la anotación morfosintáctica se ha realizado de forma aislada en ambas lenguas ya que los resultados que se obtienen en este

campo con las herramientas existentes en la actualidad son de gran calidad en ambos casos, tanto por lo que se refiere a la cobertura como a la precisión.

En cuanto al estado del proyecto, en el momento de redacción de este artículo se han traducido todos los sentidos verbales del proyecto español al catalán manualmente. Asimismo se ha efectuado la validación completa de la tarea de traducción de las formas verbales y la anotación y paralelización de oraciones aproximadamente hasta un 70%. En esta fase final del proyecto se realizará también la anotación morfosintáctica del catalán y las correcciones necesarias.

BIBLIOGRAFÍA

ATSERIAS, J., CASAS, B., COMELLES, E., GONZÁLEZ, M., PADRÓ, L. y PADRÓ, M. (2006), "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library", en *Proceedings of the fifth International Conference on Language Resources and Evaluation*, 48-55.

BAKER, M. (1993), "Corpus linguistics and translation studies: implications and applications", en Baker, M. Francis, G. y E. Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*, Amsterdam y Filadelfia, John Benjamins, 233-250.

CASTELLÓN, I., FERNÁNDEZ, A., VÁZQUEZ, G., ALONSO, L. y CAPILLA, J.C. (2006), "The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level", en *Fifth International Conference on Language Resources and Evaluation (LREC)*, 355-359.

FERNÁNDEZ, A., VÁZQUEZ, G. y CASTELLÓN, I. (2006A), "SenSem: a Databank for Spanish Verbs", en *Proceedings of the X Ibero-American Workshop on Artificial Intelligence, IBERAMIA*, Ribeirão Preto, Brasil.

FERNÁNDEZ, A., VÁZQUEZ, G. y TERUEL, D. (2006B), "Interfaz de explotación del corpus SenSem", en *Actas del Congreso de la Asociación Española de Lingüística Aplicada (AESLA)*, XXXX.

MURATA, M., Q. MA, K. UCHIMOTO, T. KANAMARU, H. ISAHARA (2006), "Japanese-to-English translations of tense, aspect, and modality using machine-learning methods and comparison with machine-translation systems on market", *Language Resources and Evaluation*, 40, 233-242.

PUSTEJOVSKY, J., P. HANKS, R. SAURÍ, A. SEE, R. GAIZAUSKAS, A. SETZER, D. RADEV, B. SUNDHEIM, D. DAY, L. FERRO y LAZO, M. (2003), "The TIMEBANK Corpus", en *Proceedings of Corpus Linguistics*, 647-656.

SAURÍ, R., M. VERHAGEN, y PUSTEJOVSKY, J. (2006), "Annotating and Recognizing Event Modality in Text", en *Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006*, Melbourne Beach, Florida.

SMITH, C. (1997), "The Parameter of Aspect", en *Studies in Linguistics & Philosopher: Volume 43*, Dordrecht, Kluwer Academic.

VÁZQUEZ, G. y FERNÁNDEZ, A. (2009), "Ampliación del Banco de Datos de Verbos del español SenSem", en Castillo Carballo, M.A y García Platero, J.M. (coords.), *La*

lexicografía en su dimensión teórica, Murcia, Publicaciones de la Universidad de Murcia, 957-969.

VILLALBA, X. (2006), "Una base de datos de construcciones en catalán y español", en *Actas del VII Congreso de Lingüística General* (CD- Rom. ISBN: 84-475-2086-8). Universitat de Barcelona. Disponible en: http://blogs.uab.cat/xaviervillalba/files/2008/11/comunicacio-clg-xaviervillalba.pdf. Acceso: 05.09.11

XIAO, R. y MCENERY, T. (2004), *Aspect in Mandarin Chinese: A corpus-based study*, Studies in Language Companion Series, 73, Amsterdam, John Benjamins P