Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 33-40 recibido 14-03-2016 revisado 19-04-2016 aceptado 03- 05-2016

Creación de un treebank de dependencias universales mediante recursos existentes para lenguas próximas: el caso del gallego

Building a UD treebank using existing resources from related languages: the case of Galician

Marcos Garcia
Grupo LyS, Dep. de Galego-Portugués, Francés e Lingüística
Universidade da Coruña
marcos.garcia.gonzalez@udc.gal

Carlos Gómez-Rodríguez
Grupo LyS
Dep. de Computación
Universidade da Coruña
carlos.gomez@udc.es

Miguel A. Alonso
Grupo LyS
Dep. de Computación
Universidade da Coruña
miguel.alonso@udc.es

Resumen: En este trabajo presentamos una nueva estrategia para crear *treebanks* de lenguas con pocos recursos para el análisis sintáctico. El método consiste en la adaptación y combinación de diferentes *treebanks* anotados con *dependencias universales* de variedades lingüísticas próximas, con el objetivo de entrenar un analizador sintáctico para la lengua elegida, en nuestro caso el gallego. Durante el proceso de selección y adaptación de los *treebanks* de origen, analizamos el impacto de propiedades de tres niveles diferentes: (i) la distancia entre las lenguas de origen y destino, (ii) la adaptación de características léxico-ortográficas, y (iii) las directrices de anotación entre los *treebanks*. Usando la estrategia propuesta, entrenamos un analizador sintáctico estadístico para etiquetar, con resultados prometedores y sin datos previos de gallego, un pequeño corpus de esta lengua. La corrección manual de este corpus, usado como *gold-standard*, nos permitió probar la eficacia del método propuesto.
Palabras clave: análisis sintáctico, *treebank*, *dependencias universales*, gallego

Abstract: This paper presents a novel strategy for creating a Universal Dependencies (UD) treebank of a low-resource language. The method consists of adapting and combining different UD treebanks from related varieties in order to train a parser for the target language. More precisely, the paper explores the influence of three different levels for the selection and adaptation of the source treebanks: (i) the relatedness of the linguistic varieties, (ii) the adaptation of features based on lexical and spelling data, and (iii) the agreement in annotation criteria between different treebanks. The proposed strategy allowed us to train a parser for analyzing, with promising results, a small Galician corpus without previous availability of labeled data for this language. After a few bootstrapping iterations, we obtained a UD gold-standard corpus, used for proving the effectiveness of the proposed method.
Keywords: parsing, treebank, universal dependencies, Galician

1 Introducción

El uso de corpus anotados sintácticamente (*treebanks*) se ha demostrado útil en diferentes áreas, como los estudios en lingüística de corpus o trabajos de análisis sintáctico automático (*parsing*), que es a su vez beneficioso para tareas como la minería de opiniones, o la traducción automática, entre otras (Socher et al., 2013; Gimpel y Smith, 2014). Con todo, la creación de este tipo de recursos es una tarea costosa, ya que implica la etiquetación manual de una gran cantidad de información lingüística de diferentes niveles. El proceso de anotación sintáctica se puede aliviar mediante la aplicación previa de un analizador automático, corrigiendo así únicamente los errores producidos por este sistema. En lenguas para las que no existen este tipo de herramientas, se han propuesto diferentes estrategias que aprovechan recursos de otros idiomas para entrenar *parsers* estadísticos. Entre estas técnicas encontramos el uso de corpus paralelos de las lenguas de origen y destino (Zeman y Resnik, 2008), a veces enriqueciendo el *parser* con reglas específicas del

* Este trabajo ha sido parcialmente financiado por el MINECO (proyectos FFI2014-51978-C2-1-R y FFI2014-51978-C2-2-R, y un contrato *Juan de la Cierva formación*: FJCI-2014-22853), y por la Xunta de Galicia (programa *Oportunius*).

idioma de destino (Hwa et al., 2005). Sin embargo, tanto diferencias lingüísticas (u otras divergencias de anotación entre los corpus) como la escasez de este tipo de recursos pueden dificultar este proceso.

En un intento de homogeneizar —en la medida de lo posible— las directrices de anotación sintáctica, el proyecto *Universal Dependencies* (UD) promueve una anotación consistente de los diferentes *treebanks* de las lenguas naturales (McDonald et al., 2013). Así, utilizando un conjunto universal de dependencias sintácticas (aunque permitiendo etiquetas diferentes para anotar fenómenos específicos de algunas lenguas), UD facilita, por ejemplo, el aprovechamiento de recursos entre varias lenguas o el análisis interlingüístico de fenómenos sintácticos.

Con el objetivo de crear un corpus con anotación sintáctica UD para gallego, en el presente trabajo proponemos una estrategia de combinación y adaptación de *treebanks* de variedades lingüísticas próximas, que permiten una anotación inicial de alta calidad. En los procesos de selección y adaptación de los *treebanks* de origen, se tienen en cuenta características de tres niveles (en relación al idioma de destino): (i) proximidad lingüística, (ii) distancia léxico-ortográfica y (iii) particularidades de anotación interlingüística.

La estrategia aquí propuesta, evaluada en $\approx$ 12.000 *tokens* corregidos manualmente, obtiene resultados prometedores en lo que respecta al aprovechamiento de recursos de lenguas próximas para la creación de un nuevo *treebank* UD, y muestra que tanto la proximidad lingüística (sintáctica y léxica) como las variaciones de anotación son relevantes en el proceso de transferencia.

Además de esta sección introductoria, el artículo se organiza de la siguiente manera. La sección 2 incluye una revisión del trabajo relacionado, mientras que la sección 3 presenta las principales características del proyecto UD y de la adaptación del corpus gallego a este proyecto. En las secciones 4 y 5 presentamos y evaluamos, respectivamente, el método de transferencia propuesto. Finalmente, la sección 6 contiene las conclusiones del estudio, así como ideas para el trabajo futuro.

2 Trabajo Relacionado

Diversos trabajos han analizado el uso de recursos sintácticos de una o más lenguas para crear un *treebank* de un idioma diferente, con resultados dispares. Así, antes de la existencia de las UD, varios trabajos utilizaron corpus paralelos para proyectar la anotación sintáctica de una lengua origen (con recursos) a la lengua de destino (Hwa et al., 2005; Ganchev, Gillenwater, y Taskar, 2009).

En Zeman y Resnik (2008) se entrena un *parser* únicamente con información sintáctica y morfosintáctica de la lengua de origen (*parser* deslexicalizado), para analizar posteriormente textos en la lengua de destino. La deslexicalización obtiene mejores resultados que el uso de información léxica en el par de lenguas evaluado (sueco–danés). Trabajos posteriores mejoraron esta técnica al combinarla con el uso de corpus paralelos y comparables, añadiendo también más de un idioma al conjunto de *treebanks* de origen (Søgaard, 2011; McDonald, Petrov, y Hall, 2011).

Utilizando las UD, McDonald et al. (2013) también evalúan el rendimiento de *parsers* entrenados para un idioma diferente del que posteriormente analizan. La estrategia de deslexicalización —con corpus paralelos— proporciona mejoras en el análisis, y la transferencia entre lenguas próximas obtiene mejores resultados que la realizada entre variedades lingüísticamente más distantes.

Con todo, las evaluaciones de Lynn et al. (2014) (para el irlandés), o de Vilares, Alonso, y Gómez-Rodríguez (2016) (donde se entrenan y evalúan varios *parsers* bilingües) sugieren que el resultado de la transferencia de recursos sintácticos entre idiomas no tiene por qué estar relacionado con la proximidad lingüística entre ellos (entendiendo la proximidad en términos de pertenencia —o no— a la misma familia lingüística).

En lo que respecta a *treebanks* de gallego, no conocemos hasta este momento ningún corpus disponible con anotación sintáctica, si bien durante el desarrollo de este trabajo la página web del proyecto UD informó sobre un *treebank* en desarrollo, que estará disponible a partir de la versión 1.3.[1]

Los trabajos sobre *parsing* para gallego tampoco son muy abundantes, aunque existen varios artículos que implementan reglas sintácticas en analizadores automáticos. Así, Gamallo Otero y González López (2011) presentan una *suite* multilingüe de análisis de dependencias que incluye un *parser* de gallego.

[1] `http://universaldependencies.org`

Por su parte, las versiones más recientes de FreeLing también disponen de un *parser* para gallego, que realiza análisis tanto de constituyentes como de dependencias sintácticas (Padró y Stanilovsky, 2012). Las dependencias utilizadas por ambos sistemas (DepPattern y FreeLing) no son UD, por lo que su utilización en el presente trabajo supondría un proceso de adaptación mayor. Además, la inexistencia de *treebanks* tampoco facilita la realización de evaluaciones empíricas de los diferentes analizadores.

Finalmente, existen algunos trabajos que —como el actual— han aprovechado la proximidad lingüística entre portugués y gallego para generar recursos de este último a partir del primero: entre otros, Malvar et al. (2010) obtienen corpus bilingües para entrenar modelos de traducción automática, mientras que Garcia y González (2012) generan, para un sistema de transcripción fonética automática, léxicos de gallego utilizando léxicos de portugués europeo.

En este trabajo analizamos el uso de recursos sintácticos de UD en español y portugués (entre otras lenguas) para el análisis de un corpus gallego, estudiando también el impacto de las características léxico-ortográficas y de anotación entre los diferentes *treebanks* de origen y destino.

3 Dependencias Universales y Corpus Gallego

McDonald et al. (2013) fueron los primeros en utilizar, en varios corpus, el conjunto de *dependencias sintácticas universales*, publicando *treebanks* de 6 lenguas diferentes. En el origen de este conjunto de dependencias está, por un lado, una versión de las etiquetas sintácticas del *parser* de inglés del NLP Group de la universidad de Stanford (De Marneffe y Manning, 2008) y, por otro lado, el conjunto de etiquetas morfosintácticas universales propuestas por Google (Petrov, Das, y McDonald, 2012). Así, el proyecto UD tiene entre sus objetivos facilitar tanto el desarrollo de analizadores multilingües y el aprovechamiento mutuo de recursos de diferentes lenguas, como el estudio interlingüístico de fenómenos sintácticos.

Como hemos referido, UD promueve una anotación (no sólo sintáctica, sino también morfosintáctica y de *tokenización*) consistente entre *treebanks* de diferentes lenguas, mediante el uso de un conjunto universal de etiquetas y unas directrices de anotación homogéneas. Sin embargo, teniendo en cuenta que existen fenómenos lingüísticos particulares, cada *treebank* puede utilizar variantes propias de las *dependencias universales* para anotar este tipo de fenómenos.

A este respecto, durante la actual etapa preliminar de etiquetación estamos definiendo unas directrices propias que, siguiendo las recomendaciones UD, nos permitan analizar satisfactoriamente los fenómenos lingüísticos específicos del gallego. Estas directrices, en su versión inicial sujeta a posibles revisiones o ampliaciones en el futuro, se basan en tres pilares básicos:

1. Utilización —siempre que sea posible— de los principios de UD
2. Uso del menor número posible de dependencias y directrices de anotación diferentes de las etiquetas universales
3. Coherencia (si es posible) con la anotación sintáctica del *treebank* portugués, en aquellos casos en los que UD permita varias soluciones de anotación

En este sentido, la principal divergencia de anotación con respecto a las directrices UD ha sido la utilización de la etiqueta *iobj* (objeto indirecto) en aquellos casos en los que el objeto directo (*dobj*) no está explícito (en estas situaciones, UD recomienda etiquetar el *iobj* como *dobj*). Esta decisión ha sido tomada porque la discriminación de estas etiquetas favorece tanto el análisis lingüístico como la extracción de información del *treebank*, dado que en el corpus gallego la preposición que introduce el objeto (normalmente *a*) aparece tanto en *dobj* como en *iobj*. La Figura 1 contiene un ejemplo de una oración del corpus gallego con dependencias UD (cuya traducción al español podría ser "La competencia le corresponderá a la RAG"), en donde se puede observar la anotación del único objeto como *iobj*, y del pronombre clítico como *expl*.

El corpus elegido para iniciar el proceso de construcción del *treebank* gallego fue el XIADA 2.6 (Rojo et al., 2015), un recurso con más de 740.000 *tokens* lematizados y con anotación morfosintáctica corregida manualmente. XIADA se compone de textos de dominio periodístico, económico y narrativo en gallego. Durante la adaptación de este corpus hemos mantenido algunas particularida-

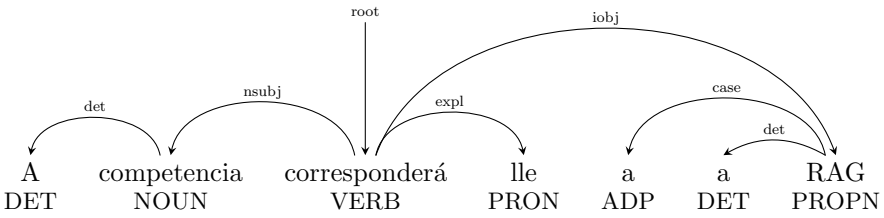

Figura 1: Oración del corpus gallego con anotación (sintáctica y morfosintáctica) UD.

des que, desde el punto de vista del proyecto *Universal Dependencies*, cabe mencionar:

***Tokenización*:** con el objetivo de preservar la *tokenización* de XIADA se ha mantenido la división original del corpus. Así, tanto los nombres propios compuestos (de más de un *token*) como algunas locuciones son etiquetadas como elementos individuales, y no separadas en *tokens* como recomienda UD.

Anotación morfosintáctica: UD usa un *tagset* universal de 17 etiquetas que incluyen las categorías morfosintácticas básicas (adjetivo, adverbio, verbo, etc.), codificando —si existe— otra información de carácter morfosintáctico (género, número, etc.) como características independientes de las categorías.

En nuestro caso, hemos extraído de las etiquetas XIADA tanto la categoría UD como las restantes características morfosintácticas, manteniendo en el corpus la información original y la extraída automáticamente.

En general, la anotación morfosintáctica del corpus no se ha modificado durante la corrección del *treebank*, salvo en aquellos casos en que se han detectado errores inequívocos de anotación.

4 Selección y Adaptación de treebanks *de origen*

La disponibilidad de *treebanks* en gallego es necesaria tanto para diferentes tareas del procesamiento computacional de esta lengua como para realizar estudios interlingüísticos con otros *treebanks* con los que se comparta anotación. Así, ante la inexistencia de recursos ya etiquetados, hemos optado por estudiar diferentes métodos para la transferencia de *parsers* desde otros idiomas.

La estrategia propuesta en este trabajo enfoca la transferencia de analizadores sintácticos de una o más lenguas de origen (que dispongan de *treebanks*) a una lengua de destino, con base en tres parámetros:

1. Proximidad lingüística —especialmente sintáctica— entre las variedades de origen y destino
2. Distancia léxico-ortográfica entre los corpus
3. Variación en las directrices de anotación (dentro del conjunto de etiquetas UD)

Proximidad lingüística: como hemos visto, las evaluaciones de diferentes trabajos no confirman firmemente que la distancia lingüística sea un factor decisivo en la transferencia de *parsers* de un idioma a otro. Con todo, diversas evaluaciones aquí realizadas (con *treebanks* UD de idiomas de diferentes familias lingüísticas) nos sugieren que algunas lenguas se pueden analizar con resultados aceptables utilizando recursos de variedades muy próximas desde el punto de vista sintáctico y léxico (véase la sección 5).

Así, con el objetivo de analizar el corpus XIADA, seleccionamos (después de evaluaciones preliminares) los *treebanks* UD de español y portugués europeo como origen. La elección de estas variedades se debe, por un lado, a que ambas lenguas tienen estructuras sintácticas muy similares a las de gallego. Además, el español coexiste con el gallego en el mismo territorio, y las interferencias sintácticas —y otras— son frecuentes entre las dos lenguas (también en el corpus XIADA). El uso del portugués como lengua origen está basado en el hecho de que tanto el gallego como el portugués provienen del mismo sistema lingüístico (*galego-português*), siendo considerados por algunos lingüistas todavía en la actualidad como variedades del mismo idioma (Cintra y Cunha, 1984).[2]

Distancia léxico-ortográfica: las diferencias léxicas entre varios idiomas propiciaron el uso de estrategias de deslexicalización,

[2]Sea como fuere, entre los dos estándares existen diferencias cuyo impacto en la transferencia tratamos de reducir usando métodos de adaptación ortográfica.

diseñadas con el objetivo de minimizar el impacto negativo al analizar idiomas con mayor distancia léxica. A este respecto, este trabajo propone como una de las estrategias de adaptación de *treebanks*, la transliteración ortográfica del corpus portugués. Para ello, hemos construido automáticamente una versión del *treebank* portugués con ortografía muy próxima a la del estándar gallego, usando la estrategia adoptada en Malvar et al. (2010).

A pesar de que este método solo es aplicable entre variedades lingüísticas muy próximas, técnicas similares (con base en diccionarios bilingües o en similitud léxica) se podrían evaluar en otros pares de lenguas.

Directrices de anotación: el proyecto UD promueve unas directrices estándar de anotación para las diferentes lenguas, pero los *treebanks* individuales pueden tener características de etiquetación propias, no sólo por el uso de dependencias específicas de un idioma, sino por decisiones particulares de los anotadores (recuérdese, por ejemplo, nuestra decisión de priorizar el uso de *iobj* sobre *dobj*, explicada en la sección 3).

Hasta el momento, el principal cambio relativo a las directrices de anotación que hemos realizado durante la adaptación de los corpus español y portugués ha sido el uso de la dependencia *expl* (expletivo). Para fortalecer la coherencia entre los *treebanks* de origen (que no utilizan la dependencia *expl*) y de destino (que sí la utiliza, de acuerdo con las directrices UD), hemos evaluado el impacto de una transformación automática de los pronombres reflexivos en español y portugués (anotados originariamente como *dobj* o *iobj*) a *expl* (véase el ejemplo de la Figura 1). Otras sustituciones automáticas, como la anotación de algunos pronombres clíticos, determinados usos de la dependencia *case* en el inicio de oraciones subordinadas, o la anotación de expresiones multipalabra están siendo estudiadas para futuros procesos de adaptación.

5 Experimentos

En la presente sección explicamos sucintamente el proceso de corrección de la versión actual del *treebank* de gallego (usado como *gold-standard*), y también evaluamos y discutimos diferentes métodos de transferencia.[3]

[3]Todos los recursos utilizados durante las evaluaciones se pueden obtener en la siguiente dirección `http://grupolys.org/~marcos/pub/sepln16.zip`

Los *parsers* utilizados durante los diferentes experimentos fueron creados con base en los conjuntos de entrenamiento de los *treebanks* de la versión más reciente del proyecto *Universal Dependencies* (1.2). Así mismo, todos los analizadores fueron entrenados con MaltParser (1.8), con la configuración por defecto (dejando, por lo tanto, margen para optimización). Todos los resultados incluyen tanto valores LAS (*Labeled attachment score*) como UAS (*Unlabeled attachment score*).

5.1 *Bootstrapping*

Para evaluar los métodos referidos hemos iniciado la anotación sintáctica del corpus XIADA del siguiente modo: los primeros ≈ 1.000 *tokens* (el número exacto varía en función de la frontera de oración) del subcorpus "xeral" (con 198.231 *tokens* de dominio periodístico general) fueron analizados con un modelo de MaltParser (Nivre et al., 2007) entrenado en una combinación de los *treebanks* UD de portugués (201.845 tokens) y español (382.436 tokens), que fue seleccionado por obtener los mejores resultados en una evaluación subjetiva (al no disponer todavía de datos anotados para la evaluación).

La anotación automática de estos ≈ 1.000 *tokens* fue corregida manualmente por uno de los autores de este trabajo utilizando la herramienta DepAnnotator (Ribeyre, 2015). Una vez finalizada la corrección, aplicamos una estrategia de *bootstrapping* para entrenar un nuevo modelo con los *treebanks* español, portugués, y las oraciones gallegas corregidas. Este proceso se repitió cada ≈ 1.000 *tokens*, hasta llegar a los 12.054 (500 oraciones corregidas), utilizando el corpus resultante como *gold-standard* para evaluar la estrategia propuesta.

5.2 Evaluación

En primer lugar, utilizamos el *gold-standard* de gallego para conocer cómo la distancia lingüística puede influir en el análisis sintáctico de una lengua diferente. La Tabla 1a contiene los resultados de aplicar directamente al gallego *parsers* entrenados sobre los *treebanks* UD de idiomas con diferente grado de distancia lingüística (sueco, inglés, francés, italiano, español y portugués). Durante el proceso de aprendizaje, se utilizó también una variante deslexicalizada (entrenada con corpus sin tokens ni lemas, únicamente con información sintáctica y morfosintáctica) de cada uno de

los *treebanks*, con el objetivo de conocer el impacto de las características léxicas en función de la distancia lingüística.

Los resultados indican que, para el análisis sintáctico del gallego, la distancia lingüística del *treebank* de origen es un factor importante (con diferencias de más de 12 % entre sueco y portugués, por ejemplo).

En relación al impacto de la información léxica en los resultados del *parsing*, los valores obtenidos en las diferentes lenguas parecen indicar que el proceso de deslexicalización es más efectivo en idiomas distanciados léxicamente de la lengua de destino (con los que, por lo tanto, comparten un menor número de palabras). Así, los modelos 'delex' de sueco e inglés obtienen mejores resultados que sus variantes lexicalizadas (entre $\approx 1\,\%$ y $\approx 2\,\%$, en función de la lengua y tipo de evaluación), mientras que en francés e italiano la mejora no es tan clara. Por último, en español y portugués (variedades más próximas al gallego), los modelos con información léxica obtienen sistemáticamente mejores resultados.

Una vez observado el impacto de la distancia lingüística (tanto sintáctica como léxica) en el proceso de transferencia, el siguiente conjunto de evaluaciones analizó (i) combinaciones de los mejores modelos individuales, (ii) la adaptación de las características léxicas —a través de la transliteración del *treebank* portugués— y (iii) la unificación de determinadas directrices de anotación entre *treebanks*.

Así, se han evaluado combinaciones lexicalizadas y deslexicalizadas de español y portugués ('es+pt'), modelos transliterados de portugués a gallego ('pt2')[4] y modelos (tanto de español como de portugués transliterado) en cuyos *treebanks* se han anotado automáticamente los pronombres reflexivos como expletivos ('expl'). Los resultados de estos experimentos se pueden ver en los diferentes bloques de la Tabla 1b.

Los valores de las combinaciones de español y portugués (tanto la variante completa como la deslexicalizada) son ligeramente superiores a los que habíamos obtenido únicamente con los modelos 'pt' y 'pt-delex', lo que sugiere que las combinaciones de recursos complementarios pueden mejorar el análisis de una lengua diferente.

En relación a la adaptación léxico-ortográfica, los resultados del modelo 'pt2' superan en casi 2 % los obtenidos por el *parser* 'pt', por lo que esta estrategia se muestra una vez más efectiva en la adaptación de recursos entre portugués y gallego.

Así mismo la adición del *treebank* español al modelo 'pt2' ('es+pt2') mejora el rendimiento de la transferencia en cerca de 2 % con relación al modelo 'pt2', y en más de 4 % (LAS) en relación al *parser* de portugués.

El último de los niveles definidos (las divergencias entre las directrices de anotación de diferentes *treebanks*) se ha evaluado a través de los modelos 'expl'. A pesar de tratarse de una conversión simple (no se han convertido todos los pronombres reflexivos y expletivos sino únicamente los anotados como "Reflex=Yes" en los corpus de origen), los resultados, tanto en modelos individuales de español y portugués como en la combinación 'es+pt2', sugieren que este tipo de adaptaciones pueden ser útiles durante el proceso de aprendizaje. A este respecto, salvo en el valor LAS del *parser* 'pt2' (con resultados $< 0,01\,\%$), las variantes 'expl' obtienen mejores resultados que aquellos que utilizan la anotación original de los *treebanks* español y portugués.

Así, los diferentes experimentos aquí presentados, realizados en función de los tres parámetros definidos en la sección 4, muestran que para el análisis sintáctico del gallego, la selección de variedades lingüísticas próximas es un factor decisivo en el rendimiento de un *parser* transferido.

Además, la adaptación ortográfica (o léxico-ortográfica, ya que la transliteración modifica directamente las palabras del corpus de origen para adaptarlas a la ortografía de la lengua de destino), es útil para aprovechar recursos sintácticos de portugués en el procesamiento del gallego.

Sobre la uniformización de ciertas variantes de anotación (incluso utilizando un mismo *tagset*, como UD), los experimentos realizados también sugieren que criterios de etiquetación más homogéneos entre las lenguas de origen y destino permiten entrenar *parsers* más precisos.

En suma, la combinación de los diferentes métodos presentados nos permite realizar un análisis sintáctico inicial de un corpus gallego con resultados competitivos con relación al *parsing* de otras lenguas con un mayor número de recursos, por lo que estamos ante un

[4] La transliteración fue realizada con *port2gal*: `http://gramatica.usc.es/~gamallo/port2gal.htm`

Modelo	LAS	UAS
sv	56,39	66,48
sv-delex	**58,24**	**67,92**
en	59,77	68,18
en-delex	**60,84**	**69,52**
fr	66,75	**75,18**
fr-delex	**67,28**	74,45
it	**69,13**	76,54
it-delex	68,98	**77,79**
es	**69,96**	**78,71**
es-delex	69,30	77,59
pt	**71,33**	**79,20**
pt-delex	69,70	76,59

(a) Resultados de modelos individuales (sueco: sv; inglés: en; francés: fr; italiano: it; español: es, y portugués: pt).

Modelo	LAS	UAS
es+pt	**74,21**	**81,65**
es+pt-delex	70,13	77,69
pt2	73,09	80,43
es+pt2	**75,45**	**81,98**
es_expl	70,92	78,82
pt2_expl	73,08	80,51
es_expl+pt2_expl	**75,85**	**82,03**

(b) Resultados de los mejores modelos combinados (líneas superiores) y modelos adaptados: portugués transliterado ('pt2') y español y portugués con conversión automática de la dependencia *expletivo* ('expl').

Tabla 1: Resultados de diferentes *parsers* lexicalizados y deslexicalizados (delex) evaluados sobre el corpus de gallego.

buen punto de partida para la ampliación de un *treebank* para esta lengua.

6 Conclusiones y Trabajo Futuro

En este trabajo hemos presentado una estrategia de combinación y adaptación de *treebanks* de lenguas próximas para el análisis sintáctico de un idioma que, hasta el momento, no disponía de *treebanks* publicados.

El método consiste en combinar recursos de idiomas similares, etiquetados con dependencias universales, y reducir las divergencias tanto léxico-ortográficas como de anotación, para incrementar la precisión de análisis en la lengua de destino.

La etiquetación de un *gold-standard* en gallego, disponible libremente, nos ha permitido probar la eficacia del método propuesto, que no necesita procesos de deslexicalización para transferir analizadores sintácticos de las lenguas origen a la lengua de destino.

Actualmente nos encontramos en proceso de ampliación y corrección del *treebank* inicial presentado en este trabajo, al mismo tiempo que revisamos las directrices de anotación. Un *treebank* de mayor tamaño (así como la publicación de otros recursos UD para gallego) nos permitirá evaluar el impacto de añadir datos propios de gallego a mejores modelos de transferencia.

Así mismo creemos necesario estudiar otras estrategias de adaptación, a través de un análisis más detallado, de recursos de otras lenguas con el fin de aumentar la precisión en los procesos de transferencia. Entre estas estrategias podría estar la realización de un mapeado de las dependencias sintácticas específicas de diferentes idiomas, o el tratamiento homogéneo de estructuras como perífrases verbales, entre otras.

Bibliografía

Cintra, L. F. L. y C. Cunha. 1984. *Nova gramática do português contemporâneo*. Sá da Costa, Lisboa.

De Marneffe, M.-C. y C. D. Manning. 2008. The Stanford typed dependencies representation. En *COLING 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, páginas 1–8, Manchester. ACL.

Gamallo Otero, P. y I. González López. 2011. A grammatical formalism based on patterns of Part of Speech tags. *International Journal of Corpus Linguistics*, 16(1):45–71.

Ganchev, K., J. Gillenwater, y B. Taskar. 2009. Dependency grammar induction via bitext projection constraints. En *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volumen 1, páginas 369–377, Singapur. ACL.

Garcia, M. y I. J. González. 2012. Automatic Phonetic Transcription by Phonologi-

cal Derivation. En H. Caseli A. Villavicencio A. Teixeira, y F. Perdigão, editores, *Computational Processing of the Portuguese Language (PROPOR 2012)*, volumen 7243 de *Lecture Notes in Artificial Intelligence*. Springer, Coimbra, páginas 350–361.

Gimpel, K. y N. A. Smith. 2014. Phrase Dependency Machine Translation with Quasi-Synchronous Tree-to-Tree Features. *Computational Linguistics*, 40(2):349–401.

Hwa, R., P. Resnik, A. Weinberg, C. Cabezas, y O. Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(03):311–325.

Lynn, T., J. Foster, M. Dras, L. Tounsi, y others. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. En *Proceedings of the First Celtic Language Technology Workshop*, páginas 41–49, Dublin. ACL.

Malvar, P., J. R. Pichel, Ó. Senra, P. Gamallo, y A. García. 2010. Vencendo a escassez de recursos computacionais. Carvalho: Tradutor Automático Estatístico Inglês-Galego a partir do corpus paralelo Europarl Inglês-Português. *Linguamática*, 2(2):31–38.

McDonald, R., S. Petrov, y K. Hall. 2011. Multi-source transfer of delexicalized dependency parsers. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, páginas 62–72, Edimburgo. ACL.

McDonald, R. T., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. B. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló, y J. Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. En *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, páginas 92–97, Sofia. Association for Computational Linguistics.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, y E. Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(02):95–135.

Padró, L. y E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the 8th edition of the Language Resources and Evaluation Conference (LREC 2012)*, Estambul. ELRA.

Petrov, S., D. Das, y R. McDonald. 2012. A Universal Part-of-Speech Tagset. En *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Estambul. ELRA.

Ribeyre, C. 2015. *Méthodes d'Analyse Supervisée pour l'Interface Syntaxe-Sémantique*. Ph.D. tesis, Université Paris 7 Diderot.

Rojo, G., M. L. Martínez, E. D. Noya, y F. M. Barcala. 2015. Corpus de adestramento do Etiquetador/Lematizador do Galego Actual (XIADA), versión 2.6. `http://corpus.cirp.es/xiada/corpus_xiada_2_6.tar.gz`.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, y C. Potts. 2013. Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, páginas 1631–1642, Seattle. ACL.

Søgaard, A. 2011. Data point selection for cross-language adaptation of dependency parsers. En *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers (ACL HLT 2011)*, volumen 22, páginas 682–686, Portland. ACL.

Vilares, D., M. A. Alonso, y C. Gómez-Rodríguez. 2016. One model, two languages: training bilingual parsers with harmonized treebanks. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin. ACL.

Zeman, D. y P. Resnik. 2008. Cross-Language Parser Adaptation between Related Languages. En *Proceedings of the Workshop on NLP for Less Privileged Language at the 3rd International Joint Conference on Natural Language Processing (IJCNLP 2008)*, páginas 35–42, Hyderabad. Asian Federation of Natural Language Processing.