Universal Dependencies for the AnCora treebanks

Hector Martinez Alonso, Daniel Zeman

Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 91-98 recibido 29-03-2016 revisado 20-04-2016 aceptado 02-02-2016

Universal Dependencies for the AnCora treebanks

Dependencias Universales para los treebanks AnCora

Héctor Martínez Alonso
Univ. Paris Diderot
Sorbonne Paris Cité,
Alpage (INRIA), France
hector.martinez-alonso@inria.fr

Daniel Zeman
Charles University in Prague
Faculty of Mathematics and Physics
Czech Republic
zeman@ufal.mff.cuni.cz

Resumen: Este artículo presenta la conversión de los treebanks AnCora del catalán y el castellano al formalismo de Dependencias Universales (UD). Describimos el proceso de conversión y estimamos la calidad de los treebanks resultantes en términos de sus resultados en análisis sintáctico automático en un esquema monolingüe, en un esquema trans-lingüístico y en un tercero trans-dominio. Los treebanks convertidos muestran un nivel de consistencia interna de anotación comparable a la de los datos originales de la distribución CoNLL09 de AnCora, e indican algunas diferencias en terminos del inventario de expresiones polilexemáticas con respecto al anterior treebank del castellano en UD. Los dos nuevos treebanks convertidos serán distribuidos con la versión 1.3 de Dependencias Universales.
Palabras clave: AnCora, treebank, catalán, castellano, Dependencias Universales

Abstract: The present article describes the conversion of the Catalan and Spanish AnCora treebanks to the Universal Dependencies formalism. We describe the conversion process and assess the quality of the resulting treebank in terms of parsing accuracy by means of monolingual, cross-lingual and cross-domain parsing evaluation. The converted treebanks show an internal consistency comparable to the one shown by the original CoNLL09 distribution of AnCora, and indicate some differences in terms of multiword expression inventory with regards to the already existing UD Spanish treebank. The two new converted treebanks will be released in version 1.3 of Universal Dependencies.
Keywords: AnCora, treebank, Catalan, Spanish, Universal Dependencies

1 Introduction

AnCora treebanks[1] (Taulé, Martí, y Recasens, 2008) are consolidated treebanks for Catalan and Spanish, and have indeed been the canonical treebanks of these languages. Their smaller, preliminary versions were used in the CoNLL 2006 and 2007 shared tasks in dependency parsing; the much larger and mature AnCora 2.0.0 was used in CoNLL 2009 (Hajič et al., 2009), henceforth CoNLL09. The native AnCora syntactic annotation is based on constituents but an in-house conversion to dependencies is also available (Civit, Martí, y Bufí, 2006).

In this article we present the conversion of the AnCora treebanks to Universal Dependencies. There is a UD Spanish treebank since release 1.0 (January 2015). This corpus is a legacy of an older universal treebank project (McDonald et al., 2013) and it is unrelated to AnCora. It is made up of web data and we refer to it as the Spanish Web UD treebank. However, according to the Web treebank documentation, its developers have made use of AnCora for lemmatization and morphological analysis. With our UD conversion of AnCora, appearing in UD release 1.3 (May 2016), the UD collection thus contains two treebanks for Spanish, as well as the first UD treebank for Catalan. Independent from our work, a Galician treebank is also scheduled to appear in UD 1.3. If we also consider the already existing Basque and Portuguese treebanks, UD 1.3 will allow parsing a great deal of the linguistic diversity of the Iberian peninsula.

Section 1.1 outlines the important characteristics of the UD formalism that we have taken into account for the conversion. Section 2 describes the conversion steps. Sec-

[1] http://clic.ub.edu/corpus/

ISSN 1135-5948

tion 3 offers quantitative evaluation of the conversion results, and finally Section 4 offers conclusions and perspectives.

1.1 Universal dependencies

Universal Dependencies (UD)[2] (Nivre et al., 2016) is a project that seeks to define cross-linguistically applicable annotation guidelines for morphology and syntax of natural languages. An integral part of this effort is frequent releases of annotated data (UD treebanks) in multiple languages, that conform to the UD guidelines and that are freely available to the research community.

UD uses a set of 17 universal POS, a set of 40 universal dependency relations, and a set of universal features to give account for lexical or grammatical information in terms of key-value pairs like *Gender=Fem*. The POS inventory is fixed across all languages—even though not all languages use all tags—whereas the formalism allows language-specific extensions of dependency relations and features.

A key aspect of the UD dependency formalism is the primacy of content over function words. While other conventions make auxiliaries the head of periphrastic verb tenses, or even determiners the head of noun phrases, content words are the preferred heads in UD. Notably, this analysis also demotes copula verbs to dependent status, and makes the attribute the head of the copula construction. This decision aims at harmonizing the representation across languages, including those that have no explicit copula. Figure 1 shows an example sentence from the original Spanish AnCora CoNLL09 corpus and its UD conversion. We describe the example in more detail in Section 2.5.

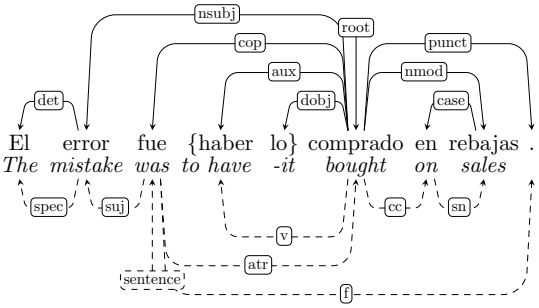

Figure 1: Example of dependency tree before and after conversion. Above: UD, below: CoNLL09 (dashed).

The main contribution of this conversion is to make the AnCora treebanks available for further research using the UD formalism, as they are treebanks that have been benchmarked for a decade. During conversion, we make the treebanks compliant with the specifications of the UD formalism, and we harmonize certain choices to increase compatibility with the other Romance languages, in particular aiming at making the choices of structure and part of speech as similar as possible between the already existing UD Spanish Web treebank and our conversion of Spanish AnCora.

2 Conversion

Recent work (Kolz, Badia, y Saurí, 2014) describes a dependency conversion of the Spanish AnCora treebank. However, it used a syntax-driven formalism, where function words are more likely to be the heads. In this present work we take the complementary stance.

Moreover, the AnCora treebanks have already appeared in another multi-lingual treebank collection, namely HamleDT (Zeman et al., 2014), which is one of the predecessors of UD.[3] HamleDT provides an automatic conversion of over 40 treebanks by first converting them all to the formalism of the Prague Dependency Treebank, and later exporting them to either the Stanford dependency formalism in the first releases of HamleDT, or to UD in the current release 3.0 (Zeman et al., 2015). However, HamleDT 3.0 does not follow some important UD guidelines, most notably those concerning tokenization. We take HamleDT as our starting point and extend the conversion to produce a fully UD-compliant release of the sibling corpora.

2.1 HamleDT conversion outline

We make use of the freely available HamletDT 3.0 conversion to incorporate the AnCora treebanks to UD. This section provides an overview of the main operations carried out during the HamletDT conversion. For further details, cf. (Zeman et al., 2014). First, HamleDT transforms the original CoNLL09 treebanks to Prague Dependencies (Böhmová et al., 2003) following these steps:

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

[3] `http://ufal.mff.cuni.cz/hamledt/`

1. Convert POS and features from CoNLL09 to UD.[4]
2. Convert the CoNLL dependency relation labels to those of the Prague Dependency Treebank. Some labels are further adjusted when the tree structure is transformed. This step is not trivial because some "dependency" relations in fact just denote leaf nodes from the original constituents, whose (dependency) relation to the constituent head is not marked. Conversion of these cases often requires also transforming the tree structure in subsequent steps.
3. The original annotation has no explicit marking of coordination. Some dependency relations such as `grup.nom` are good indicators of coordination but there are structures such as coordinated clauses, that cannot be detected this way. HamleDT searchers for larger-scope coordinated constituents joined by coordinating conjunctions, and marks them as coordination.
4. Some phrases present idiosyncratic head choices, e.g. in the Catalan *una mica* ('a bit'), the noun *mica* is attached as a dependent of the article *una*. This analysis does not correspond with the Prague guidelines, and it is also inconsistent with how determiners and nouns are connected elsewhere in the treebank. This particular analysis in CoNLL09 is in fact similar to the UD convention for multiword expressions (cf. Section 2.3).

The main steps of the conversion from Prague to Universal Dependencies are:

1. Convert Prague dependency relation labels to UD relations.
2. Convert coordination from the Prague style (headed by conjunction) to the Stanford style (headed by the first conjunct). This effectively means reverting to the approach taken in the original CoNLL09 data, except that now all coordination is explicitly and uniformly marked. For more details on the complexities of coordination conversion and the function-structure tradeoff in dependency syntax, cf. (Popel et al., 2013).
3. Invert prepositional phrases so that the nominal is the head and the preposition is attached to it using the `case` relation.
4. Invert copula constructions so that the attribute (adjective or nominal predicate) is the head and the copula is attached to it using the `cop` relation. The subject and adverbial modifiers are also re-attached to the attribute.
5. Detect controlled verbs and treat them as non-finite subordinate clauses, i.e. infinitives attached to other verbs, e.g. in *va refusar donar més detalls* ('refused to elaborate'), *donar* is attached to *refusar* with the relation `xcomp`.

[4] http://universaldependencies.org/tagset-conversion/index.html

2.2 Tokenization

Tokenization makes up the most of our adaptation from HamleDT 3.0 to full UD compliance. Notice that tokenization changes in an already-annotated treebank are not trivial, because they also imply rewriting the dependency structure of the sentences with modified tokens. We also implement other operations like feature and empty-sentence cleanup. The UD stance on tokenization is that dependency relations hold between *syntactic words*, which do not have to be identical with *orthographic words*. Surface tokens are split if their parts perform independent syntactic functions. For instance, *del* is the preposition *de* fused with the definite article *el*; in UD, the preposition and the article are independent syntactic words corresponding to separate nodes in the dependency tree.

On the other hand, UD does not allow tokens to be made up of more than one orthographic word, i.e. "words with spaces" are disallowed. If a frozen expression of multiple orthographic words behaves as one syntactic unit (where the internal syntactic structure does not exist or has become vacant), each orthographic word will have its own node in the tree and technical relations such as `mwe` (for multiword expressions) or `name` (for proper names) will connect them. The head of these two kinds of structures is the leftmost token, and all other tokens are its dependents. Figure 2 provides an example of `mwe` in UD. The expression *pel que fa a* (lit. 'for that which does to', En. 'regarding') is a single token in the CoNLL09 data joined with underscores, and becomes a `mwe` subtree.

A key difference between `mwe` and `name` subtrees is that UD guidelines treat prepositions, articles and conjunctions in proper names as such, and not as tokens with the `PROPN` POS tag. This difference implies that

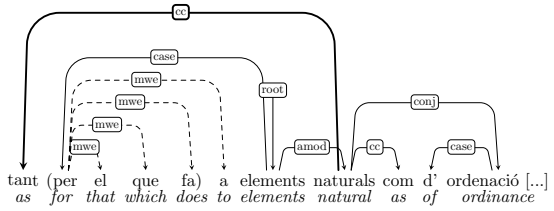

Figure 2: UD Catalan example with a multiword expression subtree (dashed) and a non-projectivity caused by coordination (thick edge) crossing another edge.

`name` subtrees like *Ajuntament* ***de*** *Vilanova* ***i la*** *Geltrú* will have more than one depth level where the words in bold will be leaves, whereas the depth of `mwe` is always one.

The CoNLL09 treebanks make elliptic subjects explicit by giving them an empty token with pronoun POS and subject function, marked with an underscore as form. Elliptic subjects are not syntactic words, and we remove them from the sentences.

Moreover, UD allows providing the original tokens of certain constructions prior to being tokenized apart into syntactic words. These *multiword tokens* are provided as an additional annotation of the sentence and play no role in the syntax, but can be used for ease of reconstruction of the original text, to train tokenizers, etc. We apply language-specific modifications to identify two kinds of multiword tokens in Catalan and Spanish, namely fused article-preposition tokens, and verbs with clitics.

2.3 Catalan-specific tokenization

The Catalan inventory of fused adposition-articles is *al*, *del*, *pel*, and their plural forms *als*, *dels*, and *pels*, for a total of six. We split these tokens into their two forming syntactic words, and provide the original token as a multiword token.

Clitic tokenization in Catalan is simple because clitics are introduced by hyphens or apostrophes. Verbs with clitic pronouns are already tokenized apart in AnCora-Catalan. We identify verb-headed spans with clitics and add the multiword-token to the annotations of the sentence. The most common possessive determiner construction is a two-word periphrasis e.g. ***la meva*** *casa* (‘**my** house’), where the first word is the definite determiner (*la*, ‘the’) and the second one is the possessive adjective (*meva*, ‘mine’), placed prenominally and consecutive to the determiner. The original AnCora tokenization treats these two words together as a one, joined with underscores. We split these pretokenized possessive constructions into their two forming syntactic words, and make them both dependent of the noun they introduce. We inform of the possessiveness of the second word using the `Poss` feature.

2.4 Spanish-specific tokenization

The Spanish inventory of fused adposition-articles is a set of two, namely *al* and *del*. We split them and keep the original fused for as a multiword token.

AnCora-Spanish does not provide split verbs with clitics like *encontrándoselas* or *abridlo*, cf. Table 1. We split away the clitics for these verbs and insert them as pronouns, with a POS tag PRON and the corresponding case, gender and number features, making them dependents of the verb. Moreover, we normalize the spelling of the verb form without clitics by removing diacritics. e.g. *encontr**á**ndoselas* vs. *encontrando*. This change in spelling is a consequence of the change of stress pattern when removing clitics from the verb. We add the original multiword token verb as a sentence annotation.

encontrándoselas		
encontrando	se	las
FIND-gerund	3-refl	3-f-p-a
abridlo		
abrid	lo	
OPEN-imperative-plur	3-m-s-a	

Table 1: Spanish clitic-verb examples.

2.5 Conversion example

Figure 1 shows the original AnCora (dashed, below) tree and the converted UD tree for the Spanish sentence *El error fue haberlo comprado en rebajas’* (En. ‘The mistake was to have bought it on sales’). We can observe how the names of the relations are all different, and the relations above use the UD inventory. In terms of the structure, the original main node *fue* is demoted to leaf status as a `cop`, a copula auxiliary of the predicate *comprado*, which is the main node in the UD tree. Note that *fue* does not license a passive reading, which would be marked as `auxpass`. The preposition *en* also becomes a leaf, in

this case of the noun *rebajas*. Moreover, the verb-clitic multiword *haberlo* ('have+it') is split in two different tokens, and *lo* is a `dobj` dependent of the verb *comprado*.

3 Results

3.1 Treebank properties

There is of course the danger that each new conversion will lose more information or introduce errors. We dedicate this section to determining the consistency of the treebanks after UD conversion. Table 2 shows the properties of the AnCora treebanks at their three distributions, namely CoNLL09, HamleDT 3.0 and UD 1.3. The columns list the number of sentences (S) and of words (W), the POS ambiguity (PA), the average edge length (EL), the average root distance (RD), and the number of sentences that are not fully projective (NP).

In terms of number of words and sentences, we keep the official test data split from the CoNLL 2009 shared task, where the training data is 80% of the corpus, and development and test data are 10% each. We can see that the number of sentences diminishes slightly as a result of empty-sentence removal, while the number of words increases in 10% as a consequence of splitting multiword expressions, fused preposition-articles, and verbs with clitics.

We measure POS ambiguity as the proportion of words that have more than one possible POS and a frequency above one. The conversion steps make PA increase. While an increase in ambiguity makes POS prediction harder using these datasets, it also indicates that the conversion process successfully applies a disambiguation of the original CoNLL09 POS inventory of 12 tags into the UD inventory of 17 tags by means of structure and feature analysis, notably incorporating the lexical verb / auxiliary verb distinction (`VERB`/`AUX`) and the common noun / proper name (`NOUN`/`PROPN`) distinction.

The average edge length (EL) increases monotonically on each conversion step, as relations across predicates are pulled up in the decision tree when e.g. a subordinate clause becomes headed by its verb and not by its subordinating conjunction. Parallel to this change, the average distance to the root node (RD) decreases because trees become flatter.

The CoNLL09 distribution of both AnCora treebanks is fully projective. However, the conversion steps incorporate non projectivities into the structure. Upon manual inspection, we find that some non-projectivities are introduced by the splitting process of large proper-name multiwords that are internally connected by determiners and prepositions such as *Les Terres de l'Ebre*. Moreover, we also find legitimate case of non-projectivity such as the Catalan example on Figure 2. The conjunction *tant* has scope over *naturals com d'ordenació*, and the intermediate word *elements*, higher in the tree, issues a crossing edge. The expression *tant ... com* is a double conjunction is a manner similar to 'as well ... as'.

3.2 Monolingual dependency parsing

Dependency parsing evaluation allows estimating the consistency of a treebank's annotations. We use TurboParser (Martins et al., 2010) for all the parsing experiments in this section. We have trained all the models using the local, arc-factored feature model for the parser. While a richer feature model would improve performance, we use the arc-factored model for speed reasons. The goal of the parsing experiments in the following sections is to assess the relative consistency of the different versions of the treebanks, and not to benchmark the parser itself. Nevertheless, arc-factored TurboParser obtains scores comparable to the best systems for Catalan and Spanish in the CoNLL09 shared task.[5]

Table 3 shows the parsing results for the three steps in the conversion of the treebanks, namely the original CoNLL09 AnCora distribution, the HamleDT 3.0 UD-compatible conversion, and our UD conversion.

These scores are not strictly comparable, given that all treebanks have different tokens, part-of-speech tags, and dependency relations. However, we provide them as an indication of the general reliability of the conversion. This approach has also been used in previous conversion works, who also report scores in the 80-85% range (Johannsen, Martínez Alonso, y Plank, ; Pyysalo et al., 2015; Silveira y Manning, 2015).

In spite of these differences, some relations are straightforward to compare. The `sentence` relation in CoNLL09 maps to the `root` relation in UD. Both the Catalan and

[5] `https://ufal.mff.cuni.cz/CoNLL09-st/results/results.php`

		S	W	PA	EL	RD	NP
Ca	CoNLL09	16,786	497k	0.10	3.66	4.60	0
	HamleDT	16,786	497k	0.13	3.86	4.07	76
	UD	16,678	547k	0.16	4.00	4.12	468
Es	CoNLL09	17,709	528k	0.11	3.69	4.76	0
	HamleDT	17,709	528k	0.13	3.90	4.22	327
	UD	17,680	569k	0.15	3.99	4.19	624

Table 2: Treebank statistics

	C09		HDT		UD	
	LAS	UAS	LAS	UAS	LAS	UAS
Ca	86.7	89.6	85.4	87.7	85.4	87.9
Es	86.1	88.9	85.0	87.1	84.9	87.3

Table 3: Labeled and Unlabeled Attachment Scores (LAS and UAs, in grey) for the three different conversion stages of the treebanks.

Spanish CoNLL09 treebanks have an average `sentence` accuracy of 93% and, and their UD conterparts have a `root` accuracy of 90%. This degradation is a result of i.a. the promotion of lexical tokens to the head of the copulas, which penalizes the general tendency to make verbs the heads of clauses.

Complementarily, the LAS of prepositions, tagged `s` in CoNLL09 and `ADP` in UD, goes from 79% to 97% after the conversion. This improvement is a result of preposition attachment becoming more local in UD, and easier to resolve, while noun attachment becomes more difficult to predict, and goes from 90% to 76%. Indeed, UD transfers most of the important relations to relations between content words, and makes function words easier to attach: auxiliaries, determiners and prepositions have all attachment accuracies above 96% in both UD AnCora treebanks.

3.3 Dependency parsing between UD languages

We use cross-lingual parsing as a way to estimate the consistency with the other treebanks, and thereby with the current state of UD as a whole. In order to do so, we apply a delexicalized transfer scenario, removing form and lemma information from the training data, and we train Turbo parser in unlabeled mode. While it is customary to also remove the feature information, we decide to keep all the morphological features of the source and target data, because they belong to the harmonized UD inventory.

Table 4 shows the results on training on Catalan or Spanish and testing on itself (*Self*), on the other converted AnCora treebank (*Sibling*), i.e. training on Catalan and testing on Spanish. The three last columns show the results on delexicalized parsing the test section of whole set of UD1.2 languages. The *All* column provides the macro-average for all 32 languages, while the *Romance* column is the macro-average score for French, Italian, Portuguese and Romanian, as well as the pre-existing Spanish Web treebank. The *Other* column shows the average results for all the non-Romance UD languages. We compare the Spanish Web and the Spanish AnCora treebank in more detail in Section 3.4.

	Self	Sib.	Rom	Other	All
Ca	84.5	81.7	66.2	49.5	52.0
Es	83.5	82.6	62.5	46.9	49.6

Table 4: mean UAS for delexicalized transfer.

The drop in UAS from full lexicalized to delexicalized is only of 5%. The high delexicalized parsing scores for Self indicate that, in spite of the chain of conversions, the AnCora UD treebanks have a well-coupled mapping between the POS tags and features, and dependency structure. The AnCora UD treebanks are internally very consistent, and the parser achieves comparatively high UAS when trained on one sibling and applied to the other. However, the results are 20 points lower in average when parsing the other Romance languages. For Italian and Portuguese the score is around 71% for both sources, while for French it is around 53%. The internal variation between the Romance language

test bench can have a linguistic basis, but it is also caused by divergences in the treebank's annotation. If we compare with the *Other* set, we observe the differences are even larger, and the drop from *Romance* to *Other* is of 13 points. Regardless of the variation in dependency annotation, there is more consistency between the new AnCora treebanks and the Romance languages, which confirms the linguistic basis for better parsing scores for e.g. Italian and Portuguese.

3.4 Dependency parsing between AnCora and Web Spanish

We assess the similarity of the Ancora Spanish conversion to UD with the preexisting Spanish Web UD using dependency parsing. If the two treebanks were as similar as possible, the differences in parsing accuracy when e.g. parsing with AnCora and testing with Web would be due to dataset size and domain change, and not to differences in dependency convention. Table 5 shows the attachment accuracies when using one Spanish treebank to parse the other (*Other*), along with intra-treebank evaluation for comparison (*Self*).

	Self		Other	
	LAS	UAS	LAS	UAS
AnCora	84.9	87.3	64.0	71.8
Web	81.7	84.5	69.6	78.7

Table 5: Labeled and Unlabeled Attachment Scores (LAS and UAs, in grey) for the four possible source/target pairs for Spanish.

There is a severe drop in performance when changing training treebank, e.g. when using Web to parse AnCora, the performance drops in about 15 points, from 84.9% to 69.6% LAS. A similar drop of 17 points appears when using AnCora to predict Web.

These differences are large enough not to be exclusively an effect of domain change. UD treebanks encode some lexical information such as multiword expressions in the dependency structure (cf. Section 2), and the inventory of multiword expressions is different across treebanks.

In AnCora UD, the elements marked as multiwords are the tokens that were underscored together in the CoNLL09 data, and that were not proper names. AnCora UD has 521 word types that are attached with a `mwe` relation, whereas Web has 113. While AnCora has a larger `mwe` inventory, it is not a perfect superset of the `mwe` expressions in Web, because it only covers 80% of the expressions in the Web corpus. Some of the common expressions are treated in the same fashion in both treebanks, such as *sin embargo*, *ni siquiera*, or *a través de* ('nevertheless', 'not even', 'through'), but the difference in multiword inventory make the parser predict mistmatching structures across datasets.

Both treebanks have a lenient definition on auxiliary verbs. Besides the auxiliares that are used to form verb tenses, namely *haber* for compound tenses, *ser* for passive forms, and *estar* for gerund forms, the treebanks also license an `AUX` reading for verbs that indicate modality *(poder, querer),* aspect *(continuar, terminar),* and also other verbs used for support constructions like *llegar* in *llegar a causar* ('come to cause'). These choices are semantically motivated and aim at promoting the lexical verb of the construction to the head of the structure (cf. Section 1.1, instead of treating light verbs as syntactic heads. However, Romance languages do not have a family of modal verbs as distributionally well-defined as Germanic languages do, and the criteria for labeling a semantically impoverished verb as `AUX` can be revised in further releases of the treebanks.

Another major difference between the two Spanish treebanks is the interpretation of the word *que* as either a subordinating conjunction or a pronoun (`SCONJ` / `PRON`), which gives a very low attachment score to pronouns (35%) in the cross-treebank (*Other*) setup, in spite of the very high accuracy in the intra-treebank setup (85%).

The AnCora Spanish corpus is largely made up of newswire from Spain, whereas the Web Spanish corpus potentially holds any of the variants of Spanish. We find sentences like *Olvidate todo, seguí tu vida* ('Forget about everything, get on with your life'), with verb usage characteristic of Rioplatense Spanish. A more detailed study on the differences between both Spanish corpora would shed light on the relevance of regional specificity in corpus choice for Spanish processing.

4 Conclusion

We have presented the conversion of the Catalan and Spanish AnCora treebanks to the Universal Dependencies formalism. We use the freely available HamleDT 3.0 as

a starting point for POS and dependency conversion, and we apply a set of operations to tune tokenization to UD, tackling language-specific phenomena like fused article-prepositions like *del* and verb-clitic multiword tokens like *encontrársela*.

We have evaluated the consistency of the resulting converted treebanks by means of dependency parsing. We obtain parsing scores comparable to other converted UD treebanks (cf. Section 3.2), and we assess the variation between UD languages in terms of typological proximity and annotation convention using delexicalized transfer parsing 3.3. The fairly large loss of unlabeled attachment score for the other languages is much less dramatic for the Romance languages.

We also analyze the differences between the Spanish Ancora UD-converted treebank and the preexisting Spanish Web treebank. While the parsing between the two Spanish UD treebanks fares above the delexicalized transfer setup, the differences in multiword treatment and some POS particularities indicate that the treebanks need further harmonization.

4.1 Further work

Universal Dependencies is a constantly improving effort, and the guidelines are refined before each release. In further releases, we expect to harmonize the two AnCora UD treebanks with regards to the treatment of auxiliary verbs across all Romance languages, revise the non-projective sentences and keep the legitimate examples, and in particular improve the comparability of the two UD Spanish treebanks, namely AnCora and Web.

Bibliografía

Böhmová, A., J. Hajič, E. Hajičová, y B. Hladká. 2003. The prague dependency treebank. En *Treebanks*.

Civit, M., M. A. Martí, y N. Bufí. 2006. Cat3LB and Cast3LB: from constituents to dependencies. En *Advances in Natural Language Processing*.

Hajič, J., M. Ciaramita, R. Johansson, D. Kawahara, M. A. Martí, L. Màrquez, A. Meyers, J. Nivre, S. Padó, J. Štěpánek, P. Straňák, M. Surdeanu, N. Xue, y Y. Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. En *CoNLL-2009*.

Johannsen, A., H. Martínez Alonso, y B. Plank. Universal dependencies for Danish. En *TLT14*.

Kolz, B., T. Badia, y R. Saurí. 2014. From constituents to syntax-oriented dependencies. *Procesamiento del Lenguaje Natural*.

Martins, A. F., N. A. Smith, E. P. Xing, P. M. Aguiar, y M. A. Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. En *EMNLP 2010*.

McDonald, R. T., J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. B. Hall, S. Petrov, H. Zhang, O. Täckström, y others. 2013. Universal dependency annotation for multilingual parsing. En *ACL*.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, y D. Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. En *LREC*.

Popel, M., D. Marecek, J. Stepánek, D. Zeman, y Z. Zabokrtskỳ. 2013. Coordination structures in dependency treebanks. En *ACL*.

Pyysalo, S., J. Kanerva, A. Missilä, V. Laippala, y F. Ginter. 2015. Universal dependencies for Finnish. En *NoDaLiDa*.

Silveira, N. y C. Manning. 2015. Does universal dependencies need a parsing representation? an investigation of English. *Depling 2015*.

Taulé, M., M. A. Martí, y M. Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. En *(LREC 2008)*.

Zeman, D., O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, y J. Hajič. 2014. HamleDT: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*.

Zeman, D., D. Mareček, J. Mašek, M. Popel, L. Ramasamy, R. Rosa, J. Štěpánek, y Z. Žabokrtský. 2015. HamleDT 3.0. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.