How Does the Granularity of an Annotation Scheme Influence Dependency Parsing Performance?

Simon Mille[1] *Alicia Burga*[1] *Gabriela Ferraro*[1] *Leo Wanner*[1,2]
[1]Universitat Pompeu Fabra Barcelona; [2]ICREA
firstname.lastname@upf.edu

ABSTRACT
The common use of a single *de facto* standard annotation scheme for dependency treebank creation leaves the question open to what extent the performance of an application trained on a treebank depends on this annotation scheme and whether a linguistically richer scheme would imply a decrease of the performance of the application. We investigate the effect of the variation of the number of grammatical relations in a tagset on the performance of dependency parsers. In order to obtain several levels of granularity of the annotation, we design a hierarchical annotation scheme exclusively based on syntactic criteria. The richest annotation contains 60 relations. The more coarse-grained annotations are derived from the richest. As a result, all annotations and thus also the performance of a parser trained on different annotations remain comparable. We carried out experiments with four state-of-the-art dependency parsers. The results support the claim that annotating with more fine-grained syntactic relations does not necessarily imply a significant loss of accuracy. We also show the limits of this approach by giving details on the fine-grained relations that do have a negative impact on the performance of the parsers.

TITLE AND ABSTRACT IN SPANISH

¿Cómo influye la granularidad de un esquema de anotación en el rendimiento de parsers de dependencia?

El uso frecuente de un único esquema de anotación estándar para crear corpus de análisis sintáctico de dependencias genera las preguntas de hasta qué punto el rendimiento de una aplicación entrenada con dichos corpus depende del esquema de anotación, y si un esquema lingüísticamente más rico implica que la calidad de la aplicación disminuya. Investigamos aquí el efecto de la granularidad de la anotación sobre el rendimiento de parsers de dependencia. Para obtener distintos niveles de granularidad, diseñamos un esquema de anotación jerárquico basado exclusivamente en criterios sintácticos. La anotación más detallada incluye 60 relaciones, y de ésta derivamos los conjuntos menos detallados. Así, las anotaciones—y el rendimiento de parsers entrenados con ellas—se mantienen comparables. Los experimentos utilizan cuatro parsers del estado del arte. Los resultados apoyan la hipótesis de que una anotación más detallada no implica una pérdida de precisión del parser. Presentamos también las limitaciones de este enfoque, ofreciendo detalles acerca de aquellas relaciones que sí tienen un impacto negativo en la calidad de los parsers.

KEYWORDS: dependencies, syntax, annotation, tagset granularity, parsing.

KEYWORDS IN SPANISH: dependencias, sintaxis, anotación, granularidad del tagset, parsing.

Proceedings of COLING 2012: Posters, pages 839–852,
COLING 2012, Mumbai, December 2012.

SYNOPSIS IN SPANISH

Para medir la precisión de los parsers en función del detalle de los tagsets, se diseñó un esquema jerárquico de anotación de relaciones de dependencia que permite expandir o contraer el número de relaciones a utilizar. La idea general tras este esquema es la aplicación de criterios sólo sintácticos (más que semánticos), más o menos finos, que permiten identificar cada etiqueta gramatical a ser introducida en la anotación, así como agrupar relaciones en una etiqueta más amplia. Así, por ejemplo, para dependientes verbales, necesitamos capturar si éstos pueden pronominalizar, si su movimiento es limitado, etc. En la Tabla 1 se muestra qué relaciones del tagset más detallado son agrupadas bajo la misma etiqueta en el siguiente, y menos detallado, conjunto. Estas agrupaciones se basan en propiedades sintácticas compartidas por un grupo de relaciones.

60 Rels	44 Rels	31 Rels	15 Rels
abs pred	abs pred	abs pred	
det	det	det	
quant	quant	quant	
compl adnom	compl adnom	compl adnom	
appos	appos		NMOD
abbrev	abbrev		
attr	attr	modif	
modif	modif		
relat	relat		
adjunct			
adv	adv		
restr			
relat expl	relat expl	adv	ADV
prolep	prolep		
adv mod			
obj copred	copred		
subj copred			
analyt fut	analyt fut	analyt fut	
analyt pass	analyt pass	analyt pass	
analyt perf	analyt perf	analyt perf	AUX
analyt prog	analyt prog	analyt prog	
modal	modal	modal	
dobj clitic	dobj clitic	dobj clitic	DOBJ
dobj	dobj	dobj	
copul	copul	copul	COPUL
copul clitic	copul clitic	copul clitic	
iobj1	iobj	iobj	
iobj2			IOBJ
iobj3			
iobj clitic1	iobj clitic	iobj clitic	
iobj clitic2			
iobj clitic3			

60 Rels	44 Rels	31 Rels	15 Rels
obl obj1	obl obj	obl obj	
obl obj2			
obl obj3			
noun compl			
agent	agent	compar	OOBJ
compar	compar		
compl1	compl	compl	
compl2			
elect	elect	elect	
subj	subj	subj	SUBJ
quasi subj	quasi subj	quasi subj	QSUBJ
compar conj	conj	prepos	PREPOS
sub conj			
coord conj	coord conj		
prepos	prepos		
coord	coord	coord	COORD
num junct	num junct		
juxtapos	juxtapos	juxtapos	
quasi coord	quasi coord		
sequent	sequent	sequent	BIN
bin junct	bin junct	bin junct	
aux phras	aux phras	aux phras	NAME
aux refl lex	aux refl	aux refl	AUX REFL
aux refl pass			
aux refl dir			
aux refl indir			
punc	punc	punc	PUNC
punc init	punc init		

Table 1: Tag groupings for a hierarchy of syntactic tags/Jerarquía de agrupación de etiquetas sintácticas (Left=top, right=bottom of table)

Para los experimentos, se utilizaron cuatro tagsets de relaciones sintácticas. El más detallado (60 relaciones) se obtuvo a partir de una adaptación, revisión y enriquecimiento de la anotación original de AnCora, desde la cual se derivaron automáticamente los otros tres tagsets (44, 31 y 15 relaciones), obteniendo así cuatro anotaciones distintas del mismo corpus. Se evaluaron cuatro parsers de referencia. Tres de ellos son los parsers con mejores resultados para español en la CoNLL Shared Task 2009: Che, Merlo y Bohnet; el cuarto es el muy conocido Malt Parser. El corpus fue dividido al azar en un grupo de entrenamiento (3200 oraciones) y un grupo de evaluación (313 oraciones). Cada parser fue entrenado con los cuatro conjuntos de relaciones y los dieciséis modelos de parsing obtenidos fueron aplicados a los correspondientes conjuntos de evaluación.

Los resultados para el Labelled Attachment Score (LAS)—es decir, la proporción de asignación de relaciones con la adecuada etiqueta y el gobernador y el dependiente correctos—se muestran en la Tabla 2. Observamos que los cuatro parsers se comportan de modo similar: su precisión

tags# >	60	44	31	15
Bohnet	81.95	84.11	84.28	84.69
Che	75.14	84.24	84.67	85.11
Malt	79.7	81.9	82.1	82.2
Merlo	82.32	84.53	84.05	84.52

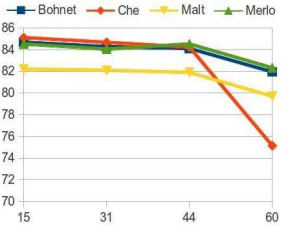

Table 2: LAS (%) of the parsers depending on tag granularity; right: graphical illustration/LAS de los parsers en función de la granularidad del tagset; derecha: ilustración

es constante de 15 a 44 relaciones, pero disminuye con 60 relaciones. Asimismo, notamos una diferencia entre las curvas de Bohnet, Merlo y Malt (prácticamente paralelas) y la de Che, que cae significativamente con 60 relaciones. Todos los parsers logran el mejor rendimiento con el tagset más pequeño y menos detallado. Sin embargo, sorprendentemente, el LAS disminuye muy poco cuando el número de relaciones se duplica, y menos aun entre 31 y 44 relaciones. Con 60 relaciones, no obstante, el LAS cae significativamente alrededor de al menos 2 puntos. También calculamos el UnLabelled Attachment Score (ULA)(ver Tabla 3). Para Bohnet, el ULA aumenta leve pero constantemente de 15 relaciones (90.27%) a 60 relaciones (90.49%). Che, en cambio, presenta la tendencia contraria, y sus resultados decrecen de 15 a 60 relaciones (habiendo una caída mayor con 60). Malt es tan estable como Bohnet, pero no presenta una clara mejora al trabajar con un número mayor de etiquetas. Asimismo, para evaluar si con 15 relaciones la calidad mejora si el parser es entrenado con un tagset más detallado, todos los outputs fueron transformados a 15 relaciones. Como vemos en la Tabla 4, en términos generales, la tendencia es la misma que para el ULA, de modo que podemos concluir que la anotación con más relaciones no parece mejorar la calidad del parser al trabajar con 15 relaciones.

Observamos que aquellas relaciones que se diferencian gracias a rasgos sintácticos muy finos (como los diferentes tipos de objetos oblicuos, completivos, o auxiliares reflexivos) son las que más influyen en la disminución de la calidad del parser. Consecuentemente, no separar estas relaciones en relaciones más finas puede ser beneficioso para el parser. Al contrario, observamos que las dependencias que implican diferentes tipos de coordinaciones entre grupos o frases se parsean mejor si no se juntan.

tags# >	60	44	31	15
Bohnet	90.49	90.39	90.31	90.27
Che	86.28	90.37	90.57	90.6
Malt	87.91	88	87.83	87.75
Merlo	90.11	90.67	90.39	-

Table 3: ULA of the parsers depending on tag granularity/ULA de los parsers en función de la granularidad del tagset (%)

tags# >	15	31→15	44→15	60→15
Bohnet	84.69	84.56	84.51	84.54
Che	85.11	84.93	84.71	77.91
Malt	82.2	82.3	82.2	82.2
Merlo	84.52	84.33	84.92	84.12

Table 4: LAS of the parsers (with 15 SyntRels) trained on fine-grained tagsets/LAS de los parsers (con 15 SyntRels) entrenados con anotaciones más finas (%)

1 Introduction

As already pointed out by some researchers (see, e.g., Kübler (2005), Rehbein and van Genabith (2007), Bosco et al. (2010), Bosco and Lavelli (2010)), the use of a single annotation scheme for treebank creation leaves the question open to what extent the performance of an application trained on a treebank depends on the annotation scheme in question. Or, in other words, whether the annotation scheme in use is the best for a given application. To answer this question, Kübler (2005) and Rehbein and van Genabith (2007) compared the performance of a PCFG parser trained on two comparable corpora of German, annotated following different annotation schemes, while Bosco et al. (2010) trained three dependency parsers on two different Italian corpora. In contrast, we are interested in a comparison of the change of the performance of a dependency parser when trained on the same corpus, but annotated with gradually more fine-grained annotation schemes, that is, with gradually more arc labels in the tagset. Our approach differs from (Bosco and Lavelli, 2010) in that we only retain functional syntax for the design of our tagsets. The background of our research is that standard annotation schemes such as the scheme underlying the dependency conversion from the Penn Treebank.[1] tend to be minimal in order to facilitate the process of annotation and to improve the readability of the resulting annotation.[2] This tendency is reinforced by the general assumption that the less fine-grained the annotation, the better the parser performance. However, this has a major drawback, namely that the parsed structure is often too poor to serve well, e.g., semantic role labeling, deep summarization, content extraction, word sense disambiguation, etc.

To the best of our knowledge, no study actually compares the performance of a dependency parser trained on annotations of varying syntactic granularity, so there are no figures that would demonstrate that it is worth to sacrifice grammatical accuracy and detail for the sake of an acceptable parser accuracy. We carried out such a study on Spanish material. We developed a hierarchical syntactic dependency annotation scheme that allows us to expand and contract syntactic relation branches into larger, more fine-grained, or smaller, more coarse-grained, annotation schemes. The results of parsing experiments demonstrate that it is possible to reach a good balance between the accuracy of a parser and the richness of the linguistic annotation. They also show that the principles that we applied when designing the hierarchical annotation schema are valid and may be used for the design of other annotation schemes in the future.

2 Hierarchical syntactic annotation scheme

The hierarchical annotation scheme in Table 1 has been developed for Spanish on a small corpus of 3513 sentences (100892 words, see (Mille et al., 2009); corpus available at UPF–TALN webpage), which constitutes a section of the Spanish corpus AnCora (Taulé et al., 2008). The general idea underlying this scheme is to apply only syntactic (rather than also semantic) criteria in order to identify each grammatical tag that is to be introduced into the scheme. Using more or less fine-grained criteria allows us to control the level of granularity of the tagset. We do not orientate our scheme towards any particular linguistic theory; the selected criteria are dictated by syntactic behaviour observed in the language in question (in our case, Spanish). For instance, for dependents of verbs, we need to capture whether they can be cliticized, promoted

[1] The dependency annotation scheme of the Penn Treebank has served as blueprint for annotation schemes of a series of treebanks in different languages and is thus a *de facto* standard. See (Marcus et al., 1993) for the original consituency annotation, and (Johansson and Nugues, 2007) for the conversion to one-word-per-line dependency representations.

[2] "Minimal" refers here not only to the number of tags, but also to the level of precision of the syntactic tags. Indeed, many corpora mix several levels of representation (e.g., syntax, semantics, lexicon, etc.) such that the number of syntactic relations does not necessarily reflect the level of idiosyncracy of the annotation.

or demoted, etc. For any kind of dependent, we need to capture the canonical order with respect to its governor, the part-of-speech of the governor, the part-of-speech of the prototypical element that appears in that paradigm, the existence or absence of some agreement between the prototypical dependent and another element of the sentence, the presence/absence and type of required features of the dependent (e.g., governed preposition, imposed finiteness or case, etc.), the possibility to remove a dependent or not without hampering sentence grammaticality, etc.; see (Burga et al., 2011) for examples and details.

The leftmost column in Table 1 represents the most detailed (and thus linguistically richest) tagset of 60 syntactic relations (henceforth *SyntRels*) we defined for Spanish: the distinction between one relation and another is, in general, very fine-grained. For instance, there are three types of oblique objects (*obl-obj1/2/3*), differentiated only by their default order of appearance in a neutral sentence; *noun-compl* is reserved for constructions in which the object cannot move to the left of its governor. The tags in this detailed tagset can be summarized under more generic tags, which would lead to a more coarse-grained, smaller tagset. The obtained more coarse-grained tagset can again be contracted, and so on. In Table 1, we illustrate this procedure for four tagsets in total. The brackets indicate which relations at one level were grouped together under the same label at the following level. Thus, in the second column (44 SyntRels), we group under the label *obl-obj* any non-agentive prepositional object which cannot be pronominalized, bringing together *obl-obj1/2/3* and *noun-compl*. In the third column (31 SyntRels), *obl-obj* and *agent* are fused into one relation *obl-obj*, defined as "prepositional object which cannot be pronominalized". Finally, in the last column (15 SyntRels), one tag *OOBJ* gathers any object which cannot be pronominalized, as opposed to *IOBJ* and *DOBJ*, which can be replaced by a dative and an accusative pronoun, respectively.

3 Experiments

3.1 Background

A number of experiments on different granularities of annotation and their impact on the performance of probabilistic parsers are known from the literature; see in particular Klein and Manning (2003) and Petrov et al. (2006), who show the benefits of splitting generic part-of-speech tags (e.g., NP, VP, etc.) into more precise subcategories for the derivation of accurate probabilistic context-free grammars (PCFG). Our proposal differs from these works in that they focus on constituency parsing and part-of-speech tags, whereas we tackle dependency parsing and edge labels.[3] But more importantly, the goals are different. Thus, they target the improvement of parsing accuracy, and for that they infer, with simple rules, from the training data (categorial) information which is more specific than what is directly available. Closer to our work, Bosco and Lavelli (2010) use an Italian corpus in which the dependency relations encode information on morphology, functional syntax and semantics. They discuss the influence of the annotation policies on the evaluation of the parsers and show that the precision and recall of hard-to-parse relations can be quite different, depending on the tag granularity in the annotation, that is, if the annotation contains or not morphological and/or semantic information. In contrast, our goal is to provide evidence that the creation of annotations that capture significant fine-grained distinctive features of the grammar (and only the grammar) of a language does not need to harm significantly the performance of the parsers. Consider as two

[3] Some other works present a hierarchical organization of grammatical relations (in particular (Bosco et al., 2000), (Briscoe et al., 2002), and (Marneffe et al., 2006)), but those hierarchies are not used to test the impact of the tagset granularity on the results of a parser.

such fine-grained distinctive features the relations *modal* and *direct-object* in the following two sentences. As indicated, only the direct object can be pronominalized by a clitic pronoun and moved before the governing verb, without that a pro-verb is needed: *Juan puede*-modal→ *venir mañana*, lit. 'John might come tomorrow' (*Juan lo puede *(hacer)*), and *Juan puede*-dobj→ *venir mañana*, lit. 'John is able to come tomorrow' (*Juan lo puede (hacer)*). If the annotation of the relations does not encode these phenomena, they are, in fact, lost.[4] Since this information is of primary relevance to applications related to natural language understanding, it would be an advantage to include it in the syntactic annotation. In the next sections, we show that its inclusion does not harm a parser's accuracy.

3.2 Setup of the experiments

In our experiments, we used the four tagsets introduced in Section 2. The annotation of the corpus with the most detailed tagset of 60 SyntRels has been obtained from the original annotation in AnCora (Taulé et al., 2008), which has been adapted, revised and enriched manually. Starting from the most fine-grained annotation, we derived automatically the other three, ending up with four different treebanks for the same corpus. Four reference parsers have been used. Three of them are the top three parsers for Spanish in the CoNLL Shared Task 2009 (Hajič et al., 2009): Che's (Che et al., 2009), henceforth *Che*, Merlo's (Gesmundo et al., 2009), henceforth *Merlo*, and Bohnet's (Bohnet, 2009), henceforth *Bohnet*. The fourth, the Malt Parser (Nivre et al., 2007), henceforth *Malt*, has been chosen because it is a very broadly used syntactic dependency parser. Malt and Merlo are transition based, while Bohnet and Che are graph based. In our experiments, all of them processed non-projective dependency trees. Each parser contains its own configuration options, which depend on the parsing approach, the learning techniques, etc. Therefore, it was not possible to apply the same setup to all parsers. Instead, we used for each parser its own default configuration, which does not guarantee an optimal performance. However, as the goal of this paper is not to compare the results of the parsers, but rather the performance of the same parser with different tagsets, optimized configurations are not needed for our purpose.

To train the parsers, the corpus has been divided randomly into a training set (3200 sentences) and a test set (313 sentences).[5] Each parser has been trained on each of the four annotations of the training set. The obtained sixteen parsing models were applied to the corresponding test sets. Also, in order to see whether or not the performance improved with respect to the smallest tagset when training with more fine-grained tagsets, we mapped the output of each parser onto the smallest tagset. The training and the test sets were the same as in the first experiment.

3.3 Results

For Malt, the assessment of the *Labelled Attachment Score* (LAS) (that is, the proportion of edges with correct governor and dependent and the right label on the edge) was carried out using the evaluation toolkit provided with the parser. For the other parsers, we used the official CoNLL'06 evaluation toolkit. The LAS figures for each parser and for each version of the annotation are

[4] One can always imagine some statistical "disambiguation" based on the context in which the construction is used, but the amount of data needed could be prohibitive—at least for Spanish—and eventually, the only way would probably be to imply human experts for the revision of the annotation.

[5] Bohnet's parser uses CoNLL'09 14-column format, while the other three need to be trained on the CoNLL'06 10-column format (Buchholz and Marsi, 2006), but the available information is exactly the same, whatever the format: word positions, word forms, PoS, lemmas, (all of which kept the same in our experiments), and dependencies.

shown in Table 2. The graphic on the right of Table 2 shows how each parser reacts to and how its performance varies with the increasing number of relations in the tagset. We can observe that all four parsers behave similarly: their accuracy is very constant from 15 to 44 SyntRels, and decreases with 60 SyntRels. We also notice that there is a significant difference between Bohnet, Merlo and Malt's LAS progressions (which are rather parallel) and the progression of Che, which drops when trained with 60 relations (see Section 4). As expected, all parsers reach the highest accuracy with the smallest tagset (15 SyntRels). But surprisingly, the LAS decreases only little with twice as many SyntRels in the tagset (namely 31 SyntRels): 0.1 for Malt, 0.41 for Bohnet, 0.44 for Che, and 0.47 for Merlo. Even more surprisingly, the drop is also rather small between 31 and 44 SyntRels (0.2 for Malt, 0.17 for Bohnet, 0.43 for Che). Merlo even gets better with 44 SyntRels, obtaining a LAS of 84.53%, comparable to that with 15 SyntRels and higher than that with 31 SyntRels. As a result, the decrease of performance from 15 to 44 tags in the tagset is surprisingly small for Malt, Bohnet and Che: 0.3 points for Malt, 0.6 points for Bohnet, 0.9 points for Che, and no decrease at all for Merlo. However, Bohnet, Malt and Merlo see their LAS drop significantly by around 2 points when trained with 60 SyntRels. Che drops by even more than 2 points. The in depth analysis of the behaviour of the parsers with respect to the groups of relations is presented in Section 4.

We also calculated the *UnLabelled Attachment* (ULA) score for all four parsers (see Table 3). For a reason beyond our control, we could not get the ULA for Merlo with 15 relations (however, even if incomplete, the ULA figures for Merlo are useful from the perspective of one of our experiments described below). For Bohnet, we observe that the ULA scores slightly but steadily increase in the range from 15 SyntRels (90.27%) to 60 SyntRels (90.49%). Opposite to this tendency, the scores for Che slightly decrease in the range from 15 SyntRels (90.6%) to 44 SyntRels (90.37%), and drop then with 60 SyntRels (86.28%). Malt is as stable as Bohnet, but does not show a regular improvement when dealing with higher numbers of tags. Note that the observed slight variation of the performance numbers of the different parsers across tagsets of varying sizes (always lower than 0.25 points, except Che with 60 relations) could be due to the small size of our training and test sets. In other words, it is possible that with more data, the parsers would give quite stable unlabeled attachment scores across tagsets of varying sizes.

In order to verify the effects of training a parser on a fine-grained tagset and using it then to parse with a coarse annotation, we took the test sets parsed with the models trained on 31, 44, and 60 relations, and mapped them to the coarse-grained tagset (15 different tags), following the hierarchy presented in Table 1. Then, we ran the evaluation of the resulting output against the gold standard of the 15-tag annotation; the results are presented in Table 4. In the first column, the figures obtained with the original 15-tag annotated test set for each parser are repeated in order to facilitate the comparison. Table 4 shows that there does not seem to be a benefit in annotating with fine-grained arc labels if one wants a coarse annotation. The only case in which a fine-grained annotation makes the parser improve significantly with 15 SyntRels (0.4 points) is the 44 SyntRel annotation for Merlo. Table 4 is actually very similar to Table 3, which contains the unlabeled attachment scores: all the figures for each parser are quite similar, with two exceptions: the fall of Che trained with 60 SyntRels, and a peak for Merlo trained with 44 relations. The correlation between ULA and LAS is obvious, but unfortunately, we cannnot explain so far those two deviations of ULA.

4 Evaluation of selected parsers with respect to specific SyntRels

In the previous section, we saw that the figures of all four parsers drop when trained on the most fine-grained tagset. In this section, we try to identify which relations particularly affect the performance of the parsers and thus obtain information on how the composition of the tagset has an impact on the figures of the evaluation.[6]

4.1 Impact of distinctive properties of SyntRels

Due to the relatively small amount of data we have at hand[7], there are only 8025 relation instances in the test set[8]. Some relations do not appear in it at all: *prolep*, *adv-mod*, *copul-clitic*, *num-junct* and *aux-refl-indir*. On the other side, it is not possible to generalize along the lines that the less a relation appears in the training set, the worse the performance of the parser on this relation is. Some relations (*compl-adnom*, *analyt-fut*, *analyt-progr*, *analyt-perf*, *compar*, *compar-conj*, and *compl1*) are scarce in the training set (<200 instances) and in the test set (<20 instances) and, in spite of this, they are parsed with a high accuracy (78%–100%) at least by one of the parsers.

Interestingly, as opposed to the example about objects and modals in Section 3, either the governor or the dependent (or both) of these relations have very distinctive features:

- *compl-adnom* implies a determiner followed by a preposition; cf. *la*-compl-adnom→*del sombrero azul*, lit. 'the of-the hat blue', 'that one with the blue hat';
- *analyt-fut*, *analyt-progr* and *analyt-perf* always presuppose the same auxiliary as governor and a governed preposition or a non-finite verb as dependent; cf. *voy*-analyt-fut→*a cocinar*, lit. 'I-will [to] cook'; *estoy*-analyt-progr→*cocinando*, lit. 'I-am cooking'; *fue*-analyt-pass→*cocinado*, lit. 'I-was cooked';
- *compar* and *compar-conj* require a comparative adjective governing a fixed conjunction, itself governing another element (*compar-conj*); cf. *mejor*-compar→*que*-compar-conj→*Juan*, lit. 'better than John';
- *compl1* requires an adjective on the right of a non-copular verb which undergoes agreement with the subject; cf. *la frase resulta*-compl1→*buena*, lit. 'the sentence$_{FEM.SG}$ ends up correct$_{FEM.SG}$.

There are also some relations that are not parsed well by either of the parsers, even if the number of their instances in the training and test sets is significant (see Table 5). There are two main explanations of the poor figures for the SyntRels in Table 5. First, the morpho-syntactic features of such relations (e.g., PoS of the head, PoS of the dependent) can vary a lot throughout the corpus: an adverbial or an adjunctive can be an adverb, a common noun, a non-finite verb, a prepositional group, etc. An appositive is usually a common or a proper noun, sometimes introduced by a preposition; an attributive can be a prepositional group or a gerund. Second, these relations also tend to share their basic syntactic configuration with other SyntRels; consider, e.g., *casa*-attr→*de Barcelona*, lit. 'house from Barcelona' vs. *hermano*-obl-obj1→*de*

[6] The problematic SyntRels were the same for all four parsers. Due to space restrictions, we chose to focus on the two graph-based parsers, since the graph-based approach becomes increasingly popular in parsing research.

[7] Still, we believe that our results are already quite reliable since the average accuracies (without tuning the parsers) get close to the accuracies obtained by the same parsers at the Shared Task 2009 with much larger data sets (http://ufal.mff.cuni.cz/conll2009-st/results/results.php).

[8] The dependencies to punctuation signs were not considered in the figures of the evaluation because they are parsed with the same (very high) accuracy whatever the tagset; considering them would boost the parser figures by 0.5% but it would not bring anything to our experiment.

	Training Set (instances)	Test Set (instances)	Bohnet (%)	Che (%)
adjunct	830	87	37.93	31.03
adv	5751	549	62.3	56.83
appos	1060	100	54	34
attr	2165	213	37.56	41
obl-obj1	3551	384	50.78	26.82

Table 5: Poorly parsed frequent SyntRels

Juan 'John's brother'. Thus, even if the two syntactic constructions seem to be the same (the governor is a noun, the dependent is a preposition, and the dependent of it is a proper noun), only the attributive dependent can be replaced by an adverb, and only the oblique objective is introduced by a preposition which cannot be changed (i.e., a governed preposition; in this case, *de* 'of'). As far as the SyntRels in Table 5 are concerned, an appositive (and even an adverbial in some cases) can also be confused with them: *nebulosa*-appos→*de Orion*, lit. 'nebula of Orion'. The other SyntRels that share the same N-Prep-N configuration are: *abs-pred*, *obl-obj2*, *obl-obj3*, and *noun-compl*; all of these SyntRels obtain poor scores in the evaluation of both parsers. Similarly, the only difference between adverbials and adjunctives is that adjunctives operate at a sentential level while the scope of adverbials is restricted to their governor: [*por ejemplo*]←adjunct-,*-funciona-*,-adv→ *con una silla*, lit. 'for instance, it-works, with a chair'. The two dependents of the verb are prepositional groups that could be found in any position of the sentence; in other words, there is no superficial clue that would differentiate one from the other.

This general absence of clear distinctive features for each particular SyntRel makes it hard for the parsers to find patterns in their learning phases. Grouping the SyntRels with similar configurations is the main factor that makes the parsers improve. In the next subsection, we give more details about the groupings made in the 60 label tagset.

4.2 Detailed analysis of the evaluations results

In this subsection, we take a close look at the SyntRels which trigger the decrease of performance of the parsers between the tagsets containing 44 and 60 labels, respectively. In order to make an adequate comparison of the tagsets, we calculate the weighted average (*WA* in Table 6) of the grouped relations and compare it with the score of the corresponding single edge label in the smaller tagset. We focus on the comparison between those two tagsets, given that the LAS variation of the parsers trained on them is higher than when trained on any other pair of tagsets. Table 6 does not show the results for the relations that have a one-to-one correspondence in both tagsets: *abs-pred*, *det*, *quant*, *compl-adnom*, *appos*, etc. This is because we observed that these relations show the same figures, or their figures only slightly improve or decrease from one tagset to another. In the end, these relations as a whole have almost no impact on the difference between the results obtained with the two tagsets. Instead, the two tables show the relations from the 60 relation tagset which are grouped together in the 44 relation tagset. Among them, only one grouping (*copred* for both parsers) does not lead to a better performance of the parser (16.67%, against 18.75% in average when separated into *obj-* and *subj-copred* for Bohnet, and 16.67% in both configurations for Che). The low number of occurrences of the relations grouped in *copred*, 25 in total, does not allow for a more profound analysis.

For all other relations in the 60 relation tagset, the weighted average in Bohnet and Che is significantly lower than the score of their corresponding group label in the 44 relation tagset:

SyntRels (60)	train #	test #	$LAS_{Boh/Che}$ (%)	$WA_{Boh/Che}$ (%)	SyntRels (44)	$LAS_{Boh/Che}$ (%)
iobj1	46	7	0/0			
iobj2	195	13	30.77/15.38	19.05/5.13	**iobj**	28.57/57.14
iobj3	1	1	0/0			
iobj-clitic1	81	5	20/40			
iobj-clitic2	262	21	76.19/61.9	62.96/55.55	**iobj-clitic**	81.48/77.78
iobj-clitic3	5	1	0/0			
obl-obj1	3551	384	50.78/26.82			
obl-obj2	662	62	20.97/8.06	52.24/26.58	**obl-obj**	71.1/73.57
obl-obj3	17	2	50/0			
noun-compl	1912	199	64.82/32.16			
compl1	141	9	66.67/77.78	50/45	**compl**	70/65
compl2	121	11	36.36/18.18			
aux-refl-pass	405	43	62.79/62.79			
aux-refl-lex	625	69	84.06/42.03	72.27/49.64	**aux-refl**	92.44/91.6
aux-refl-dir	102	7	14.29/42.86			
adjunct	830	87	37.93/31.03			
adv	5751	549	62.3/56.83	65.91/59.51	**adv**	69.64/67.71
restr	1913	194	88.66/79.9			
obj-copred	36	3	0/66.67	18.75/16.67	**copred**	16.67/16.67
subj-copred	76	9	25/0			

Table 6: Comparison between 60 and 44 SyntRels for Bohnet's and Che's parser

- *iobj1*, *iobj2*, and *iobj3* give an average weighted LAS of 19.05% and 5.13% for the two parsers, whereas when they are grouped under one single label *iobj*, the LAS reaches 28.57% and 57.14%; in other words, the LAS drops 9.52 and 52.01 points respectively when training with the most fine-grained relations relations.
- The weighted average of *iobj-clitic1*, *iobj-clitic2*, and *iobj-clitic3* is 18.52 / 22.23 points lower than when those labels are grouped under the generic label *iobj-clitic*.
- The weighted average of *obl-obj1*, *obl-obj2*, *obl-obj3* and *noun-compl* is 18.86 / 46.99 points lower than when they are grouped under the label *obl-obj*. There are 647 instances of this relation in our test set, which means more than 8% of the total number of edges. This subset of SyntRels is largely responsible for the bigger drop of Che when trained with 60 relations.
- For *compl1* and *compl2*, the drop is also important compared to when they are grouped under *compl*: exactly 20 points for both parsers;
- The different types of reflexive auxiliaries that appear in the test set (passive, lexical, and direct) also work much better as one single label *aux-refl*: when they are separated, the LAS drops 20.17 and 41.96 points.
- Finally, for the other very important group by the number of instances in the test set (more than 10% of the edges), the comparison is similar, even if the amplitude is more reduced: *adjunct*, *adv* and *restr* see their LAS 3.73 and 8.2 points inferior to the LAS of the generic label *adv*, which includes them all in the 44 label tagset. Here too the drop is more important for Che than for Bohnet and largely accounts for the global LAS as seen in Table 2.

The performance drop of the 60 relation tagset when compared to the 44 relation tagset could, actually, be expected since some relations of the 60-tagset not only have superficially identical configurations (see Section 4.1), but the properties that differentiate them are closely related to semantics: the different kinds of oblique objects, completives, or reflexive auxiliaries actually behave among each other extremely similarly at the syntactic level, but reflect very distinct

semantic realities. In fact, the number appended to the oblique object relation label not only stands for the order by default in a neutral sentence (with all the objects being present), but it also directly correlates with the slot in the valency pattern of the governor occupied by the corresponding dependent.[9] Although there is a relation between the default order of the objects and their (semantic) numbering, when several oblique objects of the same verb are used at the same time, there usually are information structure features that constrain their order. As a result, the objects are never instantiated in the same order, and the parser has almost no clue for guessing to which slot to assign an object.

From the bird's eye view of the composition of SyntRel-tagsets, it seems that grouping together SyntRels based on their syntactic properties helps the parsers. But not all relation groupings turn out to be beneficiary for the performance of the parsers. Consider the relations that connect two parallel clauses related by a coordination conjunction: *juxtapos*, *quasi-coord* and *coord*. In the 60 and 44 label tagsets, those three SyntRels are kept separated, and the average weighted LAS is 71.5% and 72.58% for Bohnet, and 61.85% and 68.63% for Che respectively. When *juxtapos* and *quasi-coord* are grouped in the 31 label tagset, Bohnet drops by more than 2 points to 70.31%, while Che slightly rises to 69.33%. However, when *coord* is also grouped with the other two under the label *COORD*, both parsers have more difficulties: Bohnet drops by one point and Che by more than six points. We believe that with these three SyntRels, the syntactic constructions at stake are too different for the parsers to be able to find strong common features: a juxtaposition involves a punctuation sign (colon or semi-colon), while a coordination involves a conjunction or a comma, and a quasi-coordination nothing but the two coordinated elements (e.g.*¡Estoy aquí*-,-quasi-coord→ *en mi cuarto!*, lit. 'I'm here, in my room!'). Therefore, we believe that even if it is tempting to annotate with a same label any coordinate structure, it is better to keep the different types annotated with different labels.

5 Conclusions

The evaluation of the performance of four state-of-the-art parsers trained on a corpus that was annotated following schemes of different granularity revealed that the loss of accuracy as a consequence of the increase of the size of the tagset, in particular, from 15 to 44 tags, is surprisingly small. This outcome supports the claim that an annotation with more fine-grained syntactic relations does not necessarily imply a significant loss in accuracy. It also supports the argumentation that it is useful to compile a detailed annotation scheme, which then allows for the derivation of a variety of more or less detailed annotations. Our study also suggests that there seems to be a limit with respect to the degree of detail of the tagset beyond which a parser's accuracy suffers significantly, and that there are some tags which provoke a drop of the LAS more than others. These are, in particular, the very fine-grained divisions which directly reflect semantic valency information. Another conclusion that can be drawn is that training a parser on a fine-grained annotation does not lead to a better performance of this parser when parsing with a coarse-grained tagset. However, it still remains unclear whether the unlabeled attachment score can improve when training on a fine-grained annotation. Experiments with more data would be necessary in order to draw more solid conclusions.

Acknowledgments

We would like to thank B. Bohnet and the anonymous reviewers for their very helpful comments.

[9] This goes along the lines of Bosco et al. (2010), who mention that semantic distinctions are problematic in their experiments, and that merging locative and temporal complements under the same label, for example, increases the f-scores of the parsers.

References

Bohnet, B. (2009). Efficient Parsing of Syntactic and Semantic Dependency Structures. In *Proceedings of CoNLL '09*, pages 67–72, Boulder, Colorado, USA.

Bosco, C. and Lavelli, A. (2010). Annotation Schema Oriented Evaluation for Parsing Validation. In *Proceedings of TLT-9*, pages 19–30, Tartu, Estonia.

Bosco, C., Lombardo, V., Vassallo, D., and Lesmo, L. (2000). Building a Treebank for Italian: a Data-Driven Annotation Schema. In *Proceedings of LREC '00*, pages 99–105, Athens, Greece.

Bosco, C., Montemagni, S., Mazzei, A., Lombardo, V., Dell'Orletta, F., Lenci, A., Lesmo, L., Attardi, G., Simi, M., Lavelli, A., Hall, J., Nilsson, J., and Nivre, J. (2010). Comparing the Influence of Different Treebank Annotations on Dependency Parsing. In *Proceedings of LREC '10*, pages 1794–1801, Valletta, Malta.

Briscoe, T., Carroll, J., Graham, J., and Copestake, A. (2002). Relational Evaluation Schemes. In *Proceedings of the Workshop at LREC '02 on Beyond PARSEVAL: Towards Improved Evaluation Measures for Parsing Systems*, pages 4–6, Gran Canaria, Spain.

Buchholz, S. and Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of CoNLL '06*, pages 149–164, New York City, USA.

Burga, A., Mille, S., and Wanner, L. (2011). Looking Behind the Scenes of Syntactic Dependency Corpus Annotation: Towards a Motivated Annotation Schema of Surface-Syntax in Spanish. In *Proceedings of DepLing '11*, pages 104–114, Barcelona, Spain.

Che, W., Li, Z., Li, Y., Guo, Y., Qin, B., and Liu, T. (2009). Multilingual Dependency-based Syntactic and Semantic Parsing. In *Proceedings of CoNLL '09: Shared Task*, pages 49–54, Boulder, Colorado, USA.

Gesmundo, A., Henderson, J., Merlo, P., and Titov, I. (2009). A Latent Variable Model of Synchronous Syntactic-Semantic Parsing for Multiple Languages. In *Proceedings of CoNLL '09: Shared Task*, pages 37–42, Boulder, Colorado, USA.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., and Zhang, Y. (2009). The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. In *Proceedings of CoNLL '09: Shared Task*, pages 1–18, Boulder, Colorado, USA.

Johansson, R. and Nugues, P. (2007). Extended Constituent-to-dependency Conversion for English. In *Proceedings of NODALIDA '07*, pages 105–112, Tartu, Estonia.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *Proceedings of ACL '03*, volume 1, pages 423–430, Sapporo, Japan.

Kübler, S. (2005). How Do Treebank Annotation Schemes Influence Parsing Results? Or How Not to Compare Apples And Oranges. In *Proceedings of RANLP '05*, pages 293–300, Borovets, Bulgaria.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marneffe, M.-C. D., MacCartney, B., and Manning, C. D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of LREC '06*, pages 449–454, Genoa, Italy.

Mille, S., Burga, A., Vidal, V., and Wanner, L. (2009). Towards a Rich Dependency Annotation of Spanish Corpora. In *Proceedings of SEPLN '09*, pages 325–333, San Sebastian, Spain.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering Journal*, 13(2):99–135.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of COLING/ACL '06*, pages 433–440, Sydney, Australia.

Rehbein, I. and van Genabith, J. (2007). Treebank Annotation Schemes and Parser Evaluation for German. In *Proceedings of EMNLP-CoNLL '07*, pages 630–639, Prague, Czeck Republic.

Taulé, M., Martí, M. A., and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the LREC '08*, pages 96–101, Marrakesh, Morroco.