Universal Dependencies for Portuguese

Alexandre Rademaker
IBM Research and EMAp/FGV, Brazil
alexrad@br.ibm.com

Fabricio Chalub
IBM Research, Brazil
fchalub@br.ibm.com

Livy Real
University of São Paulo, Brazil
livyreal@gmail.com

Cláudia Freitas
PUC-Rio, Brazil
claudiafreitas@puc-rio.br

Eckhard Bick
University of Southern Denmark, Denmark
eckhard.bick@mail.dk

Valeria de Paiva
Nuance Communications, EUA
valeria.depaiva@nuance.com

Abstract

This paper describes the creation of a Portuguese corpus following the guidelines of the Universal Dependencies Framework. Instead of starting from scratch, we invested in a conversion process from the existing Portuguese corpus, called Bosque. The conversion was done by applying a context-sensitive set of Constraint Grammar rules to its original deep linguistic analysis, which was carried out by the parser PALAVRAS, with some additional manual corrections. Universal Dependencies offer the promise of greater parallelism between languages, a plus for researchers in many areas. We report the challenges of dealing with Portuguese, a Romance language, hoping that our experience will help others.

1 Introduction

The Universal Dependencies (UD) project,[1] in its ambitious and encompassing mission of providing a single set of tags and parallel analyses common to several different languages, not only provides for a multilingual natural language processing (NLP) framework, but also allows the representation of specific features of each language and this motivates our interest in participating in the project. Since it is a well documented project, we asked ourselves to which extent the general UD guidelines were enough to represent the features of each individual language, in particular we asked ourselves whether they were enough to properly represent the grammatical features of Portuguese.

The release of the UD treebanks version 1.2, in November 2015, was the first release to include a Portuguese treebank. The `UD_Portuguese` treebank is based on the corpus Bosque, part of the Floresta Sintá(c)tica project (Afonso et al., 2002), version used in the CoNLL-X Shared Task in dependency parsing (2006); the CoNLL version was taken and converted to the Prague dependency style as a part of HamleDT (since 2011). Later versions of HamleDT added a conversion to the Stanford dependencies (2014) and to Universal Dependencies (HamleDT 3.0, 2015). The conversion path from the original Bosque still goes through the CoNLL-X format and the Prague dependencies, which may occasionally lead to loss of information. In the release 1.3 of UD, in May 2016, one additional Portuguese treebank was added, the `UD_Portuguese-BR`, a conversion of the original work of (McDonald et al., 2013), as per the description in (et al., 2016).

This paper describes the consolidation of the `UD_Portuguese` treebank in the UD Framework. For that, between September 2015 and March 2016, a set of UD conversion rules for the CG input was written, as described in (Bick, 2016), and applied to the updated version of the dependency-style Bosque (Linguateca version 7.5 of March 2016). For a team effort starting in October 2016, we were given a version of the this converted corpus, and through consistency-checking and discussion, aiming at full compatibility with UD specification, converged to a further round of manual treebank corrections and conversion rules

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

changes. The first version of our data, fully UD 1.4 compliant, was include it in the UD release 1.4 with the name `UD_Portuguese-Bosque`. Latter, motivated by the inclusion of Portuguese language on the 'Multilingual Parsing from Raw Text to Universal Dependencies' CoNLL 2017 Shared Task, we accepted the challenge to update `UD_Portuguese-Bosque` to UD 2.0 guidelines and replace the previous `UD_Portuguese` corpus. This paper describes the technical and linguistics hurdles of the conversion and of the management of the different versions of the corpus Bosque available. The Conference on Computational Natural Language Learning (CoNLL), has a long history of shared tasks in which training and test data are provided by the organizers, allowing participating systems to be evaluated and compared in a systematic way.

Many reasons supported our decision to re-use the Bosque corpus, instead of creating an entire new corpus from scratch. The Bosque corpus -– created and maintained by Linguateca [2] -– was already annotated with dependencies and was manually revised, saving us time. Besides, it was already used in previous editions of CONLL – CONLL-X Shared task on Multilingual Dependency Parsing (Buchholz and Marsi, 2006) –, and it is distributed in different versions, annotated with different tagsets and formats.[3] The existence of different versions of the same material fosters the study about different tagsest and its impacts in NLP systems. Finally, the fact that we had on the team two researchers who had already worked on previous versions of Bosque also contributed to this choice. However, the conversion to UD scheme was much more complicated than initially planned.

Different tagsets usually correspond to different reifications of grammars, which indicates different conceptualizations of a language. For this reason, a conversion of tagsets is rarely a purely mechanical task of substitution. In our improved conversion, we address both structural links (dependencies labels) and part-of-speech tagsets, following the Universal Dependencies guidelines for version 2.0. This conversion also deals with phenomena that needs manual revision, such as apposition, copular sentences and multiword expressions (MWE) structures, among others.

We first describe how and why we chose the corpus we decided to work from, then we describe the process we used to improve this data. Very many small and not so small decisions were taken along the way, and we try to recap and explain the main ones, why they are important for the specific language we are dealing with (Portuguese) and how they impact our continued plans for Portuguese NLP. We finish with preliminary conclusions on the state of this data and the the tasks ahead.

2 The Bosque versions

The Bosque corpus is a subset of the Floresta Sintá(c)tica (*syntactic forest*) treebank, first described in (Afonso et al., 2002). 'Bosque' means 'woods' in Portuguese. It consists of news running text from both Portugal and Brazil, chunked into sentences, syntactically analyzed in tree structures, making use of both automatic parsing, PALAVRAS (Bick, 2014) and fully revised by linguists.

Over its 15-year history, the corpora from Floresta Sintá(c)tica have spawned several format conversions, resulting in a somewhat complex mix of editions. The original text corpora were processed with PALAVRAS, a rule-based Constraint Grammar (CG) system (Karlsson, 1990) designed specifically for Portuguese. The parser produces deep linguistic analyses, with tags at the morphological, syntactic and semantic levels. Despite CG's native dependency tags, the first published version of the Floresta treebank opted for constituent trees.

From 2006–2008, the Floresta treebank were enriched with additional tags for cross-token morphology (e.g. definiteness and complex tenses) and some semantics, derived from a re-annotation with an improved PALAVRAS parser. The PALAVRAS native dependency annotation was retained, and aligned with the hand-corrected constituent version. The constituent version was then revised up to version 8.0 (Freitas et al., 2008),[4] while the dependency version was used for on-

[2] http://www.linguateca.pt

[3] There is the original Bosque tagset and the CONLL 2006 tagset; there is also the CG (constraint grammar, (Karlsson, 1990)) format, the AD format (phrase structure tree), the graphical and tgrep format, the Penn TreeBank and TIGER fomat. All these versions are available from `http://www.linguateca.pt/Floresta/download.html` and `http://corpora.di.uminho.pt/linguateca/FS/fs.html`.

[4] `http://www.linguateca.pt/floresta/corpus.html`

going experiments. The first UD_Portuguese treebank (published in 2006, UD 1.2) was also derived from Bosque, as said before, but it was independently converted from the constituent version 7.3 to a dependency version, and it is this version (i.e. without the later revisions in the treebank project itself) that went through a Penn treebank dependency-style conversion as part of HamleDT (2011), then Stanford Dependencies and then UD conversion (HamleDT 3.0).

For our own work, we opted to use the original Bosque treebank from Floresta, converted to UD by (Bick, 2016), rather than the existing CoNLL-U edition of the Bosque (the UD_Portuguese released in the UD 1.2), in part because we wanted to: (a) incorporate changes and additions made to the dependency version of the original treebank after 2006; (b) circumvent possible information loss due to previous conversions; and (c) because we thought that a comparison of the results of two different conversions might yield interesting insights. The most important reason, however, was methodological: we wanted to build a framework where manual revision work and consistency checks could be coordinated with automatic parser annotation and conversion rules. On the one hand, this would allow us to save work by addressing systematic errors, and thus fix them automatically, based on a few examples, rather than repeatedly fixing the same kind of error manually. On the other hand, and more importantly in the long run, we intend to enlarge the treebank, and therefore deem it important to be able to maintain a close link between live parser output and the UD conversion method. One of us is building a parser pipeline with an integrated UD conversion grammar, to support a semi-automatic system of manual revisions and consistency checks, which should allow for an efficient text-to-dependencies creation of new treebank material in the future. We also believe that having the corpus revised by native Portuguese linguists guarantees a better annotation quality, since the conversion from the original Bosque tagset to the UD tagset and relations is far from obvious.

2.1 Annotations: similarities and differences

The conversion grammar ultimately used for the first conversion of Bosque to UD contained some 530 rules. Of these 70 were simple feature mapping rules, and 130 were local MWE splitting rules, assigning internal structure, POS and features to the MWEs from Bosque. The remainder of the rules handled UD-specific dependency and function label changes in a context-dependent fashion (Bick, 2016). The main issues were raising of copula dependents to subject complements, inversion of prepositional dependency and a change from syntactic to semantic verb chain dependency. In one respect, punctuation attachment, the grammar actually went beyond conversion, identifying meaningful head tokens for commas, parenthesis etc., that all had been left unattached in the original Bosque. Figure 1 shows an example of sentence with the original PALAVRAS dependencies (top, simplified) and the resulting UD encoding after the conversion (bottom). The complete PALAVRAS annotation of the same sentence in the *niceline* format is presented below.

```
Esse [esse] <*> <dem> DET M S @>N #1->2
carro [carro] <V> N M S @SUBJ> #2->3
foi [ser] <fmc> <aux> V PS 3S IND VFIN @FS-STA #3->0
achado [achar] <vH> <mv> V PCP M S @ICL-AUX< #4->3
em [em] <sam-> PRP @<ADVL #5->4
o [o] <-sam> <artd> DET M S @>N #6->7
início [início] <temp> N M S @P< #7->5
de [de] <sam-> <np-close> PRP @N< #8->7
a [o] <-sam> <artd> DET F S @>N #9->10
tarde [tarde] <per> N F S @P< #10->8
em [em] <np-close> PRP @N< #11->10
Engenheiro Marcilac [Engenheiro=Marcilac] <civ> <*>
  <heur> <foreign> PROP M S @P< #12->11
. #13->0
```

The new UD treebank retains the additional tags for NP definiteness and complex tenses, as well as the original syntactic functions tags and secondary morphological tags, which makes it a more informative treebank. This way, the treebank keeps its original linguistic focus, but in addition it can be used for the new machine learning scenarios targeted by the CoNLL-U format. To give an example of the usefulness of having the deep, old annotations and the the new ones together, we could mention that, for instance, Bosque tags roots of sentences for their functions, such as question, command or statement. We retain these tags in our conversion. It would be very hard for a shallow dependency representation to recover these differences were they to be erased to begin with and for a question answering application these tags are very useful.

In some cases, the stored original function tags allow the user to recover a valency relation otherwise lost in the underspecified UD edge label, such as the distinction between free adverbial prepositional phrases (e.g. *trabalhar em* (ADV) 'work at' and valency-bound adverbial (e.g. *morar em* (ARG) 'live at').

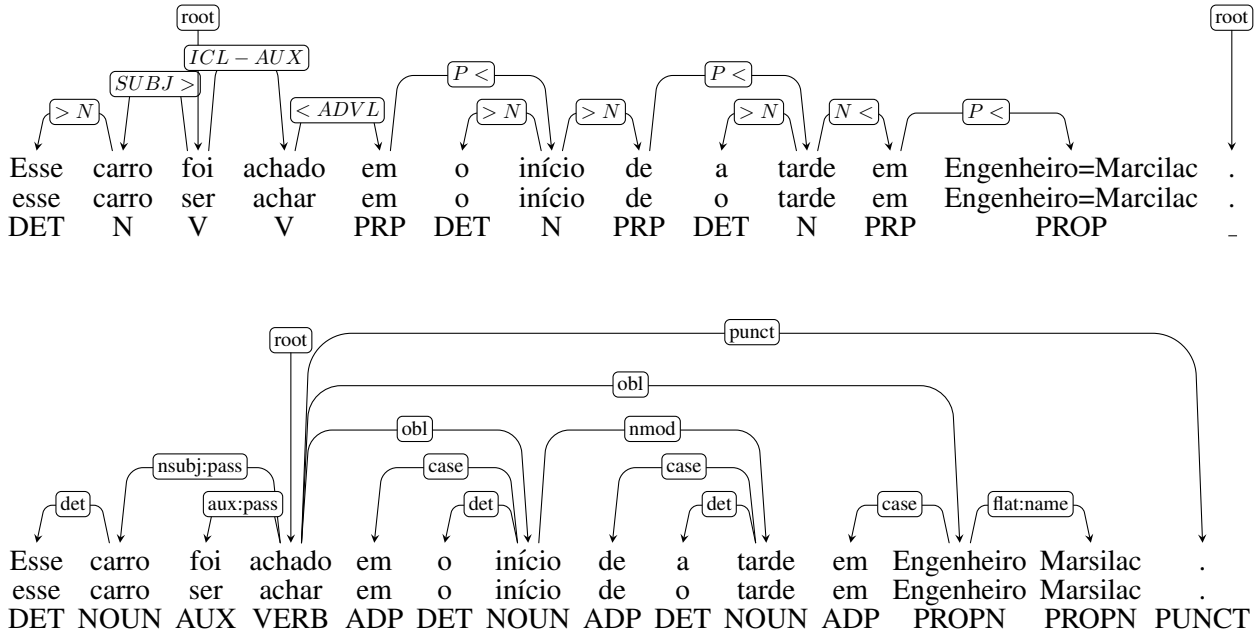


Figure 1: The sentence 'Esse carro foi achado no início da tarde em Engenheiro Marsilac/This car was found in the beginning of the evening at 'Engenheiro Marsilac' (location)' annotated with the parser PALAVRAS and UD scheme.

2.2 Improving the data

Having a version of the corpus committed to a common repository, work started on checking first basic code conventions: do we have empty CoNLL-U representations? Do we have the same number of columns for all sentences? Are we allowed to have many values for a single tag? Do all sentences have a "root" node? Can we enforce the UD requirement that representations are trees?

Then more linguistic questions began to emerge. For example, gender is one of the hallmarks of Romance languages and annotation can be complicated, as some words appear to have an underspecified gender. There are adjectives such as *grande* ('big') or *feliz* ('happy') that have only one form for both genders. So we cannot tell whether they are masculine or feminine unless we see the context they appear in. In many cases, even looking at the full sentence, one cannot tell if the word is masculine or feminine. For example, in the sentence:

CP652-3 Por enquanto, estamos *felizes* só com o reconhecimento implícito ('For now, we are happy with only the implicit recognition')

we have no way of knowing what is the gender of *felizes*. How should these expressions be annotated? After some discussion, it was decided that these cases would be annotated as 'Unsp' (for "unspecified" value) and that a similar annotation would be used for unspecified number too.

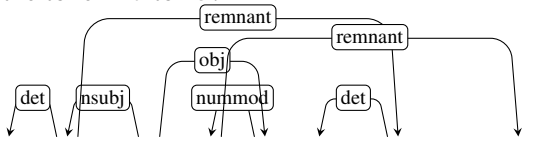
Then the first main issue with the MWEs and the different approaches to their annotation had to be tackled. The PALAVRAS annotation has MWEs tokenized as a single word, but this is not the UD recommendation. The UD version 1 guidelines proposed the dependency relations 'mwe' or 'compound', so a process of dismembering these single token MWEs and assigning each of their components a POS-tag was initiated. Things changed in UD version 2, different tags for MWE are used ('flat', 'fixed' and 'name'), but this conversion could be done automatically.

How to deal with participles was also a challenging issue. PALAVRAS tags all participles as verbs, with the 'PCP' (participle) feature. However, UD guidelines state: "Note that participles are word forms that may share properties and usage of adjectives and verbs. Depending on language and context, they may be classified as either VERB or ADJ."

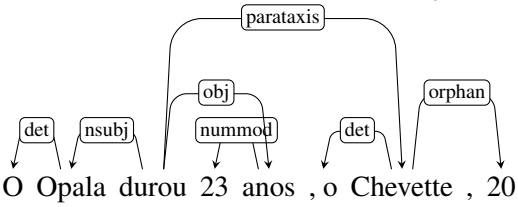
We followed the criteria discussed in (Truggo, 2016) to define participles acting as verbs or adjectives and worked on a set of linguistic rules to semi-automatically re-tag participles.

Another change from UD version 1 to 2 is the

treatment of ellipsis. In version 1, ellipsis cases were dealt with via a 'remnant' dependency relation. This relation linked the core arguments of the ellipsis clause to their corresponding arguments in the complete clause. In the sentence bellow (CF349-2 – 'Opala lasted 23 years, Chevette, 20 [. . .]'), the token *Chevette* was related to *Opala* and the token *20* to *23*.



In UD version 2.0, the 'remnant' relation was discarded and a new treatment was proposed, using a new relation 'orphan'. With this proposal, only the first core argument of the ellipsis clause is related to the main clause and the other core arguments are related to it via the 'orphan' relation. In the example above, *Chevette* is related to *durou* 'lasted' (the root of the main sentence) via 'parataxis' and *20* is related to *Chevette* via 'orphan'. All 'remnant' cases were manually fixed.



Also there were many minor discrepancies, like Bosque used 'pret' (for preterite), while UD used 'past', so we had some "unknown attribute-value pairs" to translate. Using the UD provided scripts and manual checking, the validator script was satisfied with the representations and we could start thinking about similarities and differences to the other version of the Bosque, which we discuss in the next section.

3 Portuguese annotation choices

Clear and detailed guidelines are the crucial data in annotation projects. It is reasonable to expect that the UD guidelines would be, as they are, less specific than we would like them to be. Their main motivation is to be universal, so special characteristics of the target language are to be down-played, for the sake of being able to compare features in other languages. However, this lack of specificity of the guidelines makes somewhat more explicit the interpretative dimension of linguistic analysis.

In this section we discuss some of the issues that we consider interesting, either because they were not sufficiently described in the guidelines, or because they are issues that seem mainly important for Portuguese.

3.1 Tokenization

While the first conversion grammar did convert syntactic to semantic (UD) dependencies and function-based edge labels to form-based (UD) edge labels, it did not handle UD's space-based tokenization, maintaining the original treebank's MWE (e.g. complex conjunctions, prepositions and named entities) and its - syntactically motivated - splitting of Portuguese contractions (preposition plus article/determiner/pronoun, e.g *neste* to *em* + *este* ('in this'). Linguistically, the problem in token-splitting is the need to assign (a) partial POS tags, (b) additional internal dependency links and (c) new internal hook-up points for existing outgoing and incoming dependency links. Unlike simple label conversion for, say, morphological features, this cannot be achieved with a systematic conversion table only.

Our solution was to use CG-based retokenization rules. Its most recent implementation (Bick and Didriksen, 2015) offers context-based manipulation (removal, substitution, addition etc.) of not only tags, but also of entire (annotated) tokens. We used this feature to split MWE tokens into their sub-words, while at the same time adding the missing POS, features, edge labels and dependency links to the individual parts.

This solves the problem that while the UD treatment of MWEs considers each part of an MWE as a single POS, the set of words that compose a given MWE may not contain a word that has the same POS tag as the MWE as a whole. The MWE *ao vivo* ('live'), for instance, is an ADV as a whole, while 'ao' is a contraction (ADP 'a' + DET 'o') and 'vivo' 'live' is an ADJ. Since it is clear that the most important information for the entire sentence structure is the POS tag of the whole MWE, and not the POS tag of each of its constituents, we keep a tag for the whole MWE in our representation. Then, at least for the Portuguese UD corpus, both the internal structure and the functional POS tag of a MWE are available. In the same fashion, CG rules can be used to fuse Portuguese contractions that were split in Bosque (*dos* 'of the', *pelas* 'by the', *nisto* 'in this'), assigning them a compound pos and joint external dependency links.

Another issue related to tokenization is the problem of clitics in Portuguese. As other Romance languages, Portuguese has enclisis and proclisis. Moreover, in Portuguese we have mesoclitics, that is, clitics that come inside the verb and change the verbal structure:

CP895-1 *Poder-se-á* dizer que o estilo resulta da sua profissão, fotojornalista. ('It can be said that the style results from his profession, photojournalist.')

After some discussion, we decided to follow the traditional Portuguese grammars. Mesoclitics seem to us a language specific issue that maybe each group dealing with an UD specific language corpus should manage on their own. Guidelines seem to be emerging that consider mesoclisis as two syntactic words: a verb plus a pronoun. In the example above, *poder-se-á* is *poderá*/VERB followed by *se*/PRON' ('it can' in the future plus the reflexive). The surface form *poder-se-á* is still present in the tree analysis as a multi-word token.

3.2 The particle 'se'

The analysis of the particle 'se' is well-known as a complex phenomenon in Portuguese. Traditionally, besides being a conjunction, the particle appears in:

(*a*) **reflexive and reciprocal constructions** *CF314-2 Você se acha louca?* (Do you think you are crazy?);

(*b*) **pronominal verbs** *CF340-2 O ciclista espanhol, 48, se suicidou em Caupenne d'Armagnac, no sul da França com um tiro.* (The Spanish cyclist, 48, killed himself in Caupenne d'Armagnac, south of France, with a single shot.);

(*c*) **pronominal passive voice** *CF32-2 - Primeiro aprova-se o texto enxuto e depois negocia-se a aprovação, sem prazo definido, das leis complementares e ordinárias.* (First, the short text is approved and then, without a definite deadline, the approval of the complementary and ordinary statutes is negotiated.);

(*d*) **undeterminate subject constructions** *CP263-3 Pense-se em Kingsley Amis, Malcolm Bradbury e Albert Finney.* (One can think of Kingsley Amis, Malcolm Bradbury and Albert Finney.)

The difference between (*c*) and (*d*) above, discussed in traditional grammars and textbooks, has gradually been substituted for an analysis that takes as primary the non-determination of the subject in both cases. The example in (*c*) corresponds to *Primeiro, alguém aprova o texto e depois alguém...* ('First someone approves the text and after that someone...'). This is to be compared to *Primeiro, o texto é aprovado e depois a aprovação é negociada...* ('First the text is approved and then the approval is negotiated...'). This means that we consider equivalent the analyses where 'se' assumes the function of the subject, which one cannot or does not want to make explicit. A strong argument for this interpretation is the lack of verbal concordance, the verb remaining in the singular form, even in formal registers, in some traditional examples such as *Vende-se casas* 'Houses are sold'. In this case, the verb *vender* ('sell') must be a plural form (*vendem*), to agree with the plural *casas* 'houses', but the actual use is *Vende-se casas*.

In the context of the universal dependencies this indicates that in both cases (*c*) and (*d*) we could have the particle *se* as the subject of the verb, although the subject remains non-explicit. This analysis would have the advantage of making uniform constructions that the speakers of Portuguese tend to consider the same. Nonetheless, according to UD guidelines, this analysis should be avoided: "The 'nsubj' role is only applied to semantic arguments of a predicate. When there is an empty argument in a grammatical subject position (sometimes called a pleonastic or expletive), it is labeled as 'expl'. If there is then a displaced subject in the clause, as in the English existential 'there is' construction, it will be labeled as 'nsubj'. The UD annotation creates a certain uniformity between the cases (*b*), (*c*) and (*d*). Since we consider relevant the distinction between (*b*) (which has an explicit subject) and (*c*) and (*d*) (which do not), we keep this information. Thus, to keep the additional information, cases (*c*) and (*d*) carry the label SUBJ_INDEF in the MISC field.

3.3 Additional annotations

In the corpus, we use extra fields to keep the linguistic information that we have from the parsing analysis and that we would not like to lose, even if this information is not used by the UD project presently. The CoNNL-U field MISC (miscella-

neous) is also used to keep any information that is not reported in the other fields. The indefinite subject, cited above, is one example of use of that field. Another information we keep in the MISC is the POS tags of MWE, which we had to unpack for this annotation task as described in the Section 3.1.

The indication of the POS tags in the case of 'fixed' MWEs is particularly relevant, as these expressions are crystallized in such way that their components can have completely different POS tags from the total expression. Having the information about the POS-tag of the entire MWE in the MISC field helps to justify some dependency relations. In the example already mentioned, the expression *ao vivo* is a MWE with POS-tag 'adv', although it is not composed by adverbs.

3.4 Negation

The treatment of negation has changed from UD version 1 to 2. In the earlier version, a dependency relation 'neg' was used to link a negative word, such as *não* ('not'), to its head. In the UD version 2, a polarity feature was introduced ('Polarity=Neg') to keep the negative information and the 'neg' relation was removed from the set of universal relations.

We give negation in Portuguese a different treatment than other UD corpora. In Portuguese, negation is commonly expressed with the word *não*. This word cannot be contracted and it behaves exactly like any other adverb. Traditional grammars of Portuguese state that *não* is always an adverb and we agree with this analysis. Because of this, the negation treatment we propose is slightly different from the one proposed by the universal guidelines. We understand *não* – and other words as some uses of *nada* ('nothing') – as adverbs. Therefore one should be prepared to find in the corpus fewer words tagged with the POS tag PART than in other corpora, such as the English and the French tree banks.

Another interesting aspect of negation in Portuguese is the issue of double negation, which is pervasive in Portuguese. For example in the sentence:

CP153-4 Não estava nada à espera disto. ('[I] was not waiting nothing for it.')

We tagged both the main negation, *não* in the sentence above and the second element of the negation *nada* as adverbs. Sometimes we tag the second negative in a double negation as a pronoun, depending on the kind of structure they are in. In the example above, *nada* was tagged as an adverb, since *nada* here could be replaced by another adverb, for example *pouco* ('little') or *muito* ('much'). In other cases of double negation, the second element of the negation can be seen as a direct object of the negated verb:

CP778-11 A coincidência de funerárias e queijarias na nossa circunstância não significava nada [...] ('The coincidence of mortuaries and cheesemakers in our circumstances did not mean nothing [...]')

In those cases, *nada* ('nothing') is indeed the direct object of the verb, and therefore it was tagged as a pronoun (PRON) and it has the 'obj' relation with the verb.

For those interested in double negations in Portuguese, the best way to look for them in the current `UD_Portuguese` corpus will be to check for the polarity feature ('Polarity=Neg') expressed in words that surround the verbs. We expect that the consistent use of the polarity feature in adverbs, pronouns, conjunctions, as *nem* ('neither'), and others will provide us with a full analysis of this phenomenon without loosing the surface syntactic analysis provided by the UD relations.

3.5 Appositives

In our conversion process, we have chosen – so far – to take into account the classic and comprehensive notion of appositives (non-restrictive and restrictive) (Biber et al., 1999), since a) this was already the original analysis provided by PALAVRAS; b) this is a gray area of the UD guidelines; c) in our view, the decision favors consistent analysis. According to UD guidelines, the 'appos' relation "serves to define, modify, name, or describe that noun" [5]. Combinations like *president Obama* would be 'appos' (restrictive appositive), if we agree that *Obama* describes, defines or modifies *president*. Yet UD guidelines explicit state that cases like *president Obama*, or *state senator Paul Mnuchin* should not be considered appositives, since the impossibility of the reversal

[5] It is interesting to note how this definition, essentially semantic, overlaps with the 'amod' definition ("serves to modify the meaning of the noun."). But we will not explore this point here.

(**Paul Mnuchin state senator*) indicates the presence of one and only nominal. However, guidelines also recognize that there are always borderline cases. In the sentences *I met the French actor Gaspard Ullie* and *I met Gaspard Ulliel the French actor*, the reversal indicates, in both sentences, the presence of apposition between *actor* and *Gaspar Ulliel*. It is not clear to us why *I met the president Obama* should receive a different analysis. So this cases were also tagged as 'appos' in our corpus, but we recognize the issue is still open.

4 Bosque UD in numbers

The Bosque corpus consists of 9.368 sentences and 227.653 tokens, with 18.140 unique lemmas. In Table 1 we present the frequency of all 17 UD POS tags in the corpus. The POS tag 'X' is used for foreign words. At the moment we still have 957 'dep' relations (Table 2), which we want to investigate, since this dependency is mostly used when no other relation is applicable. We also plan to check the coverage of the classes of verbs, nouns, adjectives and adverbs, against OpenWordNet-PT.[6]

5 Improving Bosque analyses

To allow us to analyze the representations and the effects of the automatically applied choices in the pipeline, we feed the result of processed sentences to the interface developed and distributed by the Turku BioNLP Group (Luotolahti et al., 2015).[7] This has been very helpful, as one can tell immediately how big the issues are within the corpus.

The UD project provides a validation script that allows us to check some basic generic facts, such as that every sentence has a root and that CoNLL representations have always the same number of fields or that there are no multiple values for the same tag. Some of these are mandatory, a corpus needs to be validated to be part of the distribution. But more sophisticated constraints, both on the level of POS tags and of dependencies, can also be checked. The Turku search tools make use of a sophisticated query language, with Boolean operators that helps ascertain whether the treebank satisfies some more semantic properties too.

In the course of the project, we have also started developing our own library for dealing with

tag	count	examples
ADJ	11560	grande, novo, primeiro, bom, último, político, pequeno, próximo, segundo, passado
ADP	36614	de, em, a, por, para, com, como, entre, sobre, sem
ADV	8742	não, mais, já, também, ainda, ontem, como, só, quando, depois
AUX	6315	ser, estar, ter, poder, ir, dever, vir, continuar, começar, acabar
CCONJ	5222	e, mas, ou, nem, quer, mais, &, tampouco
DET	35076	o, um, seu, este, todo, outro, esse, muito, algum, mesmo
INTJ	43	não, rarará, é, adeus, ah, ai, alô, basta, bem, bingo
NOUN	41353	ano, dia, milhão, país, presidente, empresa, pessoa, vez, tempo, estado
NUM	4312	um, dois, três, mil, cento, quatro, cinco, 15, 30, seis
PART	4	anti, ex, pré, pós
PRON	7236	que, se, ele, o, eu, ela, isso, quem, eles, tudo
PROPN	18984	Paulo, Portugal, Brasil, José, Porto, Governo, Nacional, Lisboa, EUA
PUNCT	29983	`,`, `.`, `...`, `(`, `)`, `,`, `:`, `?`, `;`
SCONJ	2201	que, se, porque, embora, pois, como, caso, assim, e, senão
SYM	415	%, US, R, CR$
VERB	19482	ter, fazer, dizer, haver, dar, ser, ficar, ver, ir, querer
X	136	in, pole, position, body, dream, jet, shopping, art, center, centers

Table 1: POS tags in Bosque

[6] The open wordnet for Portuguese, `http://openwordnet-pt.com/`.

[7] `https://github.com/fginter/dep_search`

rel	count	rel	count
acl	2930	flat	11
acl:relcl	2562	flat:foreign	71
advcl	2440	flat:name	5832
advmod	8461	iobj	236
amod	8732	mark	4724
appos	3272	nmod	26493
aux	2444	nmod:npmod	473
aux:pass	1125	nmod:tmod	193
case	33170	nsubj	10958
cc	5263	nsubj:pass	976
ccomp	1567	nummod	2853
compound	536	obj	8211
conj	6145	obl	4933
cop	2748	obl:agent	727
csubj	376	orphan	8
dep	957	parataxis	463
det	34942	punct	29986
discourse	13	reparandum	1
dislocated	9	root	9368
expl	948	vocative	14
fixed	607	xcomp	1900

Table 2: The dependency relations in Bosque

CoNLL-U files. The `cl-conllu` library is implemented in Common Lisp, it is open-source and freely available.[8] Since we have not yet decided in our group to use any particular dependencies editor, we also implemented an online CoNLL-U validation service. [9]

6 Comparison and Assessment

As we said in the introduction, one of the reasons for working with the same Bosque corpus, already available in UD release 1.2, was to be able to compare conversions. Some big discrepancies in numbers, as computed by the statistics script, were easy to see. For instance, it was clear that in our version had many more cases of auxiliary verbs than `UD_Portuguese` in UD 1.2. The difference is probably due to the fact that, in Portuguese, verbs like *continuar* (to continue), *começar* (to start) and *acabar* (to end) can also be seen as modal auxiliaries, and that was our decision. In the previous `UD_Portuguese` corpus from UD 1.2, such verbs were considered full verbs:

CP269-3 O soldado disparou para o ar, mas o indivíduo **continuou** a avançar e foi atingido mortalmente. (The soldier fired into the air, but the individual continued to advance and was struck deadly.)

On the other hand, we found that our version of the Bosque had many more cases of apposition dependencies ('appos'). In addition to our choice to include restrictive appositives under the tag 'appos', the main difference in numbers reflects different choices in the alignment-conversion process. In the annotation provided by PALAVRAS, the syntactic function `@N<PRED` (non-identifying apposition) can and should be converted into *appos* but, in the `UD_Portuguese` UD 1.2, all these cases were converted into 'nmod' (see Table 3). In the sentence below, there is an 'appos' relation between *diretor* (director) and *Ailton Reis*, but in the first automatic conversion, the relation was 'nmod'.

CF103-4 Os documentos foram encontrados em papel ou retirados de disquetes apreendidos em a casa de **Ailton Reis**, **diretor** da Odebrecht. (The documents were found on paper or removed from diskettes seized at Ailton Reis' house, director of Odebrecht.)

When we looked for the 'appos' relation, considering the possible cases of different POS tags pairs being related, we were surprised to find around 50 possibilities of POS tag pairs being related through the 'appos' relation.

Corpus UPOSTAG	UD PT 2.0 (appos)	UD PT 1.2 (nmod)
PROPN	234	218
NOUN	961	935

Table 3: Cases of `@N<PRED` from PALAVRAS annotation.

One relevant difference between our version and the previous `UD_Portuguese` version is that all contractions are introduced also as a multiword token, allowing one to know the surface structure of the sentence easily. The process of retokenization of these contractions made us realize many mistakes in the annotation of these contractions. For example, 'a' is a preposition but also a determiner (definite article) and, in Portuguese, two definite articles do not occur contiguously, so

[8] https://github.com/own-pt/cl-conllu

[9] https://github.com/own-pt/conll-workbench

we could easily correct, in contractions, all cases where the preposition 'a' (that should be annotated as ADP) was wrongly annotated as a determiner ('det'). Our version also keeps the raw text of all annotated sentences.

7 Conclusions

We described how we took an existing corpus, produced for us by a careful, context-sensitive conversion process using a Constraint Grammar framework, and managed to validate it, using the UD guidelines versions 1 and 2.

This required extensive work, mainly dealing with contractions (a widespread phenomenon in Portuguese) and with multiword expressions (a universal problem). We had to re-annotate many sentences and make some tough decisions. Some of these decisions are far-reaching (like the one on the treatment of negation), others are less so, but cumbersome. We had to re-annotate all proper nouns that were originally simply considered multiword expressions, to provide them with individual POS-tags and structural dependencies. This showed us how useful it would be to have a lexical resource like the English Multiword Expression Lexicons from CMU,[10] which does not exist for Portuguese, yet.

We should note that this work is not finished. While our treebank once again is syntactically validated by the UD script, we are sure that many errors remain. First because, like other treebanks, we still have so-called "semantic" failures, as described by the UD second level of validation.[11] But mostly because we know that many phenomena are not as yet susceptible of validation. Coordination, ellipsis and negation remain big issues.

References

Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 1698–1703, Las Palmas, Spain.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *The Longman grammar of spoken and written English*. Longman, London.

Eckhard Bick and Tino Didriksen. 2015. Cg-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, pages 31–39. Linköping University Electronic Press.

Eckhard Bick. 2014. PALAVRAS – a constraint grammar-based parsing system for portuguese. In Tony Berber Sardinha and Thelma de Lourdes São Bento Ferreira, editors, *Working with Portuguese Corpora*, pages 279–302. Bloomsbury Academic.

Eckhard Bick. 2016. Constraint grammar-based conversion of dependency treebanks. In *Proceedings of the 13th International Conference on Natural Language Processing (ICON)*, pages 109–114, Varanasi, India, Dec. NLP Association of India (NLPAI).

Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, CoNLL-X '06, pages 149–164, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joakim Nivre et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Cláudia Freitas, Paulo Rocha, and Eckhard Bick. 2008. Floresta sintá (c) tica: bigger, thicker and easier. In *International Conference on Computational Processing of the Portuguese Language*, pages 216–219. Springer.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 168–173. Association for Computational Linguistics.

Juhani Luotolahti, Jenna Kanerva, Sampo Pyysalo, and Filip Ginter. 2015. Sets: Scalable and efficient tree search in dependency graphs. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 51–55. Association for Computational Linguistics.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Tckström, D. Bedini, N. Castelló, and J. Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the ACL 2013*. Association for Computational Linguistics, August.

Luiza Frizzo Truggo. 2016. Classes de palavras - da grécia antiga ao google: um estudo motivado pela conversão de tagsets. Master's thesis, PUC-Rio.

[10] http://www.cs.cmu.edu/~ark/LexSem/

[11] http://universaldependencies.org/svalidation.html