Universal Dependencies for Portuguese

Alexandre Rademaker[1,3] Fabricio Chalub[1] Livy Real[6]
Claudia Freitas[4] Eckhard Bick[5]
Valeria de Paiva[2]

[1]IBM Research, Brazil

[2]Nuance Communications, USA

[3]FGV/EMAp, Brazil

[4]PUC-Rio, Brazil

[5]University of Southern Denmark, Denmark

[6]USP, Brazil

Dependencies Linguistics 2017, Pisa

Linguistic Resources are Important

- Possibly do not need to explain it here ...
- At IBM Research Brazil, many projects on language understanding and information extraction. Portuguese and English (technical domains such as O&G and Mining).
- We started and still work with the OpenWordnet-PT, a Portuguese version of Princeton WordNet, very useful in many applications.

Universal Dependencies

- Universal Dependencies offer the promise of greater parallelism between languages.
- Syntactic dependencies are not too far from semantic dependencies, useful for many applications.
- Manning's Law: grounds for individual languages, good for linguistic typology, rapid and consistent annotation, suitable for parsing with high accuracy, comprehended by non-linguist, support downstream language understanding tasks.
- In UD 1.2, first UD_Portuguese, in UD 1.3, one additional UD_Portuguese-BR (from the Google's treebanks).

The corpus Bosque

'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, and fully revised by linguists.

PALAVRAS is a rule-based Constraint Grammar CG system designed for Portuguese. It produces deep linguistic analyses, with tags at the morphological, syntactic (dependency) and semantic levels.

The Corpus Bosque

UD 1.2 version of UD_Portuguese

- UD release 1.2 was the first release to include a Portuguese treebank, **UD_Portuguese**, a treebank is based on the corpus Bosque (Floresta Sintá(c)tica project from Linguateca and VISL).
- This was based on Bosque 7.3 (AD format) converted to CoNLL-X Shared Task in dependency parsing (2006).
- CoNLL version was 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).

Bosque $\rightarrow$ CoNLL-X $\rightarrow$ Prague Deps $\rightarrow$ Stanford Deps $\rightarrow$ UD
More at `http://www.linguateca.pt/Floresta/levantamento.html`

The Corpus Bosque

In UD 1.2

> *The conversion process started from the AD format. In the end, we decided to implement a direct conversion script from AD to the CoNLL-X format instead of relying on the pipeline of Eckhard Bick's scripts. However, as far as possible, his head rules are implemented. One detail that is probably different from his rules is the linkage in case of more than one auxiliary in combination with a coordinated main verb, especially if the main verbs are accompanied by auxiliary particles. There just wasn't enough time to do this, sorry . . .*
> *In some cases, the Bosque trees contain ambiguities that the annotators could not resolve. For the training data, ambiguity was resolved by simply taking the first annotated possibility. For the test data, sentences that contain ambiguity were discarded.*

`http://www.linguateca.pt/floresta/CoNLL-X/readme.conll`

The Corpus UD

The consolidation

- 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 Mar 2016)
- We started a team effort starting in Oct 2016 and through consistency-checking and discussion, aiming at full compatibility with UD.
- First version of our data, UD 1.4 compliant, included in UD release 1.4 as UD_Portuguese-Bosque. In UD 1.4: UD_Portuguese and UD_Portuguese-Bosque and UD_Portuguese-BR.
- We accepted the challenge to update UD_Portuguese-Bosque to UD 2.0 guidelines and replace the previous UD_Portuguese corpus.

Why Bosque

Why not creating a new one from scratch?

- Besides the original tagset and the CONLL 2006 tagset; there are versions in: CG, AD (phrase structure tree), tgrep, Penn TreeBank and TIGER formats. All these are available from `http://www.linguateca.pt/Floresta/` and `http://corpora.di.uminho.pt/linguateca/FS/fs.html`.
- Different versions of the same material fosters the study about different tagsest and its impacts in NLP systems.
- We had on the team two researchers who had already worked on previous versions of Bosque
- But ... conversion to UD scheme was much more complicated than initially planned.

Why the effort?

- Incorporate changes and additions made in the original treebank after 2006;
- circumvent possible information loss due to previous conversions;
- A comparison of the results of two different conversions might yield interesting insights.
- We wanted to build a framework where manual revision work and consistency checks could be coordinated with automatic parser annotation and conversion rules. Addressing systematic errors, and thus fix them automatically, based on a few examples, rather than repeatedly fixing the same kind of error manually.
- 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. Integrate UD conversion grammar in PALAVRAS.
- Having the corpus revised by native Portuguese linguists guarantees a better annotation quality.

The CG conversion grammar

- The conversion grammar ultimately used for the first conversion of Bosque to UD contained some 530 rules.
- 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 remaining rules handled UD-specific dependency and function label changes in a context-dependent fashion.
  - Main issues were: raising of copula dependents to subject complements, inversion of prepositional dependency and the change from syntactic to semantic verb chain dependency.
  - In respect to punctuation attachment, the grammar actually went beyond conversion, identifying meaningful head tokens for commas, parenthesis etc.

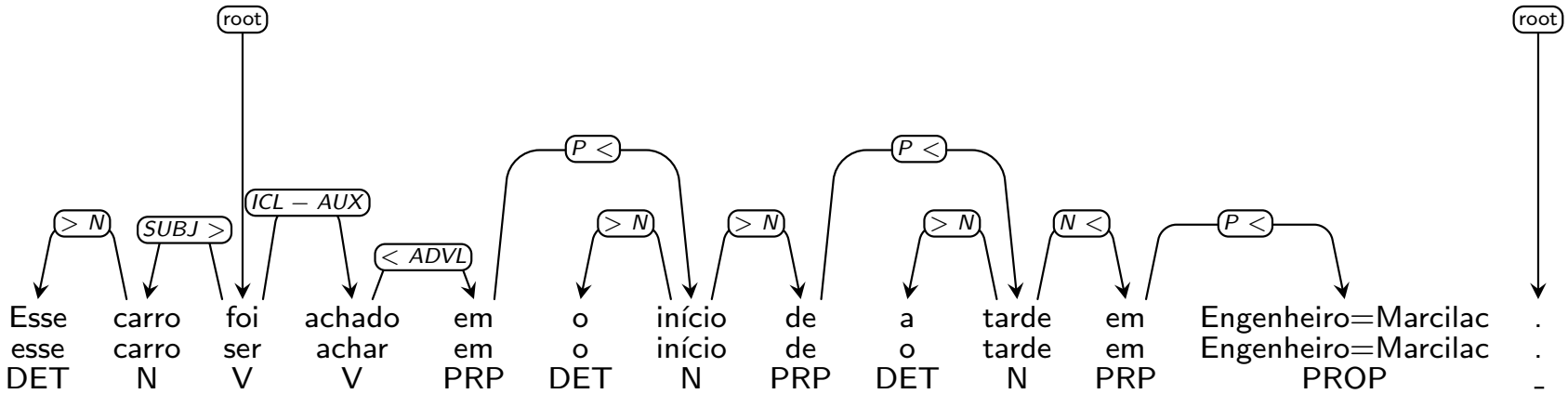
similarities and differences

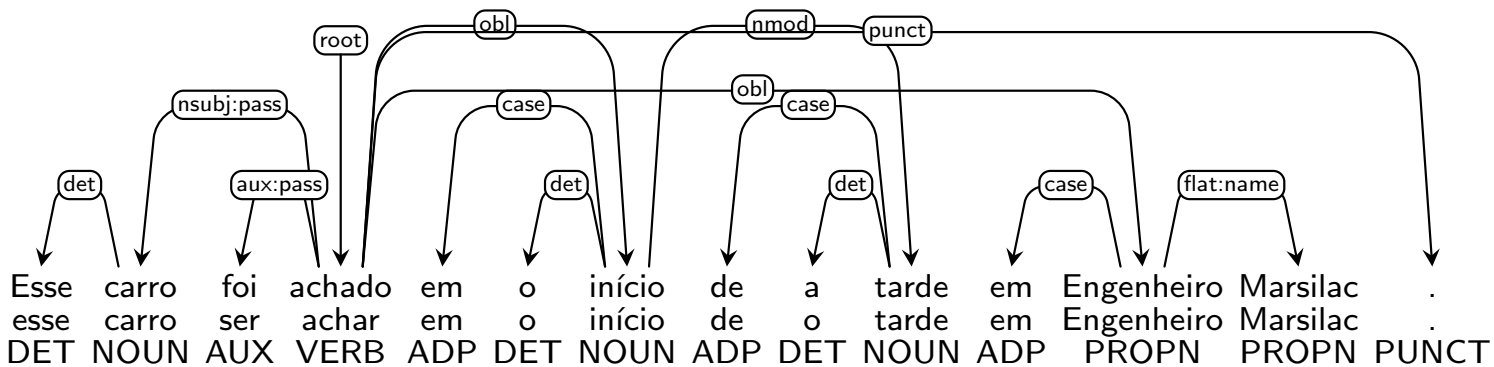
PALAVRAS niceline format.

```
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
```

similarities and differences

cont.





similarities and differences

cont.

- UD version retains the additional tags for NP definiteness and complex tenses and the original syntactic functions tags and secondary morphological tags (`xpostag` field).
- keeps its original linguistic focus, but in addition it can be used for the new machine learning scenarios
- We retain tags roots of sentences for their functions, such as question (`@FS-QUE`), command (`@FS-COM`) or statement (`@FS-STA`).
- In some cases, the stored original function tags allow 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').

Improving the data

gender

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. Sometimes we can tell by the context, sometimes not.

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

`Unsp` (for unspecified value)

Improving the data

MWEs

The PALAVRAS annotation has MWEs tokenized as a single word.

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.

UD version 2, different tags for MWE are used (`flat`, `fixed` and `name`), but this conversion could be done automatically.

Improving the data

Participles

How to deal with participles was also a challenging issue. PALAVRAS tags all participles as verbs, with the `PCP` feature.

In UD can be `VERB` or `ADJ`.

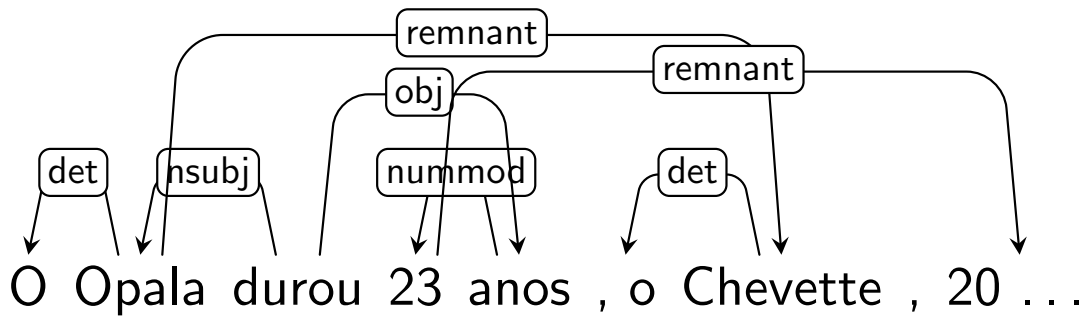
We worked on a set of linguistic rules to semi-automatically re-tag participles.

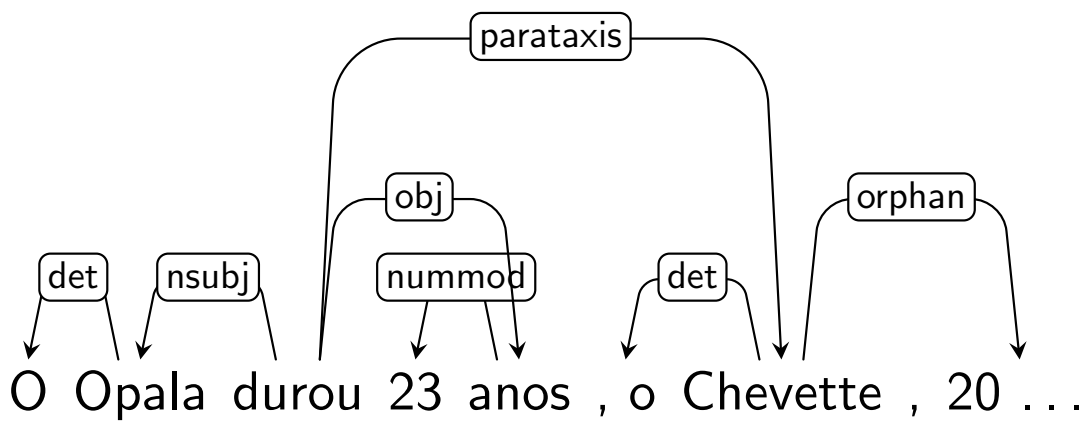
Improving the data

Ellipses

In version 1, ellipsis cases were dealt with via a `remnant` dependency relation. In version 2, the `remnant` relation was discarded and a new treatment was proposed, the relation `orphan`.

Ex: "Opala lasted 23 years, Chevette, 20 [. . .]"





Tokenization

MWE

- The first conversion did not handle UD's tokenization. Original treebank's MWE and - syntactically motivated - splitting of Portuguese contractions (preposition plus article/determiner/pronoun, e.g "neste" to "em + este" (in this).
- 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. Not a simple table conversion.
- CG3 offers context-based manipulation of not only tags, but also of entire tokens. To split MWE tokens, add POS, features, add dependency links.
- 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.
- We adopted a `MWEPOS=` in the `misc` field.

Tokenization

clitics

- Another issue related to tokenization is the problem of clitics in Portuguese. Portuguese we have mesoclitics, that is, clitics that come inside the verb and change the verbal structure:
- Ex 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, pho- tojournalist.)
- We decided to follow the traditional Portuguese grammars. In the example above, 'poder-se-á' is 'poderá/VERB' followed by 'se/PRON' (it can) in the future plus the reflexive.

The particle 'se'

1. **reflexive and reciprocal constructions** *CF314-2 Você se acha louca?* (Do you think you are crazy?);
2. **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.);
3. **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.);
4. **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 particle ‘se’

- universal dependencies this indicates that in both cases (3) and (4) we could have the particle se as the subject of the verb, although the subject remains non-explicit. “vende-se casas” (Houses are sold)
- But in UD, ‘nsubj’ role is only applied to semantic arguments of a predicate, when there is an empty argument in a grammatical subject position (a pleonastic or expletive), it is labeled as `expl`.
- UD creates a certain uniformity between the cases (2), (3) and (4). Since we consider relevant the distinction between (2) (which has an explicit subject) and (3) and (4) (which do not), we keep this information. Cases (3) and (4) carry the label `SUBJ_INDEF` in the `misc` field.

Negation

The treatment of negation has changed from UD version 1 to 2.

In the UD version 2, a polarity feature was introduced (Polarity=Neg).

We understand não – other words as some uses of nada (nothing) – as adverbs. So not many words tagged `PART`

CP153-4 **Não** estava **nada** à espera disto. ([I] was not waiting nothing for it.) both ADV. Sometimes the second is pronoun.

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 . . . ') – `obj(significa,nada)`

Appositives

So far, we used classic and comprehensive notion of appositives (non-restrictive and restrictive)

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.

'president Obama' would be appos (restrictive appositive), if we agree that Obama describes, defines or modifies president. But for UD, since it is not reversible, it is not appos.

However, there are always borderline cases.

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.

Numbers

Bosque has 9.368 sentences and 227.653 tokens, with 18.140 unique lemmas.

At the moment we still have 957 ‘dep’ relations, 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

Comparison and Assessment

Some big discrepancies in numbers between the 1.2 and 1.4/2.0 UD_Portuguese, as computed by the statistics script, were easy to see.

Our version had many more cases of auxiliary verbs than UD_Portuguese in UD 1.2. Probably due verbs like 'continuar' (to continue), 'começar' (to start) and 'acabar' (to end) can also be seen as modal auxiliaries, and that was our decision.

Ex 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.)

Comparison and Assessment

cont.

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 difference in numbers reflects different choices in the alignment-conversion.

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.

When we looked for the appos relation, considering the possible cases of different POS tags pairs being related, we found around 50 possibilities of POS tag pairs. Still need investigation.

Contributions

We implemented the cl-conllu library is implemented in Common Lisp, it is open-source and freely available.

Since we have not yet decided in our group to use any particular dependencies editor, we also implemented an online CoNLL-U validation service.

What's Next?

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.

But mostly because we know that many phenomena are not as yet susceptible of validation. Coordination, ellipsis and negation remain big issues.

A challenge: lack of editor, tabular based is easier for linguists. But for facilitate collaboration it must be web-based too.

Problems

- Some cases of `<n>` were not converter to NOUN, althouth the dependencies are right.
  "A direção do novo **semanal** será assinada por Ewaldo Ruy." (The direction of the new weekly will be assumed by Ewaldo Ruy.)
  "Pesquisadores acham que as linhas podem ser **falhas** geológicas." (Researchers believe that the lines may be geological faults.)
- Many problems with reported speech and parataxis are inconsistent annotated.
- The relation `discourse` is not consistent annotated.
- Numerals also need to me revised, cases of 'trinta e sete'(37) and 'cento e dezessei' (116) must be flat.
- Some `obl` that have PALAVRAS tag `PIV` should be `obj`.
- We are now revising the appositional modifier `appos` versus `nmod`.

Thanks!