Computational Grammars for Portuguese

Alexandre Rademaker Leonel Alencar Bruno Cuconato

2018-05-24

IBM Reseach, Getulio Vargas Foundation / EMAp, Univ Fed. Ceará

grammars

- BrGram under development in XLE, grammar development tool, within the LFG generative framework by Leonel Alencar since 2013.
- We have also started a Portuguese grammar in the Grammatical Framework (GF) by Aarne Ranta.
- HPSG - LxGram - by Francisco Costa and António Branco in Portugal (previous work?)
- PALAVRAS by Eckhard Bick (constraint grammar) implemented in VISL.

We are interested in deep parsing unrestricted standard Portuguese texts.

The Portuguese Language

- With 210 million native speakers, Portuguese is the 6^{th} or 7^{th} most spoken language of the world
- After Spanish, it is the second most spoken Romance language
- 190 million speak the Brazilian variety of the language, i.e. Brazilian Portuguese (henceforth BP)
- Between standard BP and European Portuguese (EP), there are important syntactic differences

BP versus EP

Even within the limited domain of the first 16 sentences of the ParGramBank, important syntactic differences between BP and EP emerge. For example:

In "pure" wh-questions, subject-verb inversion is mandatory in EP and disallowed in BP:

1. O que/*Que o agricultor viu? (BP)
   (the what/what the farmer saw)
2. O que/Que viu o agricultor? (EP)
   (the what/what saw the farmer)
3. What did the farmer see?

PALAVRAS

- Part of the Visual Interactive Syntax Learning (VISL) project, an initiative of the University of Southern Denmark
- Free online demo and very expensive availability otherwise
- Based on Constraint Grammar
- Outputs very flat tree structures with basic functional information
- High robustness and coverage obtained at the cost of not modeling grammatical constraints such as agreement, subcategorization, etc.

LxGram

- Parsing efforts at the University of Lisbon
- Based on HPSG
- High coverage and Focus on EP
- Publicly available for free (no free software)
- LxParser is another thing. Statistical constituency and dependency parser trained with the Stanford Parser on a treebank of EP; outputs trees conforming to a simplified X-bar format or dependency graphs with basic functional information

BrGram

- Leonel BP ParGram Grammar started in 2013 with 18 rules with 154 states, 626 arcs, and 726 disjuncts (728 DNF)
- Currently the grammar already parses the 50 first sentences from the pargrambank
- It results from one man-month effort of adapting and extending a previous small XLE-grammar of BP in order to analyze the 50 ParGramBank sentences
- This previous grammar aimed at testing hypotheses concerning the pronominal system, DP structure and variable verb agreement in Portuguese
- In 2018, an approved two-years project funded by FGV for working on the development of BrGram (focus on DHBB)

BrGram

- Tokenization and Morphology using finite-state transducers.
- Available at `https://github.com/LFG-PTBR`
- XLE is not free but a new LFG-parser is under development https://bitbucket.org/dcavar/fle/ (Free Linguistic Environment)
- First collaborative result `https://github.com/LFG-PTBR/MorphoBr` - an open source large-coverage lexical resource for morphological analysis of Portuguese (submitted PROPOR 2018)

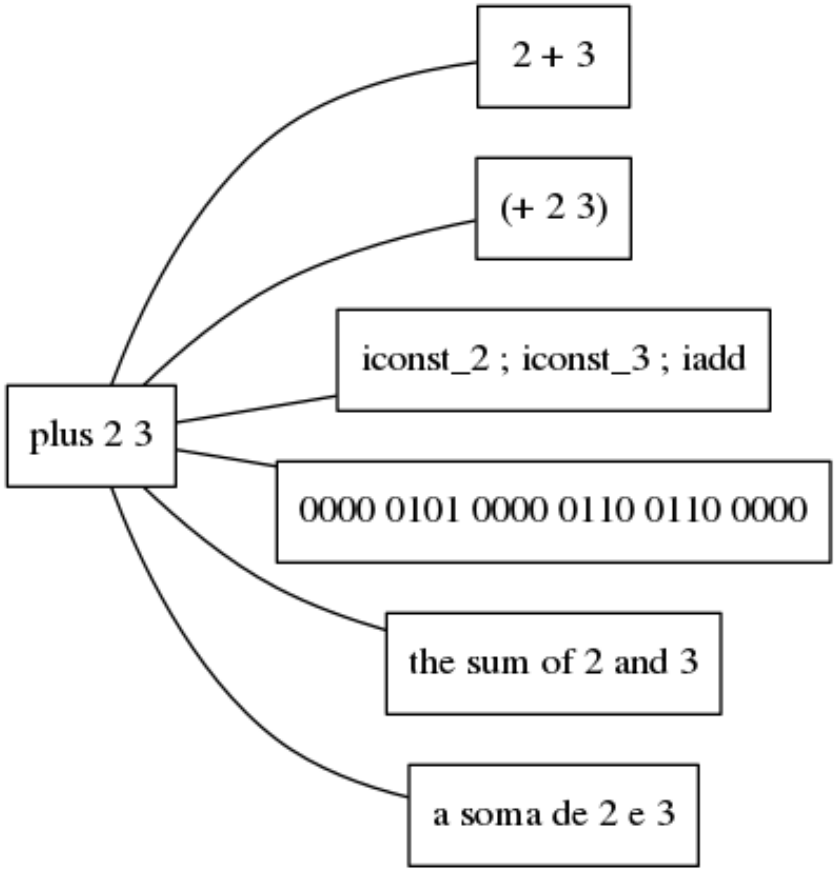
grammatical framework

A **DSL** for grammar writing.

site

`http://grammaticalframework.org`

GF and compiler theory



the GF resource grammar library

- a common abstract syntax for 30+ natural languages
  - grammatical categories
  - constructor functions
  - test lexicon

the GF resource grammar library

- a common abstract syntax for 30+ natural languages
  - grammatical categories
  - constructor functions
  - test lexicon
- resource vs. application grammars
  - specialization

GF REPL examples

```
> import -retain present/TryPor.gfo
> cc -one mkS (mkCl (mkNP these_Det \
>   (mkN "bala")) (mkA "gostoso"))
estas balas são gostosas
```

GF REPL examples

```
> import present/LangEng.gfo
> p -lang=Eng "these fish are rotten"
PhrUtt NoPConj (UttS (UseCl
  (TTAnt TPres ASimul) PPos
    (PredVP (DetCN (DetQuant this_Quant NumPl)
       (UseN fish_N))
      (UseComp (CompAP (PositA rotten_A))))))
   NoVoc
```

GF REPL examples

```
> import FoodsEng.gf FoodsPor.gf
> p -lang=Eng -tr "that pizza is delicious" \
> | l -lang=Por
Pred (That Pizza) Delicious
essa pizza é deliciosa
```

the Portuguese Resource Grammar (PRG)

- already have a functioning version

matrix MRS testsuite

- the dog was chased by Browne

  `o cachorro era por Bobi perseguido`

  `o cachorro era perseguido por Bobi`

- it bothered Abrams that Browne barked

  `incomodava Atlas que Bobi ladrasse`

- it bothered Browne that Abrams chased cats

  `incomodava Bobi que Atlas perseguisse gatos`