PorGram - The Open Portuguese HPSG Grammar

Leonel Figueiredo de Alencar and Alexandre Rademaker

UFC, IBM Research and FGV/EMAp

DELPH-In Summit 2021

Previous works on Portuguese Grammars

- LX-Gram from University of Lisbon (not free)
  - http://lxcenter.di.fc.ul.pt/tools/pt/conteudo/LXGram.html
- PALAVRAS in constraint grammar framework (not open source or free)
- The Leonel's LFG Portuguese BrGram (https://github.com/LR-POR/BrGram)
- Not sure if any from Emily Bender's courses?

Why

- UFC
    - graduate program in linguistics of the Humanities Center
    - courses on Linguistic Description and Analysis (computational linguistics)
- FGV
    - mathematics and Linguistics
    - courses on NLP
- IBM
    - practical applications with deep semantic representations (QA, knowledge acquisition for ontology construction, text entailment)

      Rob Thomas's bet on Language as one of the three pillars in IBM's AI strategy is based on a strong history of language innovation and delivery at IBM. Many are familiar with the historic 2011 moment when IBM Watson, a computer, won the game of Jeopardy against human champions.

    - resource creation for ML techniques (corpora, lexicon etc)

Studying Portuguese

- Variations of gender and number of adjectives?

    1. *Uma solução simples para vários problemas simples.*

    2. *Uma menina alegre e um rapaz triste.*

- What is the relation between the variation of gender/number of the verb and the existence or not of the subject?

    3. *Choveu e trovejou ontem*

Data driven vs knowledge driven

- Take advantage of data (UD treebanks, Lexicons: WN, MorphoBr etc)
  - UD annotation is also expensive!
- The AGGREGATION project?
- Grammar Engineering
  - MATRIX
  - LKB
  - complementary scripts and workflow (matrix, customization, test, etc)

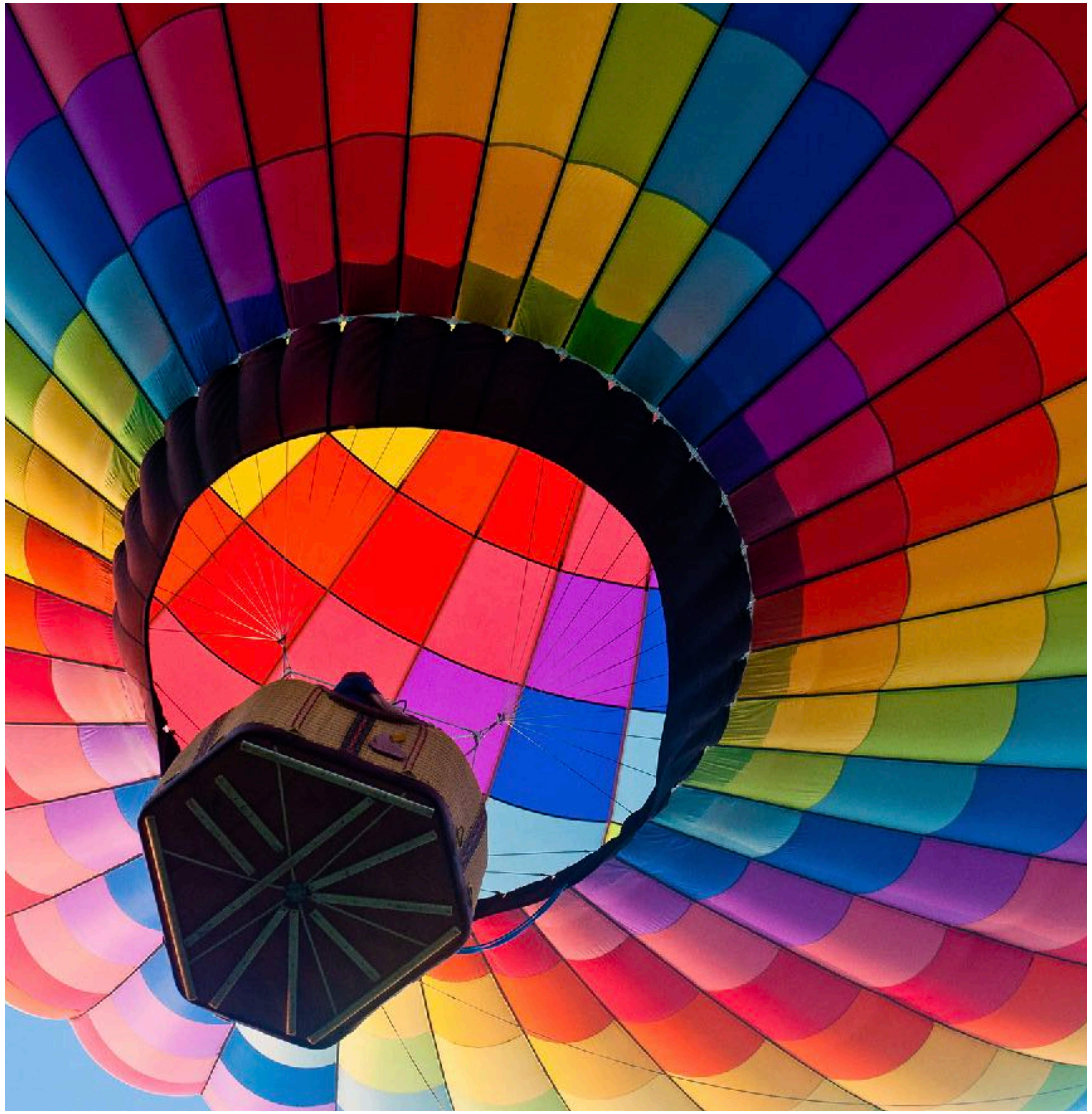
Goals

- Long-term objective (3 years): Implementation of a Portuguese grammar for deep parsing of unrestricted texts in standard language
- Medium-term objective (10 months): syntactic and semantic annotation of part of the Brazilian Historical-Biographic Dictionary (DHBB) of the CPDOC/ FGV
- Short-term objective (6 months): parsing the MRS and HP Test Suite (CSLI profile) test sets

First paper:

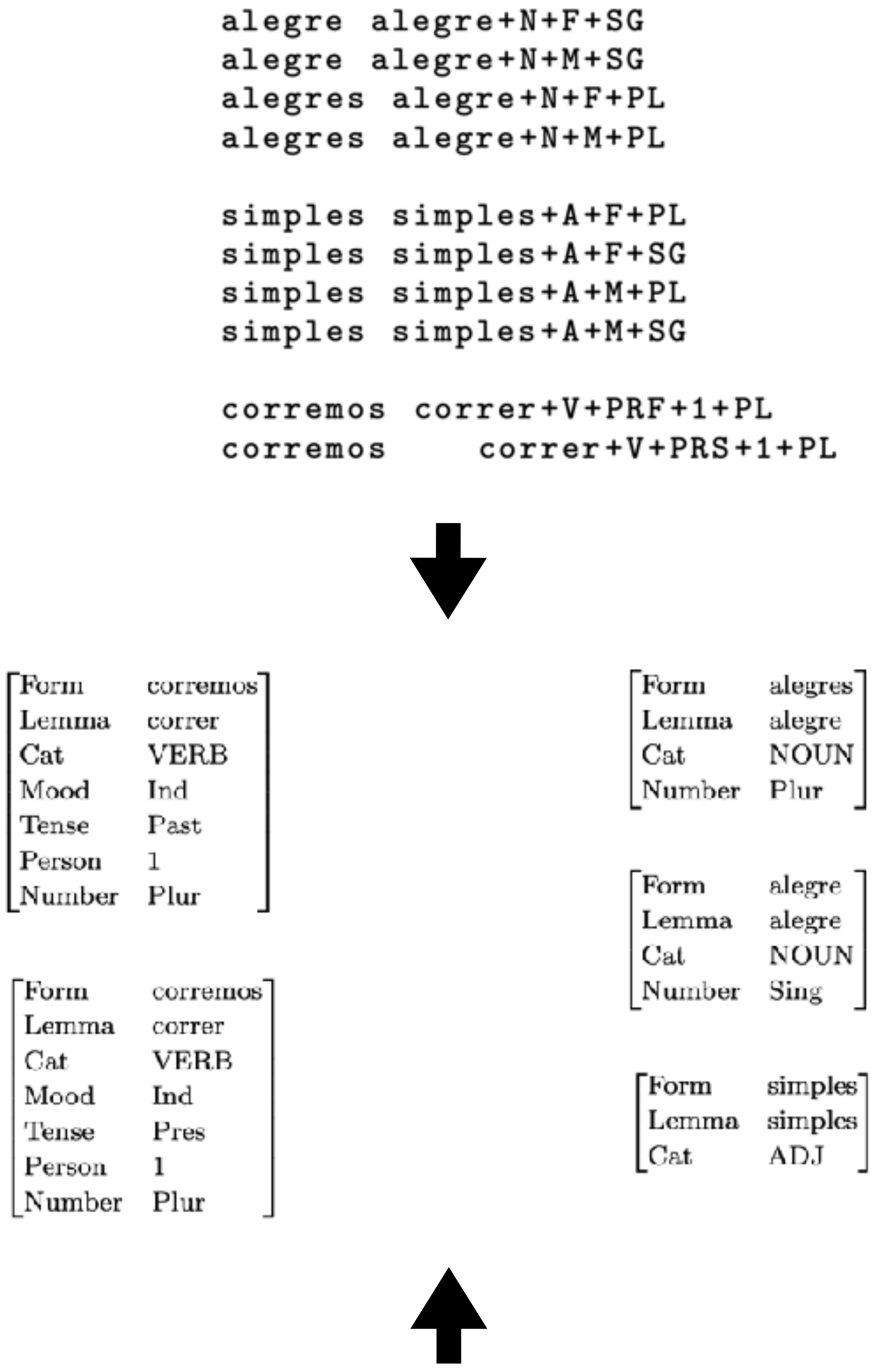
Cross-Validating Language Resources for the Development of a Large-Coverage Computational Grammar of Portuguese

Under review LREC special issue on Portuguese Processing



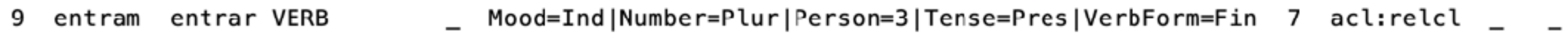
First: consolidation of resources

- diverse lemmatization of adjectives
- inconsistency of noun entries (inherent gender)
- missing forms of compounds (recém-chegado ‘newcomer’)
- Portuguese Language Orthographic Agreement
- Tokenization of clitics as MWT
- productive word-formation
    - *Manchete estréia novo* ***jornalístico*** (Manchete debut a new journalistic)
    - *foi utilizada técnica* ***mista****...* (mixed technique was used...)



```
alegre  alegre+N+F+SG
alegre  alegre+N+M+SG
alegres  alegre+N+F+PL
alegres  alegre+N+M+PL

simples  simples+A+F+PL
simples  simples+A+F+SG
simples  simples+A+M+PL
simples  simples+A+M+SG

corremos  correr+V+PRF+1+PL
corremos     correr+V+PRS+1+PL
```

```
Form    corremos
Lemma   correr
Cat     VERB
Mood    Ind
Tense   Past
Person  1
Number  Plur

Form    corremos
Lemma   correr
Cat     VERB
Mood    Ind
Tense   Pres
Person  1
Number  Plur

Form    alegres
Lemma   alegre
Cat     NOUN
Number  Plur

Form    alegre
Lemma   alegre
Cat     NOUN
Number  Sing

Form    simples
Lemma   simples
Cat     ADJ
```



```
9   entram   entrar VERB        _  Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin  7  acl:relcl  _   _
```

Cont.

Evaluatives

- augmentatives, diminutive and absolute superlative formation are extremely productive in Portugueses. Applied to adjectives, nouns, adverbs, numerals and pronouns.
- evaluative morphology are not properly modeled in UD. Most diminutives are not lemmatized to the corresponding bases in this treebank, causing discrepancies between MorphoBr and UD Bosque

    1. *Depois da peça de Leilah … os does foram **juntinhos**...* (After Leilah's play, the two went together…)

- diminutives, augmentatives, and absolute superlatives are not adjectives degree of comparison
- While most discrepancies involving diminutives show the need to correct UD Bosque, a few of the analyses provided by the corpus could be incorporated into MorphoBr.
- MorphoBr does not explicitly represent comparatives and relative superlatives, they are either not morphologically productive or are expressed analytically in Portuguese.
- The few existing synthetic forms inherited from Latin, e.g. maior (big), menor (small) and superior (high), are lexicalized, having undergone semantic extensions which have made their meaning non-compositional: *preços menores* ‘lower prices’ vs *carros menores* ‘smaller cars’

Cont.

after consolidation…

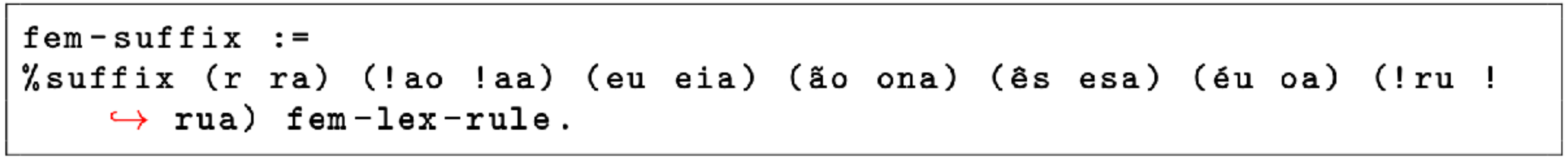
- a methodology to populate the lexicon of the grammar by reusing freely available existent resources
- Portuguese ‘*comprar*’ (to buy) has 71 forms. Full-form vs morphology rules
- Focus on adjective morphology
  - designing a hierarchy of lexical types in such a way that some types are prohibited from undergoing inflection, while others are licensed to do it.
  - In Portuguese adjectives are typically inflected for gender and number, manifesting agreement with the nouns they predicate on. The canonical inflectional paradigm for ***amarelo*** ‘yellow’
    - feminine gender is expressed substituting a for o in the lemma: ***amarela***
    - Plural number is marked suffixing s to the unmarked singular form: ***amarelos/amarelas***

Morphological rules

- Some rules cannot be modeled directly with the customization questionnaire, which only handles morphotactics (non ins/del)
- a large number of nouns and adjectives require the implementation of morphophonological rules or orthographic alternations
    - insertion of “e” before “s” when the lemma ends with "n","r", or "z", e.g. **rapazes**, plural of **rapaz** ‘boy’.
    - "*são*" versus "*sã*" ‘healthy’
    - “*europeu*" versus "*europeia*" ‘European’
- For marking plural number, similarly diverse morphophonological processes apply to numerous adjectives.
- Large group of adjective does not inflect for gender (e.g. incolor “colorless")
- We decided to model the morphophonological component of inflectional rules (handwritten `my-irules.tdl`). The morphology 72 suffixes and 335 rules (15 working days)

Cont.

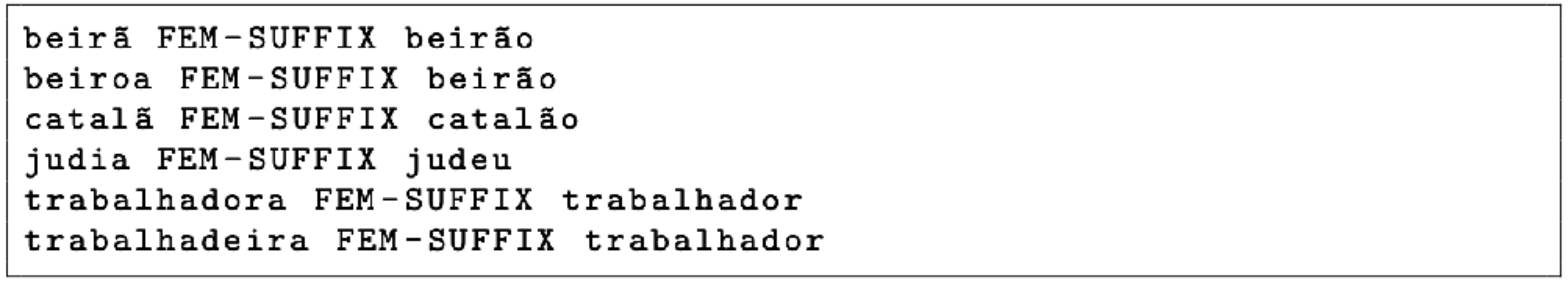
Listing 5 Feminine suffix rule



```
fem-suffix :=
%suffix (r ra) (!ao !aa) (eu eia) (ão ona) (ês esa) (éu oa) (!ru !
    ↪ rua) fem-lex-rule.
```

- *trabalhador trabalhadora ‘working’*
- *amarelo amarela ‘yellow’*
- *europeu europeia ‘European’*
- *Chorão chorona ‘crybaby’*
- *Francês francesa ‘French’*
- *ilhéu ilhoa ‘islander’*
- *cru crua ‘raw’*

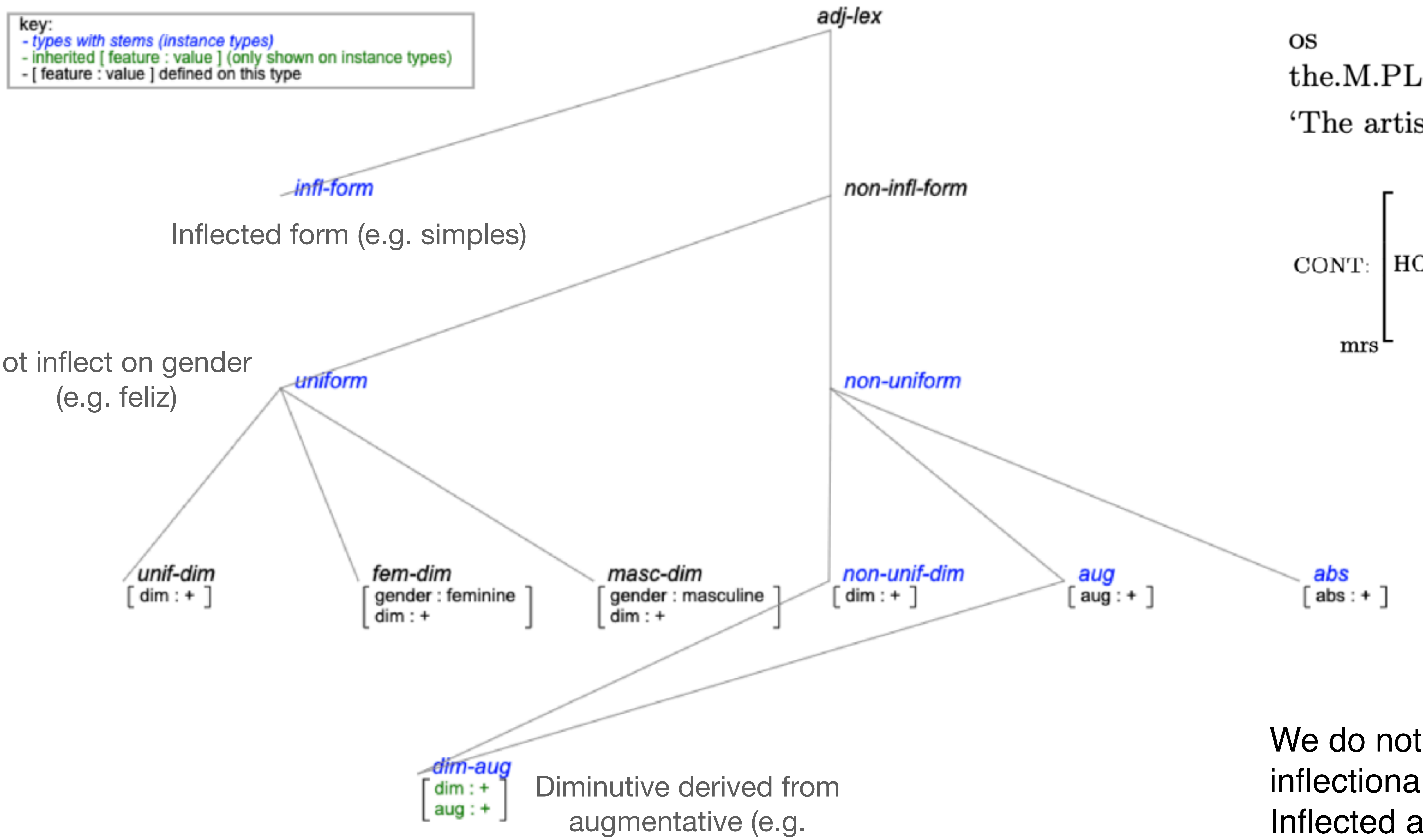
Listing 6 Irregular feminine suffixes



```
beirã FEM-SUFFIX beirão
beiroa FEM-SUFFIX beirão
catalã FEM-SUFFIX catalão
judia FEM-SUFFIX judeu
trabalhadora FEM-SUFFIX trabalhador
trabalhadeira FEM-SUFFIX trabalhador
```

Adjective types

From full-form MorphoBr to PorGram Types


key:
- types with stems (instance types)
- inherited [feature : value] (only shown on instance types)
- [feature : value] defined on this type
adj-lex
infl-form
non-infl-form
uniform
non-uniform
unif-dim
[dim : +]
fem-dim
[gender : feminine dim : +]
masc-dim
[gender : masculine dim : +]
non-unif-dim
[dim : +]
aug
[aug : +]
abs
[abs : +]
dim-aug
[dim : + aug : +]

Inflected form (e.g. simples)

Not inflect on gender (e.g. feliz)

Diminutive derived from augmentative (e.g. felizãozinho)

os	artistas	são	felizõezinhos
the.M.PL	artist.PL	be.3PL	happy.AUG.DIM.M.PL

‘The artists are very happy’

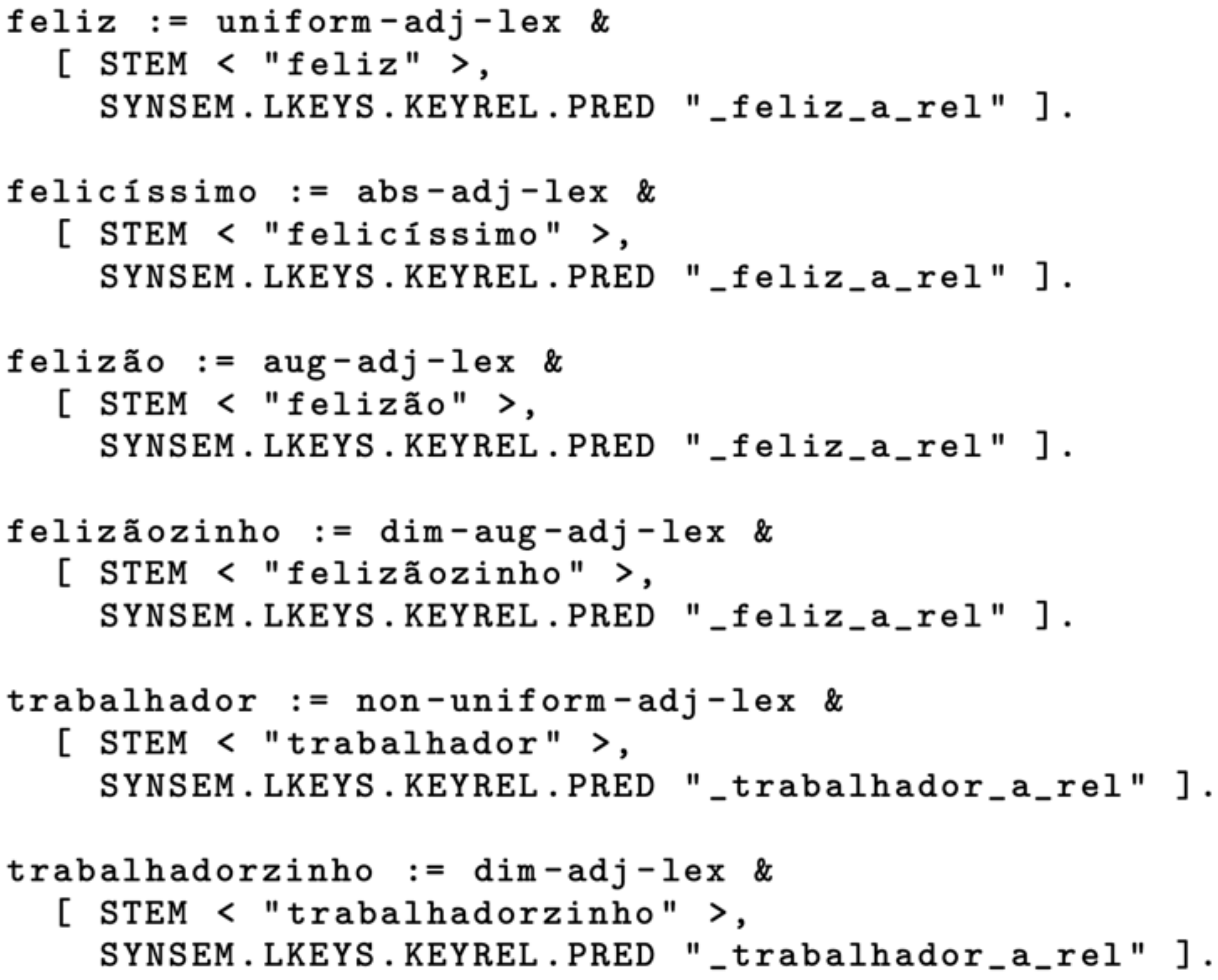
$$\text{CONT: }{}_{\text{mrs}}\begin{bmatrix}\text{HOOK }{}_{\text{hook}}\begin{bmatrix}\text{INDEX }{}_{\text{ref-ind}}\begin{bmatrix}\text{PNG }{}_{\text{png}}\begin{bmatrix}\text{DIM} & + \\ \text{AUG} & + \\ \text{NUM} & \text{plural} \\ \text{GEND} & \text{masculine}\end{bmatrix}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

We do not consider evaluative suffixation to be an inflectional process, but rather a derivational one. Inflected adjectives do undergo derivational rules, e.g., simplesinho, diminutive of simples ‘simple’.

From MorphoBr to PorGram

- Python code decideS, for each lemma, for each inflectional rule: (i) regular with no alternative irregular form, (ii) irregular, or (iii) regular with one or more alternative irregular forms
  - For each lemma a type in the adjective hierarchy by MorphoBr data
  - regular suffixations in my-irules.tdl
  - populate the table of exceptions in my-irregs.tab

In order to decide whether a form F of a lemma L is regularly derived by rule R, the suffix replacement rules of the my-irules.tdl file were implemented in Python.



```
feliz := uniform-adj-lex &
  [ STEM < "feliz" >,
    SYNSEM.LKEYS.KEYREL.PRED "_feliz_a_rel" ].

felicíssimo := abs-adj-lex &
  [ STEM < "felicíssimo" >,
    SYNSEM.LKEYS.KEYREL.PRED "_feliz_a_rel" ].

felizão := aug-adj-lex &
  [ STEM < "felizão" >,
    SYNSEM.LKEYS.KEYREL.PRED "_feliz_a_rel" ].

felizãozinho := dim-aug-adj-lex &
  [ STEM < "felizãozinho" >,
    SYNSEM.LKEYS.KEYREL.PRED "_feliz_a_rel" ].

trabalhador := non-uniform-adj-lex &
  [ STEM < "trabalhador" >,
    SYNSEM.LKEYS.KEYREL.PRED "_trabalhador_a_rel" ].

trabalhadorzinho := dim-adj-lex &
  [ STEM < "trabalhadorzinho" >,
    SYNSEM.LKEYS.KEYREL.PRED "_trabalhador_a_rel" ].
```

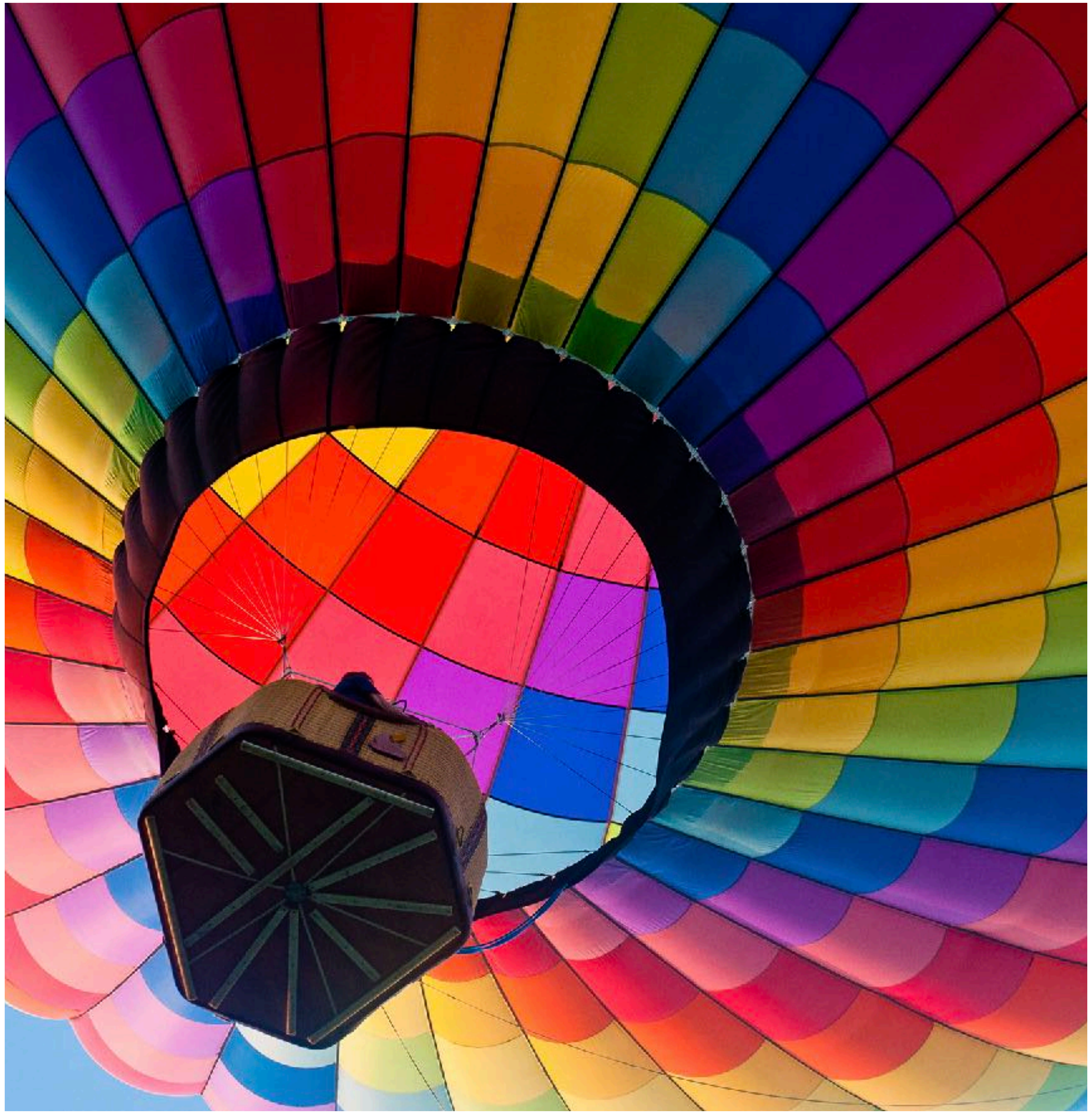
Limitation

- Evaluative morphology was not implemented yet, great variety of affixes and the complexity of the morphophonological processes involved. Current version of PorGram, diminutives, augmentatives, and superlatives have their own lemmas.
- Operations from Finite-State Morphology seems to be outside suffix replacement In LKB
    - The stress moves to the suffix "íssimo" causes the deletion of the accent of a proparoxytone base, e.g. rapidíssimo ‘very fast’, superlative of rápido ‘fast’.
    - internal inflection of words derived with evaluative suffixes beginning with z,“z-evaluative suffixes”, e.g., alemãezinhos e alemãzinhas, diminutives plural form of alemão ‘German’.
- Note that diminutives derived from augmentatives, e.g., felizãozinho, diminutive of the augmentative of "feliz", are assigned the type dim-aug-adj-lex, which inherits the features [DIM +] and [AUG +] from its parent types.

Evaluation

- a sample of 395 adjective entries from MorphoBr, all forms that map to a total of 26 lemmas, including at least one example for each of the adjective types
- An example sentence for each one of the adjective forms, following the regex pattern "the artists? (is|are) X", where X is the adjective form. The sentences were automatically built by a Python script that adjusts the gender and number of the determiner and the verb to agree with the adjective form. Due to uniform adjectives, the resulting test file exceeds the number of forms, totaling 416 grammatical sentences.
- The grammar generated by the MATRIX can analyze **none** of these 416 sentences. Equipped with the manually edited my-irules.tdl file and the exceptions table generated by the python code, the grammar could analyze 365 (87.74%) of the sentences.
- Incorrect type for diminutive persinha of persa ‘Persian’ was assigned the type dim-adj-lex, invariable for gender.
- The failures of the algorithm reside on the construction of the exceptions table. All errors and omissions in this table involve diminutive forms. Some related to MorphoBr errors.

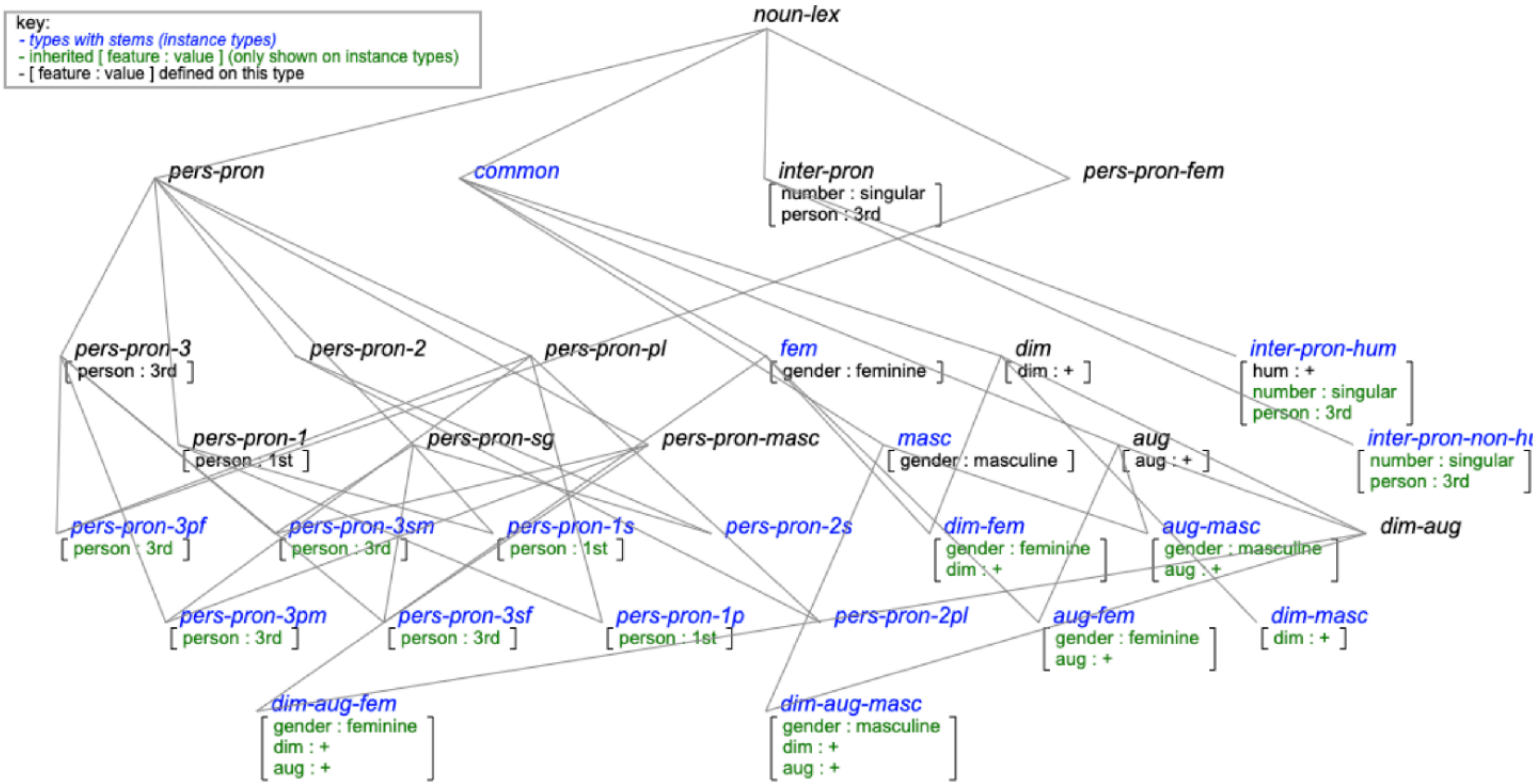
Next and current development



Current stage

- Morphology: verbs, nouns, adjectives and determinants
- Development driven by the MRS test suite

Noun Hierarchy



key:
- types with stems (instance types)
- inherited [feature : value] (only shown on instance types)
- [feature : value] defined on this type

noun-lex

pers-pron

common

inter-pron [number : singular person : 3rd]

pers-pron-fem

pers-pron-3 [person : 3rd]

pers-pron-2

pers-pron-pl

fem [gender : feminine]

dim [dim : +]

inter-pron-hum [hum : + number : singular person : 3rd]

pers-pron-1 [person : 1st]

pers-pron-sg

pers-pron-masc

masc [gender : masculine]

aug [aug : +]

inter-pron-non-hu [number : singular person : 3rd]

pers-pron-3pf [person : 3rd]

pers-pron-3sm [person : 3rd]

pers-pron-1s [person : 1st]

pers-pron-2s

dim-fem [gender : feminine dim : +]

aug-masc [gender : masculine aug : +]

dim-aug

pers-pron-3pm [person : 3rd]

pers-pron-3sf [person : 3rd]

pers-pron-1p [person : 1st]

pers-pron-2pl

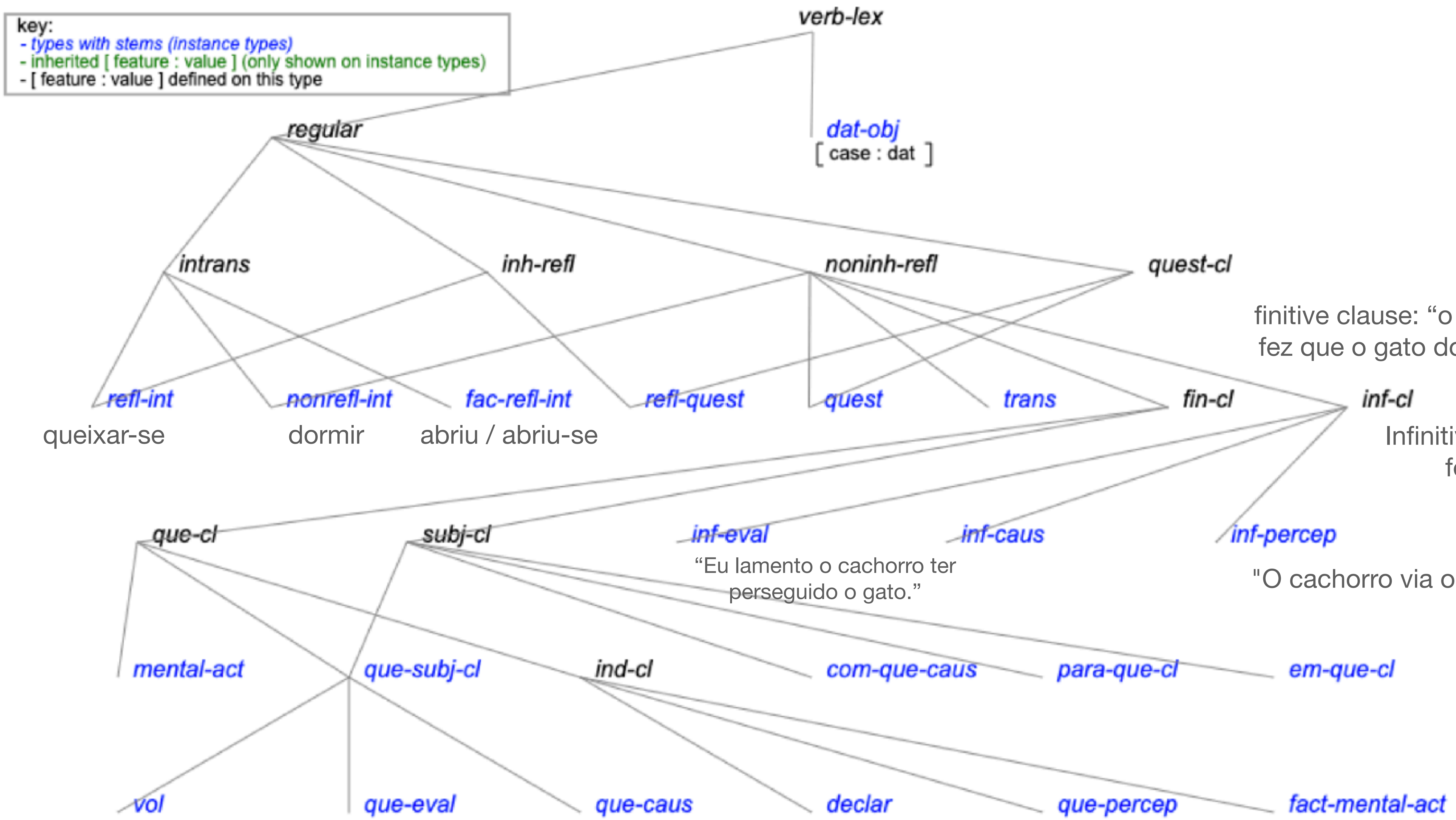
aug-fem [gender : feminine aug : +]

dim-masc [dim : +]

dim-aug-fem [gender : feminine dim : + aug : +]

dim-aug-masc [gender : masculine dim : + aug : +]

Verb Hierarchy



queixar-se

dormir

abriu / abriu-se

finitive clause: “o cachorro fez que o gato dormisse.”

Infinitive clause: “o cachorro fez o gato dormir”

“Eu lamento o cachorro ter perseguido o gato.”

"O cachorro via o gato dormir”

Can prepositions head a clause?

In Portuguese, many verbs require that complement clause be introduced by preposition

- *O cachorro insistiu* ***em que*** *o gato dormisse*
- *Ele insistiu na viagem*

How to model that the verb “insistir" requires a preposition heading a noun phrase or complementizer phrase? In the MATRIX a preposition head only a noun phrase.

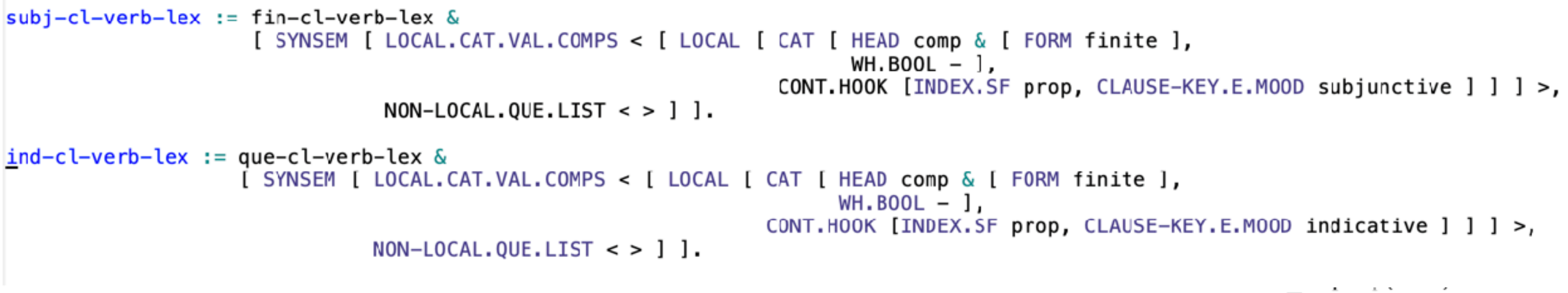
Temporary solution: composed complementizers (em-que, para-que etc)

Verb mode in completive clauses

- In the MATRIX questionnaire we can’t specify the mood of clause
- Volitive verbs require the subjunctive mode (Matthew et al., 1989, p. 273):
  - ******o cachorro* ***quer*** *que o gato* ***late***
  - *o cachorro* ***quer*** *que o gato* ***lata***
- Declarative verbs (verba dicendi) and mental activity verbs require the indicative mode (Mateus et al., 1989, p. 270-271):
  - *o cachorro* ***declarou*** *que o gato* ***late***
  - **o cachorro* ***afirmou*** *que o gato* ***lata***

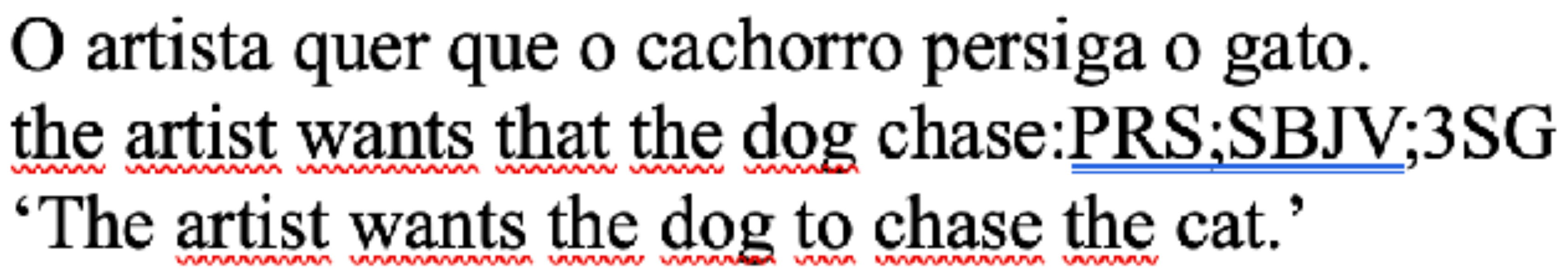
Solution: handwritten encoding

A verb that requires a finite clause with subjunctive mood



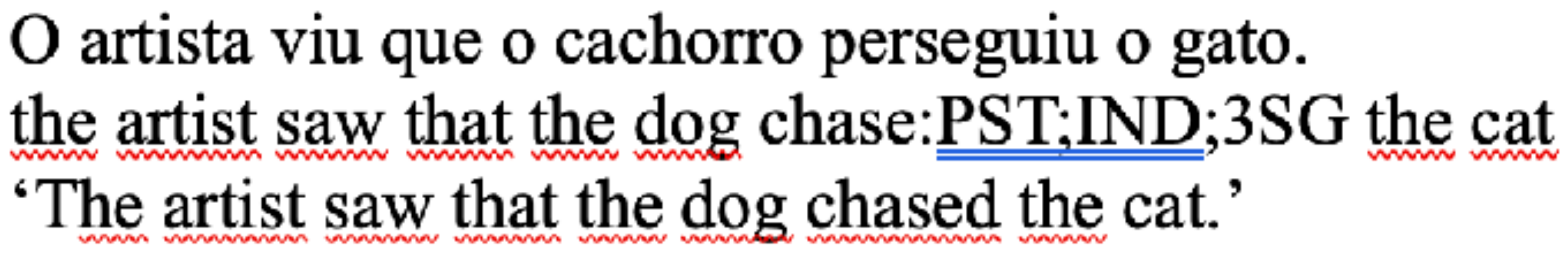
```
subj-cl-verb-lex := fin-cl-verb-lex &
                    [ SYNSEM [ LOCAL.CAT.VAL.COMPS < [ LOCAL [ CAT [ HEAD comp & [ FORM finite ],
                                                                     WH.BOOL - ],
                                                               CONT.HOOK [INDEX.SF prop, CLAUSE-KEY.E.MOOD subjunctive ] ] ] >,
                               NON-LOCAL.QUE.LIST < > ] ].

ind-cl-verb-lex := que-cl-verb-lex &
                   [ SYNSEM [ LOCAL.CAT.VAL.COMPS < [ LOCAL [ CAT [ HEAD comp & [ FORM finite ],
                                                                    WH.BOOL - ],
                                                              CONT.HOOK [INDEX.SF prop, CLAUSE-KEY.E.MOOD indicative ] ] ] >,
                              NON-LOCAL.QUE.LIST < > ] ].
```

Verb volitive

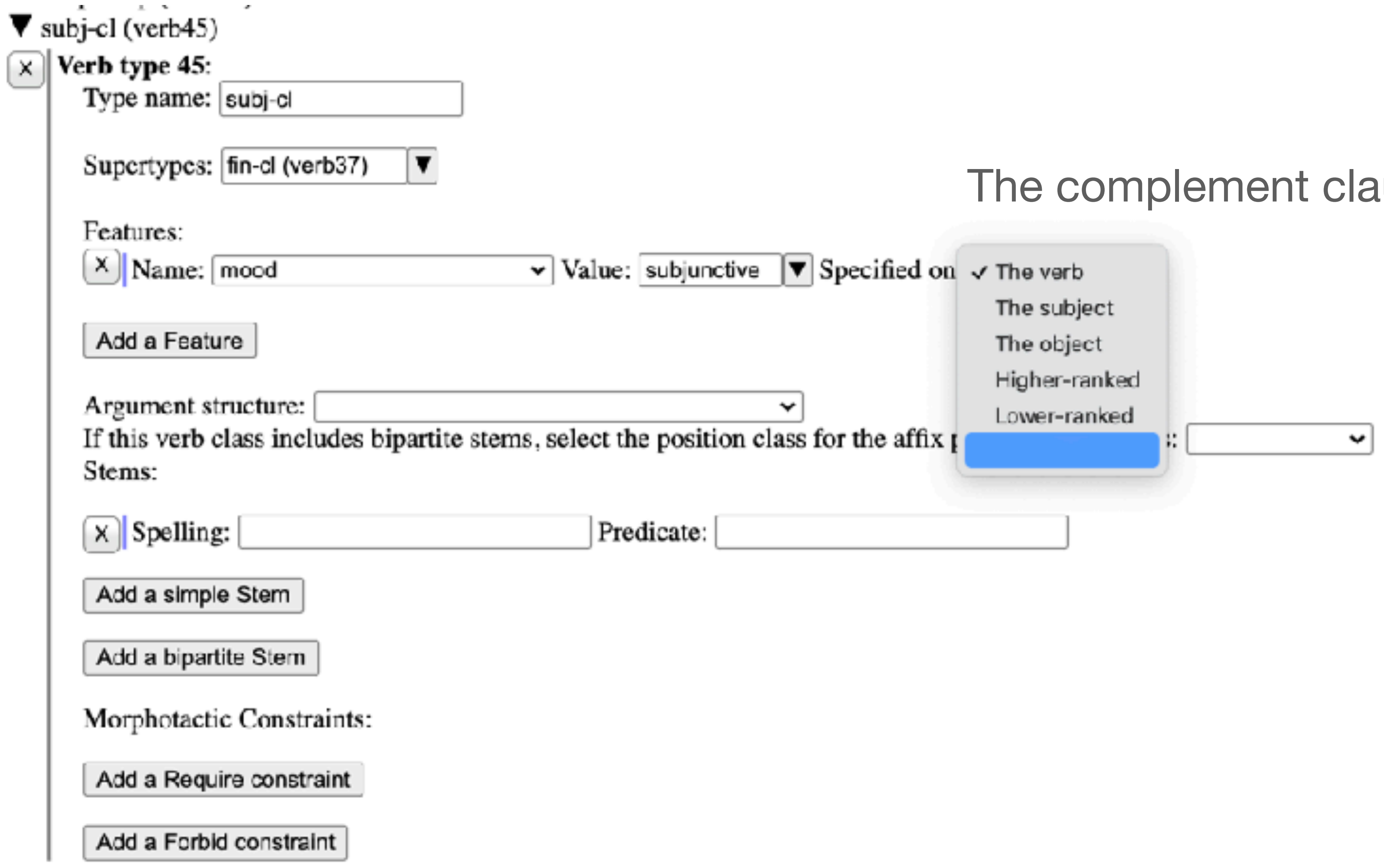


O artista quer que o cachorro persiga o gato.
the artist wants that the dog chase:PRS;SBJV;3SG
‘The artist wants the dog to chase the cat.’

Verb perception



O artista viu que o cachorro perseguiu o gato.
the artist saw that the dog chase:PST;IND;3SG the cat
‘The artist saw that the dog chased the cat.’



▼ subj-cl (verb45)

Verb type 45:

Type name: subj-cl

Supertypes: fin-cl (verb37)

Features:

Name: mood Value: subjunctive Specified on ✓ The verb

Argument structure:

If this verb class includes bipartite stems, select the position class for the affix p

Stems:

Spelling: Predicate:

Morphotactic Constraints:

The complement clause?

Matrix Limitations?

- Only divalent verbs? What about the trivalent verbs?
  1. Abrams showed Browne the office
- Subject raising in complement clauses?
  2. I asked who is sleeping “*Eu perguntei* ***quem está dormindo***”
  3. *O cachorro está dormindo* (in main clause is ok)
- object control (*promise* and *persuade* verbs)
  4. *O cachorro promoteu latir* (subject control is ok)
  5. *O cachorro persuadiu o gato* ***a*** *latir* (*gato* subj of *latir* and obj of *persuadir, insert complementizer “a latir*" *or “impedir de latir”*)
     - How to introduce those complementizers (prepositions)?

Matrix Limitation?

Many transitive verbs require their object to be introduced by a preposition, e.g.:

1. *O cachorro obedeceu ao gato* (The dog obeyed the cat)

It seems that the questionnaire doesn’t support the implementation of such cases.

Robustness

- pre-processing
  - REEP
  - POS Tagger / generic entries
  - Lexical Threads etc
- Treebanking and MaxEnt model trainning
- Performance Issues