ReGra's Lexical Database

Ronaldo Martins

Outline

- Motivation
- Warning
- The Past
- The Present
- The Future
- The Golden List
- A Checker Dictionary Commitments
- Final remarks

Motivation

- ReGra: a proofing tool for BP
  - RLP (Itautec-Philco)
  - Microsoft Office 2000, XP, .Net
- Three fases
  - 1993-1997: Local rules
  - 1997-2002: Parsing
  - 2002-2003: Modularization
- Goal
  - to emulate the behavior of a human reviser (i.e., to diagnose illegal words and constructions, to identify the source of problems, to propose acceptable alternatives and to convince the user)

NILC
núcleo interinstitucional de linguística computacional

Warning

- ReGra does not really carry out any morphological analysis but rather processes word retrieval strategies along with tokenization routines.

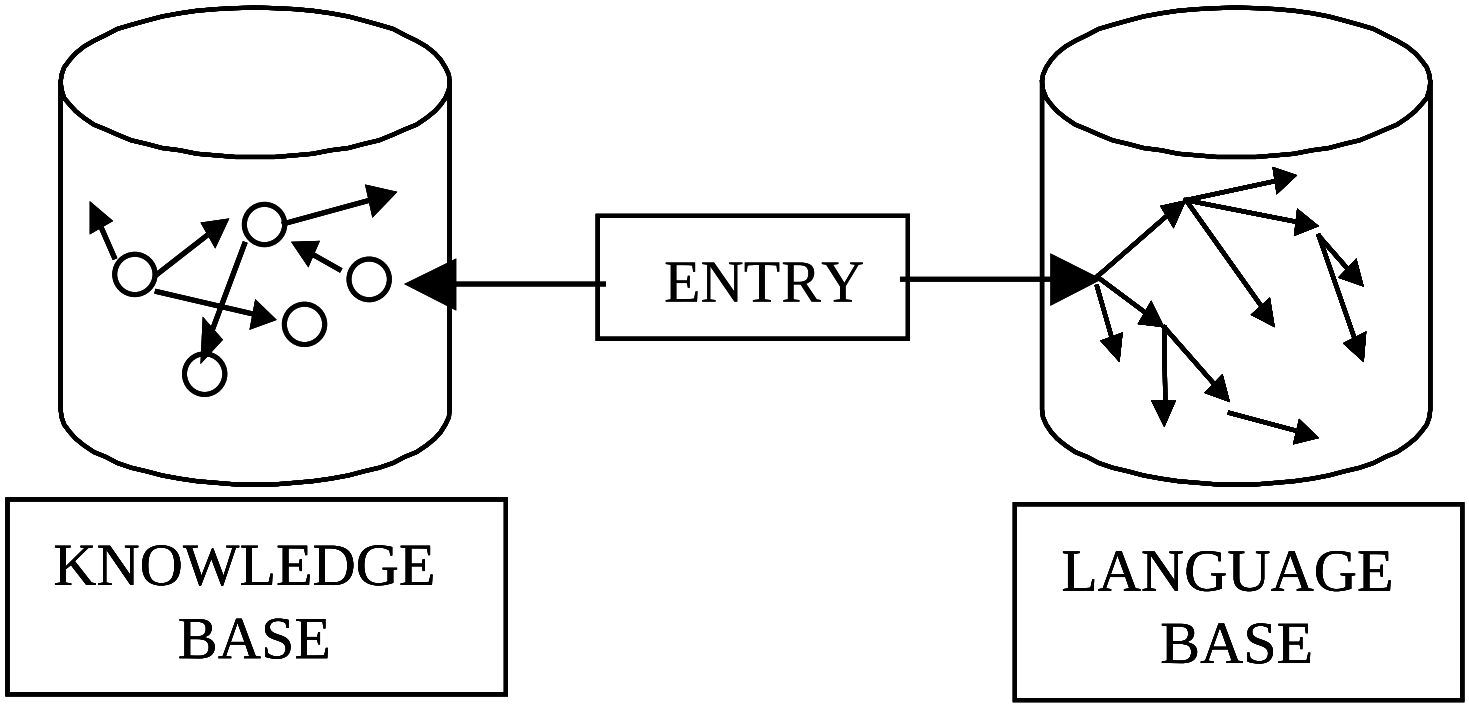
The Past

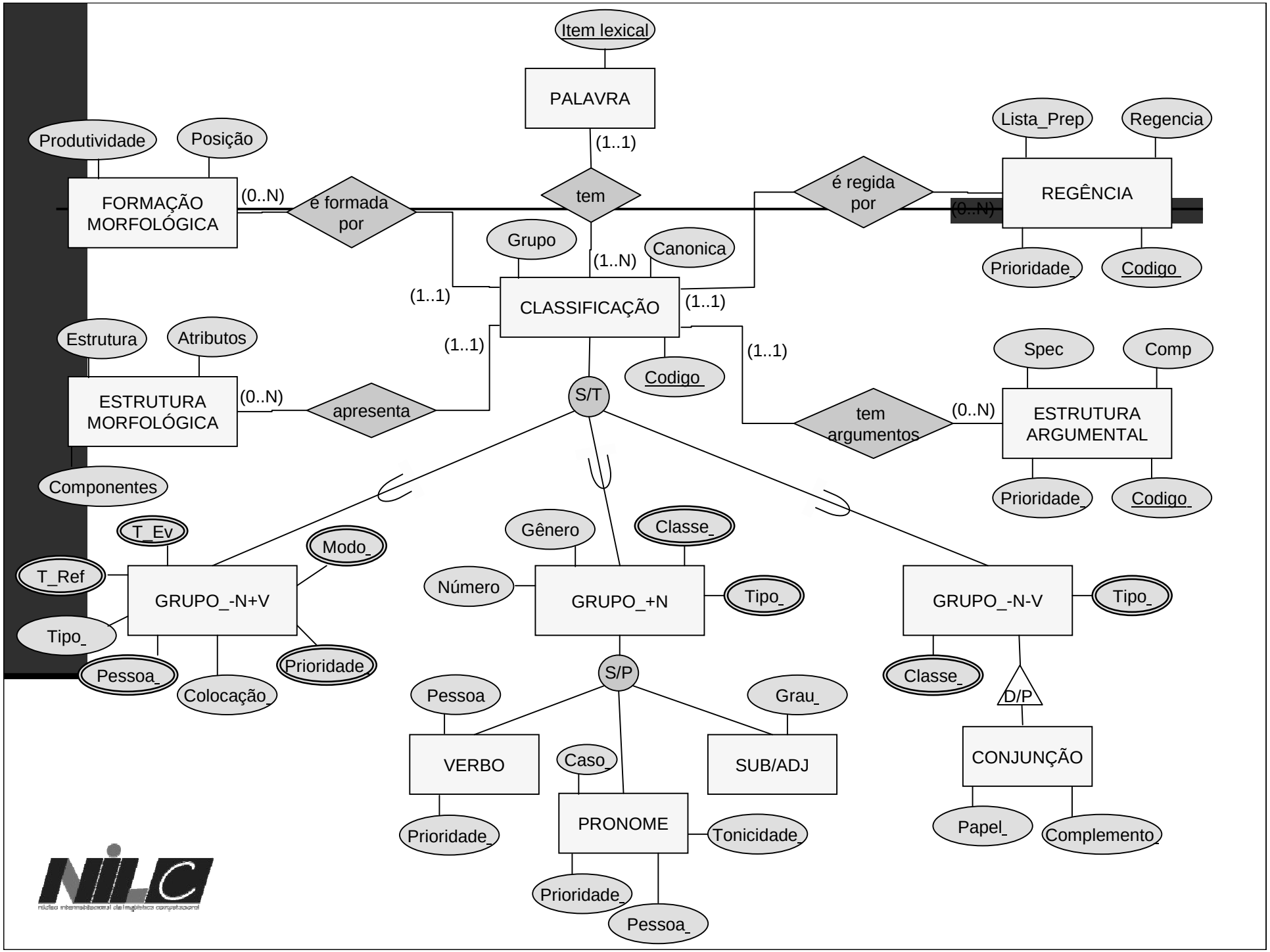
- Goal: spell, grammar and style checking
- Choices
  - full words vs. analyzed forms
  - single words vs. complex words
  - categorization
    - part-of-speech
    - morphological information
  - frequence order assignment
  - automatic generation
  - human checking

The Present

A=<ART.F.SI.DE.?.?.[o]0.#PREP.[a]0.#PRON.F.SI.3P.[
DEM.OBL-
AT.]?.?.[o]0.#ABREV.M.SI.[a]0.#S.M.SI.N.[]?.?.[a]0.>
Capitania=<S.F.SI.N.[]?.?.[capitania]0.>
da=<PREP.C.[de.a.][do]0.>
Bahia=<NOM.F.SI.[bahia]0.>
com=<PREP.[com]0.#ABREV.M.SI.[com]0.>
50=<NUMERO>
léguas=<S.F.PL.N.[]?.?.[légua]0.>
de=<PREP.[de]0.>
comprimento=<S.M.SI.N.[]?.?.[comprimento]0.>
,=<VIRGULA>

The Future




Item lexical
PALAVRA
(1..1)
Produtividade
Posição
FORMAÇÃO MORFOLÓGICA
(0..N)
é formada por
tem
é regida por
Lista_Prep
Regencia
REGÊNCIA
(0..N)
Prioridade_
Codigo_
Grupo
(1..N)
Canonica
(1..1)
CLASSIFICAÇÃO
(1..1)
Estrutura
Atributos
ESTRUTURA MORFOLÓGICA
(0..N)
apresenta
(1..1)
Codigo_
S/T
(1..1)
tem argumentos
(0..N)
Spec
Comp
ESTRUTURA ARGUMENTAL
Prioridade_
Codigo_
Componentes
T_Ev
Modo_
T_Ref
GRUPO_-N+V
Tipo_
Pessoa_
Colocação_
Prioridade_
Gênero
Classe_
Número
GRUPO_+N
Tipo_
S/P
Pessoa
Grau_
VERBO
Caso_
SUB/ADJ
Prioridade_
PRONOME
Tonicidade_
Prioridade_
Pessoa_
GRUPO_-N-V
Tipo_
Classe_
D/P
CONJUNÇÃO
Papel_
Complemento_

The Golden List

- Relative lack of convergence on the theoretical background

The Golden List

- What should stand for a lemma?
  - dimunitives (“caminha”) -> positives (“cama”)?
  - augmentatives (“abelhão”) -> positives (“abelha”)?
  - superlative (“chiquérrimo”) -> positive (“chique”)?
  - derived (“mecanicidade”) -> original (“mecânico”)?
  - ordinal (“nono”) -> cardinal (“nove”)?
  - abbreviations (“níver”) -> original (“aniversário”)?
  - etc.
- synchronic vs. diachronic criteria
- morphological vs. semantic criteria
- ReGra: synchronic + morphological (to deliver alternatives)

The Golden List

- What should stand for an entry?
  - “apesar de” vs. “apesar” and “de”
  - clitics (“referiam-se”, “reunir-se-iam”)
  - “não-violento” vs. “não-” and “violento”
  - “melhores” vs. “melhor” and “-es”
  - “desumanamente” vs. “desumano” and “-mente”
- ReGra: string of ANSI characters isolated by blank spaces

The Golden List

- What should stand for dictionary features?
  - Phonetics
  - Morphology
  - Syntax
  - Semantics
  - Pragmatics
- ReGra: problem-based category assignment

A checker dictionary commitments Phonetics

- atone vs. tonic (for hyphenization checking)
  - Ele feriu se (instead of Ele feriu-se)
- phonetic changes (for alternatives) >> spelling errors
  - phonetic transcription: caza (casa), mininu (menino)
  - phoneme addition: avoar (voar), adevogado (advogado), favore (favor)
  - phoneme subtraction: tá (está), pra (para), cantá (cantar)
  - phoneme reordering: tauba (tábua), estrupo (estupro)
  - phoneme exchange: tó/ch/ico (tó/ks/ico), ine/ks/orável (ine/z/orável), ab/r/upto (ab/x/upto)
  - accent changes: ‘rubrica (ru’brica), ca’teter (cate’ter)

A checker dictionary commitments Morphology

- Part-of-speech
  - *Ela chegou rápida
  - *Há muita pouca gente
- Structure
  - *Interviu
  - *Adequa
  - *Pãozinhos
- Number
    - *as felicidades
    - *a cócora

A checker dictionary commitments Morphology

- Gender
  - *Cerveja é boa
- Person
  - *Se você não se cuidar, a AIDS vai te pegar.
- Tense
  - *Eu queria que ela saísse.
- Mood
  - *Ele espera que eu saio mais cedo.
- Aspect
  - *Ele estava querendo sair.

A checker dictionary commitments Syntax

- Transitivy
  - *Ele custou a sair.
- Positioning
  - *Farei-o amanhã.
- Agreement
  - *Nem um nem outro irão à festa.
- Government
  - *Ele pagou o médico.

A checker dictionary commitments Semantics

- Lexical choice
  - *A mala está leviana.
  - *O médico infligiu a lei.
  - *O sangue fruía na calçada.
- Semantic anomaly
  - *Quadrados triangulares
- Contradiction
  - *Minhas idéias vão de encontro às suas: não há motivo para brigas.

A checker dictionary commitments Pragmatics

- Taboo words
- Foreign words
- Archaisms and neologisms
  - Colóquios flácidos para acalentar bovinos.
  - otimizar, maximizar, inicializar
- Clichés
  - correr atrás do prejuízo
  - a nível de

Final remarks

- As far as word formation licensing is rather historical and social, it is not possible to devise general procedures for morphological analysis capable of generating only authorized words.
  - casamento, but *casação
  - transação, but *transamento
- Is it possible (and worthwhile) to contrast error-driven lexical databases with general-purpose ones? If so, how to compare two differently-oriented lexical databases in a productive way?