Além da correção ortográfica nos editores de textos livres

William D. Colen M. Silva (colen@users.sourceforge.net)

Eng. Computação pela Escola Politécnica da USP (2006)
Mestrando Ciência da Computação – IME USP
Desenvolvedor CoGrOO desde 2004

fisl10
10° Fórum Internacional Software Livre
A tecnologia que liberta
Edição Especial

CCSL CENTRO DE COMPETÊNCIA EM SOFTWARE LIVRE
FLOSS Competence Center

IME-USP

Corretor Gramatical CoGrOO
Além da correção ortográfica nos editores de textos livres

- Agenda
  - Sobre o projeto
  - Além da correção ortográfica
  - Como funciona o CoGrOO
  - Demonstração / Módulos
  - O CoGrOO 3.1 e o CoGrOO 4.0
  - Como fazer o melhor corretor gramatical
  - O papel da comunidade
  - Além da correção gramatical
  - Conclusões

Apresentação do CoGrOO

- Primeiro e único
- Mais de 35 mil downloads diretos (contando apenas da versão 2.0 em diante)
- Estimativa: +100 mil usuários
- Usado por empresas estatais e privadas. Algumas empresas tem ele instalado em milhares de máquinas
- Reconhecido localmente como sendo um importante esforço para o desenvolvimento do BrOffice.org

Celepar – Informática do Paraná

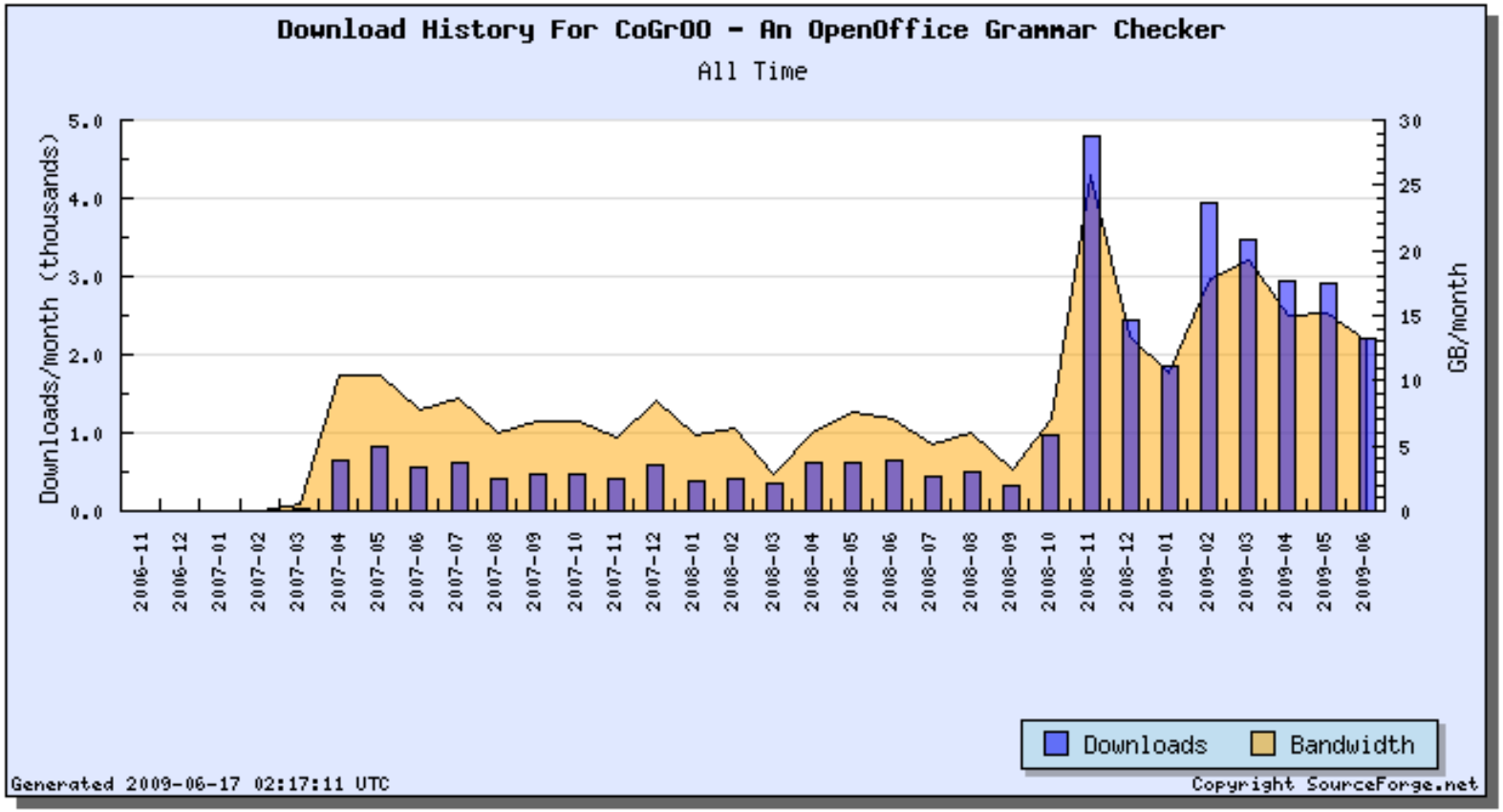

BrOffice.org
MOVIMENTO SOFTWARE LIVRE PARANÁ
Gubro-PR
SORTEIO
BONÉ
ODF
17h
BrOffice.o
Padrão
INSIGNE
ODF
CoGrOO
CELEPAR
Desktop Paraná

Banco do Brasil – 35 mil máquinas

Correios (piloto)

Petrobras (piloto) – 85 mil

Download History Statistics



Name	Relevance	Activity	Rank	Registered	Latest File	Downloads
CoGrOO - An OpenOffice Grammar Checker		99.77%	574	2006-11-13	2009-06-14	34,861

Apresentação do CoGrOO

- Foi o primeiro corretor gramatical integrado ao OOo do mundo.
- Segundo mais utilizado (perde apenas para o Language Tool, que suporta inglês).
- O projeto Golfiño (Galego) foi criado a partir do CoGrOO.
- Recebemos pedidos para criar versões do CoGrOO para português de Portugal e para o Espanhol.
- Módulos foram úteis para outros trabalhos, como por exemplo um grupo de pesquisa sobre saúde usou o CoGrOO na análise de prescrições medicas.

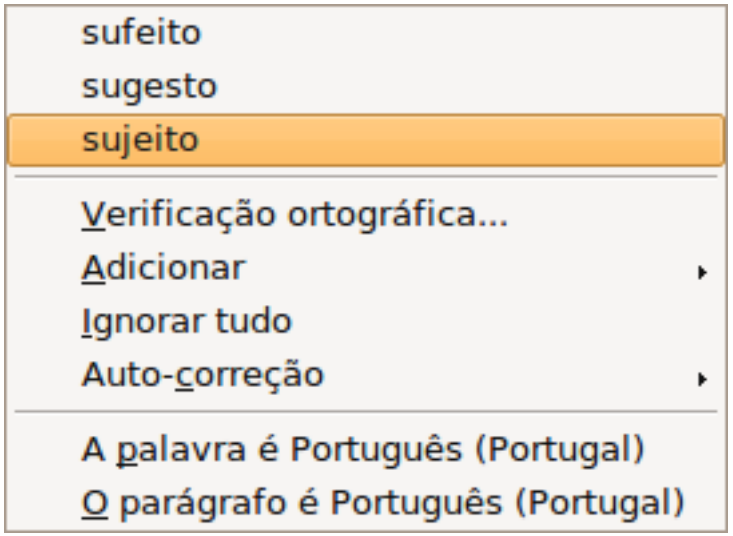
Apresentação do CoGrOO

- Hospedado pelo SourceForge
- Licença LGPL
- Fácil instalação e uso
- Atualizações frequentes
- Apoio da comunidade
- Já foi integrado com
  - Firefox plug-in (Bruno Sant'Anna)
  - AbiWord (SoC Gabriel Bakiewicz)
  - WebSevices LangBot Apoema (Bruno Sant'Anna)

Além da correção ortográfica

Qual é o sugeito da frase?

- Usuário entra um texto
- Verificador ortográfico tenta encontrar a palavra digitada em seu banco de dados
- Caso a palavra não seja encontrada ele usa algoritmos de similaridade para encontrar possíveis correções



Além da correção ortográfica

- VERO é o corretor ortográfico do BrOffice.org
  - Dicionário com 304 mil entradas
  - Descritor de afixos com mais de 25 mil linhas
  - **menino/DOPQR** → menino menina meninos meninas meninão menininho …
  - Ainda é responsável pela separação silábica
- O VERO conta com apoio da comunidade e é coordenado pelo Raimundo Moura. Frequentemente atualizado.
- Primeiro a incorporar o Acordo Ortográfico
- É o corretor oficial do BrOffice.org, e também pode ser usado no Firefox (e outros que suportarem o *Hunspell*)

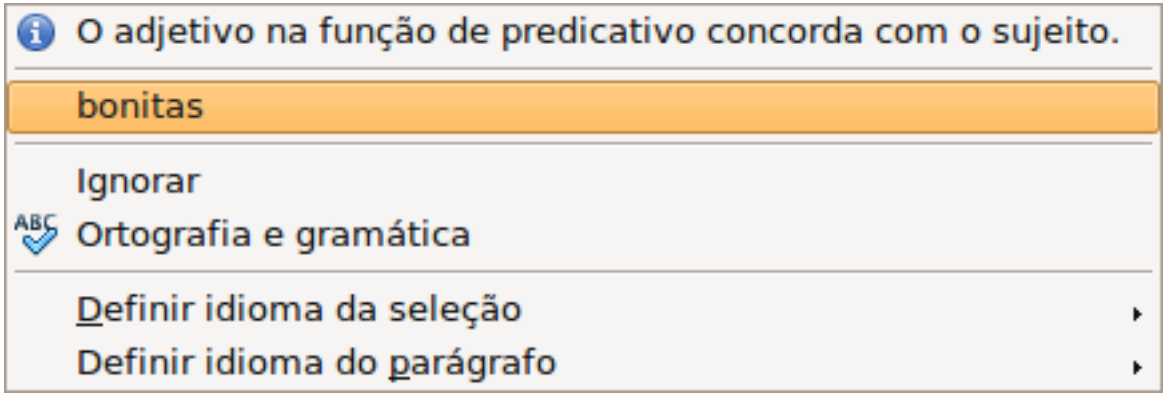
Além da correção ortográfica

- Ainda hoje a análise ortográfica é objeto de estudo na Ciência da Computação
- Existem muitos problemas que ainda não possuem solução ótima
- Existem softwares que podem tratar de diversos idiomas, por exemplo no OOo

Corretor Gramatical

As bolas são bonitos.

- Usuário entra um texto
- O verificador tenta fazer a análise gramatical do texto e na estrutura de dados gerada ele busca por padrões de erros



- O verificador tenta sugerir correções para o texto.

Corretor Gramatical

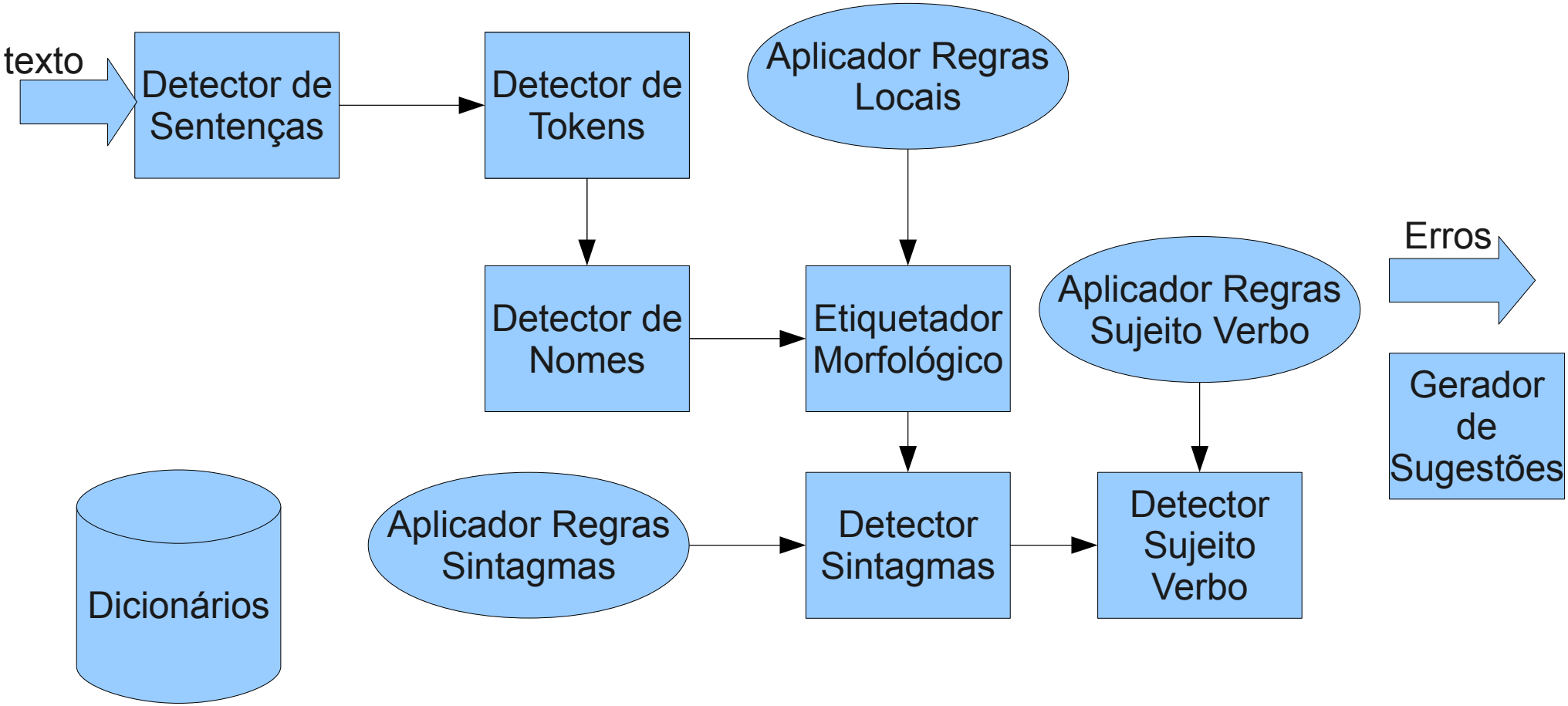
- Não existe um padrão funcional para todas as línguas (como o *Hunspell* para verificação ortográfica)
- Requer a análise detalhada do texto. É um processo que consome bastante recurso
- Existem muitos problemas ainda em aberto na análise gramatical automática, é tema de muitas pesquisas em Engenharia e Ciência da Computação (Linguística Computacional).

Como funciona o CoGrOO – Análise Gramatical

Problema Fundamental: Resolver ambiguidades

- Detecção de limites de palavras/sentenças
  - “Sr. Silva estava jogando futebol.”
  - “O computador novo custará R$ 2.500,00.”
- Ambiguidades nos sentidos das palavras
  - “Nada como voltar para **casa**!” (substantivo)
  - “Ele se **casa** na semana que vem.” (verbo)

Como funciona o CoGrOO



Dicionários

- Dicionários de palavras com classificação morfológica
  - casa: [verbo casar] [substantivo feminino singular]
- Dicionários de relacionamentos entre palavras
  - meninas → menino → menino meninos menina meninas
- Dicionário de abreviaturas

  sr.
  tel.
  r.

Separador de Sentenças

- Entrada:
  - [Ele foi procurar uma casa. Ele vai se casar com a Srta. Maria.]
- Saída:
  - [Ele foi procurar uma casa.]
  - [Ele vai se casar com a Srta. Maria.]
- Desafio:
  - Decidir se marcas de fim de linha estão separando linhas no contexto. Exemplo "Srta."

Separador de Tokens

- Entrada:
  - [A Sra. Maria, esposa do Sr. José, trouxe-nos frutas.]
- Saída:
  - [A][Sra.][Maria][,][esposa][do][Sr.][José][,][trouxe][-nos][frutas][.]
- Desafio
  - Além dos espaços muitos outros símbolos podem separar *tokens* na frase. Exemplo "José, trouxe-nos" são quatro *tokens*.

Etiquetador Morfológico

- Entrada:
  - [Ele foi procurar uma casa.]
- Saída:
  - [Ele, pronome pessoal masculino 3ª pessoa singular]
  - [**foi**, verbo ir passado 3ª pessoa do singular]
  - [procurar, verbo procurar no infinitivo]
  - [**uma**, artigo indefinido feminino singular]
  - [**casa**, substantivo feminino singular]
  - [., ponto final]
- Desafio
  - Muitas palavras de mesma grafia podem ser classificadas de diferentes formas de acordo com o contexto em que estão. Por exemplo "casa", que pode ser substantivo ou verbo (casar).

Agrupador

- Entrada:
  - [Ele, pronome pessoal masculino 3ª pessoa singular]
  - [foi, verbo ir passado 3ª pessoa do singular]
  - [procurar, verbo procurar no infinitivo]
  - [uma, artigo indefinido feminino singular]
  - [casa, substantivo feminino singular]
  - [., ponto final]
- Saída:
  - [Ele, sintagma nominal masculino 3ª pessoa singular]
  - [foi procurar, sintagma verbal 3ª pessoa singular]
  - [uma casa, sintagma nominal feminino 3ª pessoa singular]
  - [., ponto final]
- Desafio
  - Encontrar sequências que poderiam ser tratadas como elemento único. Exemplo "uma casa".

Analisador Sintático Simples

- Entrada:
  - [Ele, sintagma nominal masculino 3ª pessoa singular]
  - [foi procurar, sintagma verbal 3ª pessoa singular]
  - [uma casa, sintagma masculino feminino 3ª pessoa singular]
  - [., ponto final]
- Saída:
  - [Ele, sujeito]
  - [foi procurar, verbo]
  - [uma casa, sintagma nominal feminino 3ª pessoa singular]
  - [., ponto final]
- Desafio
  - Identificar entre os sintagmas quais compõem sujeito e verbo.

Análise da Arquitetura e do Desenvolvimento

Tipos de erros:

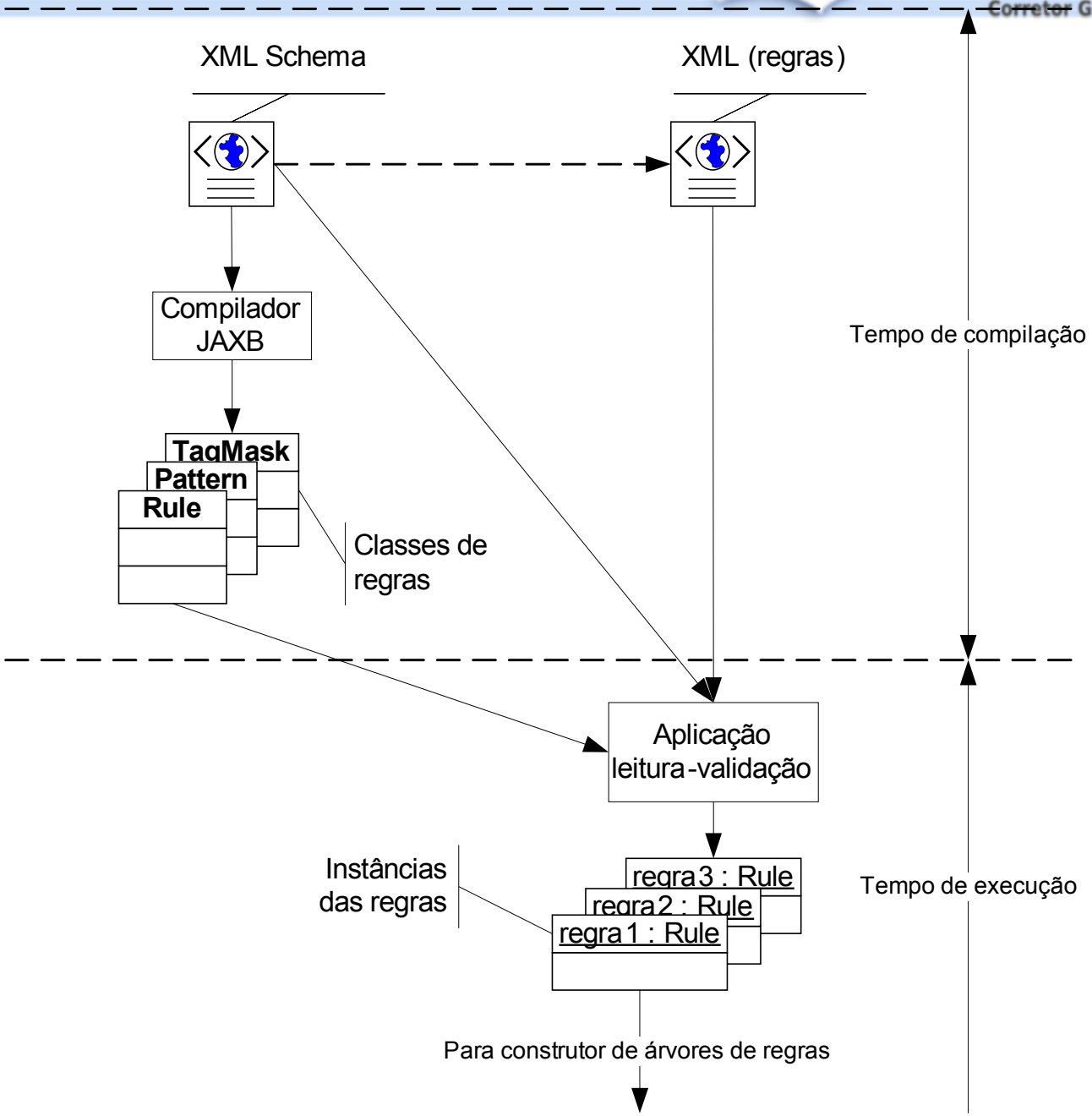
- colocação pronominal
- concordância nominal
- concordância entre sujeito e verbo
- concordância verbal
- uso de crase
- erros comuns da língua portuguesa escrita

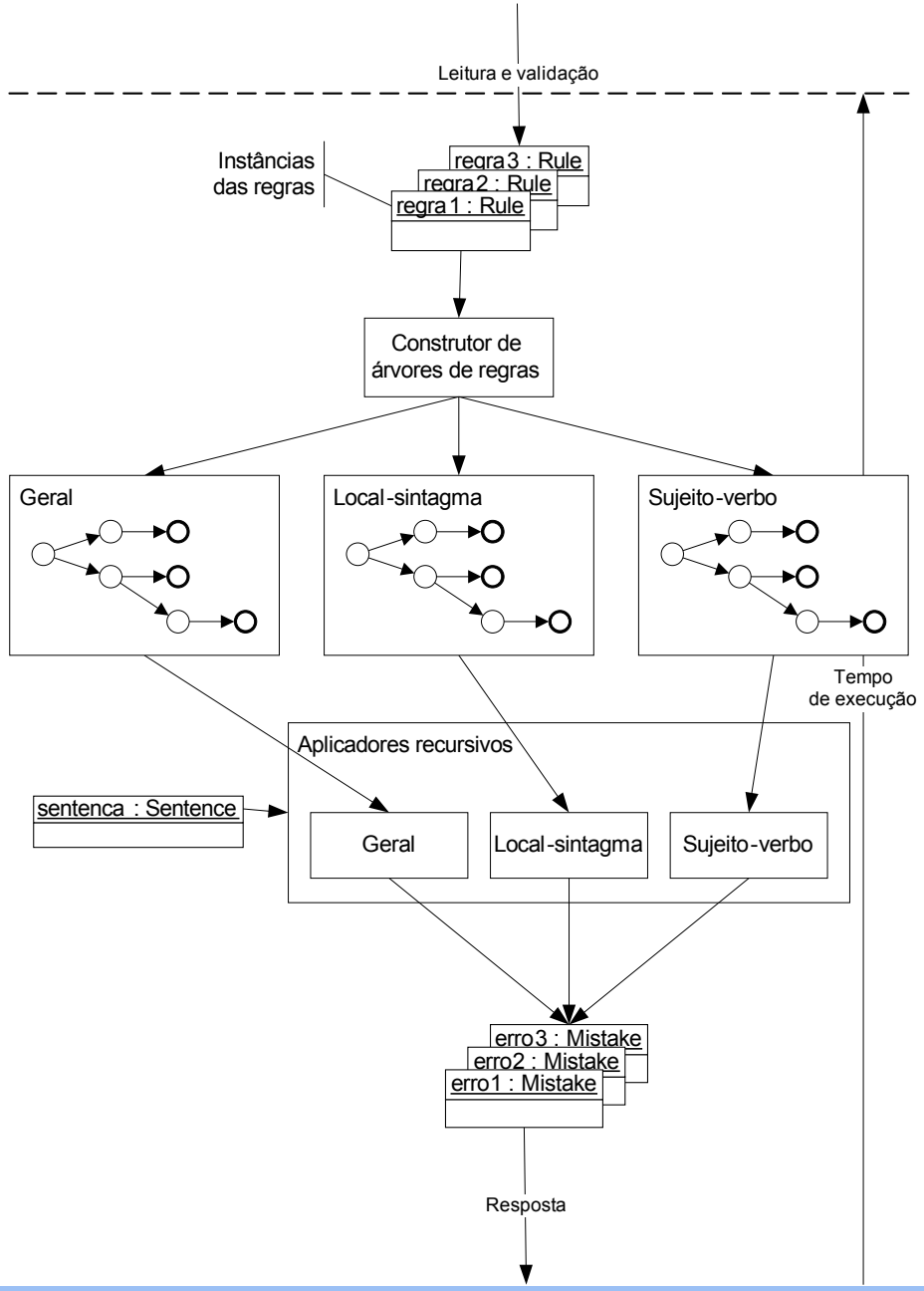
Regras

- Análise de desvios por padrões
- Estrutura de regra
  - Método
  - Mensagem
  - Padrão
    - Exemplo: artigo masculino plural + substantivo masculino singular
  - Modelos genéricos de sugestão
- Descritos em arquivo XML e validados por um XSD

Regras

- Árvores
  - Criação de árvores de busca a partir dos padrões das regras
- Aplicadores
  - São algoritmos recursivos que fazem a busca com base nas árvores e na sentença processada pelo CoGrOO


XML Schema
XML (regras)
Compilador JAXB
TagMask
Pattern
Rule
Classes de regras
Tempo de compilação
Aplicação leitura-validação
Instâncias das regras
regra3 : Rule
regra2 : Rule
regra1 : Rule
Tempo de execução
Para construtor de árvores de regras


Leitura e validação
Instâncias das regras
regra3 : Rule
regra2 : Rule
regra1 : Rule
Construtor de árvores de regras
Geral
Local-sintagma
Sujeito-verbo
Tempo de execução
Aplicadores recursivos
sentenca : Sentence
Geral
Local-sintagma
Sujeito-verbo
erro3 : Mistake
erro2 : Mistake
erro1 : Mistake
Resposta

Sistemas Desenvolvidos

- Treinamento
- Refinamento dos parâmetros de treinamento
- Teste de desempenho
- Teste das regras
- Visualizadores gráficos
- Servidor RPC
- Servidor XML
- Integração com o OpenOffice.org

Desempenho

Testes de Desempenho

<table>
<tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td></tr>
<tr><td colspan="3">Treinamento</td><td></td><td colspan="7">Treinamento</td></tr>
<tr><td colspan="4">Treinamento</td><td></td><td colspan="5">Treinamento</td></tr>
</table>

Desempenho

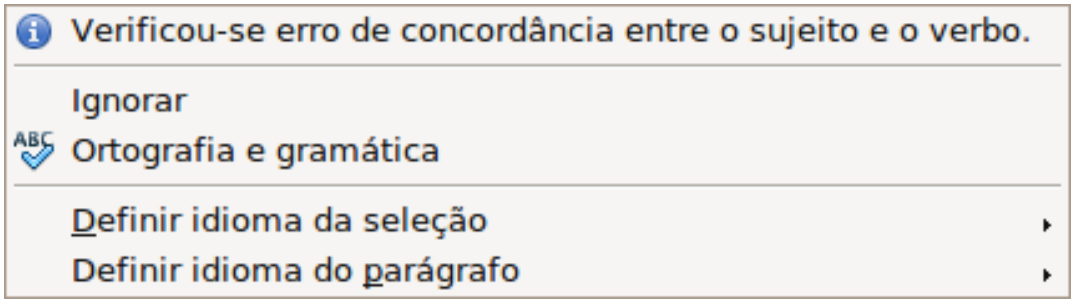
- Tokenizer - 98,74%
  - Considera a sentença
- Name Finder – 90,11%
  - Considera a sentença
- Tagger – 96,05%
  - Considerado cada token
- Chunker – 77,25%
  - Considera a sentença
- Shallow Parser – 68,80%
  - Considera a sentença

Futuro: um CoGrOO mais forte

- Expandir seu uso para atrair apoio
  - Melhorar sensivelmente o corretor para o português – conquistar mais usuários no Brasil
  - Ser suportado por outros programas (Firefox, KOffice) – atingir não usuários do BrOffice.org.
  - Disponibilizar API para processamento de linguagens naturais – apoio da comunidade acadêmica
  - Suportar outros idiomas – conquistar usuários de fora do Brasil

CoGrOO 3.1

- Novo Chunker e Shallow Parser – fim do falso erro de concordância sujeito verbo.



- Refatoração e melhorias de desempenho usando profiler
- Documentação dos módulos internos de análise gramatical, que serão disponibilizados para a comunidade
- Testes automatizados
- Incorporação do Acordo Ortográfico

CoGrOO 4.0

- Adoção de uma arquitetura padrão para processamento de linguagens naturais
- Separar o corretor gramatical da máquina de processamento de linguagens
- Sistema de plug-ins permite customizações (troca de arquivo de regras, troca de motor de análise gramatical etc)
- Suporte a múltiplos idiomas
- Meios de reportar erros diretamente do BrOffice.org

Além da correção gramatical...

E agora? Já existe um corretor gramatical... o que eu poderia fazer?

Além da correção gramatical...

- Sumarização
- Verificação de legibilidade
- Auto-texto
- Análise semântica
- Outros auxílios à redação
- Interpretação de texto e busca por referências relacionadas

Quem pode fazer o melhor corretor do mundo?

A comunidade!

Comunidade (Colaboradores Web)

- Página que possibilita experimentar o CoGrOO e seus módulos pela Web
- Página que possibilita escrever e testar regras online – regras poderiam ser submetidas para a equipe avaliar
- Página que aplica o corretor sobre textos extraídos do Wikipédia – interface permitiria que o colaborador determinasse a causa do erro (dicionário, etiquetador)
- Página que permite entrar com texto livre para cadastrar mal funcionamento do corretor

Discussão.....

http://cogroo.sourceforge.net