Um Revisor Gramatical para o BrOffice

Peter Tscherdantzew Neto[*], Fabiano da Silva Silveira[*], Marnes Augusto Hoff[*], Márcio Rocha Zacarias[*], Susan Severo de Severo[*], Tânia Maria Steigleder Costa[**], Cláudio Eduardo Farias Nunes Pereira[*], Daniel Nehme Müller[*]

*Conexum – Sistemas Computacionais Inteligentes Ltda.
Centro de Empreendimentos em Informática – Instituto de Informática
Universidade Federal do Rio Grande do Sul

**Curso de Letras
Universidade Luterana do Brasil

[*]conexum@conexum.inf.br, [**]tsteigleder@hotmail.com

Abstract. *This paper aims to present Literal grammar checker. Literal is a Portuguese grammar checker plugin to BrOffice. It was sponsored by Finep (2003) and CNPQ (2006-2007). The source code has a LGPL license and is written in C++ language with SQLite database to lexikon management. It uses UNO package to plugin installation and the interface was made in Java. Literal is a full grammar checker with concordance and regency analysis, both with nominal and verbal approaches, and lexicon with more than a thousand of words . The actual beta version has been optimized to show best recommendations to correct the text. Literal is a reality and a competitive alternative for the Portuguese written texts.*

Resumo. *Este artigo apresenta o revisor gramatical Literal, que é uma extensão para a suíte BrOffice. O desenvolvimento do Literal foi apoiado pela Finep em 2003 e pelo CNPq em 2006-2007. O código fonte possui licença LGPL e foi totalmente escrito em C++ com banco de dados SQLite para gerenciamento do léxico. O pacote UNO é utilizado para instalação e sua interface no BrOffice é feita em Java. Literal é um revisor gramatical completo, realizando análises de concordância e regência, ambas com verificação nominal e verbal, além de um léxico de mais de um milhão de palavras. A versão beta, atualmente em desenvolvimento, está sendo otimizada para apresentar melhores recomendações de correção. Literal é uma realidade e uma alternativa competitiva para textos editados no BrOffice.*

1. Introdução

A comunidade usuária do software livre até hoje não conta com um verificador gramatical eficiente para o Português brasileiro. Mas esta constatação não provoca surpresas, uma vez que, mesmo em se falando de software proprietário, encontramos apenas um editor de textos em que os verificadores gramaticais comerciais funcionam adequadamente.

Por isso, a Conexum tem desenvolvido desde 2003 um verificador gramatical para inserção na suíte BrOffice / OpenOffice. Neste sentido, projetos para desenvolvimento do verificador Literal foram apoiados em 2003 pela Financiadora de

Estudos e Projetos (FINEP) e atualmente pelo CNPq (2006-2007), através de seu programa RHAE.

Os verificadores gramaticais normalmente usam como base para seu processo de análise lingüística a avaliação de palavras de vizinhança, através de um processo estatístico, tal como é feito com os sistemas ReGra [Nunes 2000] e CoGroo [Uliano 2006]. O sistema Literal, por outro lado, realiza a análise de sintagmas e sua posterior reorganização em orações – ver [Perini 1989], [Tondo 1974] e Sautchuk [2004]. A opção pela análise a partir dos sintagmas fundamenta-se em modernas teorias gramaticais, baseadas no avanço dos estudos lingüísticos. Sintagmas são grupos de palavras que possuem um núcleo comum dentro de uma frase. Os núcleos principais de uma frase são o sujeito (um ou mais substantivos) e o verbo, e outros núcleos periféricos podem ser adjetivos ou advérbios, por exemplo. A figura 1 apresenta um exemplo de separação de sintagmas executada internamente pelo sistema Literal. Trata-se de um período composto por duas orações, portanto, cabe a análise em separado para cada oração (veja os dois sintagmas verbais – SV – apresentados).

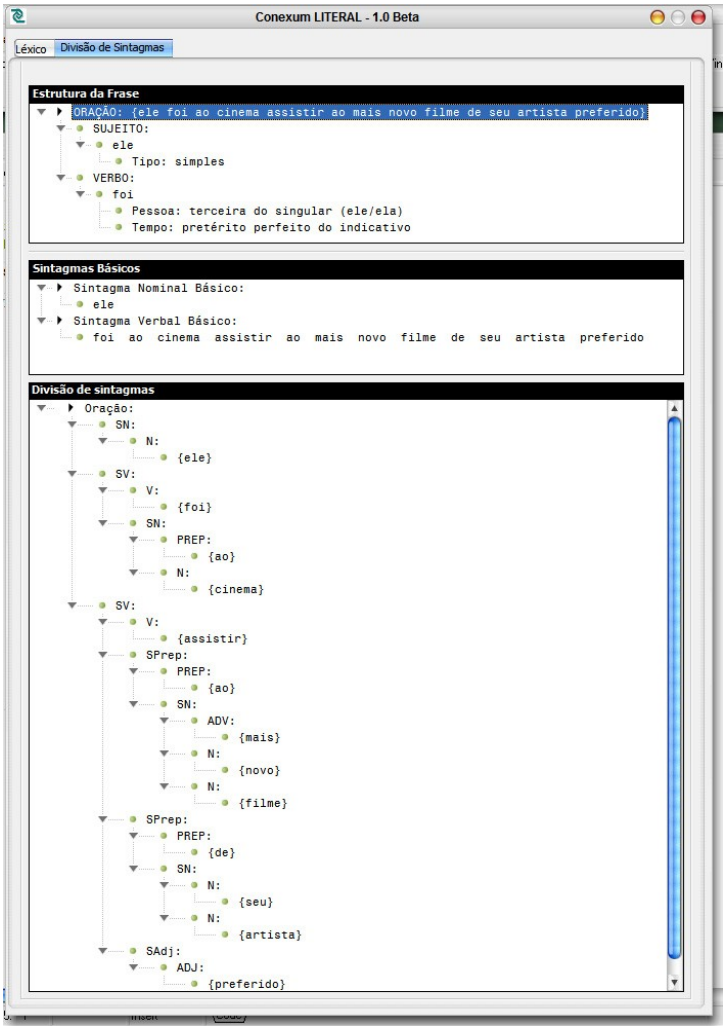

Figura 1. Exemplo de análise por divisão de sintagmas para a frase *ele foi ao cinema assistir ao mais novo filme de seu artista preferido*.

No seguimento deste artigo, será apresentado o sistema Literal, sendo descrito seu projeto e seu funcionamento básico.

2. Descrição do Sistema

O corretor gramatical Literal está sendo construído no paradigma de Orientação a Objetos, de modo a proporcionar maior flexibilidade, estabilidade e precisão ao sistema.

A tecnologia de banco de dados também está sendo aplicada para o melhor desempenho no armazenamento, atualização e manipulação dos diversos dicionários utilizados pelo sistema Literal.

As futuras versões do sistema Literal poderão ser facilmente atualizadas através da inserção de novas regras gramaticais e novas entradas no banco de dados (dicionários).

2.1. Funcionamento

Quando o usuário digita um texto e pede para corrigi-lo, esse texto é quebrado em frases e depois em palavras, que são classificadas gramaticalmente e agrupadas em sintagmas. A partir dessa organização, são feitas as análises de concordância nominal e verbal.

A concordância nominal verifica, entre outras coisas, se um artigo está coerente em gênero e número com o substantivo associado (p. ex., a expressão “os caneta” possui erros em gênero e número). A concordância verbal indica se o verbo concorda com o substantivo relacionado (p. ex., a expressão “eles foi” tem um erro de concordância verbal).

Após avaliada a concordância, são realizadas as análises de regência nominal e verbal. Basicamente elas dizem respeito às preposições que podem ser colocadas após substantivos ou verbos. Por exemplo, a frase *ele foi programado em calcular* está com a regência nominal errada, uma vez que é a preposição *para* que rege *programado*. E na frase *o monumento está situado ao topo do monte*, há um erro de regência verbal, uma vez que a preposição *em* é a correta para uso com o verbo *situar* (...*situado* ***no*** *topo*...).

Uma vez concluídas as análises, o sistema Literal gera um arquivo XML contendo os erros encontrados em cada frase. Posteriormente, o arquivo XML é lido pela interface do sistema junto ao BrOffice, que apresenta os erros ao usuário. Os passos descritos estão esquematizados no *use-case* da figura 2.

2.2. Geração de sugestões de correção

A geração de sugestões de correção gramatical é um processo muito mais complexo do que a geração para ortografia. São muitas as variáveis presentes na situação de erro. O que o sistema Literal faz é separar quais dessas variáveis estão envolvidas no erro e, a partir das informações que elas trazem, encontrar uma ou mais sugestões para corrigir o erro.

O primeiro passo é armazenar as palavras que não se enquadram na organização de um sintagma. A partir dos dados da segunda variável/palavra, procura-se no banco de dados (dicionário léxico e verbal) alguma palavra similar, mas com mesmo gênero e/ou número da segunda (que são as informações úteis – para a geração da sugestão – que essas variáveis trazem consigo). O mesmo processo é repetido para a primeira variável/palavra.

As palavras encontradas no dicionário que se enquadram nos quesitos de busca são transferidas para a frase, na posição do erro encontrado, de modo a gerar uma sugestão de correção.

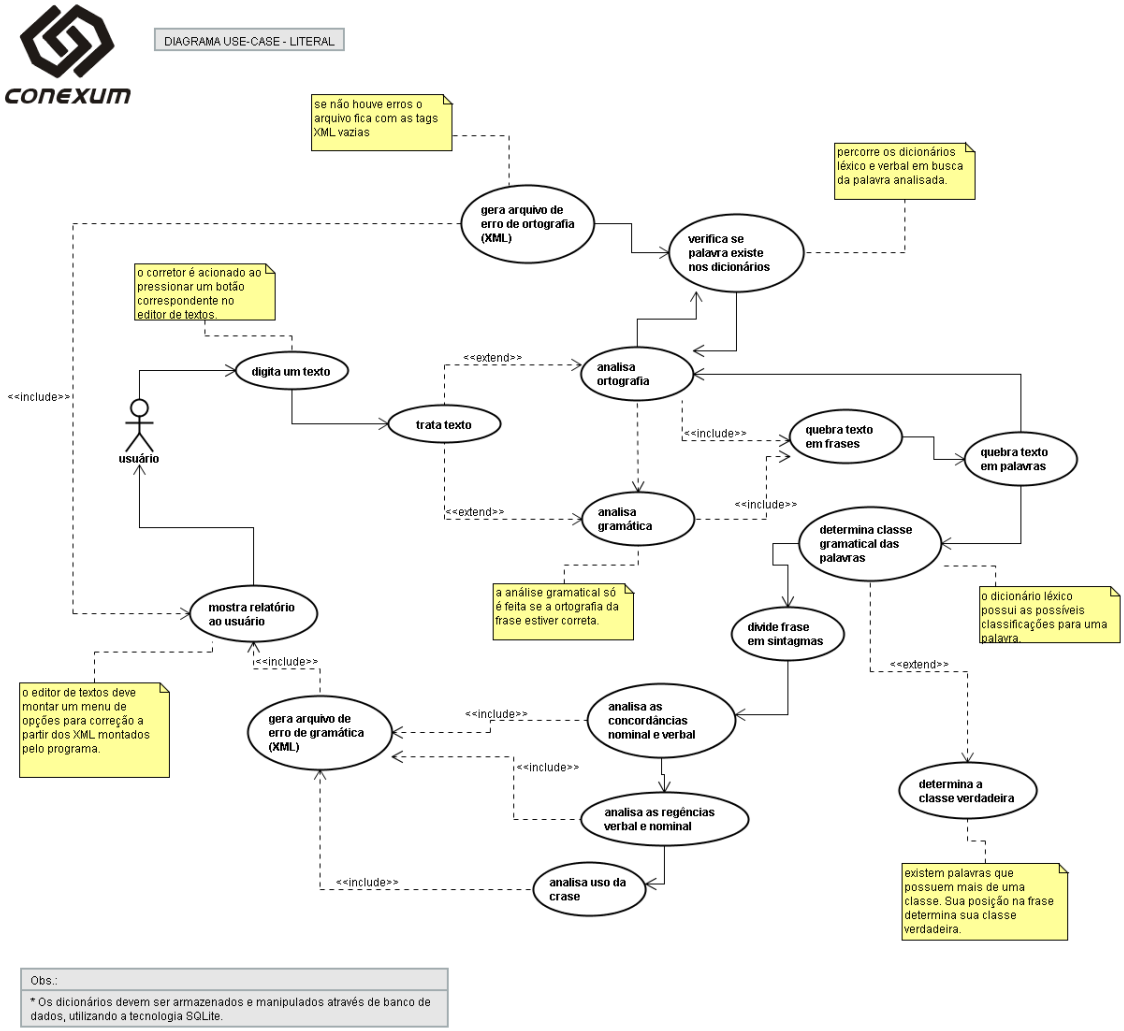

Figura 2. Diagrama *use-case* do sistema Literal.

Ao final do processo, um arquivo XML, contendo todos os erros e sugestões encontrados, é gerado. Além das sugestões, o arquivo XML contém também o tipo do erro (concordância nominal, verbal, etc...), a frase que contém o erro com seu identificador (número da frase no texto inteiro) e uma explicação sobre o erro encontrado.

Após esse processo, o arquivo XML é lido pela interface do Literal com o BrOffice. Um exemplo de janela para apresentação das sugestões de correção é apresentado na figura 3.

2.3. Classes do sistema Literal

As classes que fazem parte do grupo *função* gramatical, responsável pela análise gramatical, são:

- **clGramatica:** classe principal das rotinas gramaticais. Coordena toda a análise, desde a separação de sintagmas, análise de concordância, regência, até a análise de erros comuns do português;
- **clSintagma:** quebra a frase em sintagmas para análise futura. Cada sintagma é armazenado numa estrutura em memória;

- **clConcordancia:** coordena a análise das concordâncias nominal e verbal e ainda a análise de substantivos compostos (quando houver) na frase;
- **clRegencia:** executa a análise de regências verbal e nominal;
- **clRegrasConcV:** contém o conjunto de regras de concordância verbal. Cada método da classe equivale a uma regra;
- **clRegrasConcN:** contém o conjunto de regras de concordância nominal. Cada método da classe analisa uma regra de concordância;
- **clRegrasSubsC:** traz o conjunto de regras para o uso de substantivos compostos numa frase;
- **clRelatorio:** escreve, no formato XML, o resultado das análises gramaticais de cada frase.

Figura 3. Um exemplo de apresentação do erro ao usuário.

É importante salientar que, para se chegar à análise gramatical, é preciso, primeiramente, que a frase não apresente erros ortográficos. Não há como o sistema analisar gramaticalmente uma frase se alguma palavra for desconhecida (ou estiver errada). A identificação das palavras é importante pois, assim, sabe-se a qual classe a palavra pertence e isso é imprescindível para a análise gramatical.

Uma frase, quando quebrada em sintagmas, pode assumir 5 tipos diferentes: sintagma nominal, verbal, adjetivo, adverbial e preposicionado (ou preposicional). Sem contar ainda que o sistema Literal divide, primeiramente, de acordo com as regras da língua portuguesa, a frase em dois sintagmas básicos: sintagma nominal base e sintagma verbal base.

Cada sintagma é armazenado em uma estrutura em memória. Cada nó da estrutura possui campos para armazenar cada palavra, o número de palavras do sintagma, o código do sintagma e qual palavra é o núcleo.

Ainda há algumas classes de uso geral do sistema e que não se enquadram em apenas um dos dois grupos citados acima. São elas:

- **clCorretor:** é o núcleo do corretor e responsável por coordenar todo o processo

de inicialização do sistema e análises;

- **clDicionarios:** uma das mais importantes classes do sistema. É ela que manipula o banco de dados SQLite com os dicionários e pesquisa palavras para as sugestões de correção.

Ao todo, são tabelas no banco:

- **léxico (diclex):** palavras com as devidas classes gramaticais, gênero e número;
- **verbal (dicverb):** verbos com seus campos (pessoa, tempo verbal, transitividade e preposições – relacionadas à transitividade) ;
- **regência nominal (regnom):** dicionário de regência nominal;
- **verbos de ligação (verblig):** contém todos os possíveis verbos de ligação.

Tanto o dicionário de verbos quanto o dicionário de verbos de ligação possui os verbos conjugados em todas as formas possíveis, incluindo infinitivo pessoal, particípio e gerúndio.

3. Estado atual do sistema Literal

A versão atual do sistema está funcional tanto para o BrOffice como OpenOffice em suas versões para Windows e Linux. O sistema está no formato UNO (Universal Network Objects) e é inserido através do *Gerenciador de pacotes*, no menu *Ferramentas* do BrOffice.

O sistema Literal para BrOffice/OpenOffice atual pode ser baixado a partir do endereço http://www.broffice.org.br/ e pode ser testado online no endereço http://literal.conexum.inf.br. Como o sistema está em constante teste e evolução, quaisquer dúvidas ou sugestões podem ser enviadas para o endereço literal@conexum.inf.br.

4. Agradecimentos

O desenvolvimento do sistema Literal não seria possível sem o apoio da Finep e do CNPq, que permitiram a criação do Literal, responsável pelo seu desenvolvimento. Graças a esses órgãos governamentais, conseguimos desenvolver tecnologia nacional, além de manter e aprimorar nossos talentosos profissionais, incentivando-os a desenvolver o software livre.

Referências

Nunes, M.G.V.; Martins,R.T.; Hasegawa, R.; Haber, R.R.; Montilha, G. (2000) "Relatório dos Testes Comparativos entre Diferentes Versões do Revisor Gramatical ReGra. (NILC-TR-00-8)", http://www.nilc.icmc.usp.br/nilc/download/NILC00-8doc.zip , junho.

Perini, Mário A. (1989) Sintaxe portuguesa - metodologia e funções. São Paulo: Ática.

Sautchuk, Inez. (2004) Prática de morfossintaxe. Barueri: Manole.

Tondo, Nádia Vellinho. (1974) Uma teoria integrada da comunicação lingüística. Porto Alegre: Sulina.

Uliano, S.C.; Menezes, C.E.D.; Gusukuma, F.W. (2006) "Uma análise do CoGrOO, um Corretor Gramatical acoplável ao OpenOffice", http://cogroo.incubadora.fapesp.br/portal/down/Doc/analise , dezembro.