Instituto de Ciências Matemáticas de São Carlos

ISSN - 0103-2569

DESENVOLVIMENTO DE UM REVISOR GRAMATICAL PARA O PORTUGUÊS CONTEMPORÂNEO [1]

MARIA DAS GRAÇAS VOLPE NUNES
RICARDO HASEGAWA
SANDRA KAWAMOTO
MARIA CRISTINA FERREIRA DE OLIVEIRA
MARCELO AUGUSTO DE SANTOS TURINE
CLAUDETE MORENO GHIRALDELO
OSVALDO NOVAIS DE OLIVEIRA JR.
CLAUDIA ROSA RIOLFI
NILMARA SOARES SIKANSI
TEREZA BERENHAUSER FERNANDES MARTINS

Nº 46

RELATÓRIOS TÉCNICOS DO ICMSC

São Carlos
Set./1996

DESENVOLVIMENTO DE UM REVISOR GRAMATICAL PARA O PORTUGUÊS CONTEMPORÂNEO[1]

Convênio ICMSC/USP - ITAUTEC/PHILCO

Maria das Graças Volpe Nunes
Ricardo Hasegawa
Sandra Kawamoto
Maria Cristina Ferreira de Oliveira
Marcelo Augusto de Santos Turine
Claudete Moreno Ghiraldelo
Osvaldo Novais de Oliveira Jr.
Claudia Rosa Riolfi
Nilmara Soares Sikansi
Tereza Berenhauser Fernandes Martins

1. Introdução

Em julho de 1993, foi firmado um convênio entre a USP e a ITAUTEC/PHILCO, objetivando iniciar um projeto para o desenvolvimento de um Revisor gramatical a ser integrado ao *Redator/Windows*, o processador de textos desenvolvido e comercializado pela ITAUTEC/PHILCO. Desde então, o convênio foi renovado duas vezes, em meados de 94 e de 95, e várias ferramentas de pós-processamento de texto, além do Revisor gramatical, foram implementadas e integradas ao *Redator*. A equipe de desenvolvimento, coordenada pela Profa. Dra. Maria das Graças Volpe Nunes, inclui docentes do Instituto de Ciências Matemáticas de São Carlos e do Instituto de Física de São Carlos, além de lingüistas e profissionais de computação, em nível de graduação e pós-graduação.

As ferramentas foram implementadas inicialmente em *Borland C++*, sendo que o sistema foi posteriormente portado para o *Microsoft Visual C++*, linguagem adotada para a implementação das extensões futuras. As ferramentas já incorporadas ao *Redator/Windows*, lançado em Julho de 1995, executam diversos tipos de verificação no texto para detecção de erros mecânicos e gramaticais, além de gerar dados estatísticos referentes ao texto. Verificadores gramaticais já estão disponíveis para a língua inglesa desde a década de 80, como o GRAMMATIK e o CORRECT GRAMMAR [Reference Software, 1992, Writing Tools, 1991]. Entretanto, a iniciativa da ITAUTEC de incluir em seu processador de textos um verificador gramatical para o Português do Brasil foi pioneira. Atualmente, além do Revisor embutido no *Redator* existem outros verificadores gramaticais comerciais para o português do Brasil, como o Carta Certa e o DTS, entre outros [Folha de São Paulo 1994a, 1994b; Machado, 1995].

---

[1] Versão Julho de 1995.

Este documento tem por objetivo fornecer uma visão detalhada das ferramentas de pós-processamento de texto desenvolvidas ao longo deste projeto e incorporadas ao *Redator* em Julho de 95. Neste contexto, descreve a filosofia de desenvolvimento, a arquitetura de software utilizada e a funcionalidade implementada. A Seção 2 apresenta a arquitetura atual dos módulos de software do sistema e descreve o ANALEX, um módulo básico acessado por todos os demais. As Seções 3 e 4 descrevem os Módulos Estatístico e Mecânico, respectivamente. A Seção 5 descreve o Módulo Lingüístico, responsável pela verificação de erros gramaticais, que é também o módulo mais complexo. A abordagem utilizada para o desenvolvimento deste módulo é descrita, bem como os tipos de erros tratados e as regras de verificação gramatical implementadas até o momento. A Seção 6 descreve a interação do Revisor com o usuário do *Redator*. Finalmente, a Seção 7 discute brevemente a situação atual deste projeto (em 1996).

2. Arquitetura

A arquitetura do sistema de verificação de textos é apresentada na Figura 2.1. As ferramentas desenvolvidas estão agrupadas em 3 módulos distintos:

- um Módulo Estatístico (XESTAT), que obtém dados estatísticos relativos ao texto e calcula um *índice de legibilidade* do mesmo.
- um Módulo Mecânico (MM), que identifica e trata alguns tipos de erros facilmente identificáveis que normalmente não são detectados por um corretor ortográfico.
- um Módulo Lingüístico (ML), que identifica a ocorrência de vários tipos de erros gramaticais.

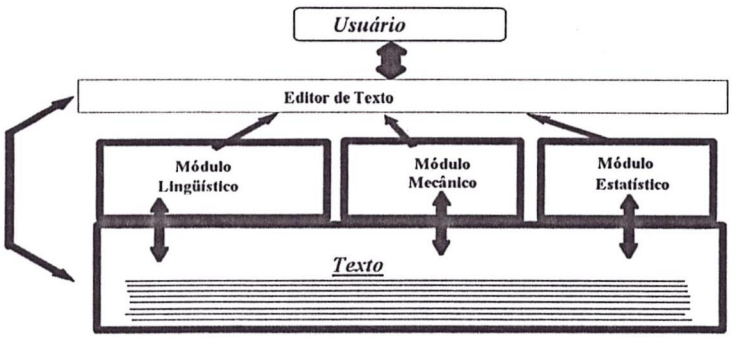

Figura 2.1: Arquitetura do Revisor Gramatical.

Os módulos Estatístico e Mecânico já estão finalizados. O módulo Lingüístico foi estendido para identificar e corrigir outras classes de erros além das que são tratadas na versão de julho de 1995. O usuário interage com os diferentes módulos do sistema através da seleção de opções apresentadas nos menus do *Redator* (ver Figura 2.2). O texto é analisado por parágrafos, sendo que cada parágrafo é passado pelo *Redator*, em uma estrutura interna, para os módulos do Revisor.

Atualmente, os novos desenvolvimentos e os testes desta versão são realizados ativando o Revisor a partir do *Redator*. A implementação ou alteração de regras requer a atualização da DLL *"rdwrev.dll"*, que deve estar disponível no diretório raiz (*rdw*) do *Redator*. Feita uma alteração no código de um dos módulos, o mesmo é compilado, *linkeditado* e a DLL é gerada.

Os três módulos interagem diretamente com um quarto, que é interno ao sistema, o ANALEX (Analisador Léxico). Este realiza a análise léxica do texto, ou seja, identifica os *tokens*, ou itens léxicos, presentes no texto, e os seus tipos (palavra, símbolo de pontuação, símbolo delimitador, abreviatura, numeral, etc.). A estrutura que contém cada parágrafo é analisada de forma seqüencial, sendo que o tipo do *token* determina as regras (mecânicas e/ou gramaticais) que devem ser aplicadas sobre ele. Também o XESTAT opera fazendo uma contagem de palavras, sentenças e parágrafos do texto, entre outros itens, com base nos *tokens* identificados pelo ANALEX. Assim, o desempenho do ANALEX é essencial para o desempenho do sistema como um todo, o que motivou um cuidado especial com a sua implementação. A versão inicial sofreu várias otimizações com vistas a aumentar a eficiência. A implementação e operação do ANALEX são descritas a seguir.

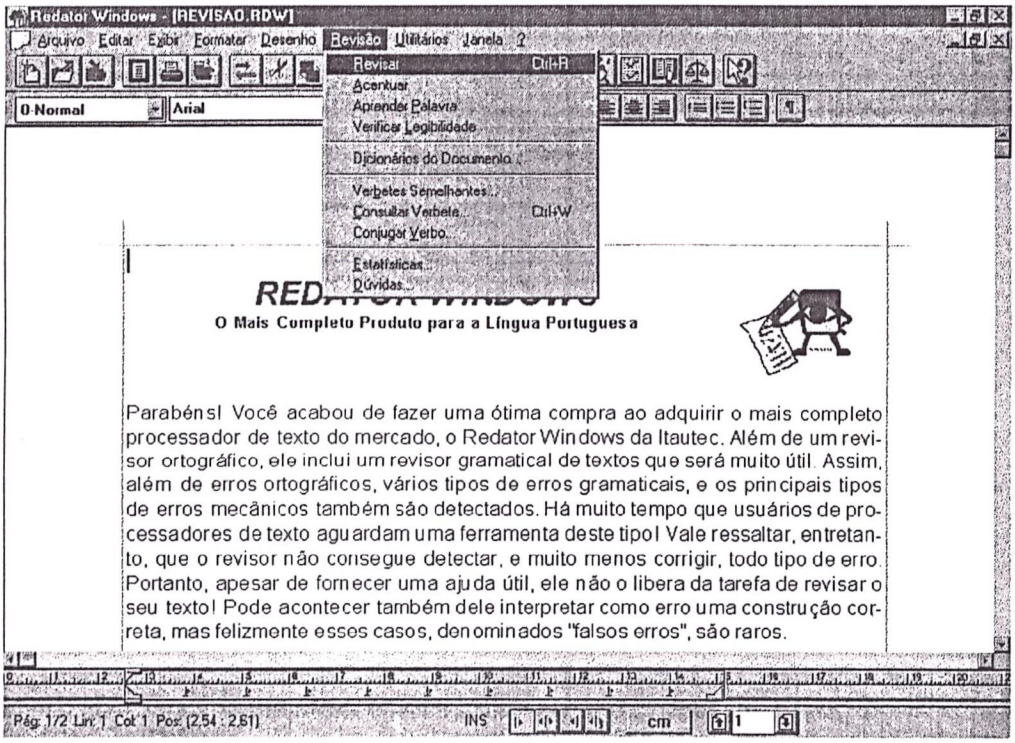

Figura 2.2: Janela do Revisor Gramatical.

2.1. O ANALEX

O código do ANALEX foi construído com base em um autômato finito que descreve a sua operação. Este módulo constrói uma estrutura descritiva associada a cada parágrafo, que contém os *tokens* do parágrafo, os seus tipos (palavra, numeral, símbolo de pontuação, símbolo delimitador, abreviação, indicador de final de parágrafo, etc.), tamanhos e a posição de cada um no parágrafo.

A identificação dos *tokens* é feita a partir da detecção de símbolos especiais no texto (símbolos de pontuação, delimitadores, espaços em branco, etc.), do acesso a tabelas específicas (o ANALEX acessa duas tabelas de abreviaturas, e também do acesso a um Léxico, ou dicionário de palavras, o *Proverb*. O *Proverb* é um sistema comercial que fornece informações sobre uma palavra - como a sua categoria lexical (se é artigo, substantivo, verbo, etc.), gênero e número - que são relevantes para a aplicação das regras gramaticais (ver Seção 5). Este sistema já era utilizado pelo *Redator* como o dicionário para verificação ortográfica do texto. Entretanto, durante o desenvolvimento do Revisor Gramatical constatou-se que o *Proverb* apresenta problemas porque muitos verbetes estão "pluricategorizados" quanto à sua categoria lexical, sendo que as categorias não estão hierarquizadas. Isso dificulta muito a realização de uma análise gramatical do texto. Esse fato motivou a equipe a iniciar o desenvolvimento de um novo léxico capaz de fornecer suporte adequado a uma revisão gramatical do texto (ver Seção 5.3).

Como a identificação correta de fim de sentença é essencial, o ANALEX trata como exceções eventuais abreviaturas presentes no texto, já que neste caso a presença do símbolo "." não indica fim de sentença. A identificação das abreviaturas é feita mantendo-se em arquivos duas listas com as abreviaturas mais comuns (uma lista com abreviaturas simples, como Sr., e outra com abreviaturas compostas, como Exmo. Sr). Sempre que é identificada uma palavra seguida de ponto final, ou duas palavras em seqüência seguidas de ponto final, estas listas são consultadas. Se a palavra não se encontra nas listas de abreviaturas, o ANALEX a classifica como palavra seguida de ponto final - o que por sua vez indica fim de sentença. No caso da palavra ser encontrada nas listas, resta ainda verificar se ela está no final da sentença.

3. O Módulo Estatístico

3.1. Estatísticas do Texto

O ME é ativado através de uma chamada ao programa denominado XESTAT. Como já foi mencionado, o XESTAT analisa o texto por parágrafos, sendo cada parágrafo passado para o Revisor pelo *Redator*. Esse módulo gera os seguintes valores numéricos sobre o texto sendo analisado: total de parágrafos, sentenças, palavras, caracteres, letras e sílabas; número médio de sentenças por parágrafo, número médio de palavras por sentença, e número médio de letras e sílabas por palavra.

O XESTAT usa o ANALEX para identificar *tokens* que delimitam sentenças e parágrafos, bem como um programa que implementa um algoritmo que faz a separação silábica de uma dada palavra, necessária para a contagem das sílabas. Utiliza também uma tabela de conversão para caracteres acentuados, uma vez que o arquivo fonte sobre o qual são aplicadas as estatísticas deve conter um texto ASCII, possivelmente com caracteres acentuados. Portanto, a ativação do módulo requer a transformação do texto escrito no *Redator* no formato do padrão ANSI *Windows*.

O algoritmo de separação silábica empregado é baseado no trabalho de Fernandes [Fernandes, 1988], que utiliza uma tabela que associa a cada possível par de letras uma ação a ser realizada durante a separação silábica. Uma ação pode ser, por exemplo, juntar as duas letras em uma única sílaba, ou separá-las em sílabas diferentes. A literatura apresenta tabelas de separação silábica para outros idiomas, como o Francês; a tabela para a Português foi construída pela equipe do projeto.

A tabela implementada inclui ainda o tratamento de todas as letras acentuadas relacionando, no total, 38 letras. O algoritmo implementado apresenta duas restrições:

- não separa corretamente prefixos. Por exemplo, a palavra "sublocar" é dividida como "su-blo-car", quando o correto seria "sub-lo-car";
- não separa corretamente hiatos não acentuados, como em "goiabada", por exemplo, que é separada como "goia-ba-da", quando o correto seria "goi-a-ba-da".

Deve-se observar que esta última restrição surge da necessidade de definir uma única ação a ser tomada para cada par de letras. No caso, encontros vocálicos podem aparecer na forma de hiatos ou ditongos. Como se acredita que na língua portuguesa a ocorrência de ditongos é maior que a ocorrência de hiatos, optou-se por considerar todo encontro vocálico como um ditongo e, conseqüentemente, está sendo ignorada a presença de hiatos. Estas restrições não são sérias no contexto do XESTAT, pois, em geral, o número de vezes que o algoritmo "erra" não é significativo para um levantamento estatístico. Além disso, não estamos interessados na divisão silábica exata da palavra, mas apenas no número de sílabas que ela contém. Assim, muitas vezes, mesmo quando o algoritmo falha na divisão, o resultado obtido é correto, como no caso de "sublocar", descrito acima.

O ME é ativado no *Redator* a partir do menu Revisão, escolhendo-se o item "Verificar Legibilidade" (ver Figura 2.2). O resultado apresenta em uma janela suspensa o índice de legibilidade calculado para o texto (ver [Nunes et al., 1995] e Seção 3.2), bem como as estatísticas referentes ao mesmo (Figura 3.1).

3.2. Índice de Legibilidade

O índice de "*legibilidade*"[2] calculado para o texto fornece uma indicação quantitativa (considerando apenas a "superfície textual") do grau de dificuldade de leitura do texto pelo público alvo. O índice pode ser calculado apenas para textos com mais de 100 palavras, e o cálculo é baseado nas estatísticas obtidas sobre o texto pelo XESTAT, descritas na Seção 3.1.

É importante mencionar que o índice obtido NÃO caracteriza de forma absoluta a dificuldade ou facilidade de compreensão do texto, pois uma análise deste tipo requer mecanismos complexos de natureza lingüística, cognitiva e pragmática que não são, de modo algum, considerados pelo Revisor. Por outro lado, o índice é útil para medir alguns quesitos relacionados à superfície textual que poderiam vir a ter influência na compreensão do texto pelo usuário. Isto é verdade, em especial, no caso de textos destinados a leitores pouco proficientes,

[2] O termo original em inglês é "*readability*". Não existe consenso na tradução para "legibilidade" (Rocco, 1981).

mais sujeitos ao efeito de "ruídos" na superfície textual. Assim, o índice pode ser usado em situações específicas para fornecer subsídios adicionais a uma análise da adequação do texto ao seu público alvo, e sua aplicação pode ajudar a identificar se o texto está escrito em um padrão acima do que seria adequado, dificultando, ou mesmo impedindo, a comunicação. É neste contexto que o índice de Flesch e outros [Flesch, 1948, Klare, 1975-1975] têm sido utilizados para a língua inglesa.

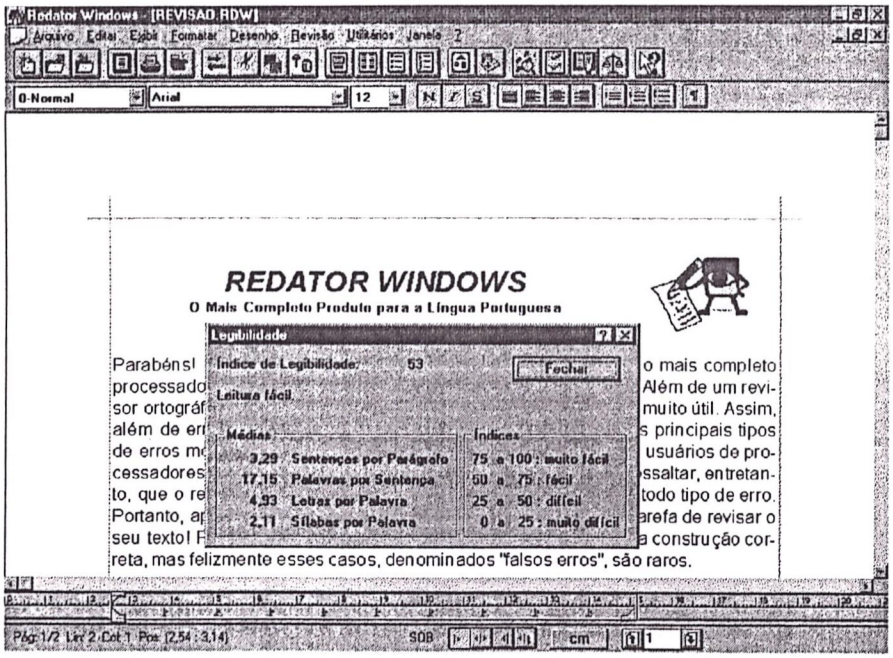

Figura 3.1: Estatísticas do texto.

Para a língua portuguesa, o trabalho realizado pela equipe foi pioneiro, e baseou-se em uma adaptação do índice proposto para o inglês por Flesch. Inicialmente foram estudados quatro índices, mas a adaptação para o português foi feita especificamente para o índice de Flesch, considerado o mais significativo. Essa adaptação resultou na identificação de quatro faixas de dificuldade de leitura, conforme indicado na Tabela I:

Pontuação do texto	**Grau de dificuldade**
75 a 100	muito fácil
50 a 75	fácil
25 a 50	difícil
0 a 25	muito difícil

Tabela I: Índice de Flesch modificado.

Textos classificados como **muito fáceis** seriam adequados para leitores com nível de escolaridade até a quarta série primária; textos **fáceis** seriam adequados para leitores com escolaridade até a oitava série, textos **difíceis** seriam adequados para leitores de nível colegial e/ou universitário, e textos **muito difíceis** seriam textos acadêmicos em áreas específicas. (ver

Figura 3.2) Testes com textos tradicionalmente dirigidos a públicos nessas quatro faixas mostraram resultados bastante satisfatórios do índice formulado (ver Relatório do projeto referente a Dezembro de 1993).

O XESTAT trabalha em três níveis: recebe a estrutura de um parágrafo, e dispara o cálculo do índice de legibilidade. Este cálculo requer a ativação dos métodos associados à sentença (para calcular as estatísticas referentes a uma sentença) e à palavra (para calcular as estatísticas referentes a uma palavra). Estas contagens são retornadas ao Revisor, que calcula o índice e exibe os resultados. Este cálculo é computacionalmente bastante eficiente. A título de ilustração, em um computador PC 486 DX com 33MHz e 8MB de memória foram necessários apenas 43 segundos para gerar as estatísticas referentes a um texto com 295 páginas e 805.983 bytes.

4. O Módulo Mecânico

O MM, módulo de tratamento de erros mecânicos, é responsável pela busca por erros que podem ser facilmente localizados, e que são independentes das estruturas sintáticas de uma frase. Geralmente, tais erros são introduzidos durante a digitação do texto, e não são detectados pelo corretor ortográfico. O conjunto de erros mecânicos tratados inclui:

i) Identificação de palavras e símbolos de pontuação repetidos: quaisquer símbolos ou palavras, encontrados em uma seqüência repetida, são localizados e indicados. É feita uma exceção para a repetição "**se se**", que é aceitável.

ii) Uso de capitalização inadequada: é verificada a utilização de palavras que abrem sentenças ou parágrafos, e cuja primeira letra seja minúscula.

iii) Ausência de delimitador de fim de sentença ("."), identificado pela ocorrência de palavra iniciada com letra maiúscula não precedida de um símbolo de fim de sentença. Nomes próprios precisam ser tratados como exceção, uma vez que sua ocorrência não caracteriza necessariamente início de sentença. Alguns nomes próprios são identificados no *Proverb*, mas a maioria é identificada exatamente pela sua ausência no dicionário do *Proverb*: se o *token* passado para o *Proverb* inicia com letra maiúscula e não está no dicionário, o mesmo é considerado nome próprio. Neste caso, a ausência de ponto final precedente não é considerada erro.

iiv) Verificação do balanceamento de símbolos delimitadores: busca de símbolos delimitadores, como parênteses, colchetes, chaves e aspas, que geralmente devem aparecer aos pares. Uma exceção é o uso de parêntesis em itemizações (como, por exemplo, em "1)", "2.2)", etc.), identificada pelo ANALEX que retorna nestes casos um *token* de enumeração, e não um numeral seguido de ")".

v) Identificação de símbolos de pontuação isolados no texto: símbolos de pontuação como vírgula, ponto, dois pontos, ponto e vírgula, exclamação e interrogação, não devem aparecer isolados da palavra anterior, ou juntos ao próximo elemento da sentença.

vi) Utilização de aspas em conjunto com outros símbolos da sentença. São considerados vários casos possíveis:

vi.1) a não utilização de um espaço em branco antes de um símbolo de abre aspas;

vi.2) vírgulas, ponto e vírgula, e ponto devem ser preferencialmente colocados após um fecha aspas;

vi.3) exclamação e interrogação devem ser preferencialmente colocados no interior de um fecha aspas.

4.1. Operação do Módulo.

O MM é ativado sobre o escopo da estrutura de um parágrafo montada internamente pelo ANALEX. Sobre esta estrutura interna que contém os *tokens*, seus tipos, e também posição e tamanho de cada um no parágrafo, é realizada a busca por erros mecânicos. A aplicação ou não de uma determinada regra de tratamento mecânico depende do tipo do item léxico analisado, ou seja, se é palavra, símbolo de pontuação, símbolo delimitador, etc. Além disso, pode-se definir, dentre as regras de tratamento mecânico, quais serão aplicadas sobre o texto em um determinado instante (Figura 4.1).

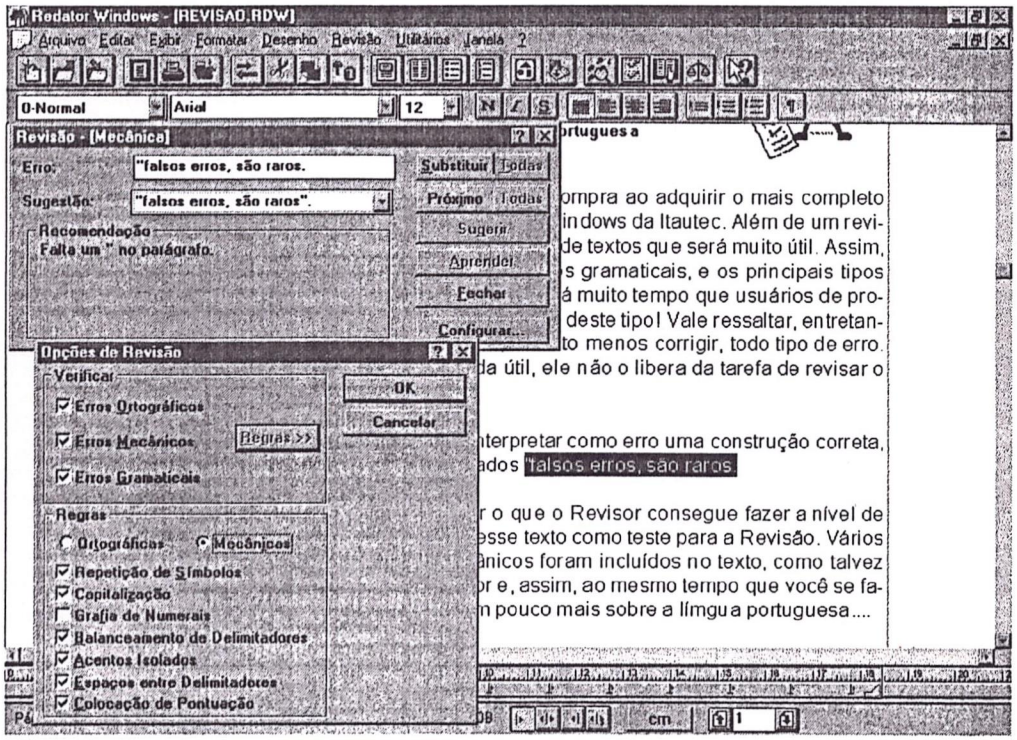

Figura 4.1: Regras de Tratamento Mecânico.

A ativação do módulo mecânico a partir do *Redator* é feita a partir do menu Revisão, opção Revisar. A revisão mecânica ocorre, a princípio, simultaneamente à revisão gramatical. No menu Arquivo, opção Preferências, o usuário pode habilitar ou não a Revisão Ortográfica, Mecânica e Gramatical. A Revisão Mecânica permite ao usuário escolher quais das regras

mecânicas serão habilitadas. Desta forma, para cada tipo de símbolo, a regra de tratamento correspondente deve estar habilitada para que seja de fato aplicada.

Assim, a estrutura contendo cada parágrafo do texto é analisada de forma seqüencial, e para cada tipo de *token* existente nesta estrutura é disparada a regra (ou regras) de tratamento mecânico associada, caso esta se encontre ativa no instante da execução. Note-se que a aplicação das regras mecânicas ocorre no âmbito de cada parágrafo do texto. Isto é importante para o caso dos símbolos delimitadores. A verificação de sua utilização balanceada será realizada dentro do parágrafo. Será considerado um erro de balanceamento a existência, por exemplo, de um símbolo de abre parênteses em um parágrafo, e o seu correspondente fecha parênteses em outro.

É importante observar que a cada marcador relevante da sentença não será necessariamente aplicada uma única regra mecânica. Sobre um mesmo *token* podem ser aplicadas várias regras. Por exemplo, a presença do símbolo de aspas requer a verificação de repetição do símbolo, de sua posição no interior da sentença com relação aos outros símbolos, e de sua utilização de forma balanceada. Os erros mecânicos identificados pelo sistema são:

- Início de sentença com letra minúscula;
- Utilização de símbolos de pontuação isolados na sentença;
- Ausência de separação (espaço em branco) entre a pontuação e o próximo elemento na sentença;
- Presença de espaço em branco após [{ (;
- Ausência de espaço em branco antes de [{ (;
- Após] }) deve ser utilizado um espaço em branco;
- Presença de espaço em branco antes de] });
- Seqüência repetida de símbolos ou palavras (com exceção da expressão "se se);
- Verificação se o fecha parênteses corresponde à presença de itens no texto (e, portanto sua ocorrência isolada não é considerada erro);
- Uso desbalanceado dos símbolos [e], { e }, (e) (ausência de [ou], { ou }, (ou));
- Uso desbalanceado dos símbolos { e } (ausência de { ou });
- Uso isolado de aspas no texto (deve ser colocada próxima a um símbolo);
- Uso de símbolos de pontuação de final de sentença (. ? !) ou de apóstrofos simples (') no interior de um par de aspas (devem ser preferencialmente colocados fora);
- Uso inválido de número no texto (deve-se preferencialmente escrever por extenso);
- Excesso de espaço em branco entre palavras, ou no início ou no final do parágrafo;
- Uso desbalanceado do símbolo de abre aspas no parágrafo;
- Uso do símbolo de abre ou fecha aspas não precedido por um espaço em branco;
- Ausência de pontuação de final de sentença, e

- Ocorrência de sinal de acentuação (~ ^ ´ ` ¨) isolado.

Na revisão ativada a partir do *Redator*, é mostrada na tela uma janela que indica o erro e sua localização na frase. Sempre que possível, a janela fornece uma recomendação e sugestão de como corrigir o problema (ver Figura 4.1). As mensagens de "Recomendação" e "Mais Informações" fornecidas pelo Revisor para o usuário durante a detecção de um erro mecânico e gramatical estão armazenadas no arquivo "erros.msg", no diretório "\rdw\dic".

O módulo de tratamento mecânico é composto por vários métodos responsáveis pela aplicação de regras para a detecção de erros classificados como mecânicos. A busca pelos erros mecânicos é feita através do método *BEMecanico()*, aplicado a cada parágrafo. Este método é o responsável pela aplicação ou não de uma determinada regra de tratamento mecânico, que depende do tipo do item léxico analisado da sentença e do conjunto de regras atualmente habilitadas.

5. O Módulo Lingüístico

5.1. Abordagem de Desenvolvimento

Como se partiu do princípio que o público alvo do *Redator* consiste, preferencialmente, de secretárias e trabalhadores de escritório, o primeiro passo no desenvolvimento do Revisor gramatical foi realizar um levantamento dos erros mais freqüentes cometidos por indivíduos de nível médio (isto é, com escolaridade em nível de segundo grau). A seguir, foi feito um estudo lingüístico para identificar possíveis formas de tratamento destes erros e também a viabilidade de implementação de heurísticas que possibilitassem a sua detecção automática. O suporte teórico para as regras foi baseado no estudo das gramáticas normativas tradicionais, sendo a adequação das mesmas avaliada através de testes em textos de diversas fontes (jornais, revistas, livros, etc.). Para tanto, foi construído um banco de textos, isto é, uma base de dados composta por um amplo conjunto de textos (ver Seção 5.2).

É importante mencionar que um fator limitante para a seleção dos problemas a serem tratados é a viabilidade de implementação computacional das regras heurísticas a serem utilizadas para a detecção e correção dos erros, uma vez que nem sempre as heurísticas propostas, resultado de um estudo lingüístico, são computacionalmente viáveis.

A implementação das regras foi acompanhada de um estudo cuidadoso sobre a detecção de **falsos erros** pelo sistema. Por **falso erro** entendemos (1) uma intervenção do sistema que induziria o usuário a um erro gramatical, através da modificação de uma estrutura lingüística originalmente correta; ou (2) uma intervenção desnecessária do sistema que induziria o usuário a alterar uma estrutura originalmente correta por uma forma (correta) alternativa. Esta última situação reflete uma preocupação em não classificar como erro as formas opcionais correntemente aceitas no uso da língua. Por exemplo, o uso da crase antes de pronomes possessivos é facultativa, de forma que as estruturas *"ele deu o presente a sua filha"* e *"ele deu o presente à sua filha"* são ambas corretas, sendo que a ausência da crase não deve ser detectada como erro.

Devido à complexidade associada ao tratamento automático de textos em ambientes "abertos" (isto é, onde todo tipo de texto, sem qualquer restrição imposta à estrutura dos parágrafos e sentenças, pode ocorrer), ainda não está sendo realizada a análise sintática dos textos, que é um objetivo do projeto em médio prazo. No sistema atual, a abordagem consistiu em implementar um conjunto relativamente amplo de regras heurísticas de correção. Estas regras tentam identificar erros cuja possível ocorrência nos textos é antecipada a partir de estudos lingüísticos empíricos. A metodologia adotada na implementação das regras de correção pode ser dividida em três etapas principais:

1) identificado um tipo de erro a ser corrigido, seguia-se um estudo detalhado em gramáticas e outras fontes da literatura que discorram sobre o uso da língua portuguesa. Mecanismos de correção eram então propostos na forma de regras heurísticas para a detecção e correção desses erros.

2) as regras propostas eram implementadas computacionalmente, e, através de testes exaustivos com o material que compõe o banco de textos, eram verificados falsos erros e erros não detectados. Quando necessário, exceções nas regras eram inseridas a fim de minimizar a ocorrência de falsos erros. Isso porque um requisito estabelecido previamente no desenvolvimento foi a minimização da ocorrência de falsos erros, pois estes diminuem a credibilidade da ferramenta junto ao usuário. Melhorias na regra eram, então, implementadas. Este processo é altamente iterativo.

3) finalmente, dava-se o acabamento final à regra, tanto do ponto de vista da otimização da implementação computacional, como da tomada de decisão com relação às mensagens a serem fornecidas aos usuários. As mensagens apresentam uma certa variedade, pois o sistema pode sugerir correções quando tiver certeza do erro, ou apenas alertar o usuário quanto ao uso de uma estrutura lingüística que pode ou não estar correta, dependendo do contexto.

5.2. O Banco de Textos

Fica implícita na importância atribuída aos testes com textos reais, a necessidade de se contar com um banco de textos (*corpus*) significativo. A coleta de textos para este banco tem sido uma preocupação constante durante o desenvolvimento. Foram coletados diversos tipos de textos, tais como científicos, de jornais, de livros, redações de universitários e vestibulandos, entre outros. Pode-se identificar 2 grupos de textos que são relevantes para o desenvolvimento do sistema: (1) textos escritos por autores experientes, revisados e corrigidos, que são utilizados na busca de padrões lingüísticos para estudos gramaticais e da língua em uso, bem como na detecção de falsos erros associados às regras; (2) textos escritos por diferentes classes de indivíduos, sem revisão ou correção, que são utilizados para detectar erros a serem corrigidos. Em outras palavras, os textos do grupo (1) servem como referência do uso correto da língua escrita, enquanto que os textos do grupo (2) fornecem uma amostra dos erros mais comuns cometidos pelos usuários da língua e de como estes erros são tratados através das regras implementadas.

O banco de textos também é útil para verificar o emprego corrente de algumas formas gramaticais. As gramáticas normativas, em geral, utilizam como exemplos textos extraídos de obras clássicas, e muitas vezes as descrições nelas apresentadas não mais refletem a realidade lingüística atual. Há, portanto, uma enorme dificuldade em se estabelecer de forma definitiva a norma a ser seguida na formulação das regras de correção, e nesse contexto o banco de textos mostra uma visão da língua que não está disponível nas gramáticas. Atualmente, o banco de textos, conta com, aproximadamente, 27 milhões de palavras (total estimado).

Os textos estão agrupados em três classes:

a) textos publicados para grande número de leitores, como jornais, revistas e livros, que são, portanto, supostamente corrigidos;

b) textos publicados para pequeno número de leitores ou não publicados, como dissertações e teses acadêmicas, relatórios e projetos científicos, documentos empresariais (cartas, relatórios, atas de empresas privadas e públicas), que são também corrigidos, porém, não passaram por uma equipe de correção, como os textos editados em livros e jornais;

c) textos não corrigidos, escritos por pessoas de nível médio de escolaridade (2º grau completo) e universitários.

Os textos para composição das partes (a) e (b) do *corpus* foram extraídos das literaturas *jornalística*, revistas e jornais de grande circulação; *técnico-científica*, das principais áreas do conhecimento; e *jurídica*, textos legais. Os textos que compõem a parte (c) são de estudantes e de candidatos ao vestibular de duas grandes universidades (USP, UFSCar). Dessa maneira, a composição do *corpus* é *heterogênea* e sob um ponto de vista *sincrônico*, já que se procura, com a diversidade de textos, delinear o português contemporâneo escrito.

Cada uma das três classes é usada diferentemente no desenvolvimento da ferramenta. Os textos das classes (a) e (b) são utilizados para indicar a freqüência de determinadas palavras e/ou construções lingüísticas, pois não há, para o português, dicionários de freqüência disponíveis. Os textos da classe (c) são utilizados para os testes das regras computacionais implementadas. São utilizados também para levantamento de erros possíveis de ocorrerem por redatores com nível médio de escolaridade.

A organização atual do *corpus* ainda não é a definitiva, pois na medida em que o tratamento computacional da linguagem se sofistica e se especializa, será necessário também modificar o *corpus* a fim de adequá-lo às exigências dos testes das ferramentas.

Para a montagem do *corpus* de textos, utilizou-se diferentes recursos, como exportação de textos de Compact Disc (CDs) e da rede *Internet*; reprodução por meio de *scanner* com os acertos necessários; digitação de textos; cópia e conversão de arquivos armazenados em disquetes. Os textos armazenados foram convertidos para o programa WORD/WINDOWS-TXT (somente texto) não contendo, portanto, formatação alguma, nem ilustrações (gráficos, tabelas, fotos) e fórmulas. A correção dos textos reproduzidos por *scanner* e dos textos digitados foi feita utilizando-se o utilitário "verificar ortografia",

do WORD, e manualmente, já que tal utilitário não é capaz de reconhecer boa parte do léxico do português, e sempre procurando ser fiel aos textos originais. Portanto, se o texto guardava algum erro gramatical, este foi reproduzido na cópia para o *corpus*. É importante lembrar que, justamente pelo fato de a correção ter sido feita manualmente, muitas palavras dos textos do *corpus* trazem erros de digitação.

Para a incorporação, no *corpus*, dos textos científicos universitários (monografia; dissertação; tese; projeto; relatório; artigo; aula), foi solicitada a autorização do(s) autor(es). Antes, porém, de eles serem anexados ao *corpus*, foram excluídos desses textos os nomes do(s) autor(es), orientador(es) ou qualquer outra pessoa que permitisse a identificação do autor(es) do documento. Foram também excluídas as partes: agradecimento, dedicatória, índice, *abstract*, ilustrações (figura, gráfico, tabela, fotografia), lista de ilustrações, bibliografia, apêndices, anexos.

O *corpus* ainda não conta com textos literários (romances, contos, crônicas). É de grande importância a incorporação, num futuro próximo, de tais textos ao *corpus*, pois neles podem ocorrer construções lingüísticas singulares. Também num futuro próximo, pretende-se incorporar transcrições de textos orais (Projeto Norma Urbana Culta do Português do Brasil - NURC, em desenvolvimento em diversas universidades brasileiras, entre elas USP, UNESP, UNICAMP), porque é comum nos textos escritos por pré-universitários e universitários aparecem, muitas vezes, formas lingüísticas próprias da oralidade. Dessa maneira, essas duas classes de textos servirão como suporte aos estudos do português escrito. Além de textos dessas duas grandes classes, pretende-se também incorporar textos de outras áreas do conhecimento, como por exemplo, culinária, humor, ficção científica, entre outras.

Dada a evolução da pesquisa, será também necessária a elaboração de um balanceamento entre os tipos de textos para levantamentos mais confiáveis de freqüência de palavras – material de auxílio ao léxico; e de expressões e/ou estruturas sintáticas – auxílio às regras gramaticais. Pretende-se também reorganizar o *corpus*, baseando-se em estudos sobre tipologia de textos que propõem classificações a partir de características intrínsecas a eles – ficção; não ficção; narrativa; textos de instruções; biografia; etc. Atualmente os textos do *corpus* estão organizados de acordo com as diversas fontes.

5.3. Implementação do Módulo Lingüístico

A implementação do Revisor gramatical seguiu três princípios básicos: i) o sistema é essencialmente *error-driven*, ou seja, tenta detectar a ocorrência de certos tipos de erros que podem ser previstos; ii) os tipos de erros "procurados" e a operação do sistema foram baseados em testes extensivos feitos sobre o *corpus* de textos reais; iii) o desenvolvimento seguiu um processo iterativo, ao longo do qual a operação das regras era verificada e melhorada com base na sua aplicação a textos do *corpus*. A implementação das regras é baseada na filosofia das ATNs (*Augmented Transition Networks*) [Woods-70]. ATN's

oferecem um modelo adequado para a análise de linguagem natural, pois podem ser adaptadas de acordo com as nuances do texto [Miller87].

Na versão corrente do Revisor, existem ATNs associadas a construções não gramaticais (a maioria), que correspondem aos tipos de erros que são previstos pelo sistema; e ATNs que correspondem a algumas construções gramaticais, associadas às regras de concordância verbal e nominal. O conjunto de regras heurísticas implementadas no ML são independentes entre si.

As regras, implementadas na forma de ATNs, são responsáveis pela detecção e correção de erros relativos ao uso impróprio da norma culta da língua portuguesa que independem de qualquer estilo[3] de escrita. A implementação das regras foi acompanhada de um estudo sistemático de detecção dos erros mais comuns e também da ocorrência de possíveis **falsos erros** do sistema. Os erros detectados e as regras implementadas para a sua detecção são descritos com mais detalhes na Seção 5.4. Os critérios utilizados para selecionar os erros a serem tratados são discutidos nos Relatórios do projeto de Fevereiro e Junho de 1994.

A arquitetura do ML é apresentada na Figura 5.1. O Gerenciador de Regras Gramaticais ativa o ANALEX, que é responsável pela identificação dos *tokens* relevantes, denominados **marcadores**. A ativação de várias regras é disparada pela presença de um certo marcador e, eventualmente, será necessário identificar se uma palavra na frase (o marcador, ou outra na sua vizinhança) é substantivo, adjetivo, conjunção, artigo, etc., e eventualmente o seu gênero (masculino, feminino) e número (singular, plural) - daí a necessidade de acesso ao léxico, no caso o *Proverb*.

Como já mencionado, o *Proverb* apresenta várias limitações em termos de conteúdo, o que motivou o desenvolvimento de um novo léxico pela equipe. Um dos problemas é que é muito comum, no português, uma mesma palavra apresentar mais de uma categoria lexical. Por exemplo, "ser" pode ser classificado como substantivo ou verbo, "a" pode ser artigo ou preposição. Para decidir qual das formas está sendo usada, é necessário verificar a vizinhança (no caso de "ser" estar precedido por um artigo, pode-se afirmar que a forma usada é substantivo). Entretanto, nem sempre é possível resolver a ambigüidade lexical desta maneira. Assim, seria interessante que o léxico contivesse informações sobre a freqüência de ocorrência das categorias lexicais de uma palavra. Numa situação ambígua, a forma mais freqüente seria a primeira a ser considerada.

Além do dicionário do *Proverb*, informações gramaticais adicionais necessárias ao funcionamento das regras são encontradas em algumas listas específicas de palavras. Estas são fornecidas em arquivos separados (extensão "LIS"), acessados diretamente pelo programa que implementa as regras. Estes arquivos estão disponíveis no diretório "\rdw\dic", localizado no diretório do *Redator*. Como o desempenho do ANALEX é essencial para o desempenho da ferramenta, no caso de palavras com alta freqüência de ocorrência no texto o ANALEX acessa

[3] Por *estilo de texto* estamos nos referindo a tipologia textual, ou seja, textos científicos, empresariais, jornalísticos, literários, etc.

estas listas, para evitar o acesso ao *Proverb*, que é lento. As listas são, portanto, usadas para acelerar o processo de classificação.

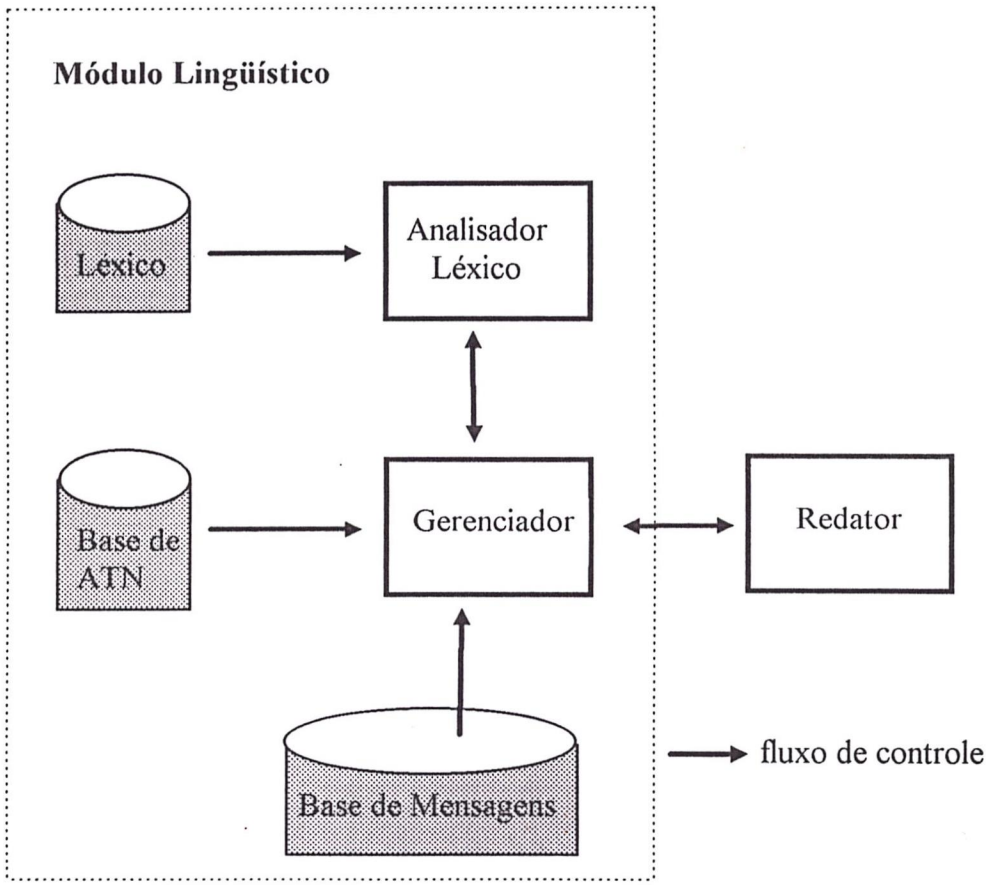

Figura 5.1: Arquitetura do Módulo Lingüístico.

Assim que as frases do parágrafo estejam classificadas, as regras para detecção de erros gramaticais disponíveis na base de ATNs são ativadas. Existem basicamente dois tipos distintos de ATNs na base. Como a implementação das regras é voltada principalmente para a identificação dos erros considerados mais freqüentes, as ATNs do primeiro tipo são associadas a construções não gramaticais. Para o núcleo de regras de concordância verbal e nominal, entretanto, foram criadas ATNs representando construções sintáticas corretas. Um exemplo de ATN é apresentado na Seção 5.4, na descrição das regras implementadas.

5.4. Conjunto de Regras

A seguir são apresentadas as classes de erros tratados pela versão do Revisor lançada em Julho/95. Alguns exemplos ilustrativos do uso das regras também são apresentados. As mensagens emitidas ao usuário são descritas no Relatório de Projeto de Julho de 1995.

5.4.1. Uso da Crase

A crase é o fenômeno de fusão de duas vogais idênticas. O caso mais conhecido de crase refere-se à fusão do *a* (preposição) e o *a* (artigo), que resulta na grafia de um só *a* marcado por um acento grave: *à*. O Revisor trata de modo bastante eficaz os erros que ocorrem com maior freqüência quando do uso incorreto da crase, a saber: antes de palavras masculinas, palavras no plural e verbos no infinitivo. Já a ausência de crase é detectada apenas em situações limitadas.

Vale observar que está em andamento um trabalho de mestrado, orientado pela Profa. Maria das Graças Volpe Nunes, que tem como objetivo atacar o problema da detecção do uso incorreto da crase usando uma abordagem conexionista, através de Redes Neurais.

Uso incorreto da crase

As regras gramaticais referentes à crase tratam separadamente dois casos: no **Caso 1**), o usuário utilizou a crase no singular e, no **Caso 2**), utilizou a crase no plural. Portanto, as regras do **Caso 1** são ativadas quando forem encontrados na frase sendo analisada os marcadores *à* ou *À*. Se estes marcadores estiverem no plural (*às* ou *Às*), é ativado o **Caso 2**.

Foi implementado um conjunto de 15 regras, enquadradas nos **Casos 1** ou **2,** para detectar o uso incorreto da crase. Devido às limitações do *Proverb*, as oito primeiras regras funcionam acessando listas de palavras, sem consulta ao dicionário do *Proverb*. Algumas dessas listas estão em arquivos armazenados do diretório "\rdw\dic" (tratamen.lis, paises.lis, paises1.lis, expres1.lis), e outras estão embutidas dentro do sistema. As sete últimas funcionam através de consulta ao *Proverb*. Deve-se mencionar que estas últimas são ativadas com maior freqüência. A seguir são apresentados alguns exemplos típicos de uso incorreto da crase detectados pelo Revisor:

- "Ela gosta muito de andar à cavalo." (Incorreto, pois "cavalo" é masculino)
- "Ele começou à andar muito cedo." (Incorreto, pois "andar" é verbo no infinitivo)
- "Chegaram a uma árvore frondosa, à cuja sombra descansaram." (Incorreto, pois "cuja" é pronome relativo que não é precedido de artigo ou preposição)
- "Esta carta diz respeito à pessoas ilustres." (Incorreto, pois "pessoas" está no plural e a crase no singular)

Ausência da Crase

A ausência de crase só é detectada em situações específicas, identificadas pela ocorrência de certos marcadores cujo aparecimento na frase, em geral, requer o uso da crase. Assim como no caso do uso indevido, as informações gramaticais necessárias ao funcionamento da regra são encontradas no dicionário de dados do *Proverb* e em algumas listas específicas de palavras. A seguir, são listados exemplos de frases nas quais a ausência da crase é detectada pelo Revisor:

- "O problema é devido a falta do material." (Incorreto)
- "Este livro pertence a professora do curso." (Incorreto)

5.4.2 Colocação Pronominal

A colocação dos pronomes oblíquos átonos está intimamente ligada à harmonia (sonora) da frase. Segundo o padrão lusitano, os pronomes oblíquos átonos devem vir depois dos verbos (na forma denominada ênclise). Essa *regra geral* é decorrente de dois fatos: (1) Os pronomes oblíquos, por serem átonos, apóiam-se no verbo para efeito de acentuação; e (2) Os pronomes oblíquos átonos constituem os complementos (objeto direto/indireto) dos verbos. Porém, em alguns casos os pronomes oblíquos átonos podem aparecer "antes" (próclise) ou no "meio" do verbo (mesóclise). Alguns exemplos de erros detectados pelo Revisor são apresentados abaixo:

- "Não se preocupe, amanhã entrego-lhe o dinheiro." (Incorreto, pois a presença do advérbio "amanhã" requer o pronome antes do verbo)
- "As respostas não pareceram-me corretas." (Incorreto, pois a presença da palavra "não" requer o pronome antes do verbo)
- "Me lembre de trazer o artigo." (Incorreto, pois o pronome não deve iniciar frases)

5.4.3 Verbos *Fazer* e *Haver*

Este conjunto de regras corrige os casos de erros nos quais os verbos *fazer* e *haver* aparecem relacionados a marcadores de tempo e, portanto, são impessoais, não devendo ser conjugados. O Revisor, ao identificar neste contexto a ocorrência de um destes verbos no plural, emite mensagem ao usuário avisando que a forma correta é no singular.

Verbo *Fazer*

As regras corrigem casos nos quais o usuário não se deu conta de que *fazer* foi usado como um verbo impessoal (sem flexão) e flexionou-o para o plural, resultando em um erro gramatical. Observe-se a diferença no uso do verbo nos dois casos:

- "Eles *fizeram duas horas* de ginástica." (Correto, pois o sujeito da oração é "eles", e *fazer* deve ser conjugado)
- "*Faz dois meses* que eu não tomo cerveja." (Correto, pois neste caso, fazer é impessoal, ou seja, não concorda com nenhum sujeito e, portanto, deve aparecer no singular)

O marcador que ativa a regra é uma ocorrência do verbo *fazer* conjugado no plural. Várias situações específicas devem ser analisadas, pois nem sempre o uso do verbo no plural, associado a tempo, ocorre no contexto impessoal. Por exemplo, está correta a frase "Eles *fizeram três dias* de caminhada" (o sujeito, determinado, é "eles"). Assim, trata-se de erro gramatical, detectado pelo Revisor, ocorrências do tipo:

- "*Fazem três dias* que não a vejo." (Incorreto)
- "*Fazem muitos anos* que eu não jogo futebol." (Incorreto)

Verbo *Haver*

A exemplo do que ocorre com o verbo *fazer*, este conjunto de regras corrige os casos em que *haver* deveria ter sido usado em sentido impessoal (como sinônimo de *existir*), mas foi conjugado pelo usuário. A seguir são apresentados alguns exemplos de frases erradas, apontadas pelo *Revisor*.

- "*Houveram* aulas ontem." (Incorreto)
- "*Houveram poucas* semanas em que eu não passei fome." (Incorreto)

5.4.4. Uso incorreto de *Há* em oposição a *a*

Esta regra detecta o uso incorreto de *há* (verbo *haver*), indicando noção de tempo, em oposição a *a* (preposição), como por exemplo, em "*A muito tempo* moro nesta casa." (Incorreto), ou em "Vou visitá-la daqui *há dois dias*." (Incorreto). Ambas as frases acima resultariam em mensagens de erro se verificadas com o Revisor.

5.4.5. Uso de *Mau/Mal*

É comum o usuário confundir *mau* com *mal*, e vice-versa, porque ambos têm classificações gramaticais muito semelhantes. Por exemplo, ambos podem ser substantivos masculinos, como em "Não deseje o *mal* ao próximo.", ou em "Todos preferem o bom ao *mau*."(ambas corretas). Além disso, *mal* é também advérbio e *mau* adjetivo, como nos exemplos: "Os negócios vão *mal*." e "Ele é um *mau* pai." (ambos corretos).

As regras implementadas no Revisor detectam os usos incorretos de *mau* e *mal*, bem como orienta o usuário para realizar a correção baseada em uma antiga noção semântica: *mal* é antônimo de *bem* e *mau* é antônimo de *bom*. Foram implementadas duas regras. A primeira é ativada quando é encontrado o marcador *mal* e a outra quando é encontrado *mau*. Se *mal* ou *mau* estiverem entre palavras repetidas a regra não é ativada (por exemplo: "Ora mal ora bem."). Exemplos de erros detectados são listados abaixo:

- "O *mal* filho não obedeceu ao pai." (Incorreto)
- "*Mau* cheguei e já tenho que sair." (Incorreto)

Observe que, dadas as limitações do estágio atual do processamento de linguagem natural, podem ocorrer problemas de ambi güidade e duplo sentido, como em "O *homem mau* podia andar depois do acidente." (correto); e "O homem *mal podia andar* depois do acidente." (também correto). Na primeira frase *mau* é uma qualidade de homem (adjetivo), enquanto que na segunda *mal* refere-se a "poder andar", ou seja, está fazendo papel de advérbio. O Revisor não tem como tratar tais situações, e vai aplicar as regras e eventualmente gerar um alerta desnecessário.

5.4.6. A Partícula *Se*

A partícula *se* pode assumir várias funções no português: conjunção ("O sucesso ocorrerá se o público quiser."), pronome ("Cobravam-*se* juros extorsivos." e "Fala-*se* de coisas novas."), pronome reflexivo ("Ele *se* atribui muito valor."), partícula integrante dos verbos

pronominais ("Todos *se* queixaram das instalações.") ou partícula expletiva ("Todos *se* foram."). Entretanto o Revisor trata, basicamente, do caso em que o *se* tem a função de partícula apassivadora, ou seja, se liga a verbos transitivos diretos para apassivá-los (por exemplo, "Cobravam-*se* juros extorsivos", equivalente a "Juros extorsivos eram cobrados"). Nesse caso, o verbo deve concordar em número com o sujeito (da passiva) que ocorre depois dele. Por exemplo, "Conserta-se sapato.", ou "Consertam-se sapatos."

Além desse caso, o corretor verifica também o uso do *se* quando sujeito de verbos infinitivos ("Pode-*se* dizer que os alunos aproveitaram bem o curso."). Nesse contexto, o verbo ao qual o *se* se liga deve sempre ficar na terceira pessoa do singular. O mesmo ocorre quando depois do *se* aparece uma das conjunções *que* ou *se* ("Considerou-*se* que o problema estava resolvido." e "Cogitou-*se* se seria possível localizar o presidente."). As regras implementadas procuram detectar o uso incorreto da forma verbal (singular ou plural) do verbo ao qual o *se* está ligado. Alguns exemplos típicos de erros detectados:

- "Tratam-se dos problemas não resolvidos." (Incorreto, pois o sujeito é indeterminado, e o verbo deveria estar no singular)
- "Conserta-se sapatos." (Incorreto, pois é uma frase na voz passiva, na qual o verbo deve concordar com o sujeito, que está no plural)

5.4.7. Concordância Nominal e Verbal

A concordância nominal é uma questão gramatical bastante relevante no Português. Em geral, um usuário de nível médio da língua conhece bem a regra de que substantivo e adjetivo devem concordar em gênero e número. Ainda assim, erros de concordância são comuns devido a duas razões principais: (i) as palavras que devem concordar entre si estão distantes na frase, o que eventualmente induz o escritor a um erro; ou (ii) o usuário não revisa o seu texto, e erros que seriam eliminados em uma revisão permanecem despercebidos. A mesma argumentação se aplica às regras de concordância verbal, sendo que sujeito e verbo devem concordar não apenas em número, mas também quanto à pessoa (primeira, segunda, terceira).

Ao contrário dos erros com alta freqüência de ocorrência discutidos até o momento, que podem ser antecipados, erros de concordância verbal e nominal podem ocorrer em contextos bastante abertos, de forma que foram geradas ATNs para representar construções sintaticamente corretas. A título de ilustração, a Figura 5.2 apresenta a ATN para uma sentença simples que segue um padrão simples formado por um sujeito composto do tipo **<Oração Substantiva 1** + *e* + **Oração Substantiva 2** + **Verbo>**. Este é um exemplo de ATN gramatical.

Orações substantivas (NP) podem ser formadas por um substantivo, um adjetivo, por bigramas do tipo **<substantivo** + **adjetivo>** ou **<adjetivo** + **substantivo>** e também incluir artigos, pronomes, etc. Portanto, NP1 e NP2 podem eventualmente incluir palavras de *List1*, *List2* e bigramas, sendo:

List1 = (*o, os,* a, *as, esse(s), essa(s), aquele(s), aquela(s), um(a), algum(a), alguns,* etc.);

List2 = (*vários, muitos, diversos, poucos,* (s), *toda*(s), *cada, meu(s), minha(s), nosso(s)*, etc.).

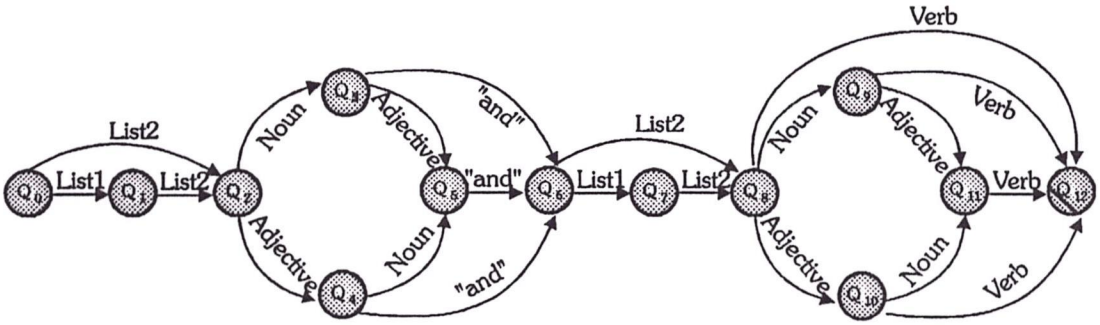

Figura 5.2: ATN que aplica uma das regras de concordância verbal. Neste caso, o verbo deve estar no plural.

Orações mais longas, subordinadas ou coordenadas, são analisadas a partir da identificação de alguns marcadores retóricos específicos (*mas*, *para*, etc.), e do estabelecimento das características sintáticas que tais sentenças possam possuir. Alguns exemplos de erros de concordância verbal detectados são apresentados a seguir:

- "Os estudantes chegou cedo mas o museu estavam fechado." (Incorreto)
- "Foi feito vários ajustes na política fiscal." (Incorreto)
- "Nós e eles discutiram muito." (Incorreto)

O penúltimo exemplo ilustra um erro de concordância verbal em uma sentença na voz passiva. Tais erros são comuns porque os escritores tendem a considerar a forma passiva singular como invariável. O primeiro exemplo ilustra uma sentença que ativaria as regras baseadas em marcadores retóricos. O aparecimento de *mas* indica que apenas certos tipos de sentenças subordinadas podem ser formadas. Através de análise de *corpus* foram criadas ATNs que permitem a identificação automática do tipo de sentença, bem como a determinação de como aplicar regras de concordância verbal.

Um estudo mais abrangente está sendo realizado para a nova versão do Revisor, em desenvolvimento, com o objetivo de criar ATNs que possam lidar com a maior parte dos tipos de sentença. Na versão atual do Revisor, existem várias limitações, associadas principalmente à presença no texto de sentenças muito longas, cuja análise é difícil, e à existência de palavras homógrafas que possuem múltiplas funções sintáticas. Problemas deste tipo podem ser minimizados investigando-se as diferenças de contexto em que tais palavras aparecem. Com relação às sentenças muito longas, a maior dificuldade está talvez na utilização de pontuação incorreta, e ao fato de que a ordem das sentenças pode variar

consideravelmente. No caso de pontuação incorreta, a análise certamente será falha, e apenas casos particulares em contextos específicos podem ser tratados.

Com relação à concordância nominal, o Revisor consegue detectar a maioria dos erros, mas alguns problemas permanecem. A maior parte destes está relacionada à concordância do artigo *a*, que também pode ser preposição, e à concordância envolvendo palavras homógrafas que podem ser verbos ou substantivos, dependendo do contexto, como *visto* (substantivo) e *visto* (particípio do verbo *ver*). Este último problema pode provocar dificuldades no processo de classificação, afetando a atuação das regras de concordância. Alguns exemplos de erros detectados de concordância nominal:

- "Eu estudei o primeiro e segundo livro". (Incorreto)
- "O livros estão na prateleira." (Incorreto).

5.4.8. Concordância Verbal de Particípio

Apesar de ser um caso de concordância verbal, este tipo de erro é tratado em separado através de um conjunto específico de regras, cujo objetivo é verificar a concordância de gênero e número que aparece nas estruturas de tempo composto com particípio e nas estruturas contendo a forma *necessário*. O tratamento em separado se justifica porque este erro é bastante comum, e, ao contrário de um erro genérico de concordância verbal, pode ser identificado a partir da ocorrência de certos marcadores. No tempo composto, o verbo no particípio ocorre junto com os seguintes verbos auxiliares: *ser*, *estar* ou *ficar* e estes apenas variam em número (singular ou plural), enquanto o particípio deve concordar também em gênero (feminino ou masculino). Situação semelhante é a da forma *necessário* que, na maioria dos casos, deve concordar em gênero e número. É importante ressaltar que essa regra apenas trata dos contextos em que o verbo em questão inicia a sentença. Essa restrição ainda é necessária, pois, no atual estágio de desenvolvimento da análise sintática automática, ainda não é possível identificar seguramente qual é o sujeito da sentença. Típicos erros detectados são ilustrados abaixo:

- "É proibido a entrada." (Incorreto)
- "É necessário a abertura de inscrições." (Incorreto)

5.4.9. Expressões Fixas

Expressões fixas são termos escritos sempre da mesma forma, independentemente do estilo utilizado, ou seja, não existe uma variação gramatical no emprego das expressões. Às vezes, entretanto, ao utilizar algumas dessas expressões os escritores cometem erros devido à influência da linguagem oral. Na versão atual, o Revisor considera três grandes grupos de regras fixas:

a) Subclasse preposição.

b) Subclasse gênero.

c) Subclasse pontuação.

Observa-se que estes conjuntos são ilustrativos dos tipos de erros relacionados a expressões fixas que podem ocorrer, e não uma lista extensiva. Estas classes devem ser

completadas através de estudos mais detalhados utilizando o *corpus*. Os erros nessas 3 classes são discutidos a seguir.

a) Subclasse Preposição

Foram separadas em duas classes de correções: as que envolvem erros gramaticais e as que envolvem um simples aconselhamento de mudança de nível de linguagem. Além disto, algumas preposições mereceram tratamento especial.

Expressões consideradas errôneas

Toda vez que a ferramenta detecta, no texto que está sendo analisado, uma das expressões erradas, é sugerido ao usuário que ele a substitua pela expressão correta correspondente. Alguns exemplos aparecem abaixo:

- "O trabalho foi feito *a nível de* iniciante. "

  (Incorreto, o correto seria "em nível de")

- "As coisas ficarão mais dificeis à *medida em que* ele cresce".

  (Incorreto, o correto seria "à medida que", ou "na medida em que")

Expressões consideradas variantes

Foram consideradas como expressões variantes aquelas que, embora não podendo ser caracterizadas como erro, afastam-se sobremaneira da norma-padrão gramatical. Desta forma, cada vez que o programa detecta uma das expressões-alvo é sugerido ao usuário que utilize a expressão substituta que mais se aproxima da norma-padrão. Exemplos:

- "Eu tenho *muito o que* fazer." (é sugerida a utilização de "Eu tenho *muito que* fazer")

Casos especiais

Foram considerados como casos especiais aqueles em que uma expressão, a depender do sentido, é usada ora de uma maneira, ora de outra. Os casos considerados são das expressões *ao ponto de*, *face a* e *para a frente*:

i) *Ao ponto de*: Exceto no caso de estar sendo usada com o sentido de "ter a capacidade de", a expressão está fora de contexto. Assim, é mostrada ao usuário a mensagem: *"Ir ao ponto de* significa "ter a capacidade de". Se a expressão estiver em qualquer outro sentido, substitua por *a ponto de.*

ii) *face:* Exceto em expressões como *face a face*, a expressão *face a* deve ser substituída por *em face de.*

iii) *para a frente:* Exceto se usada no sentido de "progredir", a expressão deve ser substituída por *para frente.* Assim, é mostrada ao usuário a mensagem: "*Ir para a frente* significa "progredir". Se a expressão foi usada em qualquer outro sentido, substitua por *para frente.*

b) Subclasse Gênero

Existem algumas palavras em português que aceitam apenas um gênero, e que causam confusões ao usuário, como por exemplo, "o champanhe". Dessa maneira, foram implementadas correções para este e outros casos.

c) Subclasse Pontuação

Estas regras conseguem corrigir a ausência da vírgula em alguns lugares em que ela obrigatoriamente deveria estar. Um exemplo de erro detectado:

- "Cheguei uns cinco minutos atrasado mas tudo bem, meu chefe também não estava lá". (Falta vírgula antes do "mas")

5.4.10. Uso Incorreto de Prefixos

O prefixo é definido como a sílaba que antecede a raiz de uma palavra, modificando-lhe o significado e formando uma palavra nova. Esse processo de formação de palavra é denominado *derivação prefixal*. Em princípio, os prefixos devem sempre se ligar diretamente ao radical:

ante + datar = *antedatar*

neo + clássico = *neoclássico*

inter + nacional = *internacional*

super + agudo = *superagudo*

No entanto, o **Formulário Ortográfico** aprovado pela Academia Brasileira de Letras em 29/01/1942 prescreve que se utilize o hífen em alguns casos específicos que serão apresentados mais abaixo. Antes de listar esses contextos, é importante apontar para o fato de que existem algumas formas livres, principalmente preposições, que se assemelham aos prefixos, dificultando a identificação, por parte da ferramenta, se naquele contexto se trata ou não do prefixo. Por exemplo, *extra* é um prefixo que quer dizer "posição exterior", "fora de", (por exemplo, *extradiscursivo*). Mas quando *extra* equivale a "extraordinário", não é um prefixo (por exemplo, *edição extra*).

Assim, os prefixos foram separados em dois grupos. No primeiro grupo estão as formas que só podem ser usadas como prefixos (como *super*, *pré*, pós, etc). Nesse caso, o Revisor verifica se o hífen está sendo usado conforme a regra em questão. Se o hífen estiver sendo usado inadequadamente, é mostrada uma mensagem para o usuário, sugerindo-lhe que escreva o prefixo junto ao radical (regra geral de prefixação).

No segundo grupo estão as formas que podem ser prefixos ou formas livres (por exemplo, *extra*, *auto*, *supra*, entre outros). Nesse caso, a ferramenta, em primeiro lugar, identifica a forma e, depois, aplica a regra de uso do hífen. Caso este esteja sendo usado inapropriadamente, ela junta a forma à palavra que a segue e verifica se a palavra resultante existe no dicionário do *Proverb*. Caso exista, uma mensagem é apresentada ao usuário sugerindo a junção das duas formas. Se a palavra não existir no *Proverb*, então se tem duas situações: ou a

palavra resultante de fato não existe e, portanto, a forma que apareceu nesse contexto não é um prefixo; ou a forma em questão é um prefixo, mas a palavra resultante não consta na listagem do *Proverb* por limitação deste.

Para solucionar esse problema, optou-se por apresentar ao usuário uma mensagem que explica o significado daquela forma enquanto prefixo e sugere que, se a forma está sendo usada como prefixo, então deve ser escrita junto à palavra que a segue (regra geral de prefixação). Caso não seja um prefixo, a forma deve ser mantida separada da palavra que a segue.

Uma observação importante é que o *Proverb* apontará como erro ortográfico uma palavra que não consta de seu dicionário, mesmo que essa palavra exista e que a derivação esteja correta. Como é impossível esperar que o dicionário do *Proverb* contenha **todas** as formas possíveis na língua, é necessário estabelecer um mecanismo que faça com que o Revisor aceite as formas que condizem com as regras de derivação prefixal, por exemplo. O procedimento de separação dos prefixos em dois grupos diminuiu sensivelmente as chamadas para que o usuário esclareça se, num determinado contexto, a forma em questão estava ou não sendo usada como prefixo, o que agilizou consideravelmente o uso da ferramenta.

As regras, além de verificar o uso do hífen, detectam a ocorrência das formas listadas como prefixos para tentar checar se o usuário está, indevidamente, escrevendo o prefixo separado de seu radical. Caso a ferramenta identifique algum erro, este é apontado ao usuário através de uma mensagem. Alguns exemplos de erros detectados são ilustrados a seguir.

- "Gire no sentido antihorário". (O correto seria "anti-horário")
- "Ela é uma pessoa super-importante." (O correto seria "superimportante")
- "Ninguém consegue ser um superhomem." (O correto seria "super-homem")

6. Interface com o Usuário

A interface com o usuário segue o padrão adotado pelo *Redator Windows*. Coube à equipe de desenvolvimento do Revisor a definição das mensagens de erro a serem apresentadas, bem como a discussão do leiaute para as janelas geradas durante o processo de revisão. Durante todo o desenvolvimento do Revisor, teve-se como filosofia oferecer ajuda apenas em casos de erro realmente identificados, filosofia reforçada pela tentativa de minimizar ao máximo a ocorrência de falsos erros. Ou seja, erros que poderiam ser tratados, mas cuja detecção pode gerar a ocorrência de muitos falsos erros, foram ignorados.

Ainda de acordo com essa filosofia, o Revisor apresenta mensagens objetivas e curtas, assertivas apenas quando se tem certeza do erro (ver Figura 6.1). Um botão na janela de revisão oferece informações adicionais, que desta forma são apresentadas apenas quando solicitadas pelo usuário (ver Figura 6.2). As mensagens de erro e o conteúdo apresentado como *Mais Informações* associados a cada regra são apresentados no Relatório de Projeto referente ao mês Julho/95.

A inclusão de novas mensagens de erros associadas às regras é feita através de sua inserção no arquivo "erros.msg", disponível no diretório "\rdw*dic*" do *Redator*. A

formatação das mensagens é especificada através de sua descrição de acordo com um *template* pré-definido.

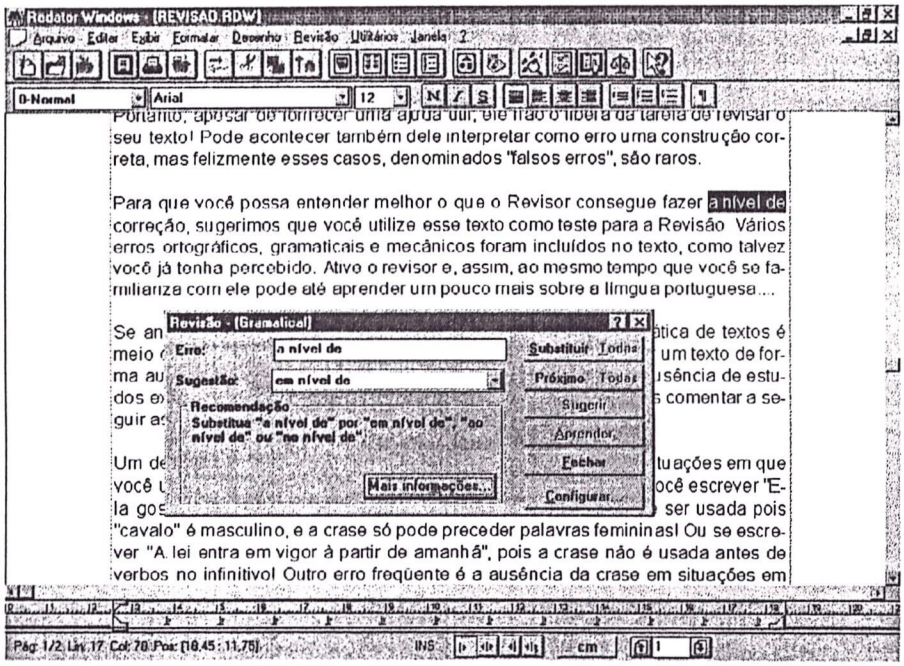

Figura 6.1: Exemplo de uma mensagem "Recomendação".

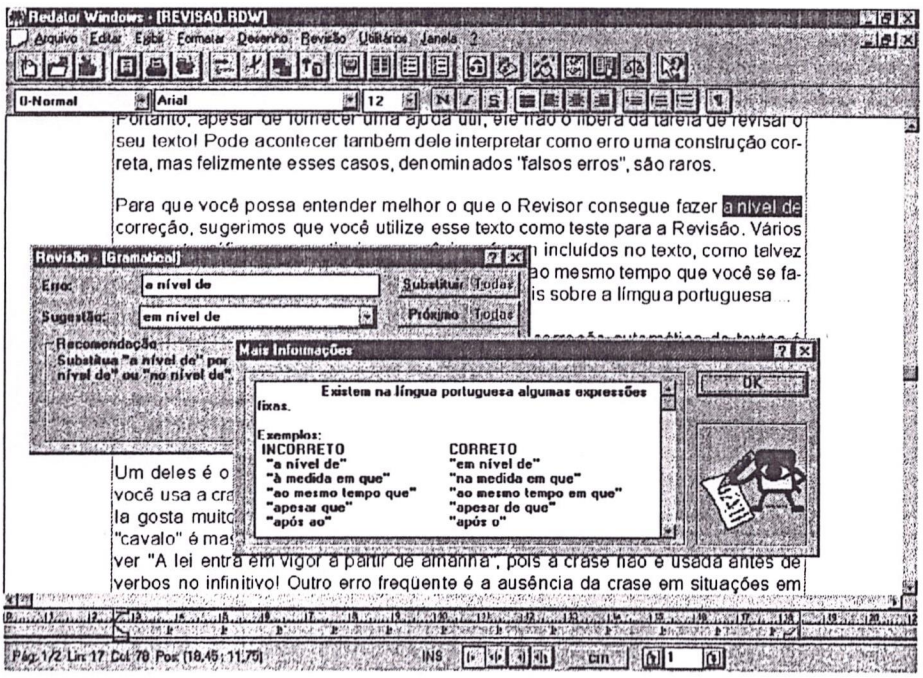

Figura 6.2: Exemplo de uma mensagem "Mais Informações".

7. Conclusões e Desenvolvimentos Futuros

Computacionalmente, a implementação de aproximadamente 200 regras na forma de ATNs mostrou-se bastante eficiente. Executado no modo não interativo, o *Revisor* é capaz de verificar uma tese de doutorado de 130 páginas (aproximadamente 17000 a 20000 palavras) em apenas 5 minutos. Em média, para esta quantidade de texto ocorrem menos de 10 falsos erros, um valor bastante satisfatório. Não temos disponível, infelizmente, uma estimativa dos erros não detectados.

A versão atual do Revisor funciona muito bem em casos específicos, e certamente pode ser bastante útil a um usuário comum, corrigindo possíveis deslizes nos seus textos. Entretanto, ainda há um grande número de erros que não podem ser corrigidos. Talvez os mais aparentes em um texto sejam os de concordância verbal em orações longas, nas quais a determinação do sujeito e dos complementos verbais é extremamente difícil. O tratamento de elipses é uma dificuldade adicional a ser considerada. Obviamente, existem as limitações inerentes ao fato de que é impossível fazer um tratamento em nível semântico do texto. Não é difícil imaginar que, no estágio em que se encontra o processamento de linguagem natural para o português, erros advindos de ambigüidades e relacionados ao contexto situacional não possam ser detectados automaticamente. Detectar desvios da norma gramatical vigente já é um desafio monumental.

É importante ressaltar também que, durante o desenvolvimento da ferramenta, têm sido feitos esforços no sentido de estabelecer uma metodologia de trabalho adequada através de um estudo sistemático da língua escrita em uso no Brasil. Uma séria dificuldade é a ausência de estudos e informações sobre o português contemporâneo, diferentemente do que ocorre com a língua inglesa, por exemplo. A título de ilustração, para a formulação de várias regras seria importante dispor de um estudo que mostrasse, em contexto, os vários usos de uma mesma palavra, para então estabelecermos freqüência de uso. É por isso que o nosso Grupo está agora empenhado em pesquisas em lexicografia e lexicologia, além da implementação da análise sintática automática que permitirá tratar com maior eficiência principalmente os erros de concordância.

A nova versão do Revisor, em desenvolvimento, incorpora técnicas para realizar a análise sintática do texto. Dessa forma, ao invés de um conjunto de regras o Revisor passa a ter uma abordagem integrada para a análise do texto, o que só se tornou possível devido ao desenvolvimento simultâneo de um léxico que mantém as informações necessárias, como categoria lexical, gênero, número, etc., para realizar a análise sintática.

Do ponto de vista da teoria computacional, temos adotado basicamente uma abordagem simbolista ao implementar as heurísticas, mas é provável que, a partir dos estudos em andamento, adotemos um sistema híbrido simbolista-conexionista. Um projeto de mestrado cujo objetivo é tratar os problemas relacionados à crase de forma coneccionista está em andamento.

Referências

Fernandes, C.T. Um algoritmo de Hifenização Multilíngue. Monografias em Ciência da Computação, 17/88, Departamento de Informática, PUC-Rio, Outubro de 1988.

Flesch, R.F. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, 32, pp. 221-233.

Folha de São Paulo, (6/7/1994a). Soft Corrige Erros Gramaticais. Caderno de Informática.

Folha de São Paulo, (11/7/1994b). Carta Certa Windows ganha Revisor Gramatical. Caderno de Informática.

Klare, G.R. (1974-1975). Assessing Readability. *Reading Research Quarterly*, N°1, X/1, pp. 62-102.

Machado, J. (15/3/1995). O Papai-e-mamãe caiu no Processador 'Word'. Folha de São Paulo, Caderno de Informática.

M.G.V. Nunes, R. Hasegawa, S. Kawamoto, M.C.F. de Oliveira, M.A.S. Turine, C.M. Ghiraldelo, O.N. Oliveira Jr., C.R. Riolfi, N.S. Sikanski e T.B. Martins (1995); "Style and Grammar Checkers for Brazilian Portuguese". Apresentado na *VIII Annual Conference on Writing and Computers*, Londres, Setembro de 1995. Disponível como Notas do ICMSC, n.25, Série Computação, Maio de 1996.

Reference Software International (1992). The #1 Grammar & Style Checker Gram.mat.ik 5 (for Windows) User's Guide. San Francisco, CA, USA.

Rocco, M.T.F. Texto e Discurso: Uma caracterização da Linguagem Escrita de Candidatos a Vestibular. Faculdade de Educação, USP-SP, 1981 (Tese de doutoramento).

Writing Tools Group, Inc.(1991). *Using Correct Grammar for Windows*. Sausalito, CA, USA.