O processo de desenvolvimento do revisor gramatical ReGra

Maria das Graças Volpe Nunes
Instituto de Ciências Matemáticas e da Computação
ICMC/USP - São Carlos
mdgvnune@icmc.sc.usp.br

Osvaldo Novais de Oliveira Jr.
Instituto de Física
IFSC/USP - São Carlos
chu@ifsc.sc.usp.br

Núcleo Interinstitucional de Lingüística Computacional - NILC
ICMC/USP - São Carlos
Caixa Postal 668
13560-970 São Carlos, SP

O processo de desenvolvimento do revisor gramatical ReGra[1]

Resumo

A parceria USP-Itautec/Philco possibilitou a colocação no mercado de um revisor gramatical para o português do Brasil, avaliado como o de melhor desempenho na categoria, com qualidade similar à dos revisores gramaticais para o inglês. Trata-se, portanto, de um raro exemplo em que uma ferramenta de software específica para a língua portuguesa está no mesmo nível das do inglês. O outro efeito visível da parceria foi a formação de uma equipe multidisciplinar, a maior já criada no Brasil para pesquisas em processamento de linguagem natural (PLN) do português, dando origem ao Núcleo Interinstitucional de Lingüística Computacional (NILC), que hoje ocupa lugar de destaque na comunidade de PLN. Sobretudo, o sucesso da parceria é demonstração da possibilidade de uma colaboração efetiva entre universidades e setor produtivo, que ao mesmo tempo em que gera riquezas também produz trabalho científico de qualidade. Por fim, mencione-se que o projeto de P&D da parceria foi agraciado com dois prêmios de Inovação Tecnológica em 1998 e 1999.

Abstract

The partnership between USP and Itautec/Philco has generated a commercially available grammar checker for Brazilian Portuguese, which performs best among the systems of this category and has comparable quality to the checkers available for English. It is therefore a rare example of a software tool for Brazilian Portuguese that is at the same level of its English counterparts. Another important result of the partnership was the formation of a multidisciplinary team, the largest ever in Brazil for natural language processing (NLP) of Portuguese. This has given origin to the Núcleo Interinstitucional de Lingüística Computacional (NILC) which plays a leading role in the NLP community for Portuguese. Above all, this successful partnership is demonstration of the possible, effective collaboration between university and industry, which may generate businesses as well as quality fundamental research. The importance of the partnership has been recognized through two technological innovation prizes with which it was awarded in 1998 and 1999.

1. Introdução

As ferramentas de pós-processamento de textos em língua portuguesa são recentes, ao contrário das ferramentas para o inglês, que são largamente utilizadas há bastante tempo (p.e. CorrectGrammar, Grammatik, e as embutidas em processadores de textos, como o MS Word). O revisor de textos aqui apresentado está entre as primeiras ferramentas de revisão gramatical do português brasileiro que surgiram no mercado nacional. Trata-se de um sistema que, embutido em um processador de textos -- no caso, o MS Word ou o Redator (Itautec) -- promove a revisão ortográfica e gramatical de qualquer texto escrito em português. Enquanto a revisão ortográfica objetiva detectar palavras faltantes do léxico em questão e sugerir alternativas válidas a ela, a revisão gramatical procura detectar desvios gramaticais cometidos pelo usuário, tais como de concordância nominal ou verbal, pontuação, regência nominal ou verbal, uso de pronomes, além de problemas mais simples, porém bastante freqüentes.

[1] Este trabalho tem o apoio de Itautec-Philco, FAPESP, CNPq, MCT/PADCT e Finep.

Neste artigo pretende-se enfocar alguns aspectos da parceria feita entre a USP e a Itautec-Philco, como a formação de uma equipe multidisciplinar e o impacto da atuação da equipe em outros projetos de pesquisa. Aspectos técnicos, tanto lingüísticos como computacionais, do revisor gramatical não serão abordados em detalhe, uma vez que tal material já foi divulgado em outras oportunidades (Oliveira et al., 1995, Nunes et al., 1996a, 1996b, Martins et al., 1998).

2. A Parceria

O revisor gramatical ReGra é fruto de uma parceria entre USP e Itautec-Philco que teve início em 1993. A história da parceria começou com a disposição da Itautec/Philco em investir em ferramentas de software para a língua portuguesa, algo extremamente incipiente em 1993, quando o convênio de parceria foi celebrado. Tal disposição em investir em projetos de P&D, cujo retorno na época era bastante incerto, surgiu do idealismo de líderes da empresa que acreditavam ser importante para a Itautec/Philco, na condição de maior empresa nacional de informática, contribuir para que o Brasil não ficasse totalmente dependente de companhias e iniciativas estrangeiras para o desenvolvimento de ferramentas de auxílio à escrita da nossa própria língua. O projeto apresentava alto risco não apenas do ponto de vista mercadológico, pois não se podia prever com precisão nem o tempo de gestação nem o custo de um produto minimamente aceitável para ser colocado no mercado, mas também da perspectiva da viabilidade tecnológica. Ocorre que apesar de as pesquisas em processamento de linguagem natural (PLN) de português terem se iniciado muito antes da década de 1990, praticamente nada havia sido feito que visasse à criação de uma ferramenta robusta e de uso genérico, que requer recursos lingüísticos e computacionais de grande monta. As indefinições e incertezas características de uma inovação tecnológica visando ao desenvolvimento de um sistema complexo como um revisor gramatical fizeram com que os docentes da USP, então convidados para participar da parceria, assumissem o compromisso inicial apenas de um estudo exploratório, sem a responsabilidade de ter que gerar algo que obrigatoriamente levasse a um produto comercial. Apesar dessas salvaguardas, era necessária ousadia para assumir um projeto de características inéditas, com as dificuldades que podiam ser antevistas tanto para a geração de um sistema robusto como para a produção de trabalho acadêmico. As primeiras dificuldades estavam relacionadas à possível inadequação da tecnologia de PLN disponível na época para atingir uma ferramenta de qualidade para o português, principalmente por causa da necessidade de volumosos recursos lingüísticos como dicionários, bancos de texto, e sistemas de análise sintática automática. Já as dificuldades para a produção acadêmica de alto nível eram representadas pela necessidade do trabalho cooperativo e coordenado de pesquisadores de áreas muito distintas do conhecimento, a saber, ciência da computação e lingüística.

A constatação dessas dificuldades acabou por se transformar em incentivo adicional para que o convite para a parceria fosse aceito, uma vez que fez crescer a percepção de que avanços significativos em PLN do português só seriam alcançados a partir do trabalho de uma equipe multidisciplinar que pudesse dispor de recursos lingüísticos de grande porte. O projeto de P&D em parceria com uma empresa privada era a única possibilidade para que estes requisitos fossem alcançados, especialmente porque nosso sistema universitário não contempla a formação de pesquisadores em Lingüística Computacional que pudessem desenvolver esse tipo de trabalho no âmbito de projetos acadêmicos. Tal percepção acabou sendo vindicada, na medida em que o trabalho multidisciplinar com desenvolvimento de

recursos lingüísticos é atualmente considerado essencial por toda a comunidade de PLN. Além disso, o apoio incondicional oferecido pelos Institutos da USP aos quais pertencem os docentes, tem refletido a importância que atualmente a universidade brasileira dá para os projetos que possam efetivamente transferir tecnologia para o setor produtivo.

O projeto de parceria permitiu a formação do Núcleo Interinstitucional de Lingüística Computacional, que hoje conta com pesquisadores docentes da USP de São Carlos, da Universidade Federal de São Carlos e da Unesp de Araraquara, coordenando uma equipe de cerca de 20 pesquisadores de ciência da computação e lingüística. Pode-se afirmar que a criação do NILC é paradigmática para o PLN do português, pois o impacto de seu trabalho de P&D serviu de demonstração que resultados de valor acadêmico podem ser gerados a partir do tratamento de problemas à primeira vista específicos para uma aplicação particular, como a revisão gramatical (Martins et al., 1998). Isso decorre primordialmente da possibilidade de trabalhar com contextos reais a partir da utilização de recursos lingüísticos volumosos, e não mais domínios circunscritos aos quais eram limitados os sistemas de PLN desenvolvidos até então. Este sentimento vem enfatizando o sinergismo entre pesquisas fundamentais e aplicadas.

O aparecimento de programas de inovação tecnológica, financiados a fundo perdido pela FAPESP[2] e Finep (PADCT)[3], permitiu que uma série de avanços fosse conseguida, primordialmente no que tange ao desenvolvimento de pesquisas mais fundamentais, cujo efeito no produto final poderia não ser sentido imediatamente. No caso do projeto USP-Itautec/Philco, esta possibilidade acabou se transformando em enorme catalisador, uma vez que – embora a viabilidade tecnológica do ReGra já estivesse confirmada – havia pouco retorno dos investimentos maciços da Itautec/Philco. Encontrávamos em 1997, ano no qual o primeiro projeto de inovação foi aprovado pela FAPESP, numa situação de impasse, pois apesar de a parceria ser considerada extremamente bem-sucedida pelas equipes envolvidas, tanto do NILC como da coordenação técnica da Itautec/Philco, a aparente falta de perspectiva de viabilidade comercial quase fez a empresa desistir da empreitada. O apoio da FAPESP serviu como um aval para a qualidade do trabalho de P&D do NILC, com efeitos psicológicos consideráveis. Durante a realização desse projeto de parceria Universidade/Empresa, da Fapesp, além da criação de acessórios ao revisor gramatical, como a gramática online, não existente em revisores similares, mesmo para o inglês, o projeto como um todo atingiu maturidade. As perspectivas de viabilidade melhoraram consideravelmente, e a celebração de um acordo entre a Itautec/Philco e a Microsoft garantiu o uso em larga escala do revisor gramatical. A parceria foi novamente avalizada com a concessão de um projeto de Desenvolvimento Tecnológico, do PADCT, com recursos da Finep e contrapartida da Itautec/Philco. Neste projeto, novas versões do revisor gramatical estão sendo geradas, empregando resultados de pesquisas fundamentais que incluem análise semântica da tipologia de falhas do sistema de revisão. E já não é somente uma parceria da USP, pois agora envolve todo o NILC, com a participação de docentes da Universidade Federal de São Carlos e da Unesp de Araraquara.

Até o momento focalizamos a parceria da perspectiva do NILC, que realizou o trabalho de P&D especificamente para o sistema de revisão gramatical automático. Obviamente, este é apenas um componente de um sistema muito mais complexo, que envolve correção ortográfica e recursos de compactação dos dados do dicionário, algo que foi obtido numa

[2] Programa de Inovação Tecnológica da FAPESP, Proc..# 97/02608-1
[3] Programa PADCT-III(CDT), Conv. # 8.8.98-0591/00

parceria entre a Itautec/Philco e o Instituto de Computação da Unicamp. Todo o trabalho de engenharia de software e engenharia de produto, para incorporar os sistemas de revisão gramatical e ortográfica aos produtos da Itautec/Philco, foi realizado por uma empresa, Techno Software, de Ribeirão Preto. O sucesso da parceria foi fruto, assim, de uma cooperação eficaz entre membros do NILC, do Instituto de Computação da Unicamp, da Techno Software e da Itautec/Philco. Além disso, a administração do projeto na USP vem sendo efetuada pela Fundação de Apoio à Física e à Química (FAFQ), ligada aos Institutos de Física e de Química da USP de São Carlos.

3. O Revisor Gramatical ReGra

Chamamos de ReGra o sistema de correção gramatical, não incluindo as rotinas para detecção de erros ortográficos, embora a base lexical que suporta o corretor ortográfico tenha sido compilada para o projeto de correção gramatical. O ReGra é constituído por três módulos principais: i) o módulo estatístico, ii) o mecânico e iii) o módulo gramatical. As rotinas para compactação e acesso aos dados do léxico foram desenvolvidas pela equipe do Prof. Tomasz Kowaltowski, do Instituto de Informática da Unicamp (Kowaltowski & Lucchesi, 1993).

O módulo de tratamento estatístico realiza uma série de cálculos, fornecendo parâmetros físicos de um texto sob análise, como o número total de parágrafos, sentenças, de palavras, de caracteres, etc. O componente mais importante desse módulo, entretanto, é o que fornece o "índice de legibilidade", uma indicação do grau de dificuldade da leitura do texto (Martins et al., 1996). O conceito de índice de legibilidade surgiu a partir do trabalho de Flesch (Flesch, 1948) para a língua inglesa e busca uma correlação entre tamanhos médios de palavras e sentenças e a facilidade de leitura. Não inclui aspectos de compreensão do texto, que requereriam tratamento de mecanismos complexos de natureza lingüística, cognitiva e pragmática. O índice Flesch, assim como outros similares, tem sido empregado para uma grande variedade de línguas, mas o trabalho do NILC foi o primeiro para a língua portuguesa. Através de um estudo comparativo de textos originais em inglês e traduzidos para o português, verificou-se que a equação que fornece o índice Flesch precisaria ter seus parâmetros adaptados para o português, pois as palavras desta língua são em média mais longas, em termos do número de sílabas, do que em inglês.

Textos classificados como **muito fáceis** seriam adequados para leitores com nível de escolaridade até a quarta série do ensino fundamental; textos **fáceis** seriam adequados a alunos com escolaridade até a oitava série do ensino fundamental; textos **difíceis** seriam adequados para alunos cursando o ensino médio e/ou universitário, e textos **muitos difíceis** em geral seriam adequados apenas em áreas acadêmicas específicas. Por se tratar de um dado estatístico, o índice de legibilidade só é calculado para trechos com mais de 100 palavras. Testes realizados com textos tradicionalmente dirigidos a públicos dessas quatro faixas mostraram resultados bastante satisfatórios. Por exemplo, jornais de grande circulação como a Folha de São Paulo e o Estado de São Paulo têm em seus cadernos principais índices de legibilidade que correspondem a textos adequados a leitores com escolaridade equivalente ao final do ensino fundamental. Textos de cadernos infantis, por outro lado, apresentam índices de Flesch modificados na faixa de muito fácil, ou seja, podem em princípio ser acompanhados por crianças que ainda não completaram os quatro primeiros anos do ensino fundamental.

O segundo módulo do ReGra, o mecânico, detecta erros facilmente identificáveis que não são percebidos por um corretor ortográfico. Exemplos desse tipo de erro são: i) palavras e

símbolos de pontuação repetidos; ii) presença de símbolos de pontuação isolados; iii) uso não balanceado de símbolos delimitadores, como parêntesis e aspas; iv) capitalização inadequada, como o início da sentença com letra minúscula; v) ausência de pontuação no final da sentença.

O primeiro passo para a elaboração do módulo gramatical foi o levantamento de erros (ou inadequações) mais comuns entre usuários de nível médio, como secretárias e profissionais de escritório em geral, e alunos cursando o ensino médio ou ingressando a universidade. O termo "erro", aqui, refere-se ao que os gramáticos normativos consideram como forma desviante da norma culta. Como talvez pudesse ser esperado, o levantamento apontou erros de ortografia, de concordância e relacionados à crase como os mais freqüentes, seguidos de erros associados a escolhas léxicas inadequadas, principalmente por influência da oralidade. O objetivo foi o de implementar uma ferramenta voltada aos interesses de potenciais usuários. Essa escolha pressupõe a tomada de importantes decisões acerca dos itens lexicais a serem incluídos na base, a definição do que será considerado “erro” ou “inadequação”, e a elaboração de regras para detectar tais erros. Uma preocupação importante é a de minimizar o número de falsos erros (falsos positivos), ou seja, uma intervenção do sistema que pode induzir o usuário a um erro gramatical, por meio da modificação de uma estrutura lingüística originalmente correta; ou uma intervenção desnecessária do sistema que pode levar o usuário a alterar uma estrutura originalmente correta por uma outra forma correta.

3.1 Exemplos de Erros Detectados

Os erros detectados pelo ReGra incluem, entre outros, aqueles decorrentes de:

- uso incorreto ou da ausência de crase
  - Costumava ler o evangelho durante **às** refeições.
  - A boa safra começa **à** partir de julho deste ano.
  - Eu vou **as** duas horas ao encontro marcado. .
  - O carro parou devido **a** falta de combustível.
- colocação pronominal imprópria
  - Eu **darei-te** todo o auxílio que puder.
  - Nunca **vi-a** tão gorda.
- uso inadequado de pronomes
  - **Eu** fiquei fora de **si**.
  - **As visitas** bateram à porta. Mande **elas** entrar. .
- falta de concordância verbal de particípio
  - **Foram aprovado todas** as alunas.
  - **Foi detectado uma** pane no sistema de ar-condicionado.
- falta de concordância nominal ou verbal
  - **A maioria dos corredores chegaram** ao fim da prova.
  - **Deu** três horas no relógio da matriz. .
  - Há uma semana, **acabou as férias**.
  - **Começou as aulas** no novo colégio de ensino médio e fundamental da cidade.
  - **Tudo é** flores. .
  - Devemos nos manter**mos** em pé.
- inadequações no uso dos verbos Fazer e Haver
  - Ele chegou **a** dois anos.
  - **Fazem** dois meses que não tomo cerveja.
  - **A** muito tempo moro nesta casa. .

Vou visitá-la daqui **há** dois dias. .
Naquele ano **houveram** poucos acontecimentos que valem a pena recordar.

- uso inadequado da partícula "se"
  Vende-se casas.
  Precisam-se de funcionários.
- regência verbal ou nominal
  Eu assisto **o** jogo, **o** filme e **a** novela.
  Aonde você está? .
  Onde você vai? .
  Prefiro escrever **do que** falar. .
- uso inadequado do particípio regular/irregular
  Os criminosos **foram pegados** pela polícia.
  A polícia **tem pego** criminosos.
- pontuação inadequada
  O interesse no trabalho informal no Brasil, cresceu a partir dos anos 90.
- vícios de linguagem
  Ele reincidiu **de novo** no erro.
  Ele subiu **para cima** do palco.
- inadequação lexical
  A rua é **melhor iluminada**. .
  A moça comprou **duzentas** gramas de ameixas.
- emprego de mau/mal
  O **mal** filho não saiu de casa.
  Mau cheguei e já tenho que sair.

3.2 Implementação do Revisor

Nas primeiras versões do ReGra, os erros eram detectados através de regras heurísticas implementadas na forma de redes de transição estendidas (“augmented transition networks”) (Woods, 1970), numa abordagem que se poderia chamar de "*error-driven*". O paradigma que direcionou a construção do corretor gramatical baseou-se fortemente, portanto, no estudo da língua em uso, com testes em textos reais. Para tanto, foi compilado um *corpus*, que contém textos de várias áreas do conhecimento, e inclui tanto textos já corrigidos e editados que servem como referência do uso corrente da língua escrita quanto textos escritos por algumas classes de usuários comuns, sem correção. Pertencem à primeira classe dissertações e teses, jornais e livros. A segunda classe de textos inclui redações de vestibular e monografias, que fornecem uma amostra dos erros cometidos pelos usuários da língua.

Alguns tipos de erros podem ser detectados a partir de contextos lingüísticos bastante específicos, limitando-se à identificação, na sentença, de combinações lexicais (*patterns*) que configuram formas desviantes. Corrige-se, dessa forma, uma série de desvios, comuns para usuários inexperientes da norma-padrão da língua portuguesa escrita, como o uso indevido de crase diante de palavras masculinas e verbos, uso de ênclise nas formas do futuro, uso de ênclise e mesóclise em prejuízo de palavras atrativas, etc. Existem, também, regras de estilo, que detectam o uso de uma palavra ou expressão que não se configura como um erro gramatical, mas que é considerado impróprio para o estilo de escrita selecionado pelo usuário. Por exemplo, o uso de coloquialismos é inadequado em um texto formal, ainda que aceitável em um texto

jornalístico. O mesmo conjunto de regras (gramaticais e de estilo) pode ser aplicado a qualquer estilo. O usuário pode, durante a análise do texto, optar por desabilitar algumas regras.

O emprego de uma metodologia consistente e sistemática para a implementação de cada regra de correção foi essencial para o desenvolvimento do revisor. Foram identificadas três etapas principais:

1) identificado um tipo de erro que se deseja corrigir, é feito um estudo extensivo de gramáticas e fontes da literatura que discorram sobre o uso da língua portuguesa. Mecanismos de correção são então propostos na forma de regras para a detecção e correção desses erros.
2) as regras propostas são implementadas computacionalmente, e através de testes exaustivos com o material que compõe o corpus, são verificados falsos erros e erros não detectados. Melhorias na regra são então implementadas. Este processo é altamente interativo.
3) é dado o acabamento 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.

As primeiras versões do ReGra apresentavam vários benefícios do ponto de vista da implementação computacional: agilidade, especificidade, rapidez, portabilidade, e disponibilidade de memória. Entretanto, seu escopo de atuação era muito limitado: problemas envolvendo itens lexicais não contíguos e estruturas recursivas não podem ser atingidos pelas estratégias heurísticas normalmente desenhadas por abordagens *error-driven*. Para prover a essas insuficiências, optou-se por analisar sintaticamente as sentenças do usuário, antes de operar a revisão propriamente dita. Isso permite aplicar regras que apontam desvios nas relações entre núcleos e adjuntos, entre núcleos e modificadores, entre regentes e regidos. A realização de análise sintática automática obviamente requer que todos os itens lexicais estejam categorizados apropriadamente. Para tanto, realizou-se em paralelo a construção do léxico, que envolveu a compilação exaustiva das palavras da língua portuguesa e a hierarquização das categorias dos itens lexicais morfologicamente ambíguos.

Uma vez que alguns erros em contextos lingüísticos específicos ocorrem independentemente de desvios sintáticos, na versão atual do ReGra convivem as duas abordagens mencionadas acima. Ou seja, além de realizar análise sintática automática, muitas das regras heurísticas da primeira versão foram mantidas, como as de correção de erros de crase.

3.3 Avaliação do Desempenho do Revisor

O desempenho do Revisor, quanto a tempo de execução, pode ser considerado ótimo, uma vez que as mensagens de erro são apresentadas ao usuário praticamente instantaneamente. As limitações do Revisor, entretanto, estão localizadas nas intervenções indevidas e nos erros não detectados (omissões). A maior parte destes problemas advém da impossibilidade de serem previstas todas as estruturas sintáticas desviantes que podem ser empregadas por usuários médios. Embora o sistema conte, hoje, com mais de 600 produções, ainda aparecem estruturas para as quais nenhum casamento (*matching*) é obtido. Além disso, as dificuldades advindas de pluricategorização de alguns itens lexicais, principalmente nos casos de homonímia, precisarão ser tratadas em casos especiais, o que certamente demandará grandes esforços de pesquisa. Em algumas situações, a inserção de conhecimento semântico no léxico é indispensável, sendo essa uma meta da nossa equipe para o futuro próximo.

Apresentamos a seguir alguns resultados obtidos recentemente de um teste comparativo entre as diferentes ferramentas de revisão ortográfica e gramatical de português brasileiro que estão comercialmente disponíveis: o ReGra, nas suas versões do Office 2000 e a mais recente, da versão 9.0 do RLP (Redação Língua Portuguesa) da Itautec; a versão 3.0 do DTS e a versão 1.0 da Gramática Eletrônica (abreviada aqui por GE). Trata-se de testes realizados sob a perspectiva do usuário, isto é, as ferramentas de revisão são utilizadas de forma a simular a atividade de revisão automática promovida por pessoas sobre textos escritos em português do Brasil.

Os testes foram baseados num conjunto de textos (corpus) recuperados da base textual do NILC (vide seção 4). A fim de torná-la representativa, a amostragem de análise reúne 15 textos para cada um dos corpora, resultando num total de 45 textos. As fontes diversas de cada gênero textual determinaram a abrangência da amostragem. Podemos encontrar no nosso corpus textos jornalísticos, acadêmicos e redações livres (diversidade de gênero textual), bem como textos da revista *Istoé*, do jornal *Folha de São Paulo*, redações do vestibular da *Fuvest* e também da *Unicamp*. Na amostragem, os textos de cada corpus mantêm a média de 3 páginas e 1.400 palavras.

As intervenções das ferramentas foram avaliadas segundo a seguinte classificação:

1) Intervenções Indevidas: quando não há erro e a ferramenta intervém equivocadamente no texto do usuário. Também chamadas de falsos erros ou falsos positivos, cujo número procura-se minimizar.

   a) *intervenção mecânica indevida*: por exemplo, quando há abertura de aspas em determinado trecho do texto e o fechamento das mesmas ocorre após um ponto final. Se a ferramenta pede o fechamento das aspas antes de sua ocorrência, há uma intervenção indevida no módulo mecânico.

   b) *intervenção gramatical indevida*: por exemplo, quando a ferramenta sugere a concordância nominal com o nome mais próximo, sendo que ela é feita corretamente com uma conjunção anterior. Ex.:"palavras e símbolos de pontuação repetidos" e a ferramenta pede a concordância entre "repetidos" e "pontuação", então há uma ocorrência de intervenção gramatical indevida.

   c) *intervenção ortográfica indevida*: por exemplo, quando uma palavra grafada com hífen (ex.: "extra-lingüístico") é tomada como palavra inválida (ou inexistente no léxico). Nesse caso, não se trata de palavra inexistente na língua, mas de erro na ortografia da mesma (extralingüístico) e, portanto, uma ocorrência de intervenção ortográfica indevida.

2) Intervenções devidas: quando existe o erro e a ferramenta detecta o problema. Quanto maior esse número, melhor o desempenho da ferramenta.

3) Omissões: quando o erro está manifestado, mas a ferramenta não é sensível a ele. Quanto menor o número de omissões, melhor a ferramenta.

   a) *omissão mecânica*: por exemplo, símbolos como parênteses, aspas, apóstrofes, quando usados, devem aparecer imediatamente concatenados, à esquerda ou direita, às palavras. Quando não detectado pela ferramenta, constitui uma ocorrência de omissão mecânica.

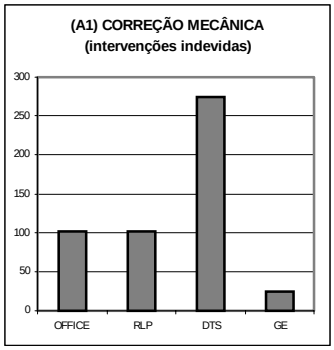
   b) *omissão gramatical*: por exemplo, quando o usuário omitiu uma crase necessária e a ferramenta não detecta esse erro.

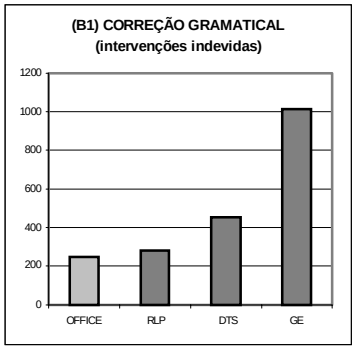
   c) *omissão ortográfica*: por exemplo, quando uma palavra deve iniciar obrigatoriamente com letra maiúscula (ex.: "Brasil") e a ferramenta não aponta o erro quando a encontra começando com minúscula.

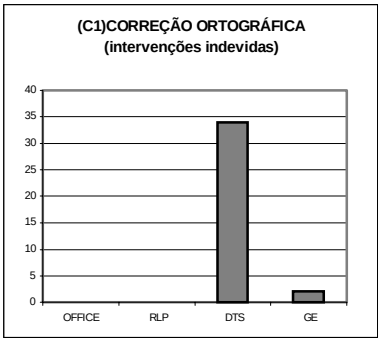
Apresentamos a seguir os gráficos de desempenho das ferramentas, separados pela categoria do erro — mecânico (A), gramatical (B), ortográfico (C) — e tipo do fenômeno avaliado — intervenção devida (1), indevida (2), omissão (3).

Quadro Comparativo por módulos

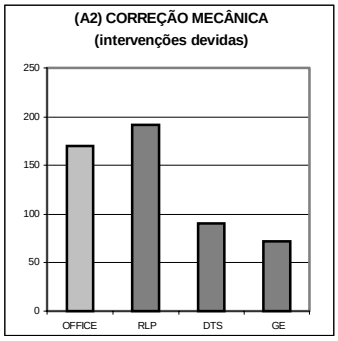
NO CONJUNTO TOTAL DOS CORPORA = 45 TEXTOS

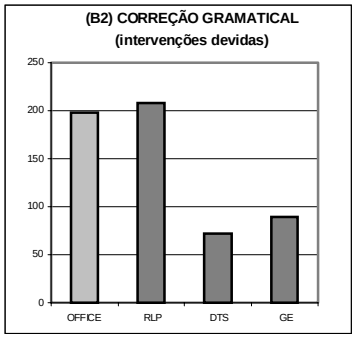


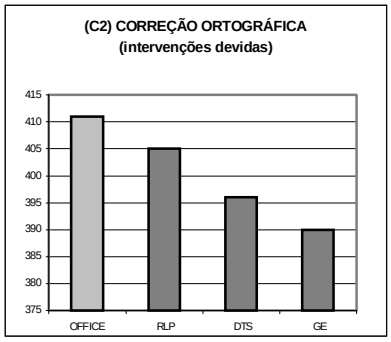




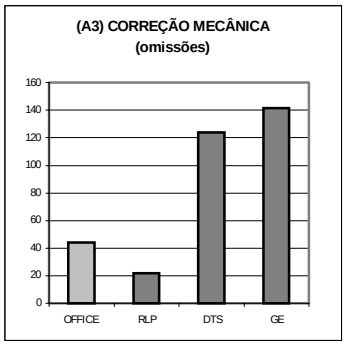
Obs.: quanto mais BAIXO o bloco, melhor o desempenho da ferramenta

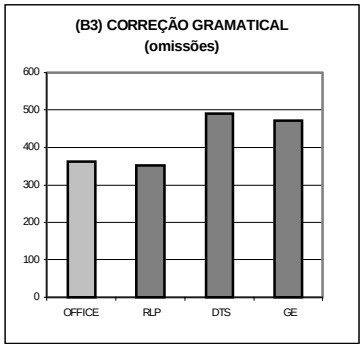


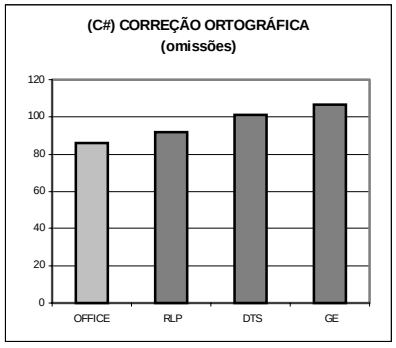




Obs.: quanto mais ALTO o bloco, melhor o desempenho da ferramenta







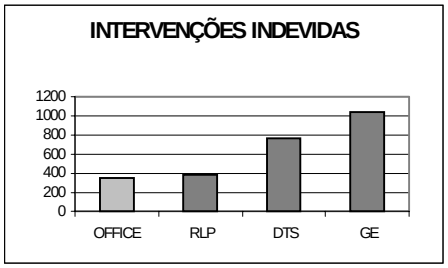
Obs.: quanto mais BAIXO o bloco, melhor o desempenho da ferramenta

Os gráficos a seguir resumem a comparação entre as ferramentas quanto aos três fenômenos avaliados. Juntamente com outras análises detalhadas em (Montilha & Nunes, 2000), eles nos levam a concluir, entre outras questões que não cabem aqui, que: (a) quaisquer das versões do

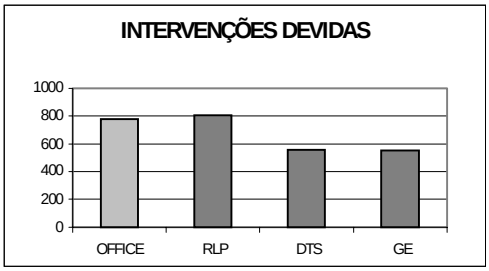
ReGra (Office ou RLP) apresentam um desempenho geral superior sobre seus concorrentes diretos (DTS e GE), especialmente com relação ao comportamento gramatical; (b) as omissões, especialmente as gramaticais, continuam freqüentes em todas as ferramentas de revisão, o que justifica o investimento em pesquisa que se continua fazendo para melhorar o desempenho do ReGra.

Quadro Comparativo (comportamento geral)

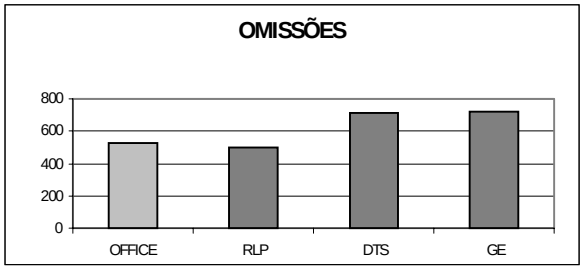
NO CONJUNTO TOTAL DOS CORPORA: 45 textos



Quanto mais BAIXO o bloco, melhor o desempenho



Quanto mais ALTO o bloco, melhor o desempenho



Obs.: quanto mais BAIXO o bloco, melhor o desempenho

4. Benefícios Indiretos do Projeto

O Projeto ReGra, além de permitir o desenvolvimento de um revisor gramatical para o português cujo desempenho é o melhor do mercado, propiciou uma série de estudos que culminaram na produção de recursos lingüísticos de apoio e em projetos completamente independentes de um revisor gramatical. Isso se deu não só pela disponibilidade de recursos reutilizáveis, mas também devido à experiência de uma equipe multidisciplinar que se formou a partir do Projeto ReGra. Destacamos os seguintes recursos e ferramentas:

(a) Um léxico da língua portuguesa. O léxico compilado no projeto de colaboração com a Itautec/Philco serviu para os revisores ortográfico e gramatical. Uma descrição detalhada da compilação desse léxico pode ser encontrada em (Nunes et al., 1996a). Para o revisor ortográfico, o léxico deve ser o mais abrangente possível, contendo inclusive nomes próprios, siglas, abreviaturas, etc. Já para o módulo gramatical, as palavras do léxico precisam ser categorizadas quanto a sua classe gramatical, o que dificulta a manipulação de grandes massas de dados requeridas pela abrangência do revisor ortográfico. A compilação de um conjunto de palavras para a formação de um léxico é conceitualmente simples, apesar do enorme volume de trabalho envolvido. De fato, a compilação do presente léxico tomou praticamente um ano de trabalho de três lingüistas e dois informatas, dedicando-se respectivamente 30 e 20 horas semanais. Além disso, o que em princípio parecia um trabalho mecânico, ainda que exaustivo,

acabou mostrando facetas interessantes com perspectivas de uma nova gama de pesquisas em lexicografia.

Partindo-se de um conjunto de aproximadamente 120 mil palavras normalmente encontradas em dicionários impressos, o maior trabalho consistiu em expandir o conjunto com: a) as conjugações dos verbos, b) as flexões de gênero, c) as flexões de número, d) as derivações de grau. Essas tarefas foram todas feitas automaticamente, a partir de algoritmos formulados pelos lingüistas. Para a maioria dessas tarefas foi necessária uma revisão "manual" cuidadosa para a detecção de malformação de palavras. Ressalte-se que a decisão de se construir um léxico cujas entradas são palavras (no máximo, palavras compostas hifenizadas) deveu-se a duas razões básicas: 1) o suporte ao revisor ortográfico não permitiria ou, pelo menos, dificultaria o uso de um léxico baseado em regras de aglutinação de morfemas, uma vez que, ao permitir construções morfologicamente válidas, estaríamos permitindo o uso de palavras que configurariam erros de fato (p.ex. *imexível*); 2) o conjunto inicial de verbetes foi extraído de um dicionário eletrônico, o que certamente economizou tempo e esforços. Testes do revisor ortográfico empregando o léxico (parcial) construído a partir do conjunto de verbetes inicial mostraram um desempenho insuficiente. Por isso, adicionalmente às formas previstas, um grande trabalho de verificação de formas faltantes foi feito utilizando-se um corpus (descrito a seguir) e o léxico foi expandido, contando atualmente com cerca de 1.500.000 lexemas gerados a partir de aproximadamente 100.000 lemas.

(b) Um corpus de textos autênticos. Com o papel de servir como fonte para extração de padrões sintáticos válidos e inválidos, e como base de testes para avaliação do revisor (quanto a omissões, intervenções devidas e indevidas), foi compilado um banco de textos autênticos (corpus) que conta hoje com cerca de 37 milhões de palavras, divididas em textos ditos corrigidos (livros, jornais, revistas), semicorrigidos (relatórios, dissertações, teses acadêmicas, contratos) e não corrigidos (redações de vestibular, monografias). Enquanto textos corrigidos, ou revisados, têm a função de indicar padrões sintáticos comuns e acusar intervenções indevidas do Revisor, os demais textos servem ao propósito de apontar erros gramaticais freqüentes e de validar o Revisor através de testes exaustivos sobre o corpus. O corpus do NILC tem sido utilizado por diversos grupos de pesquisa em PLN do Brasil e de Portugal (www.nilc.icmc.sc.usp.br/tools).

(c) Um etiquetador morfossintático (*tagger*). A etiquetação morfossintática é uma tarefa bastante conhecida em PLN. Ela consiste em se atribuir a categoria gramatical (e talvez outros atributos referentes a cada categoria como, por exemplo, gênero e número de substantivos) a cada palavra de um texto, escolhida dentre todas as listadas no léxico. Essa anotação é útil para várias manipulações subseqüentes do texto, pois fornece: a) uma forma de abstração das palavras do texto quando desejamos processar todas as palavras que pertençam a certa categoria de alguma forma específica, por exemplo, levantar todos os nomes próprios do texto; b) um grau de desambigüização que é crucial para os níveis subseqüentes, como o *parsing*. A existência do léxico e do corpus do projeto ReGra facilitou muito a construção de um *tagger* híbrido, que combina diferentes classificadores, atualmente em fase de testes (Aluísio & Aires, 2000; Aires et al., 2000). Essa ferramenta, além de ser utilizada para etiquetar o corpus do NILC, gerando assim um outro recurso de imenso valor, será útil para outras aplicações de processamento de língua portuguesa, especialmente aquelas que requeiram análise sintática.

(c) Um Thesaurus do Português. Um thesaurus eletrônico tem a função principal de sugerir a um usuário conjuntos de sinônimos e antônimos de palavras da língua. Para isso, é preciso também operacionalizar meios de integração entre o thesaurus e os mais variados aplicativos como processadores de texto, planilhas eletrônicas, programas de correio eletrônico e navegadores da Internet. Com essa funcionalidade, disponibiliza-se para o usuário uma ferramenta de auxílio à escrita que lhe permite encontrar com eficiência a palavra que procura enquanto escreve seu texto. Através de um trabalho exaustivo e altamente qualificado, o NILC está envolvido na construção de um thesaurus que deve conter, em breve, cerca de 30 mil palavras do português semanticamente relacionadas entre si (Dias-da-Silva et al., 2000).

Outros recursos e ferramentas construídos que, por limitação de espaço, deixamos de descrever aqui, incluem: a minigramática *online* que acompanha o revisor e pode ser consultada pelo usuário quando notificado de um erro pelo revisor, um sistema de compactação de léxicos baseado em autômatos finitos determinísticos (Jesus & Nunes, 2000), uma base lexical para consulta, construída à semelhança da Wordnet para o inglês, entre outros.

Além disso, benefícios podem ser sentidos em outros projetos. A experiência adquirida pelo NILC no projeto ReGra deu origem à sua participação no projeto *Universal Networking Language* (UNL), coordenado pelo Instituto de Estudos Avançados da Universidade das Nações Unidas (UNU), sediado em Tóquio. Sensível à enorme barreira da língua para a comunicação entre os povos, a UNU resolveu patrocinar um projeto de longa duração - 10 anos - para o desenvolvimento de ferramentas de software para vencer essa barreira. Nesse projeto, ao invés de tradução de uma língua para outra, faz-se a codificação do conteúdo de um texto em uma dada língua natural em uma língua artificial, a UNL, criada especificamente para textos escritos no ambiente da Internet. O texto já codificado em UNL pode então ser decodificado para a língua destino. O projeto UNL é de âmbito mundial. A partir de 1997, estão sendo desenvolvidos codificadores e decodificadores para cerca de 15 línguas, incluindo chinês, russo, alemão, francês, italiano, japonês, inglês, hindi, espanhol e português. No futuro espera-se atingir praticamente todas as línguas oficiais das Nações Unidas. As ferramentas para o português estão sendo desenvolvidas no NILC e a coordenação do projeto UNL-Brasil é do Prof. Tadao Takahashi.

Atualmente vários trabalhos acadêmicos e ferramentas e aplicativos de PLN estão em desenvolvimento no NILC[4], como conseqüência direta do envolvimento da equipe em projetos de cooperação. A natureza interdisciplinar da área faz também com que a formação de recursos humanos altamente especializados seja talvez a maior contribuição oriunda desses projetos.

5. Comentários Finais

Pesquisar com o propósito específico de atingir uma inovação tecnológica ainda não faz parte da tradição da universidade brasileira, talvez porque projetos de P&D sejam, em geral, menos pródigos do que os puramente acadêmicos para a obtenção de contribuições científicas que levem a publicações, mormente em periódicos internacionais, como hoje é exigido em muitas áreas do conhecimento. Além disso, grande é o tempo normalmente requerido para se concluir pela viabilidade tecnológica de uma inovação, período no qual não se deve esperar

[4] Para mais detalhes, visite http://www.nilc.icmsc.sc.usp.br/projects.html.

contribuições científicas originais, já esgotadas na etapa de concepção da mesma. Um projeto de parceria de P&D deve ser concebido, portanto, com perspectivas de longo prazo, em que no decorrer do projeto seja possível acumular conhecimento e instrumentos que possam levar a saltos qualitativos em uma determinada área. Este é o caso de áreas altamente multidisciplinares em que a focalização dos esforços em torno de uma inovação, ou produto, pode ser a alternativa mais viável para atingir objetivos que requerem volumosos recursos humanos em tarefas aparentemente rotineiras, mas que exigem formação específica. Recursos financeiros em geral não estão disponíveis para a realização dessas pesquisas, pois o sistema acadêmico e de pós-graduação privilegia, como não poderia deixar de ser, o trabalho altamente criativo e original. Em processamento de linguagem natural, são exemplos a compilação de grandes bases lexicais, corpora e o desenvolvimento de analisadores sintáticos. Portanto, num primeiro momento utiliza-se dos recursos advindos do desenvolvimento tecnológico para uma aplicação específica, mas os resultados podem ser posteriormente essenciais para a realização de pesquisas, que podem até estar completamente desvinculadas de aplicações tecnológicas.

O Projeto de parceria USP/Itautec-Philco relatado aqui, a nosso ver, atendeu à expectativa do sinergismo entre pesquisa básica e aplicada, na medida em que propiciou a formação de um Núcleo multidisciplinar (NILC). A empreitada bem-sucedida serve de demonstração que parcerias deste tipo podem beneficiar a Universidade, ao gerar pesquisa e formação de pessoal, e a sociedade como um todo, através da transferência efetiva de tecnologia para o setor produtivo.

Agradecimentos

Os autores agradecem a colaboração inestimável de todas as pessoas (pesquisadores, desenvolvedores, lingüistas, informatas, consultores e coordenadores) que participaram, ao longo desses 7 anos, do projeto ReGra.

Referências

Aluísio, S.M.; Aires, R.V. (2000) Etiquetação de um Corpus e Construção de um Etiquetador de português. Relatórios Técnicos do ICMC. NILC-TR-00-2. Março 2000, 18p.

Aires, R.V.; Aluísio, S.M.; Nunes, M.G.V. (2000). Em busca de um SuperTagger para o Português do Brasil. Submetido ao V Worskhop de Processamento Computacional do Português Escrito e Falado. Atibaia, Novembro 2000.

Dias-da-Silva, B.C.; Oliveira, M.F.; Moraes, H.R.; Hasegawa, R.; Amorim, D.; Paschoalino, C.; Nascimento, A.C. (2000) A Construção de um Thesaurus Eletrônico para o Português do Brasil. Submetido ao V Worskhop de Processamento Computacional do Português Escrito e Falado. Atibaia, Novembro 2000.

DTS Software Ltda. 1995. Manual do Revisor DTS Windows. Rio de Janeiro.

Flesch, R. (1948) A new readability yardstick, J. Appl. Psychology, 32, 221-233.

Gramática Eletrônica, v.1.0. 1997. Lexicon Informática, Rio de Janeiro.

Kowaltowski, T.; Lucchesi, C.L. (1993) Applications of finite automata representing large vocabularies. *Software-Practice and Experience*, 23(1), 15-30.

Martins, R.T.; Hasegawa, R.; Nunes, M.G.V.; Montilha, G.; Oliveira Jr., O.N. (1998) Linguistic issues in the development of ReGra: a Grammar Checker for Brazilian Portuguese. *Natural Language Engineering.* Volume 4 (Part 4 December 1998): p287-307; Cambridge University Press.

Martins, T.B.F.; Ghiraldelo, C.M.; Nunes, M.G.V.; Oliveira Jr., O.N. (1996) *Readability Formulas Applied to Textbooks in Brazilian Portuguese*. Notas do ICMSC-USP, Série Computação, nro. 28, 11p.

Montilha, G.; Nunes, M.G.V. (2000) Avaliação Comparativa de Revisores Gramaticais. Relatório Técnico NILC-TR-00-6.

Nunes, M.G.V.; Ghiraldelo, C.M.; Montilha, G.; Turine, M.A.S.; De Oliveira, M.C.F.; Hasegawa, R.; Martins, R.T.; Oliveira Jr., O.N.(1996a) Desenvolvimento de um sistema de revisão gramatical automática para o português do Brasil. In: Simpósio Brasileiro de Inteligência Artificial, 13. Encontro para o Processamento Computacional do Português Escrito e Falado, 2., Curitiba, Outubro, 1996.

Nunes, M.G.V., Ghiraldelo, C.M.; Hasegawa, R.; Kawamoto, S.; De Oliveira, M.C.F.; Oliveira Jr., O.N.; Schiabel, H.; Turine, M.A.S.; Riolfi, C.R.; Sikanski, N.S.; Martins, T.B.; Nascimento, L.H.R.(1996b) Style and grammar checkers for the Brazilian Portuguese, VIII Conference on Writing and Computers, Londres, Inglaterra, Setembro, 1995. Disponível como Notas do ICMSC-USP, Série Computação, 25.

Nunes, M.G.V. et alii. (1996c) A Construção de um Léxico para o Português do Brasil: Lições Aprendidas e Perspectivas. Anais do II Encontro para o Processamento Computacional do Português Escrito e Falado. CEFET-PR, Curitiba, pp. 61-70.

Oliveira Jr., O.N.; Nunes, M.G.V.; De Oliveira, M.C.F.; Ghiraldelo, C.M.; Schiabel, H.; Martins, T.B.; Sikansi, N.S.; Riolfi, C.R.; Turine, M.A.S.; Hasegawa, R; Kawamoto, S.; Nascimento, L.R.R. (1995) Desenvolvimento de um corretor gramatical para o português do Brasil, Anais do II Congresso de Informática e Telecomunicações do Mato Grosso, Cuiabá.

Woods, W.A.(1970) Transition networks grammars for natural language analysis, CACM, Special Issue on Computational Linguistics, 13 (2) 591-606.