



## ASPECTOS DA CONSTRUÇÃO DE UM REVISOR GRAMATICAL AUTOMÁTICO PARA O PORTUGUÊS\*

(Some issues on the development of an automatic grammar checker for Brazilian Portuguese)

Lucia Helena Machado RINO (DC-CCET/UFSCar & NILC – São Carlos)

Ariani DI FELIPPO (NILC – São Carlos)

Gisele M. PINHEIRO (NILC – São Carlos)

Ronaldo T. MARTINS (NILC – São Carlos)

Vanessa M. FILLIÉ (NILC – São Carlos)

Ricardo HASEGAWA (NILC – São Carlos)

Maria das Graças Volpe NUNES (DC-ICMC/USP-São Carlos & NILC – São Carlos)

**ABSTRACT:** *We describe below the “ReGra” grammar checker for Brazilian Portuguese, focusing upon its electronic lexicon and grammar. We also pinpoint some of the ReGra’s problems concerning representational and operational complexity, reinforcing that, in spite of those, ReGra is currently the best grammar checker available in Brazil.*

**KEYWORDS:** Grammar checker; Natural Language Processing; Automatic resources for Brazilian Portuguese

### 1. Introdução

Embora a construção de revisores gramaticais automáticos seja, atualmente, uma área bem consolidada na Linguística Computacional, ela é uma tarefa complexa, dependente de restrições severas de robustez (para contemplar tarefas em tempo real) e abrangência (para servir a qualquer tipo de requisição). As principais razões da complexidade de P&D de revisores gramaticais provêm: a) da necessidade de se representar eletronicamente grandes repositórios de informações lingüísticas (léxicas, sintáticas, semânticas e/ou pragmáticas) e b) da incorporação de processos computacionais que dêem conta, sistemática e eficientemente, das principais características do problema em foco. Considerando-se, em particular, o contexto do processamento automático do português, ela se reveste de importância maior, devido à escassez de ferramentas computacionais. Apresentamos aqui uma ferramenta automática especial, o REvisor GRAMatical (ReGra) para o português do Brasil, desenvolvida por pesquisadores da USP-São Carlos e ainda em constante aprimoramento, atualmente com a colaboração de pesquisadores da UFSCar e da UNESP-Araraquara<sup>1</sup>.

A fim de contextualizar este trabalho, algumas premissas fundamentais do ReGra devem ser ressaltadas: a) seus usuários-alvo são indivíduos com suposta proficiência no português em nível secundário ou superior; b) seu objeto de análise são textos prosaicos com predomínio da função referencial da linguagem; e c) seu objetivo é a identificação de desvios à disciplina gramatical da língua portuguesa, ainda que esses desvios possam vir a representar hipóteses significativas do usuário em relação à linguagem e possam constituir, na variedade lingüística do usuário, estruturas admissíveis. A seguir, descrevemos a arquitetura do ReGra (Seção 2), enfatizando sua organização lexical (Seção 3) e seu processamento sintático (Seção 4), seguidos do relato de testes de desempenho (Seção 5) e considerações finais (Seção 6).

### 2. A arquitetura do ReGra

\* Este trabalho conta com o apoio da FAPESP (Proc. Nro. 00/10688-0) e FINEP-Itautec/Philco (Processo FINEP/PADCT RC: 3.1.3-0012/98).

<sup>1</sup> Os quais formam, atualmente, o NILC – Núcleo Interinstitucional de Linguística Computacional de São Carlos.

O ReGra conta hoje com os seguintes recursos operacionais (Figura 1)<sup>2</sup>: um léxico, com mais de um milhão e meio de formas do português; 2) um conjunto de regras de desambiguação lexical; 3) um verificador mecânico, constituído de um conjunto de regras para identificação de problemas de digitação (excesso de espaços em branco, balanceamento de delimitadores, etc.) e de seu correspondente aconselhador mecânico; 4) um verificador ortográfico<sup>3</sup>, constituído dos correspondentes conjunto de regras e aconselhador, o qual sugere as formas lexicalizadas próximas daquelas não constantes do léxico; 5) um verificador gramatical, constituído de (Martins et al., 1998): a) uma gramática com mais de 600 regras de reescrita que contemplam a maior parte das possibilidades sintáticas de períodos simples do português; b) um *parser*, ou analisador sintático automático, capaz de derivar representações arbóreas (hierárquicas) a partir de listas de palavras do português, segundo as estruturas previstas na gramática; c) um conjunto de regras de verificação gramatical, subdivididas em distribucionais (com a consideração dos contextos mínimos, à esquerda e à direita) e estruturais (dependentes da identificação das funções sintáticas); d) um aconselhador gramatical, capaz de sugerir correspondentes corretos para formas ou estruturas consideradas inadequadas e e) um conjunto de mensagens de revisão, associadas aos padrões derivados da verificação gramatical; 6) uma minigramática, indexada como hipertexto às mensagens de revisão, para consulta.

Além desses, foram incorporados ao ReGra outros recursos lingüísticos, para maior apoio aos usuários, como um módulo de consulta à gramática, acessível diretamente no ambiente WORD (menu “Ferramentas”). Um *thesaurus*, em fase de conclusão, deverá ser analogamente disponibilizado. Com esses recursos, o sistema passou a ter um grande potencial para funcionar também como uma ferramenta de ensino da língua (cf. Nunes et al., 1999), servindo também a várias outras aplicações de processamento de línguas naturais (PLN), voltado especialmente ao português. Ainda, o próprio léxico e um *corpus* de textos eletrônicos constituem, hoje, os maiores recursos disponíveis eletronicamente (Nunes et al., 1996). A construção de um etiquetador de textos em português (Aires, 2000), para inúmeras investigações lingüísticas ou lingüístico-computacionais, também foi motivada pelos recursos do ReGra.

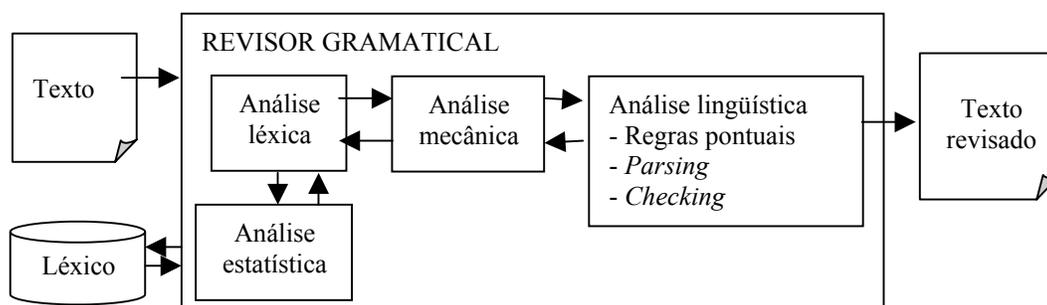


Figura 1: Arquitetura do ReGra

### 3. O léxico do ReGra[MSOffice1]

O léxico do ReGra é um repositório de formas lexicais válidas para o português contemporâneo, acrescidas de informações morfosintáticas e de formas canônicas. Assim, ele se diferencia dos léxicos descritos nos últimos tempos, na literatura especializada. Por um lado, ele se distancia da noção de “dicionário” porque se apresenta como uma lista limitada, embora bastante abrangente, de palavras do português. Por outro lado, ele não consiste de uma base de dados, pois não admite mecanismo dinâmico algum de acesso a suas informações. Ainda, ele não simula o chamado “léxico mental”. A especificidade da tarefa à qual ele se destina limita suas funções, caracterizando-o objetivamente pelas restrições de “computabilidade lingüística”: ele se apresenta como um recurso à revisão gramatical, servindo também de apoio ao módulo ortográfico, para validar ou rejeitar as construções vocabulares dos usuários.

Em relação à sua estruturação, a ênfase está no aspecto organizacional para o acesso automático, e não no aspecto morfológico, implicando quívamento exaustivo das formas do português, sem considerar, p.ex., o aproveitamento de sufixos flexionais e derivacionais da língua. Essa organização resultou também no armazenamento das lexias simples, i.e., das formas de palavras de radical único (p.ex., *pé*) ou das formas hifenizadas (p.ex., *pé-de-moleque*), decisão determinada por questões de compactação do léxico, operação que, ao menos por enquanto, rejeita o tratamento de lexias complexas, tais

<sup>2</sup> Processament realizado de forma cooperativa entre a análise léxica, estatística e mecânica, a aplicação de regras pontuais e o *parsing* e a verificação gramatical (*checking*).

<sup>3</sup> As ferramentas de compactação do léxico e as rotinas de aconselhamento ortográfico foram propostas e implementadas por T. Kowaltowski, C. Lucchesi e J. Stolfi, do Instituto de Computação da Unicamp.

como as locuções (p.ex., ‘na medida em que’, ‘à distância’) ou palavras compostas (p.ex., ‘boca de uma’, ‘estado civil’). Esse tratamento, no momento, é feito como pré-processamento lexical associado às regras de refinamento do *parser*, em um intercâmbio entre a revisão ortográfica e a gramatical, em detrimento da independência do processamento lexical, conforme defendem as propostas teóricas. A definição lexical exaustiva tem, ainda, a desvantagem de multiplicar as formas homônimas, resultando em extrema redundância formal do sistema, que se tornará mais séria se informações sobre o significado forem incluídas no léxico, para, p.ex., resolver problemas complexos como a ambigüidade lexical verificada atualmente, decorrente, em grande parte, da própria opção de organização e armazenamento lexical. Por exemplo, informações gramaticais distintas para uma mesma forma implicam entradas lexicais diversas (*canto*<sub>1</sub> [verbo], *canto*<sub>2</sub> [substantivo]).

A constituição do léxico deu-se de maneira quase que exclusivamente automática, conforme descrevem (Hernandez, 1997; Zavaglia, 1997). É preciso salientar, entretanto, que seu refinamento é preocupação constante do NILC. Testes sistemáticos revelaram, desde o início, que muitos problemas de registros lexicais foram gerados pela inserção automática das palavras, cuja resolução exige um grande esforço humano, altamente especializado. Entretanto, novos desafios, como a representação de informações morfossintáticas e a possível expansão dessa estrutura informacional têm se apresentado, principalmente devido à necessidade de tratar os fenômenos complexos da língua, tais como a homonímia, os quais ainda não são devidamente contemplados no ReGra. Estudos nesse sentido (Rino et al., 2001; Pinheiro et al., 2001) têm apontado, p.ex., a necessidade de se representar valores semânticos ou a estrutura argumental dos verbos. Nesse caso, seria necessário definir um processo diferenciador dos traços gramaticais particulares, recuperáveis do léxico, para a desambigüização lexical. No entanto, essa opção esbarra na opção representacional original do léxico, de: a) representar-se as formas lexicais exaustivamente e b) utilizar-se a frequência de uso como parâmetro de organização lexical, segundo critérios intuitivos do especialista lingüista<sup>4</sup>. Notadamente, tal limitação não permite contemplar a desambigüização categorial. Embora o ReGra conte com um sistema de regras específicas para o tratamento da preferência categorial para alguns casos da língua, não é possível garantir a consistência lingüística necessária para cobrir a maioria das ocorrências possíveis, como relatamos no Projeto FAPESP/IT (Proc. Nro. 00/10688-0).

A ambigüidade lexical, da qual a ambigüidade categorial é um tipo, pode ser considerada, sem sombra de dúvida, um grande desafio para o aprimoramento do ReGra, pois interfere em sua eficiência e robustez. As alterações de representação lexical citadas implicam profunda alteração do modelo representacional adotado, principalmente se considerarmos a necessidade de associar às palavras sua caracterização semântica, por si só uma tarefa já altamente complexa na lingüística computacional e, especialmente, no contexto de revisão gramatical irrestrita, como é o caso do ReGra. Por sua vez, comprovando-se a possibilidade de alteração do léxico, há implicações profundas no próprio modelo de processamento. Em especial, a revisão do analisador sintático se faz obrigatória, questão discutida a seguir.

#### 4. A gramática do ReGra

São três as etapas em que se subdivide o processo de revisão gramatical: a) o pré-processamento da cadeia de entrada, com a etiquetagem dos itens lexicais e a conseqüente desambigüização da informação lexical; b) o processamento gramatical propriamente dito, com a verificação da correção das escolhas gramaticais feitas pelo usuário; e c) o pós-processamento sintático, com a aplicação de procedimentos de aconselhamento gramatical para readequação da cadeia de entrada aos padrões perseguidos pela ferramenta.

A desambigüização lexical (caso a) torna-se pertinente porquanto, como já foi referido, a homografia em língua portuguesa não é fenômeno raro, induzindo freqüentemente a erros de diagnóstico uma ferramenta que, diferentemente do ser humano, e na ausência de estratégias de representação da estrutura semântico-pragmática dos enunciados lingüísticos, não pode contar senão com a forma, com a frequência de ocorrência e com a distribuição das palavras nas sentenças para proceder ao processamento. A frequência de ocorrência aparece, no ReGra, como principal mecanismo de desambigüização lexical, apesar de, em muitos casos, a ferramenta pouco aproveitar a prioridade de classificação, por não serem exatamente nítidas as diferenças entre as probabilidades de ocorrência de itens lexicais muito utilizados em mais de uma acepção gramatical. Neste caso, é necessária a consideração de critérios distribucionais de diferenciação entre o uso das palavras, com a delimitação dos contextos de ocorrência dos itens lexicais, principalmente na situação da vizinhança imediata que se estabelece no intervalo de um a cinco *tokens*. A adoção dessas duas estratégias torna o processo de etiquetagem lexical eminentemente probabilístico, com uma evidente zona residual de ambigüidade que escapa às atuais possibilidades de decisão do revisor. Fossem conservadas as taxas de imprecisão observáveis nos etiquetadores disponíveis para o português (da ordem de 97%, cf. Aires, 2000), o problema não seria talvez tão relevante. No entanto, é forçoso considerar que os

---

<sup>4</sup> Não ignoramos o trabalho de Biderman (1992). No entanto, este dicionário da língua geral contemporânea contém um número baixo de verbetes para nossos interesses.

etiquetadores concorrentes partem de um pressuposto estranho e inverso ao do ReGra: supõem serem corretas as sentenças de entrada. Por este motivo, a margem de indefinição do revisor é consideravelmente mais expressiva do que a verificada em outros contextos do PLN, sendo muitas das decisões lexicais adiadas para o processamento sintático, nem sempre hábil o suficiente para capturar a correta categorização gramatical das palavras. Para evitar a explosão combinatória devida à indefinição lexical, restringimos o escopo de atuação do revisor à oração simples, isolando para processamento, sempre que possível, as coordenações e subordinações. Com a inevitável perda das relações de dependência entre as várias orações de um período composto, o espectro de problemas a serem resolvidos também foi restrito. Entretanto, na análise do desempenho final, essa contingência revelou-se preferível à proliferação das intervenções indevidas, com riscos à credibilidade do ReGra.

A etapa de processamento gramatical, propriamente dita (caso b), diretamente relacionada à verificação da adequação gramatical da sentença de entrada, está subdividida em três subcomponentes: a verificação de boa-formação sintática da sentença, a etiquetagem sintática de seus itens lexicais durante o *parsing* e a verificação da consistência das relações de dependência entre os mesmos, com correspondente aconselhamento ao usuário. O formalismo gramatical, definido para reconhecer sentenças bem-formadas, segue uma gramática livre de contexto, cujo conjunto de regras de produção torna-se diretamente operacional com as redes sintáticas (ATNs), responsáveis por derivar a sentença de entrada, processo realizado de forma *top-down*, da esquerda para a direita, a partir do símbolo inicial da gramática, com o auxílio de uma janela de geração de dimensão variada, para evitar ciladas sintáticas frequentes, particularmente devidas aos problemas de desambiguação lexical. O *parser* apresenta como resultado uma única possibilidade de estruturação sintática: a mais provável, segundo escolhas gramaticais realizadas também com base na prioridade de frequência (de suas regras), definida a partir da análise de fragmento expressivo do *corpus* do NILC.

Há dois subconjuntos de regras de revisão: as pontuais, que acompanham padrões fixos de distribuição sintática, pouco ou nada dependentes do *parsing*, e as genéricas, que operam diretamente sobre a saída do *parser*. As primeiras, numericamente expressivas, são extremamente eficientes, mas possuem alcance limitado, pois se restringem a problemas de adequação lexical e ao uso de padrões lingüísticos já registrados nas gramáticas normativas do português (p.ex., uso de crase diante de palavras masculinas). As segundas, muito menos numerosas e muito mais abrangentes que as anteriores, investigam as relações de concordância e regência entre os itens lexicais etiquetados sintaticamente. Como a saída do *parser* é falível, devido a indeterminações lexicais ou a outros acidentes de processamento, o conjunto dessas regras induz frequentemente a ferramenta a erro.

O pós-processamento gramatical (caso c) consiste na identificação precisa do desvio da norma, para o aconselhamento correspondente. Constatada, por meio de qualquer um dos tipos de regras de revisão, a má formação de alguma estrutura utilizada pelo usuário, dispara-se imediatamente a medida de revisão correspondente, que indicará a existência do problema e proporá nova estrutura, a ser aceita ou rejeitada pelo usuário, para que se possa persistir na verificação da sentença. Diferentemente do aconselhamento ortográfico, que é feito pela recombinação sistemática dos caracteres da palavra não-encontrada no dicionário, o aconselhamento gramatical depende profundamente do processamento nos níveis anteriores, razão pela qual o aconselhador retroalimenta a revisão gramatical.

## 5. Testes comparativos: objetivos e resultados

O NILC realizou dois tipos de testes comparativos de desempenho do ReGra (Nunes e Oliveira, 2000): a) sob as perspectivas do usuário, entre versões distintas do ReGra, o revisor DTS e a Gramática Eletrônica (GE) (Montilha e Nunes, 2000) e b) sob as perspectivas de processamento (Nunes et al., 2000). Os testes visando o uso foram baseados num conjunto de textos autênticos recuperados da base textual do NILC, que é tripartida em *corpus corrigido*, *semicorrigido* e *não-corrigido*, para manter o perfil textual da escrita. Para cada *corpus*, compilaram-se 15 textos, resultando num total de 45, cujos gêneros diversos – jornalístico, acadêmico e redações livres – determinaram a abrangência dos testes. A diversidade desse *corpus* foi significativa para a observação da qualidade das intervenções em cada uma das modalidades de análise (mecânica, gramatical e ortográfica) e da qualidade do léxico, as quais foram classificadas como intervenções indevidas, intervenções devidas e omissões. Os testes visando o processamento foram baseados em um *corpus* formado por 9.808 sentenças, das quais 7.627 foram consideradas corretas e 2.181, erradas<sup>5</sup>. Para efeito de análise, a atuação do ReGra foi dividida em quatro categorias de intervenção: verdadeiros negativos (a ferramenta não deveria intervir e não intervém), verdadeiros positivos (deveria intervir e intervém), falsos negativos (deveria intervir e não intervém) e falsos positivos (não deveria intervir mas intervém). Esses testes determinaram que: a) a revisão gramatical e ortográfica automática é não apenas possível, mas viável e útil, já que a ferramenta identifica 60,52% dos problemas praticados pelo usuário (na situação de teste) e se omite em apenas 17,84% dos casos; e b) o preço pago para a identificação de mais da metade dos inadequações praticadas pelo usuário ainda é alto (o número de falsos positivos ainda é maior do que o número de verdadeiros positivos) e que as estratégias utilizadas para

---

<sup>5</sup> A proporção de uma sentença errada para quatro corretas corrobora o que se observa normalmente em textos de usuários inexperientes no uso do registro da língua escrita.

reduzir o comportamento inadequado da ferramenta não têm produzido os efeitos esperados, já que as diferenças entre as várias versões são bem pouco significativas, se considerado o conjunto de sentenças que compõem o corpus de teste. Estabelece-se, portanto, como diretriz futura, mais do que a ampliação do escopo de atuação da ferramenta (para o incremento de número de erros identificados), uma revisão mais radical nas regras já existentes, para que se reduza o número, ainda excessivo, de falsos positivos. Numa avaliação global, testes de ambos os tipos demonstraram que: a) qualquer versão do ReGra apresenta um desempenho geral superior aos dos seus concorrentes; b) o DTS é o que mais intervém desnecessariamente; c) a GE apresenta o pior desempenho na correção gramatical; d) o DTS e a GE apresentam um léxico reduzido, em comparação com o léxico do ReGra; e) dentre as versões do ReGra, Beta3 é a melhor quanto às omissões.

## 6. Considerações finais

Muito embora o desempenho do ReGra (textos revisados à taxa de, aproximadamente, 200 palavras por segundo) comprove a robustez de seu analisador sintático e seus recursos lingüísticos, disponíveis eletronicamente, sejam muito importantes para o processamento do português de um modo geral, seu aprimoramento consiste um grande desafio, do ponto de vista de modelagem, de manipulação automática de informações lingüísticas e de auxílio a tarefas que o usuário realiza no ambiente computacional. Entendemos que os recursos hoje disponíveis e as estratégias atualmente utilizadas no ReGra atingiram, nos dois últimos anos, o ponto de exaustão. Embora seu desempenho possa ser ainda melhorado, qualquer investimento realizado no modelo atual revela um custo bastante superior ao benefício obtido. Basta notar que, em suas últimas cinco versões, o ReGra avançou apenas 1,04% na redução de intervenções indevidas e apenas 3,27% na ampliação das intervenções devidas. A produção de taxas significativas de evolução do desempenho do revisor depende, pois, não apenas da ampliação do material existente, mas principalmente da reorganização e da revisão das estratégias empregadas. Além disso, o público-alvo do revisor, constituído por usuários do português brasileiro com nível de escolaridade médio, tem-se revelado insatisfeito com a restrição do escopo de intervenção do ReGra aos problemas ortográfico-sintáticos, exigindo a consideração de problemas de natureza semântico-pragmática, quase sempre relacionados a níveis maiores do que a oração. É necessário, p.ex., que tratemos problemas de estilo, muito mais vinculados ao propósito da eficácia comunicativa do que ao da adequação à norma gramatical. Do ponto de vista mercadológico, são necessárias contínuas atualizações, além da ampliação do próprio escopo de atuação, para que o ReGra continue sendo, no cenário dos aplicativos para o português brasileiro, a mais eficiente, a mais completa e a mais satisfatória ferramenta de revisão gramatical.

**RESUMO:** *Descrevemos o revisor gramatical eletrônico ReGra, para o português brasileiro, contemplando seu léxico e sua gramática. Apontamos, ainda, alguns problemas que remetem à complexidade de representação e manipulação lingüístico-computacional, ressaltando que, apesar deles, o ReGra é o melhor revisor, hoje, no mercado brasileiro.*

**PALAVRAS-CHAVE:** Revisor gramatical automático; Processamento automático de línguas naturais; Recursos automáticos para o português do Brasil.

## REFERÊNCIAS BIBLIOGRÁFICAS

- AIRES, R. *Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil*. Tese de Mestrado. ICMC-USP, São Carlos, Outubro, 2000.
- BIDERMAN, M.T.C. *Dicionário contemporâneo da língua portuguesa*. Vozes, Petrópolis, 1992.
- HERNANDEZ, J.M.C. O tratamento computacional dos erros mais frequentes manifestados em textos produzidos por alunos universitários e de segundo grau. *Anais do XXVI Seminário do Grupo de Estudos Lingüísticos (GEL)*, pp. 169-174. Campinas-SP, 1997.
- MARTINS, R.T.; HASEGAWA, R.; NUNES, M.G.V.; MONTILHA, G.; OLIVEIRA Jr., O.N. Linguistic issues in the development of ReGra: a grammar checker for Brazilian Portuguese. *Natural Language Engineering*, Volume 4 (Part 4), pp. 287-307. Cambridge University Press, December, 1998.\*
- MONTILHA, G. e NUNES, M.G.V. *Testes comparativos entre revisores gramaticais de português*. Rel. Técnico NILC-TR-00-9, 32p. Junho 2000.\*
- NUNES, M.G.V. et al. 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*, pp. 61-70. CEFET-PR, Curitiba. Outubro, 1996.\*
- NUNES, M.G.V.; KUHN, D.; MARCHI, A.R.; NASCIMENTO, A.C.; ALUÍSIO, S.; OLIVEIRA Jr., O.N. Novos Rumos para o ReGra: extensão do revisor gramatical do português do Brasil para uma ferramenta de auxílio à escrita. *Anais do IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99)*, pp. 167-182. Évora, Portugal, 1999.\*

- NUNES, M.G.V.; OLIVEIRA Jr., O.N. O processo de desenvolvimento do Revisor Gramatical ReGra. XXVII SEMISH. *Anais do XX Congresso Nacional da Sociedade Brasileira de Computação*, Volume 1, p.6 (resumo). Artigo Completo na Versão em CD-Rom. PUC-PR, Curitiba, Julho 2000.\*
- NUNES, M.G.V.; MARTINS, R.T.; HASEGAWA, R.; HABER, R.R.; MONTILHA, G. *Relatório dos Testes Comparativos entre Diferentes Versões do Revisor Gramatical ReGra*. Rel. Técnico NILC-TR-00-8, 8p. Junho 2000.\*
- PINHEIRO, G.M.; MARTINS, R.T.; RINO, L.H.M.; DI FELIPPO, A.; FILLIÉ, V.M.; HASEGAWA, R. Projeto TraSem: A investigação empírica sobre o problema da ambigüidade categorial. Tech. Rep. NILC-TR-01-2. São Carlos, Abril, 2001.\*
- RINO, L.H.M.; MARTINS, R.T.; MARCHI, A.R.; KUHN, D.C.S.; PINHEIRO, G.M; PARDO, T.A.S.; DI FELIPPO, A. *Projeto TraSem: A investigação teórica sobre o problema da ambigüidade categorial*. Tech. Rep. NILC-TR-01-1. São Carlos, Abril, 2001.\*
- ZAVAGLIA, C. A Construção de um Dicionário de Máquina como Suporte para Revisores Ortográfico e Gramatical. *Anais do XXVI Seminário do Grupo de Estudos Lingüísticos (GEL)*, pp. 157-162. Campinas-SP, 1997.

\*Trabalhos disponíveis no site <http://nilc.icmc.sc.usp.br/Publications.html>