Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

# Projeto TraSem: A investigação teórica sobre o problema da ambigüidade categorial

**Lucia Helena Machado Rino (Responsável)**
**Ronaldo Teixeira Martins**
**Ana Raquel Marchi**
**Denise Campos e Silva Kuhn**
**Gisele Montilha Pinheiro**
**Thiago Alexandre Salgueiro Pardo**
**Ariani Di Felippo**
**Maria das Graças Volpe Nunes**

**NILC-TR-01-1**

Abril, 2001

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

## ÍNDICE



## FIGURAS



## TABELAS

# Resumo

Este relatório descreve a investigação teórica do problema proposto inicialmente no subprojeto "Especificação dos Traços Semânticos dos Itens Lexicais", cujo trabalho se iniciou em Março/99, estendendo-se até meados de 2000, com a implementação de um protótipo visando o aprimoramento do ReGra. A decisão pela prototipagem teve por objetivo, inicialmente, buscar caminhos para a avaliação da adoção de uma estratégia de aprimoramento baseada nas teorias investigadas. Após alguns meses de desenvolvimento do protótipo e sua conseqüente avaliação, decidiu-se adotar outra metodologia de investigação, agora empírica, com base na exploração do corpus do NILC. Além de descrever a abordagem teórica, apresentamos aqui as razões para a mudança de abordagem.

## 1. Introdução

Este relatório apresenta a abordagem fundamental adotada no projeto "Especificação dos Traços Semânticos dos Itens Lexicais" (Processo FINEP/PADCT RC: 3.1.3-0012/98) – Projeto TraSem – durante seu primeiro ano de desenvolvimento.

De acordo com a proposta do Projeto TraSem, de aprimorar o ReGra de modo a torná-lo mais abrangente e eficiente durante a revisão gramatical, o foco do trabalho está no problema da ambigüidade categorial, para o qual não há resolução adequada pelo ReGra, devido à falta de correspondência biunívoca entre o significante (a forma) e o significado (o conteúdo) das palavras da língua. Embora isso não constitua problema para o falante humano, revela-se impedimento dos mais sérios para o processamento automático das línguas naturais (PLN) e, conseqüentemente, compromete a qualidade do revisor.

Ilustraremos, neste relatório, vários casos em que a ambigüidade lexical se estabelece como problema de muitas faces e inúmeras repercussões, cuja resolução exige uma representação mais rica e estruturada do conteúdo lexical, resultando em duas questões fundamentais: a identificação correta das classes gramaticais e a identificação correta da acepção de componentes polissêmicos. Para resolvê-las, investigamos, primeiramente, a utilização de um módulo semântico junto ao módulo sintático (o *parser*) do ReGra, buscando integrá-los de tal modo que a troca de informações em diferentes níveis de processamento viesse aumentar a confiabilidade da análise sintática e, portanto, da própria revisão gramatical. A conseqüência natural dessa busca de integração implica, por sua vez, a reavaliação (ou redefinição) do léxico do ReGra, cuja reestruturação se baseia em um modelo léxico-semântico, implicando, por fim, a investigação teórica de casos de representação em todos os níveis lingüísticos relacionados à revisão: o léxico, o sintático e o semântico, conforme relataremos.

Exemplos dos problemas de inadequação, compilados a partir da execução do ReGra, são apresentados na Seção 2, juntamente com considerações sobre as possíveis causas e sugestões para solucioná-los, as quais foram elaboradas por lingüistas. Uma síntese dos problemas apontados nessa seção é apresentada na Seção 3, fazendo-se uso da análise de alguns casos estereotipados. Essa análise persegue dois objetivos: um, visando à representação do conteúdo lexical; outro, visando à eleição de um modelo para implementação no escopo do Projeto ReGra, de identificação das acepções dos componentes polissêmicos, conforme apontamos acima. O relato da busca dessas metas encontra-se, respectivamente, nas Seções 4 e 5. Finalmente, na Seção 6 discutimos a proposta de integração entre sintaxe e semântica para o aprimoramento do ReGra, apontando, a seguir, as perspectivas de continuidade do projeto (Seção 7).

## 2. Ilustrações de casos de inadequação do ReGra

Ilustramos, nesta seção, problemas de duas naturezas distintas: aqueles resultantes da ausência (ou incompleteza) de informações lingüísticas no ReGra e aqueles resultantes do processamento automático. O primeiro caso remete à sub-especificação lingüística de possíveis ocorrências no português; o segundo, além de poder ser também decorrente dessa sub-especificação, remete ao modo como o modelo computacional foi idealizado, o qual interfere nas decisões de revisão.

### 2.1. Problemas de ausência de informação lingüística

A partir de um exemplário gerado automaticamente, os problemas de ocorrência de falsos erros ou de omissões do ReGra[1] foram divididos em três classes, de acordo com as causas de ambigüidade ou com a necessidade de novas regras de correção: a ambigüidade lexical, a

[1] Refira-se a (Nunes, 1999), para detalhes sobre essa terminologia.

ambigüidade estrutural e a incompleteza do elenco de regras de revisão. São apresentados aqui exemplos assim classificados, para esclarecer a natureza dos problemas a serem discutidos. Segue, ainda, uma compilação detalhada de ocorrências de processamento, subdivididas em falsos erros e omissões.

### 2.1.1. Exemplos de ambigüidade lexical

*A espada feriu fundo.*

*'fundo'* é categorizado como adjetivo, em vez de advérbio (caso de falso erro).

*Ela saiu feito uma doida*.

*'feito'* é categorizado como adjetivo, em vez de conjunção (caso de falso erro).

### 2.1.2. Exemplos de ambigüidade estrutural

*A menina tinha desengonçados os braços e as pernas*.

O ReGra não reconhece a inversão frasal do adjetivo *'desengonçados'*, relativo a *'os braços e as pernas'*, tomando-o como parte integrante do particípio passado. Vale lembrar, no entanto, que ambas as formas (particípio passado e adjetivo) são possíveis no português e, portanto, o ReGra deveria, provavelmente, aceitar a forma apresentada como adequada (caso de falso erro).

### 2.1.3. Exemplos de conjunto incompleto de regras de correção, introduzeindo casos de omissão

*Elas têm alguma coisa de bonitas*. (em vez de "*Elas têm alguma coisa de bonito*.")

A ferramenta não é capaz, aqui, de reconhecer uma expressão fixa e considera a sentença correta. A inserção de uma nova regra para esse tipo de expressão fixa resolveria este problema.

*Se caso você chegar antes, abra a porta.* (em vez de "*Se acaso você chegar antes, abra a porta"*).

A questão apresentada aqui se relaciona a um problema de ortografia (*'caso'* e *'acaso'*), que seria resolvido por uma regra que recusasse a co-ocorrência de 'se' e 'caso', por expressar a redundância do caso condicional.

### 2.1.4. Falsos erros

Do conjunto completo de sentenças com diagnóstico problemático, 113 sentenças constituíram, inicialmente, o exemplário de falsos erros (ou falsos negativos), que foi aumentado posteriormente, buscando-se um número de sentenças que fosse mais significativo (em termos quantitativos, para cada problema em foco) e mais abrangente (em termos qualitativos, em relação à variedade de problemas detectados). Foram coletadas 3.563 novas sentenças do corpus não corrigido, assim como dos corpora corrigido e semi-corrigido do NILC. Dentre essas, 297 sentenças apresentaram algum tipo de intervenção inadequada do ReGra, de falso erro. A Tabela 1 mostra esse resultado, em função do tipo de corpus sob análise.

**Tabela 1: Distribuição sentencial com inadequações de revisão nos corpora do NILC**

| Corpus do NILC | Nro. total de sentenças | Nro. de intervenções inadequadas |
|---|---|---|
| Corrigido | 1.546 | 165 |
| Semicorrigido | 862 | 86 |
| Não corrigido | 1.155 | 46 |

Nessa compilação, não foram levadas em conta as intervenções devidas à inexistência das palavras correspondentes no léxico ou a algum tipo de incorreção ortográfica, que resulta também na incapacidade de identificação lexical. Não se considerou, ainda, o número de intervenções adequadas, ou acertos, para a revisão de sentenças dos corpora semi-corrigido e não corrigido.

Exemplos de falsos erros são apresentados abaixo, juntamente com sugestões de correção, as quais, apesar de terem se baseado em normas lingüísticas e/ou em diagnósticos gerados automaticamente, ainda precisam ser adequadamente refinadas, para a modelagem computacional.

*Flores até pouco tempo atrás trazia acento circunflexo.*

Problema: O ReGra sugere "Flor até pouco tempo atrás trazia acento circunflexo." ou "Flores até pouco tempo atrás traziam acento circunflexo."

Natureza do problema: Não reconhece o uso metalingüístico da palavra 'flores', que remete ao uso singular deste componente sentencial.

Sugestão de correção: Se considerarmos que ocorrências dessa natureza são bastante freqüentes, o ReGra deveria incorporar o nível metalingüístico de processamento. No entanto, qualquer sugestão dessa natureza é questionável, à medida que se faz necessário recorrer ao contexto extragramatical para interpretar a metalinguagem no uso do português. Isto implicaria: a) priorizar acepções de sentido, em relação a acepções "congeladas", que seriam correspondentes ao uso metalingüístico em questão; b) dotar o sistema de um nível de processamento semântico altamente refinado, para que ele conseguisse reconhecer contextos em que se emprega o uso metalingüístico das palavras. A opção (a) não resolveria o problema, pois possíveis interpretações recairiam nas sugestões que hoje o ReGra já oferece. A opção (b) implicaria associar um "verificador" extragramatical a cada possibilidade de uso metalingüístico, recaindo no problema de se determinar exaustivamente todos esses contextos. Assim, tanto em termos de representação do conhecimento, quanto em termos da eficiência do processamento, a adoção de um nível de processamento metalingüístico implica o aumento considerável da complexidade do revisor.

*Como é bom cerveja gelada no verão!*

Problema: O ReGra sugere "Como é boa cerveja gelada no verão!", exigindo a concordância nominal do predicativo 'bom' com 'cerveja'.

Natureza do problema: Não reconhece o verbo elíptico da sentença ("Como é bom [tomar/beber] cerveja gelada no verão.").

Sugestão de correção: Qualquer sugestão dessa natureza exige que se considere a resolução de elipses ou b) algum tipo de classificação semântica que permita recuperar expressões cuja flexão de gênero deva ocorrer somente se houver artigo precedendo o substantivo, após o adjetivo.

*É permitido permanência de veículos neste local.*

Problema: Sugere "É permitida permanência de veículos neste local.", exigindo a concordância nominal do predicativo 'permitido' com 'permanência'.

Natureza do problema: Embora similar ao caso anterior, esta sentença não apresenta uma elipse verbal, mas, sim, uma expressão invariável, que deveria ser tratada, p.ex., como uma colocação (uso de expressões fixas). Este é um caso particular, em que a flexão de gênero deveria ocorrer somente se houvesse um artigo precedendo o substantivo.

Sugestão de correção: Expressões invariáveis, ou fixas, deveriam ser consideradas lexias complexas e, assim, deveriam ser tratadas computacionalmente como 'colocações' (*collocations*) e, portanto, a sugestão é modificar o léxico do ReGra e o *parser*, para reconhecer expressões fixas dessa natureza. Isto implica a) o estudo das possíveis colocações relevantes, no contexto do uso irrestrito da língua e b) a alteração do modelo de léxico atual do ReGra, para admitir frases, adicionalmente a suas entradas simples.

*Ternos azul-marinho*.

Problema: O ReGra sugere "Terno azul-marinho" ou "Ternos azuis-marinhos", classificando 'ternos' como adjetivo (acepção de meigo, afetuoso ou brando, suave) e 'azul-marinho' como substantivo.

Natureza do problema: Segundo a gramática normativa, adjetivos compostos de nome de cor + substantivo são invariáveis (p.ex., farda verde-oliva/ fardas verde-oliva). Neste caso, por tomar a acepção de adjetivo para 'ternos', imputa a desinência plural ao nome que este adjunto modifica.

Sugestão de correção: refinar a resolução da ambigüidade lexical[2].

*A espada feriu fundo.*

Problema: O ReGra acusa um erro, mas não sugere construção gramatical alternativa.

Natureza do problema: Segundo as normas gramaticais implementadas no sistema, a análise sintática produz as seguintes acepções para os componentes sentenciais:

A ⫢ ART
espada ⫢ SUBST
feriu ⫢ VERBO[ferir] ⫢ Intrans TransDir
fundo ⫢ ADJ

Tais acepções, em particular, a categorização de 'fundo' como adjetivo, evidencia sua inadequação lexical atual: neste exemplo, ele ocupa a função de adjunto adverbial de modo – 'com profundidade' – e, portanto, a resolução automática deveria privilegiar sua categoria adverbial.

Sugestão de correção: Elaborar um estudo detalhado dos adjetivos nessas condições, para complementar (ou refinar) o léxico, e, ao mesmo tempo, averiguar o refinamento das regras gramaticais, assim como o refinamento dos mecanismos automáticos de escolha de uma regra, em particular. Neste caso, faz-se necessário analisar a ordem de aplicação das regras, assim como a freqüência de ocorrência das acepções adverbiais, em comparação com as acepções adjetivais. Em suma, é necessário buscar algum tipo de refinamento lexical, como no caso anterior, além de se averiguar a possibilidade de melhorar os mecanismos de escolha automática durante o *parsing*.

*O sonho de todo astronauta é voltar à Terra.*

Problema: O ReGra sugere "O sonho de todo astronauta é voltar a Terra."

Natureza do problema: 'Terra', no léxico, tem a acepção de "terra firme" (em oposição, p.ex., a 'mar'). Neste caso, a norma gramatical indica que não se use crase.

[2] Nesta etapa do estudo, ainda é vaga a noção de refinamento lexical, para a resolução da ambigüidade. Por esse motivo, não foi dada uma sugestão mais precisa sobre qualquer procedimento de correção no âmbito do ReGra. Outros casos similares serão relatados adiante.

Sugestão de correção: Introduzir a acepção de 'Terra' como 'planeta', no léxico. Entretanto, isto implica definir um modelo de mundo, que remete não só a um modelo semântico, como também a um modelo pragmático, com evidentes reflexos na natureza do processamento automático.

*Deve ser de conhecimento amplo, tanto interna como externamente.*

Problema: O revisor acusa erro de concordância nominal, sugerindo que *'interna'* (advérbio, com valor de *'internamente'*) concorde em gênero com o adjetivo *'amplo'*.

Natureza do problema: de ambigüidade lexical, atribuindo a categoria de adjetivo ao advérbio.

Sugestão de correção: Inserir uma regra (sintática) de resolução de locuções adverbiais que admitam a ocorrência de adjetivos como advérbios.

*O sistema monitora todas as atividades realizadas pelo estudante, o qual apenas responde às questões formuladas pelo sistema*.

Problema: também relativo à concordância nominal, o ReGra sugere incorretamente que se troque *'monitora'* (verbo) por *'monitor'* (adjetivo), ou seja, considera o verbo como sendo adjetivo.

Natureza do problema: de ambigüidade lexical ou de ordem de escolha de aplicação de regra gramatical, durante o *parsing*.

Sugestão de correção: refinar a resolução da ambigüidade lexical e/ou rever o elenco de regras gramaticais, para possível alteração de sua ordem.

*Por que três provas juntas?*

Problema: o revisor aponta erro de concordância verbal, sugerindo *'três provas juntam'* (acepção verbal de *'juntar'*).

Natureza do problema: O ReGra não reconhece a elipse do verbo (*'fazer'* ou *'dar'*): casos de elipse estão previstos no ReGra como última opção de análise.

Sugestão de correção: É raro, para casos dessa natureza, não haver no ReGra uma regra gramatical que, aplicada antes, venha a identificar o item lexical como verbo. Qualquer sugestão para a correção desse tipo de problema implica a revisão, com possível alteração, da ordem de aplicação das regras. No entanto, qualquer inversão pode, também, piorar o desempenho do sistema em outros casos. Logo, não temos, no momento, sugestões plausíveis para resolver esse tipo de falso erro.

*A história do Brasil no século XVI foi em grande parte a história das relações entre índios e europeus*.

Problema: o revisor sugere crase em "*foi em grande parte a história*".

Natureza do problema: ambigüidade lexical, pois o revisor assume a acepção do verbo *'ir'* (passado) para *'foi'*, em vez de assumir a do verbo *'ser'*.

Sugestão de correção: refinar a resolução da ambigüidade lexical do verbo e/ou rever o elenco de regras gramaticais, para possível alteração de sua ordem.

*Primeiro, submeter a alta nobreza ao controle da Coroa.*

Problema: o revisor sugere crase em '*submeter a alta nobreza'*.

Natureza do problema: o verbo 'submeter' está sendo usado no sentido de 'dominar' ou 'vencer' e, portanto, com transitividade direta. Entretanto, o revisor opta pela categorização verbal de 'submeter' igual a 'transitivo indireto'.

Sugestão de correção: refinar a resolução da ambigüidade lexical do verbo.

*A Ásia e a África, conheciam-nas já os europeus de fins do século XV através dos autores clássicos greco-romanos, e do relato de alguns viajantes, como Marco Pólo.*

Problema: O revisor não reconhece o sujeito posposto ao verbo, *'os europeus de fins do século XV'*, e toma *'Ásia e a África'* como sujeito da oração.

Natureza do problema: de pontuação, cuja maior ocorrência no exemplário se dá para construções em que o ReGra entende os componentes sujeito e verbo separados por vírgula, emitindo a mensagem *"não se separa, com vírgula, o sujeito e verbo"*

Sugestão de correção: Revisar as regras de pontuação.

*A escrita, até então monopólio quase exclusivo da Igreja, adquiriu novas funções.*

Problema: *'Igreja'*, em vez de *'A escrita'*, é considerado o componente na posição de sujeito de *'adquiriu novas funções'.*

Natureza do problema: de pontuação, similar ao caso anterior. O ReGra ignora as vírgulas de um período, não reconhecendo orações subordinadas apositivas. Isto se deve à assunção inicial de projeto, de que os dados de entrada do revisor são considerados errados: o ReGra sempre desconfia da pontuação original, acarretando uma ambigüidade estrutural, de difícil resolução na versão atual do ReGra.

Sugestão de correção: analisar cuidadosamente os casos de ambigüidade.

Os problemas de falsos erros apontados pelo revisor durante o processamento do exemplário completo são sintetizados na Tabela 2. Como essa tabela indica, a grande maioria dos falsos erros detectados está relacionada aos casos de concordância nominal e verbal e ao uso de crase.

Tabela 2: **Diagnóstico de falsos erros**

| **Tipo de problema** | **Nro. de intervenções inadequadas** |
|---|---|
| Concordância nominal | 149 |
| Concordância verbal | 89 |
| Uso de crase | 58 |
| Pontuação | 30 |
| Uso do particípio passado | 1 |
| Colocação pronominal | 1 |
| Regência | 2 |

A Tabela 3 apresenta a distribuição dos falsos erros nas sentenças sob análise segundo as classes gramaticais dos componentes sentenciais. As classes assinaladas correspondem àqueles componentes que são os causadores diretos da inadequação de diagnóstico do ReGra.

Tabela 3: **Falsos erros apontados pelo ReGra**

| **Classe gramatical** | **Total de exemplos** |
|---|---|
| Adjetivo | 73 |
| Verbo | 73 |
| Substantivo | 68 |
| Crase | 56 |
| Pontuação | 25 |
| Advérbio | 4 |
| Preposição | 3 |
| Conjunção | 2 |
| Locução Adjetiva | 1 |
| Locução Prepositiva | 1 |
| Expressão Fixa | 1 |
| TOTAL | 307 |

### 2.1.5. As omissões

Omissões (ou falsos acertos, falsos positivos) incluem os casos em que o ReGra considera que as sentenças sob análise são gramaticais e, portanto, aqueles cuja análise automática é bem sucedida. A partir do conjunto de sentenças com diagnóstico problemático, os casos de omissão do ReGra foram identificados com a interferência de especialistas humanos, já que eles não permitem a recuperação automática de qualquer informação. Apresentamos abaixo alguns exemplos dessa natureza, com correspondente parecer dos especialistas lingüistas.

*Agora a súplica dos olhos e a melancolia deles eram mais intensos e puramente voluntários*.

Problema: gênero incorreto de ambos os adjetivos.

Natureza do problema: especificação na base lexical: o léxico categoriza o item 'deles' como 2a. pessoa do verbo 'delir', sendo esta a primeira acepção escolhida durante a análise sintática. A segunda opção é a contração "de + ele", que seria a adequada, neste caso.

Sugestão de correção: Refinar a resolução da ambigüidade lexical, pela alteração da base lexical do ReGra.

*Abraçou-se emocionados e visivelmente contidos pai e filho*.

Problema: Flexão verbal incorreta.

Natureza do problema: Dificuldade de interpretação da forma SE, conforme já foi amplamente discutido (p.ex., em Martins et al., 1999). O revisor entende que: a) 'emocionados e visivelmente contidos pai e filho' ocupam a posição de sujeito composto e b) 'pai e filho' ocupam a posição de núcleos do sujeito, o que está gramaticalmente correto. No entanto, ele também entende que: c) o verbo 'abraçar' é transitivo direto, em vez de pronominal e d) o pronome SE exerce a função de objeto indireto, em vez de partícula integrante de verbo pronominal.

Sugestão de correção: Várias medidas foram sugeridas, dentre as quais apontamos as seguintes:

1. Especificar, por meio de regras sintáticas, o fato de que, quando um verbo tiver como atributo o traço "pronominal" e, seguido a ele, aparecer a forma SE, a acepção adequada será a de "verbo pronominal" e, portanto, ao SE não se atribui qualquer função sintática.
2. Inserir, como traço semântico de 'abraçar', características selecionais do tipo *humano*, para preenchimento de suas posições argumentais. No contexto desta sentença, isso resolveria a referência verbal ao sujeito composto e, portanto, seria possível estabelecer a concordância de número necessária.

*Casamento e mortalha no céu se talha*.

Problema: Flexão verbal incorreta.

Natureza do problema: Ambigüidade estrutural: o revisor entende que há um sujeito composto, 'casamento e mortalha no céu'[3]. Porém, a acepção eleita para o verbo 'talhar' é a de verbo pronominal, em vez de ser intransitivo. Como resultado, admite-se que a acepção escolhida foi a da forma SE como pronome sem função sintática alguma, quando deveria estar apontando para a sua função de "apassivador verbal", permitindo, assim, que o ReGra admitisse o SE como objeto direto e descartasse a concordância com um sujeito elíptico[4].

[3] Apesar de "no céu" ser interpretado como adjunto do sujeito, isso não interfere na correção gramatical.

[4] Conforme se discute em (Martins et al., 1999), esta acepção é questionável. Porém, a forma *default* considerada adequada, no ReGra, para o uso da forma SE pronominal como a ilustrada, implica a concordância verbal com o objeto (que, no exemplo, está anteposto).

Sugestão de correção: Admite-se a hipótese dessa ocorrência se dever a uma falha da regra gramatical especificada para os verbos pronominais, pois mesmo nos contextos em que o pronome reflexivo exerce a função sintática de objeto direto, deve haver a concordância entre sujeito e verbo pronominal. Entretanto, ainda não se chegou a uma conclusão sobre a melhor forma de tratar os casos que envolvem o SE, pois esta é uma das construções cujo esforço de tratamento computacional esta'sendo questionado. O exemplo seguinte ilustra uma estrutura em que há a presença de um verbo estritamente pronominal.

*Homens e mulheres no céu se arrepende*.

Problema: Flexão verbal incorreta.

Natureza do problema: Merecem atenção alguns dados curiosos a respeito da análise dessa sentença: 1) se inserirmos o expletivo 'é que' antes do verbo, nada se altera, ou seja, a necessidade da concordância não é reconhecida. A ferramenta entende o 'é que' como um delimitador, indicando um problema bastante sério de representação no ReGra, uma vez que o uso do expletivo é bastante comum na língua portuguesa; 2) se retirarmos o segmento 'no céu' o revisor pede a concordância verbal. Isto é notável, se considerarmos que tanto a interpretação morfológica quanto a sintática são a mesma, de acordo com os resultados sintáticos indicados pelo *parser*. Vale notar aqui que a ferramenta entende 'no céu' como um adjunto adnominal, quando deveria compreendê-lo como adjunto adverbial; 3) se inserirmos o determinante 'os' no início da oração, a ferramenta também pede a concordância verbal. Entretanto, da mesma forma que no caso anterior, o resultado da análise sintática é idêntico ao dos casos em que não há qualquer problema sintático.

Sugestão de correção: Como medida mais imediata, sugeriu-se adotar a criação de regra específica para os verbos pronominais, independentemente da presença de adjuntos na sentença.

A Tabela 4 sintetiza os diagnósticos de omissão do ReGra. Para esses problemas, de natureza lingüística, observamos que as classes gramaticais mais representativas de falsos erros e omissões são, respectivamente, [adjetivo, crase, substantivo, verbo] e [preposição, verbo:pessoa, adjetivo, crase]. Uma infinidade de problemas de ordem semântica devido à ausência de informação lingüística no ReGra, como os ilustrados acima, foram registrados durante esta etapa do trabalho, a qual, embora realizada por especialistas, requer maior aprofundamento, pois qualquer alteração de ordem sintática pode acarretar problemas diversos e mais sérios, em relação a outros tipos de construção gramatical. Sugestões dessa natureza constituem, portanto, uma das grandes dificuldades ainda a superar neste projeto. Além disso, há fatores não lingüísticos, mais precisamente, operacionais, que interferem nos diagnósticos, tais como a escolha da ordem de aplicação de regras gramaticais durante o *parsing* ou a impossibilidade de escolha de uma regra adequada, devido à incompleteza do conjunto de regras de revisão gramatical, como descrevemos na seção seguinte.

Tabela 4: **Omissões do ReGra**

| **Classe de palavras** | **Total de exemplos** |
|---|---|
| Verbo: pessoa[5] | 130 |
| Crase | 90 |
| Adjetivo: concordância | 71 |
| Preposição | 66 |
| Pronome | 34 |
| Advérbio | 29 |
| Verbo: modo | 14 |
| Artigo | 9 |
| Pontuação | 7 |
| Expressões fixas | 6 |
| Conjunção | 6 |
| Locução prepositiva | 5 |
| Há/a | 4 |
| Verbo: falta | 4 |
| Locução adverbial | 4 |
| Substantivo: singular/plural | 2 |
| Verbo: tempo | 2 |
| Verbo: excesso | 1 |
| Substantivo: feminino/masculino | 1 |
| Verbo: voz | 1 |
| TOTAL | 487 |

### 2.2. Problemas de implementação

Todas as análises realizadas pelo ReGra são acessíveis por meio de um arquivo gerado durante o *parsing*, chamado "saida.sin", que contém a distribuição sintática de cada sentença analisada. A Figura 1 ilustra uma instância dessa análise, evidenciando a inadequação de interpretação da forma SE na sentença

> *Aos portugueses com residência permanente no País, **se** houver reciprocidade em favor dos brasileiros, serão atribuídos os direitos inerentes ao brasileiro nato, salvo os casos previstos nesta Constituição.*

Neste caso, o ReGra produz, como interpretação da forma SE, a categoria gramatical 'oda' (objeto direto anteposto), em vez da forma correta conjunção. Isto se deve ao fato de o verbo 'haver' ser classificado no léxico como pronominal ou transitivo direto, implicando a invocação de uma rotina de desambigüização que, ao reconhecer a partícula SE antes do verbo, classifica-o como pronominal, tomando sempre a acepção de pronome (PPOA – pronome pessoal oblíquo átono - ou PPODA – pronome pessoal oblíquo dativo átono) em primeiro lugar. Considerando, assim, o verbo 'haver' como pronominal, o ReGra aplica a regra[6]

OD_ANTEPOSTO + verbopro + [ADJ_ADV] + SUJ_POSPOSTO

[5] 'verbo: pessoa' indica a ocorrência de falta de gramaticalidade devido à discordância de flexão verbal em relação a pessoa. As outras ocorrências do tipo 'X:Y' têm interpretação similar: discordância da categoria lexical X em relação à característica Y.

[6] objeto direto anteposto + verbo pronominal + adjunto adverbial optativo + sujeito posposto

quando, na verdade, deveria aplicar a seguinte regra[7]:

[verboaux + [verboaux]] + vtd + [ADJ_ADV] + OD

Essa inadequação se deve, sobretudo, à falta de mecanismos para dar conta de toda a informação lingüística transmitida no contexto da sentença, que permita ao ReGra recuperar, do contexto, a acepção correta. Em outras palavras, com mais informações o ReGra poderia aplicar outra regra gramatical, em vez da regra que termina por apontar um erro a partir da acepção pronominal atribuída ao SE. Esse problema, por sua vez, é decorrente do determinismo na aplicação das regras gramaticais durante o *parsing*.

Seria possível resolver o tipo de ambigüidade introduzido pelo SE no exemplo acima caso se permitisse o *backtracking* no sistema. Entretanto, a utilização dessa regra de exceção não ajudaria em nada, se não fossem incluídas mais informações lingüísticas, pois o ReGra continuaria sendo incapaz de determinar o momento em que o *backtracking* se faria necessário, já que ambas as regras continuariam sendo válidas. A solução para esse tipo de problema está em fornecer mais recursos para o ReGra, para resolver ambigüidades de natureza operacional, visando a resolução das ambigüidades de natureza lingüística.

"§ 1º Aos portugueses com residência permanente no País, **se** houver reciprocidade em favor dos brasileiros, serão atribuídos os direitos inerentes ao brasileiro nato, salvo os casos previstos nesta Constituição."

[periodo_composto] = <aos portugueses com residência permanente no país , se houver reciprocidade em favor dos brasileiros , serão atribuídos os direitos inerentes ao brasileiro nato>
[periodo_simples] = <aos portugueses com residência permanente no país , se houver reciprocidade em favor dos brasileiros ,>
[adj_adv] = <aos portugueses com residência permanente no país>
[predicado] = <, se houver reciprocidade>
**[oda] = <se>**
[nucleo] = <houver>
[principal] = <houver>
[sujeito_simples] = <reciprocidade>
[nucleo] = <reciprocidade>
[adj_adv] = <em favor dos brasileiros ,>
[periodo_simples] = <serão atribuídos os direitos inerentes ao brasileiro nato>
[predicado] = <serão atribuídos os direitos inerentes ao brasileiro nato>
[nucleo] = <serão>
[principal] = <serão>
[pred_sujeito] = <atribuídos>
[sujeito_simples] = <os direitos inerentes ao brasileiro nato>
[adj_adn_esq] = <os>
[nucleo] = <direitos>
[adj_adn_dir] = <inerentes ao brasileiro nato>

**Figura 1: Estrutura sintática resultante do *parsing* realizado pelo ReGra**

Os problemas apontados acima demonstram a necessidade de agregar vários níveis de conhecimento ao ReGra, para seu aprimoramento. Foi mencionada, particularmente, a necessidade de se contemplar: 1) o problema de representação lexical, pela especificação de

[7] verbo auxiliar + verbo auxiliar opcional + verbo transitivo direto + adjunto adverbial opcional + objeto direto

traços semânticos mínimos associados aos itens lexicais do léxico do ReGra; 2) a resolução do significado a partir da distinção entre categorias lexicais, no contexto sintático e, portanto, a necessidade de se buscar a especificação de procedimentos semânticos para integrar sintaxe e semântica, conforme necessidades impostas pelo contexto; 3) a análise e possível alteração do próprio *parser*, para resolver (1) e (2). Buscando aprofundar o conhecimento sobre tais questões, apresentamos algumas considerações sobre o problema que ora se coloca, seguidas de uma proposta para contemplá-las.

## 3. Síntese dos casos de inadequação do ReGra

Por meio dos seguintes exemplos sintetizamos abaixo os problemas apontados indevidamente pelo ReGra, discutindo os principais aspectos lingüísticos do comportamento da ferramenta:

(1) Se eu desse vida nova a ele, eu morreria. (Sugere ‘dessa vida’)
(2) A espada feriu fundo. (Sugere ‘funda’)
(3) Ela falava áspero. (Sugere ‘áspera’)
(4) Tu deves fazer o que achas indicado. (Sugere ‘achas indicadas’)
(5) As gueixas nunca mais foram as mesmas. (Sugere ‘às’)
(6) Bethay recebe e hospeda em sua casa pastores e pessoas ligadas à Igreja para reuniões. (Sugere ‘sua casa pastora’)
(7) Canto sempre linda, disse a soprano. (Sugere ‘canto sempre lindo’)
(8) Devido às suas propriedades termomecânicas e viscoelásticas especiais oriundas de estruturas química e morfológica bem distintas dos demais sólidos em geral, esses materiais ocupam um espaço muito importante na indústria de transformação, de bens de consumo, e mesmo automobilística e aeroespacial. (Acusa-se erro ortográfico em ‘termomecânicas’ e ‘viscoelásticas’)
(9) Eu moro na Marquês de Sapucaí. (Sugere ‘no Marquês’)
(10) Itens nunca teve acento gráfico. (Sugere ‘Itens nunca tiveram’)

A consideração das sentenças assinaladas permite perceber que a ambigüidade lexical pode ser classificada em quatro grupos, evidenciando a necessidade de representação mais rica e estruturada do conteúdo lexical, a saber: a) morfológica; b) sintática; c) semântica; e d) pragmática.

A ambigüidade morfológica envolve quase sempre a possibilidade de análises mórficas concorrentes para uma mesma palavra. Neste sentido, confunde-se com a ambigüidade sintática (porque análises mórficas divergentes conduzem freqüentemente a diferentes classes gramaticais) e com a ambigüidade semântica (porque o significado da palavra pode ser interpretado como uma composição dos significados dos morfemas que a compõem). Como exemplo desta última situação pode ser citado o caso de ‘interminável’, que pode ser interpretada como “aquilo que não pode terminar” ou “aquilo que não pode ser minado entre”, conforme tenha sido analisado, respectivamente, como ‘in-termin-á-vel’ ou ‘inter-min-á-vel’. A ambigüidade semântica deriva, neste caso, da diferente delimitação dos elementos derivacionais (o prefixo e o radical) que compõem a palavra. O mesmo, de certa forma, ocorre para as sentenças (1) e (3), mas aqui a ambigüidade morfológica produz, principalmente, problemas sintáticos, porque relacionados à concordância. Em (1), a forma ‘desse’ foi analisada como ‘dess-e’ (do pronome demonstrativo de 2ª pessoa) e não como ‘d-e-sse’ (do verbo ‘dar’). Em (3), a forma ‘achas’ foi interpretada como ‘acha-s’ (do substantivo ‘acha’) e não como ‘ach-a-s’ (do verbo ‘achar’). A rigor, não se pode dizer, pelo menos nesses dois casos, se a ambigüidade é morfológica, dela derivando repercussões sintáticas (a mudança de classe gramatical), ou se a ambigüidade é, na verdade, sintática, permitindo diferentes análises

morfológicas. A origem do problema importa pouco, porém. O fato é que a ambigüidade morfológica não pode ser resolvida independentemente do contexto. A própria palavra não fornece, em nenhum dos casos citados, pistas suficientes para decisão. Devemos, portanto, recorrer ao contexto sintático – caso de (1) e (3) – ou ao contexto semântico – caso de 'interminável' – para selecionar a correta estrutura morfológica da palavra, o que, em inúmeros pontos, fragiliza a importância de se detectar (e mesmo pensar) sobre a ambigüidade morfológica. A utilidade da análise mórfica ficaria, pois, restrita ao reconhecimento dos processos produtivos de formação de palavras não dicionarizadas, caso de 'termomecânicas' e 'viscoelásticas', presentes em (8). Embora perfeitamente possíveis na língua portuguesa, essas duas formas foram interpretadas como erro ortográfico, dada a ausência de estratégias de composição dos radicais existentes no dicionário ('termo-mecânic-a-s' e 'visco-elástic-a-s').

A ambigüidade sintática envolve, principalmente, as relações de dependência que a palavra pode ou não estabelecer dentro da sentença. Neste caso, é possível pensar em duas subdivisões: a ambigüidade de classe gramatical e a ambigüidade de estrutura temática. No primeiro caso, trata-se da possibilidade de uma mesma forma vir a desempenhar, em diferentes contextos, diferentes papéis gramaticais. Um exemplo particularmente ilustrativo é o da palavra *se*, extremamente freqüente nos textos de língua portuguesa, que pode ser, segundo a gramática normativa: a) substantivo (*Nenhum se aparecia naquele texto*); b) pronome reflexivo (*A menina olhou-se no espelho*); c) pronome reflexivo recíproco *(Pai e filho se abraçaram*); d) parte integrante do verbo (*Ninguém jamais se esquece do primeiro amor*); e) palavra expletiva ou de realce (*Passaram-se os anos*); f) partícula apassivadora *(Vendem-se casas*); g) índice de indeterminação do sujeito (*Falou-se de democracia durante muito tempo*); h) conjunção subordinativa integrante (*Não nos disseram se tudo estava resolvido*); i) conjunção subordinativa condicional (*Se você não estudar, não será aprovado*); e j) conjunção subordinativa causal (*Se está com tanta pressa, não precisa me esperar*). Entre os exemplos citados, além de (1) e (3), já referidos, este é o caso também de (2) e (3) (em que 'fundo' e 'áspero' foram erroneamente interpretados como adjetivos e não como advérbios), de (6) (em que 'pastores' foi interpretado como adjetivo e não como substantivo), e de (7) (em que 'canto' foi tomado como substantivo e não como verbo). Em todas essas situações, as formas acumulam, no dicionário, várias funções sintáticas, com predomínio daquelas que, no contexto das sentenças, acabam por se revelar inadequadas. O fato é tanto mais perturbador quando se percebe que nenhum desses contextos sintáticos autoriza a formulação de regras de desambigüização genéricas.

O segundo caso de ambigüidade léxico-sintática remete à estrutura temática. Estão envolvidos, nesta situação, principalmente os verbos de transitividade ou de regência variável. É o caso, por exemplo, de 'assistir', que pode significar: "estar presente a" ou "caber", quando transitivo indireto regendo 'a' (*Não assisto a novelas; Este direito não assiste ao aluno*); 'socorrer', quando transitivo direto (*O médico assistiu o paciente*); 'morar', quando intransitivo (*O presidente assiste em Brasília*). Este também é, de certa forma, o caso da forma 'foram', presente em (5), que pode ser interpretada como derivada de 'ir' (caso em que seria intransitiva) ou como derivada de 'ser' (caso em que funcionaria como verbo de ligação). Como o contexto sintático não fornece elementos suficientes para diferenciar um ou outro uso da mesma forma, a ferramenta faz a opção pela estrutura intransitiva, exigindo uma preposição que não deveria.

A terceira modalidade de ambigüidade lexical, a semântica, está diretamente associada à idéia de arbitrariedade do signo lingüístico. Sendo arbitrária a relação entre significante e significado, é natural que, para um mesmo significante, possam concorrer diferentes significados ou que, para um mesmo significado, possam concorrer diferentes significantes. O exemplo, já clássico, é o de homônimos perfeitos, como a palavra 'manga', que pode significar "parte da camisa" ou "espécie de fruta" sem que essas diferentes acepções estejam

necessariamente associadas a diferentes possibilidades de análise mórfica ou a diferentes papéis presentes na sentença. Mas este também é o caso de certos fenômenos semânticos específicos, de deslizamento de sentidos, como a metonímia presente em (9): *Moro na (rua) Marquês de Sapucaí*. Ou da distinção entre uso e menção, presente no uso da palavra 'itens' em (10). Nesses dois casos, a ausência de estratégias de representação desses movimentos semânticos acaba por produzir erros sintáticos (de concordância).

Por fim, a ambigüidade pragmática concerne à ambigüidade de uso ou de referência de certas palavras ou expressões da língua. O caso mais célebre é, sem dúvida, o dos dêiticos (pronomes pessoais, advérbios de lugar e tempo, por exemplo), cuja referência ultrapassa o co-texto lingüístico. O significado de 'eu', de 'aqui' e de 'agora' não dependem do contexto imediato, mas extrapolam o texto, estabelecendo relações de referência exofórica. No entanto, podem ser também classificados como casos de ambigüidade pragmática todos aqueles que, como em (7), encontram fora da sentença os elementos necessários à desambigüização. A origem da incorreção de (7) está menos na ambigüidade sintática de 'canto' do que na cláusula 'disse a soprano' que não a permite.

A identificação dessas quatro classes de problemas relativos à ambigüidade lexical, somada aos problemas de processamento durante a revisão, levou-nos ao recurso à análise de dependência entre os constituintes sentenciais, pela verificação das normas lingüísticas para o português. Especialmente, averiguamos os casos de regência verbal e de seus complementos, verbais ou nominais, considerando o contexto do próprio ReGra, ou seja, de seu modelo de ATNs (Woods, 1970) para a revisão gramatical. Antes, porém, de abordar essas questões (na Seção 5), tecemos algumas considerações a respeito da necessidade de crítica sobre a representação lexical e a gramática do ReGra.

## 4. Considerações sobre a ambigüidade categorial em função da representação do conteúdo lexical e das normas gramaticais em uso

### 4.1. O léxico

Eleger o significado lexical como objeto de estudo pressupõe 1) uma delimitação clara do conceito de "palavra", ou pelo menos de "lexia", e 2) a crença na pertinência do significado lexical em extratos de significação mais abrangentes, como, por exemplo, na constituição do significado dos sintagmas e das sentenças. Embora essas duas assunções possam parecer, à primeira vista, triviais, cabe dizer que, nos dois casos, há material empírico disponível que, se não diretamente as contradiz, fragiliza-as sobremaneira.

Em primeiro lugar, cabe pensar se são palavras formas como 'léu', 'toa' e 'apesar', que não aparecem senão em contextos pré-determinados: 'ao léu', 'à toa', 'apesar de'. Se considerada a questão da autonomia semântica, é forçoso admitir que a todas elas podem ser atribuídos significados bem delimitados. Mas, ao mesmo tempo, nenhuma delas conserva autonomia sintática. O mesmo pode ser observado em construções verbais com mesóclise: as três formas em 'amá-la-ei' têm autonomia semântica, mas nenhuma possui autonomia sintática. Já o inverso ocorre nos clichês e expressões fixas, como 'bater as botas', em que, embora dotadas de autonomia sintática, as formas não participam de forma autônoma da composição do significado da expressão (o significado de 'bater as botas' não é a composição dos significados individuais de 'bater', 'as' e 'botas'). Por tudo isso, a representação do significado lexical exige, de início, que seja definido o que é a palavra e qual seu papel na constituição do significado da sentença.

Na trajetória de construção da base de dados lexicais que serve ao Projeto ReGra, 'palavra' foi definida como qualquer seqüência de caracteres pertencente à língua portuguesa (no sentido de poder ser encontrada em textos brasileiros), isolada por espaços em branco ou sinais de pontuação. Neste sentido, foram dicionarizados, não morfemas, mas formas inteiras,

não analisadas, independentemente de sua autonomia sintática ou semântica. O único critério efetivamente utilizado foi a autonomia gráfica no registro da escrita, por facilitar o processo de análise e reconhecimento lexical. Como se trata de uma ferramenta conservadora, que deve alertar o usuário para o uso de verbetes que não constam nos grandes dicionários da língua, não caberia aqui o esforço de analisar e representar os processos de criação lexical, por envolverem o risco da produção de formas regulares que pudessem ser desautorizadas pelas autoridades gramaticais (caso de 'casação', em vez de 'casamento'; ou 'atribuimento', em vez de 'atribuição').

No entanto, ao tratar as palavras como blocos monolíticos, sem a consideração de seus processos de formação, e ao admitir uma representação lexical independente de informação semântica, a qual, na verdade, inexiste na versão atual do léxico, acabou-se por reduzir todos os casos de ambigüidade referidos na seção anterior à ambigüidade de classe gramatical. A qualidade das informações disponíveis no dicionário não permite descrever a tipologia da ambigüidade e optar por estratégias diferenciadas em cada um dos diferentes casos em que uma determinada forma pode ser ambígua. O único tipo de ambigüidade abordável tornou-se, pois, aquele em que uma determinada forma acumula informações contraditórias (referentes à classe gramatical, ao gênero e ao número, nomeadamente) no dicionário.

Para tentar contornar o problema da co-indexação de categorias contraditórias, decidiu-se que o léxico que serve de apoio ao ReGra deveria conter uma ordenação da freqüência de ocorrência das partes do discurso para cada palavra ambígua, freqüência esta intuitivamente determinada pelos lingüistas do Projeto. Embora a iniciativa tenha tido o mérito de isolar os arcaísmos, os regionalismos e outros casos de uso localizado (e bastante incomum) das palavras, é necessário reconhecer que, na maior parte dos casos, a freqüência de uso não funciona como critério razoável de desambigüização lexical. Outros mecanismos de desambigüização lexical revelaram-se necessários e esta necessidade conduziu à elaboração de um conjunto de regras que permitisse calcular a classe da palavra a partir do seu contexto de ocorrência. Trata-se, pois, de uma tentativa de retirar do dicionário parte da responsabilidade pela decisão, deixando-a também a cargo da distribuição da palavra na sentença, com a consideração dos contextos mínimos, à esquerda e à direita, junto aos quais ela aparece. No entanto, a combinação das duas estratégias de desambigüização – a freqüência do uso e o contexto de ocorrência – ainda que tenha permitido eliminar alguma parte da ambigüidade lexical, mostrou-se também insatisfatória (cf. apresentamos no relatório anterior). A aplicação em textos reais fez perceber que, não apenas os itens lexicais, mas também os contextos (formados, obviamente, por outros itens lexicais) são predominantemente ambíguos. E a análise combinatória que nos permitiria selecionar uma entre as várias possibilidades categoriais de uma dada palavra provou ser capaz apenas de restringir o conjunto de categorias elegíveis: se antes havia cinco ou quatro possibilidades de categorização, agora elas seriam três ou duas, e não formariam ainda um conjunto unitário. A ambigüidade, portanto, persiste; e a eficácia das regras anteriores fica restrita a contextos unívocos, bastante específicos, de resto raros no tratamento das línguas naturais.

Mais do que a persistência da ambigüidade, porém, incomoda o fato de terem sido adotadas, para diferentes casos de ambivalência, as mesmas estratégias. Se a freqüência de uso e o contexto de ocorrência podem ser efetivamente úteis para equacionar problemas de ambigüidade relativos à classe gramatical, pouco têm a acrescentar para a solução dos problemas relativos à estrutura temática, a fenômenos semânticos (como a metonímia) e à pragmática. Como resultado, observa-se que, em quase todas essas situações, a ferramenta é induzida a erro pelo comportamento exageradamente sintático das estratégias de desambigüização, o que nos obriga à reconsideração do modelo gramatical até agora adotado.

### 4.2. A gramática

O conceito de gramática é historicamente derivado da crença na existência de uma unidade sob a diversidade dos fatos lingüísticos. Sob a aparente heterogeneidade de forma das sentenças das línguas naturais se esconderia uma regularidade que pode ser tornada acessível à consciência. Haveria, pois, um conjunto finito de regras capaz de gerar o conjunto infinito das sentenças pertencentes a uma determinada língua. Em última análise, a formulação de uma gramática nos ofereceria uma definição intensional do conceito de "sentença", para além da experiência extensional que a língua a todo momento nos oferece.

Há, aparentemente, ganhos teóricos óbvios na hipótese gramatical. Sem ela, desconfia-se que as sentenças das línguas naturais, caracterizadas pela produtividade e pela diversidade de superfície, dificilmente poderiam constituir nível válido de análise lingüística (cf. Saussure, 1916). Conseqüentemente, haveria um sem-número de fenômenos lingüísticos, como a concordância, a regência e a colocação, entre outros, para os quais não haveria nem poderia haver resposta.

Do postulado da gramática derivam pelo menos dois axiomas, bastante difundidos na teoria lingüística contemporânea: o primeiro afirma que é possível formular um conjunto fechado, definitivo, de regras gramaticais a um só tempo suficientes e necessárias para a geração de todas e apenas das sentenças de uma determinada língua. Em outras palavras: a sintaxe poderia ser completamente exaurida, e todos os fenômenos sintáticos poderiam ser descritos por referência à gramática. Não haveria, neste caso, fenômenos sintáticos irregulares – a irregularidade seria sobretudo o estado do que ainda não foi (mas que futuramente será, sem dúvida) regularizado. O segundo postulado prevê que os fenômenos sintáticos podem ser descritos sem referência aos estados anteriores da língua. Ou que o estudo da sintaxe pode ser feito, sem prejuízo de seu poder descritivo, apenas da perspectiva da sincronia. Reconhece-se que a variedade lingüística em uso é depositária de uma série de alterações (inclusive sintáticas) da língua ao longo do tempo, mas que o conjunto de regras é, em qualquer momento de sua história, autoconsistente.

Esses axiomas têm presidido as tentativas de construção de gramáticas formais para as línguas naturais. O primeiro axioma tem sido utilizado como critério de validação (ou de escolha entre gramáticas concorrentes); o segundo tem inspirado a busca por princípios universais, de validade interlingüística. Em ambos os casos, no entanto, responde-se apenas parcialmente aos problemas oferecidos pela busca da gramática. Outros desafios presentes na tarefa têm merecido respostas menos convergentes e serão alguns deles os objetos de discussão deste texto. Trata-se, nomeadamente, do problema da natureza desse conjunto de regras e da maneira pela qual ele pode ser posto em uso, de forma a permitir a análise automática de qualquer sentença da língua. No primeiro caso, estaremos falando a respeito de modelos de gramática e formalismos gramaticais; no segundo, de estratégias de *parsing*. Embora vinculados, são problemas diferentes, e serão abordados aqui separadamente.

O consenso em torno da possibilidade de formulação de uma gramática raramente extravasa para a especificação da gramática ótima, aquela cujos resultados coincidiriam, ponto por ponto, com as ocorrências sintáticas efetivamente verificáveis para uma dada língua. Admite-se que esta gramática deve consistir em um conjunto finito de regras, mas não há convergência quanto à delimitação deste conjunto ou quanto ao formato e à função dessas regras.

Há um dissenso original entre as concepções atuais de gramática da teoria lingüística e da teoria lingüístico-computacional que é interessante desde já assinalar. Para a lingüística-computacional, a gramática deve ser definida como um conjunto de regras (ou de instruções) de natureza matemática ou, mais especificamente, algébrica. A lingüística teórica não tem este compromisso e, freqüentemente, afirmará que a gramática das línguas naturas não é redutível à álgebra. A favor desta última hipótese pode ser citada a indeterminação que não raro marca a

ocorrência dos fenômenos lingüísticos: as expressões lingüísticas reais, efetivamente praticadas pela comunidade de falantes, principalmente no registro da fala, são marcadas por hesitações, falsos começos, lapsos de língua, atos falhos, topicalizações, interrupções, sobreposições, ritmo desordenado, silêncios e outros fenômenos marcados quase sempre por alguma imponderabilidade. Mesmo quando higienizadas pelo lingüista, através da postulação de um falante-ouvinte ideal, o uso conotativo (através de metáforas e metonímias, por exemplo) das expressões lingüísticas autorizará extensões de difícil matematização, principalmente porque a língua não é lógica: na língua, nem toda tautologia é redundante (*Quero morrer uma morte pacífica*), nem toda contradição é contraditória *(Maria foi e não foi*.), nem toda negação é uma negação (*Eu não sei nada ≠ Eu sei alguma coisa*), nem sempre mais de um significa plural (*O pessoal foi ao cinema*), nem sempre uma proposição afirmativa é verdadeiramente afirmativa (considere-se, a este respeito, a ironia), não apenas veiculam-se idéias, mas fazem-se coisas (*Eu prometo que vou voltar*).

A falta de correspondência exata entre a linguagem e o pensamento (lógico) tem inviabilizado, em inúmeros pontos, o programa da semântica filosófica e pode bem ser que a gramática padeça do mesmo mal. Poderia até haver um conjunto de regras que governaria a produção dos enunciados lingüísticos, mas dificilmente estas regras partilhariam a natureza determinística das construções algébricas. No entanto, esta versão não-determinística das regras gramaticais dificilmente serviria à máquina, que ainda não dispõe de informação extralingüística para processar as línguas naturais. E este é o recorte e o mérito da perspectiva formal, ou algébrica, da linguagem: permitir provisoriamente alguma automação e prometer permanentemente a automação completa dos processos lingüísticos. A matematização da linguagem, embora muito possivelmente simplificadora, não se revela produto de uma opção, mas antes de uma falta de opção, diante de um estoque de categorias teóricas e metodológicas muito mais restrito, ou de natureza muito menos ambígua, do que o que hoje convém à teoria lingüística. Se se poderá enquadrar a linguagem a partir desse repertório muito reduzido de categorias ou se teremos antes que aguardar a ampliação (ou a reformulação) desses conceitos de partida e em que direção esta ampliação ou reformulação deverá ser feita, apenas a praxe lingüístico-matemática poderá dizer.

Envolvidos no cenário de produção de ferramentas computacionais para o processamento automático da língua portuguesa, estaremos limitados aqui à exploração dos principais expoentes da primeira vertente gramatical assinalada no parágrafo anterior. Trata-se daquela interpelada pelo filtro do "realismo computacional", mais conhecido como computabilidade. Serão comparados quatro modelos gramaticais: 1) a gramática transformacional (*Transformational Grammar*, doravante TG), em três versões: a) a TG original (Chomsky, 1957), b) a Teoria Padrão, ou *Standard Theory* (ST – Chomsky, 1965) e c) a Teoria Padrão Estendida, ou *Extended Standard Theory* (EST – Chomsky, 1972); 2) a gramática léxico-funcional, ou *Lexical-Functional Grammar* (LFG – Bresnan, 1982); 3) a gramática de estrutura de constituintes generalizada, ou *Generalized Phrase Structure Grammar*, (GPSG – Gazdar and Pullum, 1982); e 4) a gramática de adjunção de árvores, ou *Tree Adjunct Grammar*, (TAG – Joshi and Shabes, 1992). Além destes, há pelo menos três outros modelos gramaticais importantes, que não foram ainda explorados: a HPSG, ou *Head-Driven Phrase Structure Grammar*, que é uma subespecificação da GPSG (Pollard and Sag, 1994); a FUG, ou *Functional Unification Grammar* (Kay, 1984); e a CUG, ou *Categorial Unification Grammar* (Haddock et al., 1987).

Um primeiro dissenso entre as gramáticas assinaladas parece envolver duas posições: (a) a gramática possui uma conformação holística, em que o conjunto de regras é homogêneo, ou seja todas as regras têm a mesma natureza (e, por extensão, a mesma sintaxe); ou (b) a gramática é modular e o conjunto de regras é heterogêneo, envolvendo dois ou mais subconjuntos de regra (ou componentes gramaticais). Um exemplo da primeira postura é a

gramática de constituintes imediatos, ou *Phrase Structure Grammar* (PSG) apresentada nos capítulos iniciais de (Chomsky, 1957). Exemplo da segunda postura é a TG apresentada nessa mesma obra e particularmente desenvolvida em (Chomsky, 1965). Em uma PSG, todas as regras possuem a forma $\alpha \rightarrow \beta$, em que $\alpha$ e $\beta$ têm, em princípio, comprimento variável e correspondem aos símbolos terminais (itens lexicais) ou não-terminais (categorias funcionais) da língua[8]. No modelo da TG apresentado em (Chomsky, 1965), prevêem-se três tipos diferentes de regra: as regras de reescrita categorial (do tipo $\alpha \rightarrow \beta$, em que $\alpha$ e $\beta$ correspondem apenas a símbolos não-terminais), as regras de inserção lexical (que respeitam o princípio da subcategorização) e as regras transformacionais (subdivididas em obrigatórias e opcionais, no modelo de 1957; e apenas obrigatórias em 1965), que promovem alterações na estrutura da sentença. No caso do modelo da TG, a gramática estaria subdividida em duas componentes: o componente de base (formada pela subcomponente categorial e pelo subcomponente lexical) e o componente transformacional, cada uma das quais dotada de regras com funções e sintaxe específicas.

A grande diversidade que hoje se observa em relação aos modelos gramaticais deriva principalmente da especificação dos módulos (ou componentes) da gramática e, em cada módulo, do formato postulado para as regras. O formato das regras transformacionais da versão da TG de 1965, por exemplo, é bastante diferente do adotado na versão da TG de 1972. Na primeira versão, havia um conjunto bastante numeroso de regras transformacionais (apassivação, movimento dos afixos, topicalização, etc.) que foram reduzidas a apenas uma regra (*move*-$\alpha$, em que $\alpha$ corresponde a qualquer categoria funcional) no modelo posterior. Os outros modelos gramaticais (LFG, GPSG, TAG) simplesmente não prevêem o componente transformacional.

Um outro ponto de divergência entre os modelos parece envolver a relação pressuposta entre a gramática e os outros módulos que regem a faculdade da linguagem. Aqui, novamente, parecem ser duas as posições: (a) a gramática é independente, e (b) a gramática é dependente de informações que provêm dos outros módulos lingüísticos (o módulo fonológico e o módulo semântico, nomeadamente). A primeira posição é a pedra-de-toque do programa da TG. Na versão de 1957, além do módulo sintático, estavam previstos, fora da gramática, os módulos fonológico e semântico. Esses módulos eram, no entanto, secundários, visto que a relação intermodular era unilateral (sempre do módulo sintático para os demais módulos) e que o componente fonológico e o componente semântico operavam diretamente sobre as saídas dos subcomponentes do módulo sintático. Assim, a saída do componente transformacional estava relacionada ao componente fonológico (a partir do qual era gerada a representação fonética) e a saída do componente de base, ao componente semântico (responsável pela interpretação semântica da sentença). Na versão de 1965 houve uma alteração: o componente semântico passou a operar diretamente sobre a saída do componente transformacional, mas manteve-se a tese da autonomia da sintaxe. Exemplo da posição contrária (ie, de gramática dependente) pode ser retirado da reformulação que a própria TG sofreu em 1972. Na EST foi incorporado ao módulo sintático o componente forma lógica, de natureza semântica, com ingerência sobre a sintaxe (através do critério temático, ou critério-$\theta$). Fragiliza-se, dessa forma, a independência do componente sintático. Os outros modelos aqui explorados acompanham a tese da dependência da gramática.

Os níveis de representação da estrutura sintática constituem outro ponto de desacordo entre vários modelos gramaticais. O modelo da TG, em suas várias versões, postulava a existência de dois níveis de representação da estrutura sintática: a estrutura superficial (mais

[8] O formato que $\alpha$ e $\beta$ podem assumir subclassifica a gramática na hierarquia proposta em (Chomsky, 1959). Assim, as gramáticas do tipo PSG poderiam ser regulares, livres-de-contexto, sensíveis-ao-contexto e irregulares, dependendo da configuração de suas regras.

tarde apelidada de *S-Structure*) e a estrutura profunda (ou *D-Structure*). Nem sempre houve consenso sobre a natureza dessa separação proposta por Chomsky, derivada da hipótese transformacional. Assim, o modelo da LFG, embora também postule a existência de dois níveis de estruturas gramaticais, o faz em outra direção, não-transformacional: separa-se o nível dos constituintes e o nível funcional, que não correspondem exatamente às estruturas do modelo TG. O mesmo, de certa forma, acontece com o modelo da TAG, que também prevê dois níveis de representação da estrutura sintática: o nível das árvores centrais e o nível das árvores adjuntas. Mas nem mesmo em relação ao número de níveis parece haver consenso: o modelo da GPSG postula um único nível de descrição sintática: o nível de constituintes, sobre o qual se aplicarão as meta-regras.

O nível de granularidade das regras envolve menos divergência. Tanto a TG (pelo menos nas versões posteriores a 1965), quanto a LFG, a GPSG e a TAG, utilizam, como símbolos terminais, pares atributo-valor. São os chamados traços de categoria (*features*), que podem referir-se a propriedades intrínsecas ou extrínsecas dos itens lexicais. Por propriedades intrínsecas entendemos aquelas não relacionadas ao contexto de ocorrência dos itens lexicais. São informações inerentes às palavras, constituindo quase um prolongamento de sua classificação sintática. Assim, os substantivos são subclassificados em [+animado] ou [-animado], [+contável] ou [-contável], etc. As propriedades extrínsecas referidas nos traços de categoria dizem respeito à distribuição dos itens lexicais no sintagma, especificando a) os contextos categoriais em que o item lexical pode ocorrer e b) os traços de categoria que os ocupantes desses contextos categoriais devem ter. No primeiro caso, fala-se em subcategorização, ou seja, na propriedade de o item lexical prever a conformação do sintagma em que pode ocorrer. Verbos transitivos (como *matar*, por exemplo) somente podem ser gerados dentro de um sintagma verbal que subcategoriza um objeto. No segundo caso, fala-se em traços de seleção, ou seja, nas condições que devem ser satisfeitas para o preenchimento das posições sintáticas subcategorizadas pelo item lexical. A inserção do verbo *matar* depende não apenas do preenchimento da posição de objeto, mas de que este preenchimento seja feito por um item subclassificado em [+animado].

Alguns modelos (LFG, por exemplo) vão ainda mais além e propõem um conjunto ainda mais complexo de traços de categoria, que envolve também a representação da estrutura temática dos itens lexicais, produzindo um aninhamento de traços. Ao prever o caso paciente, a seleção temática do verbo *matar* subsume, por exemplo, a seleção sintática (objeto) e a seleção semântica ([+animado]) referidas anteriormente. O conjunto de traços de categoria para um dado item lexical não constitui, portanto, apenas uma lista, caracterizando antes uma estrutura que pode comportar variáveis e índices e que será projetada para o nível sentencial. A esse processo de projeção está normalmente associado um princípio uniforme de ordenação (*merging*), que é normalmente referenciado (pelo menos no caso de LFG, GPSG e TAG) como "unificação", de onde serem conhecidas, todas elas, como "gramáticas de unificação".

Entre as gramáticas de unificação existem outras diferenças (mais tópicas ou computacionais) que não serão por ora consideradas. Elas dizem respeito, principalmente, ao princípio de unificação utilizado e à estrutura dos tipos de categoria. Mais importante agora é considerar critérios de escolha entre os vários modelos concorrentes. Perrault (1985) sugere algumas idéias a respeito do assunto. Para o autor, a escolha deveria ser presidida cumulativamente por três critérios fundamentais: a) decibilidade; b) capacidade gerativa e c) complexidade, desde que os modelos comparados compartilhassem pelo menos as mesmas assunções relativas ao escopo da sintaxe (que parece ser o caso das abordagens aqui adotadas). Pelo critério da decidibilidade, modelos determinísticos seriam preferidos aos modelos não-determinísticos. Pelo critério da capacidade gerativa, os modelos deveriam ser poderosos o suficiente para contemplar todos os casos de gramaticalidade previstos na língua e fracos o bastante para não contemplar também casos de agramaticalidade (ou seja: o modelo deveria

apenas gerar, e não sobregerar). Pelo critério da complexidade, seriam mais interessantes modelos que conduzissem à elaboração dos algoritmos de menor custo (ou seja, cujo processamento envolvesse a menor necessidade de tempo e espaço).

A esses critérios, podem-se somar outros dois igualmente pertinentes: d) custo de desenvolvimento e e) capacidade de reutilização de recursos. No primeiro caso, devemos sempre preferir modelos gramaticais cuja especificação possa ser feita no menor tempo pelo menor número de pessoas (o que, necessariamente, implica a economia do volume de anotações que se deseja imprimir aos itens lexicais ou às regras gramaticais). No segundo caso, é sempre conveniente que a gramática não seja restrita à aplicação e que possa ser capaz de gerar e analisar sentenças de domínios irrestritos.

Esses cinco pontos constituem um roteiro de comparação bastante interessante, que deveria ser utilizado na eleição do modelo gramatical. No entanto, pelo menos um desses critérios, o critério (c), torna já necessária a consideração do processador sintático (*parser*), a ferramenta que, fazendo uso da gramática, procede à recuperação da estrutura sintática da sentença. Essa relação não é, porém, direta e responde a diferentes questões, que não serão aqui consideradas. Entretanto, muitas delas são indicadas quando tentamos eleger um modelo de implementação de um protótipo do ReGra, como veremos a seguir.

## 5. Especificação de traços semânticos mínimos e de procedimentos semânticos para um modelo computacional de integração entre sintaxe e semântica no ReGra

Apontamos, na seção anterior, vários modelos de organização minimamente ambivalente do repertório de itens lexicais, para entender mais profundamente as razões de inadequação do ReGra e para selecionar casos que pudessem ser contemplados na construção de um protótipo de uma nova versão do revisor.

Evidenciamos, ainda, a necessidade de considerar, para uma revisão gramatical mais elaborada, não só o processamento sintático da língua portuguesa, como também a utilização de informações de cunho semântico que tornem o processo sintático mais informativo e, assim, a revisão mais refinada. A inclusão desse tipo de informação, por sua vez, acarreta a necessidade de se avaliar e modificar os processos envolvidos no *parsing*, quer pela inclusão, exclusão ou alteração de módulos de processamento ou de dados já existentes e, logo, já considerados na versão atual da ferramenta. Notadamente, isto remete à necessidade de revisão, já mencionada, da independência do módulo sintático em relação aos demais componentes do ReGra, tendo, como conseqüência, a revisão da própria natureza das regras gramaticais e das restrições de computabilidade.

Com base no enfoque assinalado, de interdependência entre diversos fatores que afetam o desempenho da ferramenta, para uma possível alteração do *parser* do ReGra foram adotados métodos analíticos e estatísticos, visando a seleção e manipulação de informações semânticas que levassem à otimização dos processos correspondentes. Consideramos, em particular, a recuperação de escolhas sintáticas e a análise da ordem de processamento, a partir dos problemas apontados nas seções 2 e 3. Descrevemos, a seguir, uma proposta para a integração entre sintaxe e semântica durante o processamento do ReGra, com base nas abordagens fundamentais e estatísticas.

### 5.1. Investigação fundamental

Paralelamente à análise dos dados (cf. ilustrada na Seção 2 e sintetizada na Seção 3), os estudos fundamentais tomaram como base a subcategorização lexical, com o objetivo de a) aprofundar o entendimento da natureza do problema, b) buscar alternativas de representação lexical; c) verificar mecanismos para se traçar (ou corroborar) as relações semânticas interlexicais; d) verificar como integrar significado e contexto, pela investigação da relação

entre palavras, seu significado e seu uso. Referimo-nos, especialmente, 1) à crítica do comportamento do *parser*, que implica a necessidade de crítica da própria revisão gramatical; e 2) às definições semânticas que se fazem necessárias, estas implicando a consideração de uma ontologia vinculada ao léxico.

Quanto a (1), a análise realizada para os casos do uso da forma 'se' no português (Martins et al., 1999), p.ex., sugeriu subsídios valiosos para tal crítica, pois o 'se' possui diversas classificações (substantivo, pronome, parte integrante do verbo, partícula de realce, partícula apassivadora, índice de indeterminação do sujeito ou conjunção) que interferem na determinação de sua função sintática na sentença, podendo, inclusive, não ocupar função sintática alguma. Exploraram-se aqui diversas teorias semânticas (p.ex., Abrahão e Lima, 1996; Katz and Fodor, 1963; Marcus, 1980; Marrafa, 1996; Pustejovsky and Boguraev, 1996; Silva e Lima, 1996) e métodos computacionais, particularmente os voltados para o *parsing* (p.ex., Beesley and Grefenstette, 1996; Church and Mercer, 1993; Lucchesi and Kowaltowski, 1993; Pacheco et al., 1996). Esse estudo levou à sugestão do modelo de subcategorização da Gramática Léxico-Funcional (Kaplan and Bresnan, 1982), para a inserção e manipulação de conhecimento semântico no ReGra. Como vimos na seção anterior, nesse nível de granularidade contemplamos traços de categorias (*features*) que podem ser representados por pares atributo-valor. Veremos, mais adiante, que tais pares podem ser expressos por estruturas predicado-argumento (P-A), cujo modelo faz do verbo o elemento responsável por determinar seus argumentos com base em seus próprios traços semânticos.

Quanto a (2), i.e., às definições semânticas dos itens lexicais, assim como a uma possível representação ontológica vinculada ao léxico, a investigação fundamental remeteu a obras de referência no campo da lingüística e da psicolingüística, merecendo destaque a Teoria de Traços (Katz and Fodor, 1963), o traçado das relações semânticas interlexicais (Fillmore, 1968; Lyons, 1977) e alguns modelos de integração entre significado e contexto, como, p.ex., (Jackendoff, 1990; Hirsh-Pasek et al., 1993). Especialmente no que diz respeito à representação ontológica, a idéia seria investigar a adequabilidade de uma ontologia para complementar decisões durante o *parsing* e, assim, servir como um recurso adicional de suporte às escolhas lexicais, para a desambigüização categorial.

Averiguamos, ainda, como os sistemas de representação do significado, ou mesmo os sistemas de *parsing*, tratam os problemas de representação léxico-semântica (p.ex., Atkins and Zampolli, 1994; Pustejovsky, 1995; Pustejovsky and Boguraev, 1996), de representação do conhecimento e desenvolvimento de ontologias para representação de conceitos do mundo (p.ex., Bateman et al., 1990; Viegas and Raskin, 1998) e de técnicas de *parsing* (Bunt and Tomita, 1996). Merecem destaque, nesse contexto, o modelo do léxico gerativo de Pustejovsky, cujo material semântico é tomado como tema central, para a "computabilidade do significado", e o modelo ontológico proposto no Projeto Mikrokosmos (Viegas and Raskin, 1998).

Pustejovsky relaciona 'semanticalidade' a 'gramaticalidade', postulando a capacidade humana de julgar tanto unidades de significado quanto unidades isentas de significado, ou seja, de tratar estruturas mentais, mesmo quando há ausência de estruturas sintáticas válidas. Pustejovsky admite, assim, a inexistência de um léxico enumerativo, imputando o sucesso do processamento do significado a estruturas semânticas cujas informações são agrupadas, organizadas ou até herdadas de modo a permitir que o ser humano produza e interprete a significação dos objetos e fatos do mundo. A partir dessa hipótese, seria possível definir um mecanismo rico e expressivo por meio do qual os diversos sentidos seriam construídos com base em seus diversos contextos. No entanto, considerar essa visão de léxico gerativo no contexto do ReGra significa admitir uma profunda reestruturação do léxico atual, bem como do seu *parser*, pois a semântica da sentença emergeria dos nomes e esses deveriam ser exaustivamente especificados com base num amplo estudo da sua estruturação sintática que, por sua vez, seria determinada pelo uso. Seria necessário, por exemplo, especificar as estruturas

de argumentos, de eventos, as estruturas *qualia* e, ainda, as estruturas de herança lexical, para dar conta do modelo de Pustejovsky. Esse estudo, por sua vez, não poderia prescindir do estudo de corpora, os quais deveriam ser marcados com as informações sintáticas subjacentes, para permitir a apreensão dos valores semânticos dos itens lexicais correspondentes. Desse modo, essa proposta se torna de difícil implementação, considerando a arquitetura atual e os objetivos do ReGra.

Opostamente à proposta de Pustejovsky, o modelo de Jackendoff (1983; 1990), que também sugere a formalização do processamento do significado e aborda a questão da ambigüidade categorial, consiste em um modelo mais simples, passível de incorporação ao protótipo pretendido. Optamos, neste caso, por incorporar desse modelo de subcategorização somente as características que pudessem direcionar a implementação do componente semântico.

O modelo ontológico proposto no Projeto Mikrokosmos, de cunho implementacional, tem uma boa fundamentação teórica, tratando dos problemas da significação relacionados à semântica lexical (representação semântica, organização do léxico, resolução da ambigüidade) em um ambiente de tradução automática e efetivamente encontrando uma saída computacional para essa questão. Por essa razão, demonstrou ser um importante parâmetro para as nossas investigações. Além disso, nesse modelo convergem as idéias de Pustejovsky, assim como as de Fillmore (1968). A linha-mestra da representação do conhecimento necessária ao projeto está na construção da base ontológica que, combinada ao léxico, determina o processamento semântico e, conseqüentemente, a resolução de boa parte dos problemas de ambigüidade, tanto estrutural quanto lexical. Essa alternativa parece apontar os elementos necessários para a incorporação de mecanismos léxico-semânticos ao ReGra. No entanto, ela envolve a complexidade de definição e representação do conhecimento, além da complexidade de inserção e manipulação de tal conhecimento pelo ReGra. Desse modo, adotamos os modelos fundamentais de subcategorização de Jackendoff e Fillmore, que podem ser satisfatoriamente representados pela gramática léxico-funcional, não levando em conta, no momento, qualquer especificação explícita de natureza ontológica, questão esta a ser investigada com mais profundidade no futuro.

Para incorporar o modelo de subcategorização do português ao ReGra, procedemos à prototipagem: foi realizada a seleção de casos particulares de inadequação, cuja descrição se encontra na subseção 5.2. A partir dessa seleção, foi realizada a especificação preliminar dos traços semânticos mínimos necessários para a integração entre sintaxe e semântica, cuja base metodológica é descrita na subseção 5.3.

Os exemplos ilustrados nas seções anteriores deste relatório não foram contemplados em sua íntegra, muito embora espelhem os tipos de problemas que ora se apresentam no ReGra. Optou-se, ao contrário, por um método de desenvolvimento "nuclear" durante a prototipagem: buscou-se a crítica do funcionamento do protótipo com base nos casos escolhidos, a partir da qual a estratégia poderia se repetir, expandindo gradativamente a potencialidade de representação e resolução do protótipo, para contemplar, no futuro, potencial de manipulação lingüística similar ao da versão atual do ReGra. O modelo do protótipo, assim como a adequação dos casos selecionados para o funcionamento do mesmo, são apresentados na subseção 5.5.

### 5.2. Método estatístico para a seleção de casos de inadequação do ReGra

Segundo o modelo de subcategorização proposto, foram realizadas, novamente, a coleta e análise de sentenças, visando somente as construções envolvendo os verbos plenos de sentido, i.e., aqueles pertencentes ao domínio das palavras de classe aberta do português. Utilizamos o WordSmith (http://www.liv.ac.uk/~ms2928/wordsmit.html) para as medidas estatísticas sobre o

corpus do NILC, de textos corretos (aproximadamente, 37 milhões de palavras)[9]. Buscamos, primeiramente, a distribuição de freqüência e a ocorrência no contexto de alguns verbos plenos de sentido que pudessem ser computacionalmente tratáveis. A escolha do verbo como elemento central deste trabalho está ligada ao fato de que o verbo é o mais representativo predicador (P) da língua, ou seja, o responsável por determinar seus argumentos (As) através de traços semânticos.

Com base nessa distribuição, escolhemos exemplos de ocorrências do ReGra para compor o "conjunto de prototipagem", de construções gramaticais problemáticas. A seguir, foi realizamos um levantamento de suas propriedades sintático-semânticas, a fim de identificar seus traços argumentais mais ocorrentes, segundo a Teoria de Valência descrita por Borba (1996)[10]. As combinações argumentais dos verbos se basearam no dicionário de Borba (1990), o único dicionário disponível, para a língua portuguesa, que apresenta os verbos descritos segundo uma teoria sintático-semântica.

Selecionamos, primeiramente, os verbos *ferir* e *falar* nos seguintes contextos e na seguinte estrutura sintática:

**Contextos:**

A espada feriu fundo.

A menina fala alto.

**Estrutura sintática:**

Suj + VI + [ Advérbio] (adjetivo "adverbalizado")

A primeira sentença já foi discutida anteriormente (Seções 2 e 3). Assim como este, no segundo caso o revisor sugere que haja concordância de gênero entre o sujeito e seu predicativo: *A espada feriu funda./A menina fala alta.*

Como resultado do *parsing*, são atribuídas as seguintes estruturas sintáticas durante a análise de cada uma dessas sentenças:

**Estrutura 1:**

[periodo] = <a espada feriu fundo>
[periodo_simples] = <a espada feriu fundo>
[sujeito_simples] = <a espada>
[adj_adn_esq] = <a>
[predicado] = <feriu fundo>
[nucleo] = <feriu>
[principal] = <feriu>
[pred_sujeito] = <fundo>

**Estrutura 2:**

[periodo] = <a menina fala alto>
[periodo_simples] = <a menina fala alto>
[sujeito_simples] = <a menina>
[adj_adn_esq] = <a>
[predicado] = <fala alto>
[nucleo] = <fala>
[principal] = <fala>
[pred_sujeito] = <alto>

[9] Disponível no site (http://cgi.portugues.mct.pt/acesso/), por meio do *IMS corpus query tools*, da Universidade de Stuttgart.

[10] Outras obras de referência incluem, p.ex., (Almeida, 1995; Borba, 1991; 1996; Dubois et al., 1988).

Diante dessa análise, nota-se que, para o revisor, os itens 'fundo' e 'alto' são adjetivos e, conseqüentemente, realizam a função sintática de predicativo do sujeito. Isso acontece porque adjetivos tais como os ilustrados, ao incidirem diretamente sobre o verbo, têm, em função adverbial, a mesma forma que o adjetivo:

| Como advérbio: | Como adjetivo: |
|---|---|
| feriu *fundo* | poço *fundo* |
| fala *alto* | moço *alto* |

Uma vez identificados como adjetivos e, portanto, como predicativos do sujeito, o revisor acusa a necessidade da concordância em gênero entre adjetivo e sujeito, produzindo o falso erro já discutido anteriormente. Neste caso, a escolha determinística da regra [Suj+VI+PredSuj] exclui, *a priori*, qualquer consideração posterior de outras regras igualmente aceitáveis para os mesmos contextos de análise. Entretanto, a regra adequada seria a que assume a categoria de advérbio, ou circunstante de modo, para 'fundo', cuja seleção somente se faz possível ao se fornecer informações suficientes para a desambigüização lexical em tempo real de revisão automática. O método analítico descrito a seguir busca essa especificação mais refinada.

### 5.3. Método analítico para a especificação semântica do protótipo

A Teoria de Valência ilustrada acima une as propostas sintáticas de Harris e Tesnière à gramática de casos, de Fillmore. Dessa união, ocorre a adoção de uma teoria de predicados, ou teoria argumentativa, amplamente explorada na lingüística computacional como forma de representação do conhecimento profundo das línguas naturais. Essa teoria se apóia em três princípios básicos, como seguem: (i) na estrutura semântica das línguas naturais só há duas classes de unidade: predicado e argumento; (ii) na superfície, elas se fazem representar por expressões predicativas ou por expressões argumentativas; (iii) as combinações de predicado e de argumento formam estruturas P–A simples e complexas, cuja representação superficial pode assumir expressões de diferentes formas. Isto quer dizer que uma mesma estrutura P–A pode ser representada, na superfície, por um conjunto não unitário de estruturas formais ou mórficas, cujo elo central, que estabelece as relações fundamentais com seus argumentos, é o verbo. A valência[11] corresponde, assim, ao número de argumentos implicado pelo significado de um item lexical; o argumento, a cada um de seus traços pertinentes à sua especificação lexical.

Segundo essa teoria, é no esquema profundo, portanto, que as relações básicas ou fundamentais são definidas, tornando possível a interpretação semântica das frases, indicando a conexão e o números de constituintes, a ordem e as relações semânticas fundamentais. A combinatória de morfemas (concordância e regência), por sua vez, é definida no esquema superficial ("frase realizada" propriamente dita).

Seguem alguns exemplos de definição de estruturas P–A:

**Exemplo 1:** verbo 'ferir', acepção *ação-processo* ou *mudança de estado*: valência dois ($V_2$), i.e., dois argumentos obrigatórios; representação estrutural: *ferir(Agente,Paciente)*, como em

*A polícia feriu o assaltante durante a fuga.*
Argumento 1: polícia
Argumento 2: assaltante
Circunstante de tempo: "durante a fuga"
Estrutura P-A (simplificada): feriu(polícia, assaltante)

[11] O termo *valência* recobre o termo mais tradicional *transitividade*.

**Exemplo 2:** verbo 'ferir', acepção *processo*: verbo de valência um ($V_1$), i.e., tem um argumento obrigatório; representação estrutural: *ferir(Agente)*, como em

***A <u>vaca</u> feriu-se na cerca.***
Argumento 1: vaca
Circunstante de lugar: "na cerca"
Estrutura P-A: feriu(vaca)

Os circunstantes, segundo a Teoria de Valência, não fazem parte da valência do verbo, como indicam ambos os exemplos, pois não são indispensáveis para a interpretação semântica do conteúdo objetivo que se quer comunicar. O pronome 'se', no segundo exemplo, é chamado *índice de processividade*.

Como ilustrada, a construção da estrutura P-A na Teoria de Valência adota uma estratégia distinta, p.ex., da análise de constituintes, que decompõe o enunciado por regras de reescrita, indicando como os constituintes se hierarquizam.

A partir de informações dessa natureza, assim como dos contextos compilados pelo WordSmith, determinamos as informações semânticas para o léxico do protótipo, assim como para o *parsing*. Adotamos o seguinte procedimento:

1) A estrutura argumental dos verbos 'ferir' e 'falar' foi recuperada do dicionário de Borba (1990);
2) As tabelas correspondentes, que apresentam as valências semântica e sintática dos verbos, além dos papéis temáticos de seus argumentos e suas classes sintático-semânticas, foram geradas. A Tabela 5, p.ex., descreve a estrutura argumental de 'ferir', cujos contextos foram extraídos do corpus do NILC ou do próprio dicionário (que contém exemplos não discriminados)[12].

Segundo Borba, embora os circunstantes (de lugar, modo ou tempo) não estejam previstos nas combinações argumentais do verbo estabelecidas no esquema profundo, eles aparecem nos esquemas superficiais, especificando as condições ou circunstâncias nas quais se dá o que se indica na relação SN+SV[13] (sendo sua presença opcional). Tais circunstantes não estão sujeitos a restrições de co-ocorrência, sendo sua distribuição na sentença livre, ao contrário da distribuição dos complementos verbais, que é restrita, por serem eles elementos regentes particulares.

[12] As abreviaturas estão listadas no Apêndice 1.

[13] Sintagma nominal + sintagma verbal

Tabela 5**: Estrutura argumental do verbo 'ferir'**

| |
|---|
| **Valência semântica:** [ anim/ ação- processo/ anim ] **Sentido:** provocar chagas; contundir<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç- proc/ Ob ]<br>**Exemplos:** *1. Soldados israelenses feriram com disparos de bala de borracha nove manifestantes palestinos. (corpus)*<br>*2. A polícia feriu o assaltante durante a fuga. (corpus)* *(1)* |
| **Valência semântica:** [inanim; con/ ação- processo/ anim ] **Sentido:** provocar chagas; contundir<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [Ca/ aç-proc/ Ob]<br>**Exemplo***: Foguetes Katyusha lançados pelo Hesbolá feriram 36 civis na cidade israelense de Kiriat Shmora.(corpus)* *(2)* |
| **Valência semântica:** [ anim/ ação- processo/ inanim; con ] **Sentido:** provocar marcas em<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç- proc/ Ob ]<br>**Exemplos:** *1.O gato feria o sofá com as unhas.*<br>*2. [Paulo] desbastava cuidadoso os ramos laterais para não ferir as fibras dos nós. (VC, 47);*<br>*3. Rufino levemente feria as cordas do violão; (Borba, 753)* Sentido: tocar levemente *(3)* |
| **Valência semântica:** [ inanim; con/ ação- processo/ inanim; con ] **Sentido:** provocar marcas em<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [ Ca/ aç- proc/ Ob ]<br>**Exemplo:** *Um ramo de primavera feria o muro recém- pintado do cemitério; (Borba, 754)* *(4)* |
| **Valência semântica:** [ anim; hum/ ação- processo/ anim; hum ] **Sentido:** magoar; irritar<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç-proc/ Ob ]<br>**Exemplo:** *O deputado feriu os eleitores com um discurso provocativo. (corpus)* *(5)* |
| **Valência semântica:** [ inanim; abs/ ação- processo/ anim; hum ] **Sentido:** magoar; irritar<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [ Ca/ aç- proc/ Ob ]<br>**Exemplos:** *1. Esse verbo me feria. (Borba,754/ BA, 8)*<br>*2. O silêncio da noite me feria. (corpus)* *(6)* |
| **Valência semântica:** [ inanim; abs/ ação- processo/ inanim; abs ] **Sentido:** violar<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [ Ca/ aç- proc/ Ob ]<br>**Exemplos:** *1. A reforma agrária ( ) feriu de morte a propriedade rural (Borba, 754/Mayer- 0,4)*<br>*2. Embargo a Cuba fere livre comércio. (corpus)* *(7)* |
| **Valência semântica:** [ anim; hum/ ação- processo/ inanim; abs ] **Sentido:** violar<br>**Valência sintática:** SN (A1) + V + SN (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç- proc/ Ob ]<br>**Exemplos:** *1. Os militares, transformados em políticos, feriram a hierarquia e a disciplina*<br>*2. Todos nós ferimos a Constituição.(corpus)* *(8)* |
| **Valência semântica:** [ anim/ processo ] **Sentido:** machucar<br>**Valência sintática:** SN (A1) + V ( V+pron) **Classe do verbo e papéis temáticos:** [ Pa/ proc ]<br>SN (A1) + V (Aux+V)<br>**Exemplos:** *1. A vaca feriu-se na cerca (Borba, 754/ DM, 33)*<br>*2. Moreira César foi ferido nos Canudos (Borba, 754/ CJ, 73)*<br>*3. Uma moradora da rua Gustavo Sampaio, no Leme, também foi ferida na porta de casa. (corpus)* *(9)* |
| **Valência semântica:** [ inanim; abs/ processo ] **Sentido:** realizar-se<br>**Valência sintática:** SN (A1) + V (V+pron) **Classe do verbo e papéis temáticos:** [ Pa/ proc ]<br>**Exemplos:** *1. A discussão feriu-se ontem. (Borba, 754/ FM- 14.2.50, 14)* *(10)* |
| **Valência semântica:** [ inanim; con/ ação- processo] **Sentido:** machucar<br>**Valência sintática:** SN (A1) + V **Classe do verbo e papéis temáticos:** [ I/ aç- proc ]<br>**Exemplo:** *A espada feriu fundo. (corpus NILC)* *(11)* |

Vamos mostrar, agora, como os circunstantes são "aceitos" pelo revisor no caso das sentenças em foco: *A espada feriu fundo* e *A menina falou alto*.

Sabemos que, no esquema superficial, os circunstantes aparecem ligados ao verbo na forma de SAdv (sintagma adverbial), podendo corresponder a um adjetivo "adverbalizado", um advérbio, propriamente dito, ou um SPrep (sintagma preposicional). No entanto, dentre as várias realizações dos circunstantes de modo ligados ao verbo 'ferir', o ReGra detecta erro nas ocorrências de adjetivos adverbalizados, reconhecendo-os, sintaticamente, como adjetivos, como foi explicado anteriormente. O problema, aqui, é que o item lexical 'fundo', com categoria de adjetivo e desempenhando a função de predicativo do sujeito (PredSuj), não faz parte das combinações argumentais desse verbo, o que deveria levar o ReGra a reconhecer a

regra gramatical selecionada durante a revisão, qual seja, [Suj+VI+PredSuj], como inválida e, assim, excluí-la do rol de regras aplicáveis para a sentença sob análise.

Para a segunda sentença, *A menina falou alto*., ou para contextos sintáticos envolvendo, analogamente, o adjetivo/advérbio *grosso*, com categoria de circunstante, ocorre problema similar, no que diz respeito ao desempenho do ReGra, pois este elege inadequadamente também a regra [Suj +VI + PredSuj], acusando falso erro para tais contextos.

A Tabela 6 mostra as possíveis variações de 'alto' e 'grosso', segundo a análise lingüística com base na Teoria de Valência.

Vale salientar que os verbos 'ferir' e 'falar' foram escolhidos apenas como exemplo. Entretanto, os testes realizados até o momento indicam a possibilidade de se distinguir duas classes de verbos intransitivos: os que "aceitam" a estrutura sentencial Suj + VI + PredSuj e os que aceitam somente a estrutura Suj + VI + [Adv]. Essa distinção tornaria factível a desambigüização lexical. Porém, é prematuro afirmar que o problema se encontra, de um modo geral, na caracterização verbal, pois a raiz do problema pode estar no inter-relacionamento entre os próprios argumentos do verbo e, neste caso, estudos mais profundos se fazem necessários.

De resto, vale ressaltar, ainda, que a escolha de um conjunto de prototipagem teve, como condição primordial, o reconhecimento dos casos tratáveis computacionalmente dos não tratáveis, sendo que estes foram diretamente eliminados de nossa análise e correspondem àqueles para os quais sequer há comportamento previsível por parte do leitor humano (devido, p.ex., à dificuldade de compreensão ou, mesmo, da própria desambigüização). Dentre os casos tratáveis, o trabalho lingüístico deveria apontar aqueles cuja perspectiva de resolução fosse viável, quer pelo volume ou complexidade da especificação lingüística a ser incorporada ao revisor, quer pela complexidade computacional envolvida. A partir desses dados, passamos à descrição da especificação computacional do protótipo.

Tabela 6: **Estrutura argumental do verbo 'falar'**

| |
|---|
| **Valência semântica:** [ anim/ ação] **Sentido:** articular os sons de uma língua<br>**Valência sintática:** SN (A1) + V **Classe do verbo e papéis temáticos:** [ Ag/ aç]<br>**Exemplos:** *1. A menina fala alto. (com ou sem especificador)*<br>*2. É bom que Zagalo fale grosso com os jogadores. (corpus)* *(1)* |
| **Valência semântica:** [ anim; hum/ ação] **Sentido:** exprimir-se usando sons<br>**Valência sintática:** SN (A1) + V **Classe do verbo e papéis temáticos:** [ Ag/ aç ]<br>**Exemplo:** *Deixei o jogador falar. (corpus)* *(2)* |
| **Valência semântica:** [ anim; hum./ ação/ "a/com"+ hum.] **Sentido:** dirigir a palavra<br>**Valência sintática:** SN (A1) + V + Sprep (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Dest ]<br>**Exemplo:** *Eles falaram com o chefe. (corpus)* *(3)* |
| **Valência semântica:** [ anim; hum./ ação/ "a favor de/ contra"+ abs.] **Sentido:** argumentar<br>**Valência sintática:** SN (A1) + V + Sprep (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ]<br>**Exemplo:** *Os deputados falaram a favor da reeleição. (corpus)* *(4)* |
| **Valência semântica:** [ anim; hum./ ação/ "de/em/sobre"+ inanim] **Sentido:** discorrer<br>**Valência sintática:** SN (A1) + V + Sprep (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ]<br>**Exemplos:** *1. Falamos sobre sexo. (corpus)*<br>*2. Estava falando da guerra.*<br>*3. E já que falamos em campeonato, nada menos atraente do que este que se inicia.* *(5)* |
| **Valência semântica:** [ anim; hum./ ação/ "de/em/sobre"+anim] **Sentido:** Discorrer<br>**Valência sintática:** SN (A1) + V + Sprep (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ]<br>**Exemplo:** *Foi por isso que não falamos de Antônio Alves taxista. (corpus)* *(6)* |
| **Valência semântica:** [ anim; hum./ ação/ "de/em"+"Oinf"] **Sentido:** manifestar intenção de<br>**Valência sintática:** SN (A1) + V + Sprep (A2) **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ]<br>**Exemplo:** *E os dirigentes falam em criar uma bolsa de jogadores. (corpus)* *(7)* |

| | |
|---|---|
| **Valência semântica:** [ anim; hum./ ação/ "por"+ hum] | **Sentido:** usar a palavra em nome de |
| **Valência sintática:** SN (A1) + V + Sprep (A2) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ] |
| **Exemplo:** *O porta voz falou ontem pelo presidente. (corpus)* | *(8)* |
| **Valência semântica:** [ anim; hum./ ação/ inanim] | **Sentido:** dizer |
| **Valência sintática:** SN (A1) + V + SN (pronome) (A2) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ] |
| **Exemplo:** *Ninguém fala isso. (corpus)* | *(9)* |
| **Valência semântica:** [ anim; hum./ ação/ "Oconj"] | **Sentido:** dizer |
| **Valência sintática:** SN (A1) + V + Oconj (A2) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ] |
| **Exemplo:** *Falaram que o processo não daria nada. (corpus)* | *(10)* |
| **Valência semântica:** [ anim; hum./ ação/ "DD"] | **Sentido:** dizer |
| **Valência sintática:** SN (A1) + V + O (A2) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ] |
| **Exemplo:** *Falei: Obrigado. (corpus)* | *(11)* |
| **Valência semântica:** [ anim; hum./ ação/ inanim/"para/ao"+hum] | **Sentido:** dizer |
| **Valência sintática:** SN (A1) + V + SN (A2) + Sprep (A3) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob/ Dest ] |
| **Exemplo:** *Não falaram nada para os companheiros. (corpus)* | *(12)* |
| **Valência semântica:** [ anim; hum./ ação/ "Oconj."/"para/ao"+hum] | **Sentido:** dizer |
| **Valência sintática:** SN (A1) + V + Oconj (A2) + Sprep (A3) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob/ Dest ] |
| **Exemplo:** *Eu só falei para ele que a bala sobe e desce. (corpus)* | *(13)* |
| **Valência semântica:** [ anim; hum./ ação/ "DD"/"para/ao"+hum] | **Sentido:** dizer |
| **Valência sintática:** SN (A1) + V + O (A2) + Sprep (A3) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob/ Dest ] |
| **Exemplo:** *Falaram aos estudantes: _ Estamos em greve! (corpus)* | *(14)* |
| **Valência semântica:** [ anim; hum./ ação/ "para"+ "Oinf."] | **Sentido:** Dizer |
| **Valência sintática:** SN (A1) + V + Sprep ( prep. + O/Inf) (A2) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob ] |
| **Exemplo:** *Zé falou para levar você daqui. (corpus)* | *(15)* |
| **Valência semântica:** [ anim; hum./ ação/ "Oconj.inf"/ "com"+hum] | **Sentido:** Conversar |
| **Valência sintática:** SN (A1) + V + Oconj.inf (A2) + Sprep. (A3) | **Classe do verbo e papéis temáticos:** [ Ag/ aç/ Ob/ Dest ] |
| **Exemplo:** *Falei com o pai da moça que ia casar no fim do ano. (corpus)* | *(16)* |
| **Valência semântica:** [ anim./ação/ inanim] | **Sentido:** Tornar público |
| **Valência sintática:** SN (A1) + V + SN (A2) | **Classe do verbo e papéis temáticos:** [ Ag/ aç-proc/ Ob ] |
| **Exemplo:** *Ela falou o nome do rapaz sete vezes. (corpus)* | *(17)* |
| **Valência semântica:** [inanim./ação- processo/ "a"+hum] | **Sentido:** Impressionar |
| **Valência sintática:** SN (A1) + V + Sprep (A2) | **Classe do verbo e papéis temáticos:** [ Ca/ aç-proc/ Dest ] |
| **Exemplo:** *A honestidade fala à todos. (corpus)* | *(18)* |
| **Valência semântica:** [ inanim./estado] | **Sentido:** Ter valor |
| **Valência sintática:** SN (A1) + V | **Classe do verbo e papéis temáticos:** [ Ex/estado] |
| **Exemplo:** *Os gestos da menina falaram por si. (corpus)* | *(19)* |
| **Valência semântica:** [anim; hum/ estado/ "designativo de língua"] | **Sentido:** Ter capacidade |
| **Valência sintática:** SN (A1) + V+ SN (A2) | **Classe do verbo e papéis temáticos:** [ Ex /estado/ Ob ] |
| **Exemplo:** *Ele fala vários idiomas.(corpus)* | *(20)* |

### 5.4. A forma de representação semântica no protótipo

O dicionário de verbos utilizado descreve, em geral, os verbos sob a forma de classes sintático-semânticas e seus argumentos sob as formas de casos (agentivo, experimentador, beneficiário, objetivo, locativo, instrumental, causativo, meta, origem, resultativo, temporal, comitativo) e traços semânticos. Os casos representam papéis ou funções temáticas resultantes da relação que se processa ao associar as estruturas conceituais de um verbo e de um nome. O seguinte verbete do dicionário ilustra tal descrição (Borba, 1990, p. 753):

> FERIR: I. Indica ação-processo. 1. Com sujeito **agente** expresso por nome **animado** ou com sujeito **causativo** expresso por nome **concreto não-animado.** (…)

Segundo a Teoria de Valência, é somente sob as formas de classificação de casos e classes sintático-semânticas que a dinâmica das relações lexicais pode ser contemplada. Entretanto, na resolução computacional, utilizou-se apenas os traços semânticos dos itens lexicais (substantivos, adjetivos) que, na estrutura argumental, constituem os traços dos argumentos do verbo e a estrutura argumental dos verbos, excluindo-se a representação explícita dos *casos* (que seriam correspondentes a cada termo da própria estrutura argumental). Por exemplo, ao item lexical 'espada' foram relacionados os traços semânticos "inanimado|concreto", sendo que estes, por sua vez, constituem um dos argumentos do verbo 'ferir' em uma de suas combinações argumentais: "inanimado|concreto, processo". A combinação argumental para termos como 'espada', por sua vez (*A espada feriu fundo.*), poderia ser representada sob a forma de casos, como, p.ex., "instrumental, processo".

Já a incorporação das classes sintático-semânticas dos verbos (ação, ação-processo, processo e estado) em detrimento de uma subcategorização em traços semânticos, como "atividade, evento", entre outros, não acarreta implicações sérias ao processamento, uma vez que os responsáveis pela desambigüização dos itens lexicais são os traços semânticos dos argumentos, nesse modelo. Cabe ressaltar também que, apesar de não estar descrita no dicionário, a valência sintática dos verbos foi estudada e incluída nas tabelas ilustradas para que as classes gramaticais que preenchem os argumentos pudessem ser analisadas, servindo de apoio para a descrição semântica.

A partir da Teoria de Valência, adotamos a seguinte nomenclatura para a representação semântica do protótipo:

| **Para os traços semânticos** | **Para as classes sintático-semânticas** |
|---|---|
| concreto | ação |
| abstrato | ação-processo |
| animado | processo |
| inanimado | estado |
| humano | |
| não humano | |
| modo (para advérbios de "modo") | |
| lugar | |
| serve para cortar | |

Os verbetes do léxico, acrescidos de informações semânticas, para a sentença-problema *A espada feriu fundo*. são ilustrados abaixo:

'espada': inanimado|concreto|serve para cortar

'fundo' (como adjetivo): inanimado| abstrato| serve para cortar
'fundo' (como substantivo): inanimado|concreto|lugar
'fundo' (como advérbio): inanimado|abstrato |modo

'ferir': animado, ação|processo, animado
'ferir': animado, açã|processo, animado
'ferir': animado|humano, ação|processo, animado|humano
'ferir': inanimado| abstrato, ação|processo, animado|humano
'ferir': inanimado| abstrato, ação|processo, inanimado|abstrato
'ferir': animado|humano, ação|processo, inanimado| abstrato
'ferir': animado, processo
'ferir': inanimado| concreto, processo

Das onze possíveis combinações argumentais regidas pelo verbo 'ferir' (cf. Tabela 5), apenas oito foram implementadas no protótipo. As três combinações descartadas nesta etapa de testes foram: (3) [anim/ ação- processo/ inanim; con]; (4) [inanim; con/ ação- processo/ inanim;

con] e (10) [inanim;abs/ processo]. A não incorporação dessas combinações argumentais se deveu ao fato de não haverem sido encontradas ocorrências representativas dessas acepções no corpus do NILC.

### 5.5. Método computacional sugerido para a prototipagem

A prototipagem consistiu, basicamente, da incorporação das informações semânticas ao léxico delimitado a partir do léxico original do revisor e da implementação do módulo de processamento semântico, a ser realizado em conjunto com o *parser* (analisador sintático) atual do ReGra, com base nos exemplos. O protótipo mantém a arquitetura original inalterada, como ilustra a Figura 2.

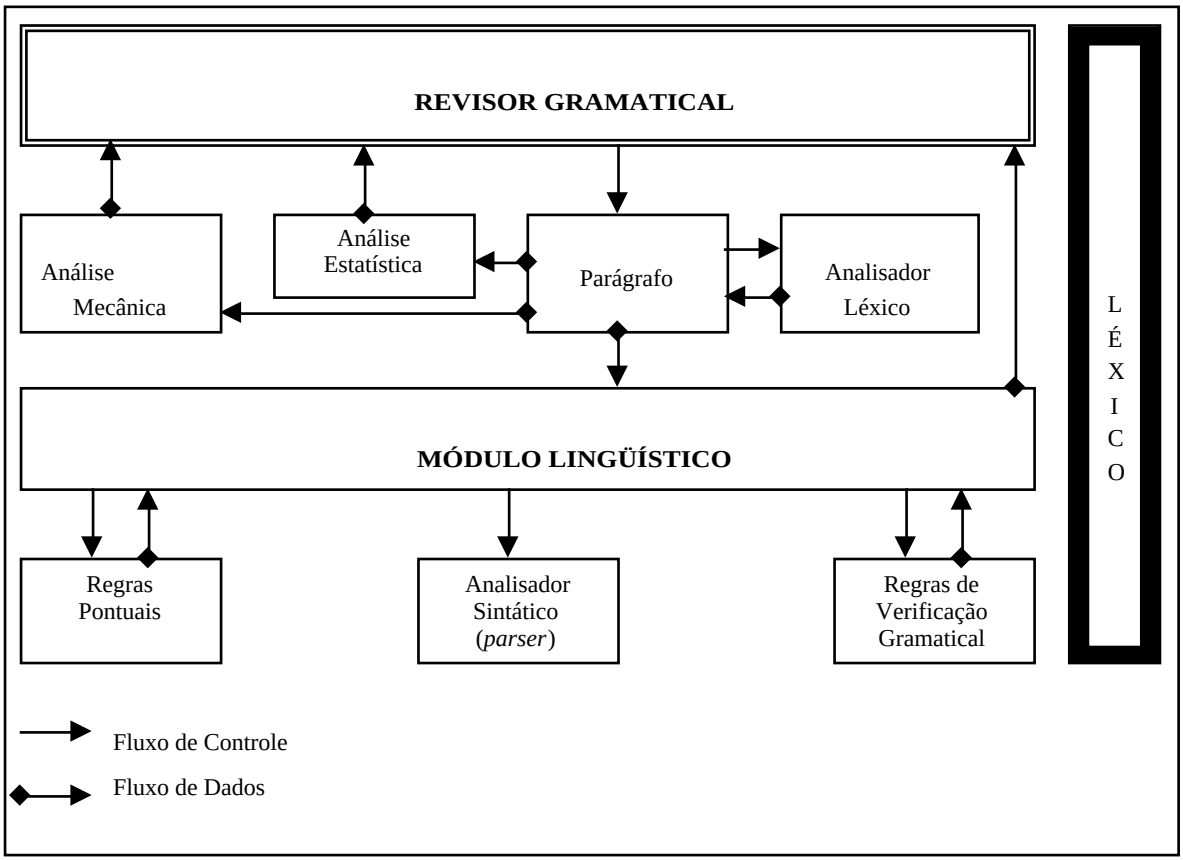


**Figura 2: Arquitetura do ReGra**

Segundo essa arquitetura, a revisão gramatical é realizada na seguinte seqüência[14]: 1) análise léxica; 2) análise estatística; 3) análise mecânica; 4) aplicação de regras pontuais e, finalmente, 5) análise sintática e verificação gramatical. Para o protótipo, o léxico é alterado em seu conteúdo, pelo acréscimo dos traços semânticos aos itens lexicais do conjunto de prototipagem (referentes às Tabelas 5 e 6). O principal módulo de processamento, o analisador sintático, é refinado pela incorporação do módulo semântico, a qual foi elaborada em seis fases, a maioria delas já descritas:

- Coleta e análise de sentenças contendo verbos plenos de sentido;
- Seleção das sentenças de interesse, com base em medidas estatísticas dos verbos a serem tratados;
- Determinação das respectivas estruturas sentenciais a serem manipuladas;

[14] Para detalhes sobre o sistema, refira-se a (Martins et al., 1998).

- Identificação dos traços semânticos relevantes para a representação lexical, segundo a Teoria de Valência;
- Atualização das entradas lexicais de interesse, pela introdução das características semânticas correspondentes;
- Especificação e desenvolvimento do protótipo.

Para a prototipagem, decidimo-nos pela implementação de *demons* (programas disparadores de outras rotinas, cf. Tanenbaum, 1995). Esse procedimento é similar ao sugerido por Viegas and Raskin (1998). Para sentenças do tipo S-V-O, p.ex., o processo baseado na estrutura argumental do verbo deve se dar do seguinte modo:

1) A análise de uma sentença deve ser realizada da esquerda para a direita (análise *left-to-right*, ou LR);
2) Qualquer tipo de processamento semântico deve ser retardado até que o verbo principal da oração seja encontrado. Neste caso, informações sobre os componentes sentenciais já processados devem ser reservadas para processamento posterior, incluindo-se, aqui, informações dicionarizadas já recuperadas;
3) Ao se recuperar as informações dicionarizadas do verbo principal, um *demon* verificará se sua estrutura argumental é passível de preenchimento pelos componentes já processados, indicando a continuidade do *parsing* para a ativação de *demons* subseqüentes. Desse modo, cada *demon* guiará a determinação das regras de *parsing* aplicáveis e das acepções lexicais apropriadas.

Esse procedimento sugere *parsing* e análise semântica aninhados (*interleaved*), como mostra a Figura 3.

Considerando verbos plenos de sentido, a implementação de *demons* para o desenvolvimento do protótipo pareceu-nos adequada para testes de viabilidade e eficiência do novo sistema: *demons* diversos, para verificação e complementação semântica, foram introduzidos em regras sintáticas (ou ATNs) específicas, que contemplam os exemplos selecionados para a prototipagem. Assim, durante a aplicação de uma regra gramatical, *demons* ficam responsáveis por buscar uma correspondente semântica que satisfaça as restrições da estrutura argumental do verbo. Neste caso, a revisão gramatical de *A espada feriu fundo.*, que, atualmente, é realizada com sucesso pela regra [Suj + VI + PredSuj] (muito embora produzindo um falso erro), passa a não ser mais licenciada pelo protótipo, pois a existência de *demons* de verificação semântica inibe a aplicação dessa regra, buscando a próxima que seja aplicável. A próxima regra, no elenco de regras aplicáveis do *parser*, é [Suj + VI + ADV [opcional]][15] que, agora, admite a compatibilidade semântica entre os argumentos verbais. Caso análogo ocorre para a outra sentenção escolhida, *A menina fala alto.*

[15] [Sujeito + Verbo Intransitivo + [Adjunto Adverbial Opcional]]

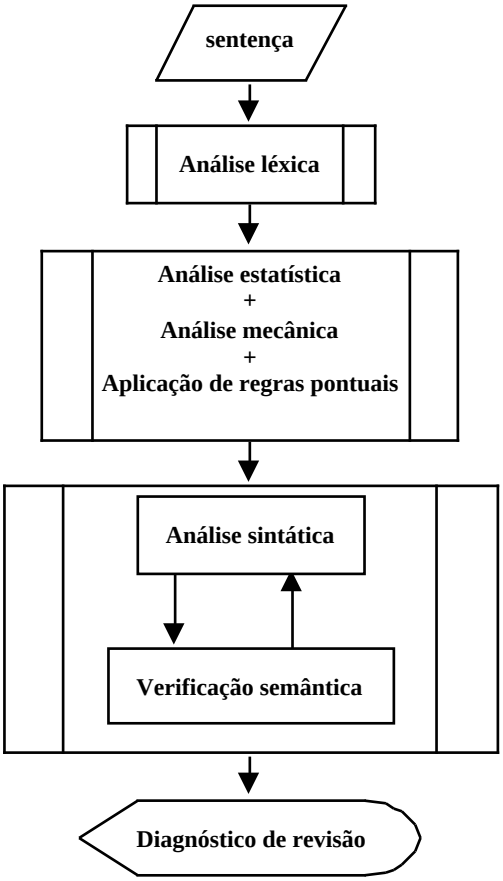


**Figura 3: Arquitetura do protótipo de revisão gramatical**

De um modo geral, para orações simples do tipo S-V-O, a análise pode ser realizada com a ativação de um *demon* após o reconhecimento do verbo principal e após o reconhecimento de um marcador de fim de sentença, como ilustra a Figura 4. O primeiro *demon* restringe o número de possíveis sentidos do verbo de acordo com o contexto, através do preenchimento de sua estrutura argumental em função das informações do sujeito. Já a esta altura, é possível que o *parser* realize uma seleção mais apurada das regras sintáticas aplicáveis à oração em questão. O segundo *demon* completa a especificação argumental do verbo com as informações semânticas relativas ao objeto da oração, ao fim do *parsing*.

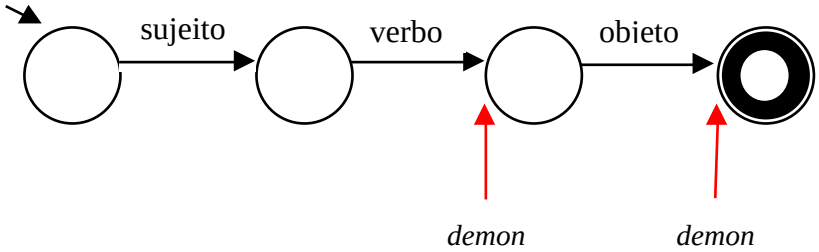


**Figura 4: Inserção de *demons* em uma ATN do *parser***

O caso de *A espada feriu fundo.* difere desse, por não caracterizar S-V-O, mas sim a ocorrência de um verbo intransitivo. No entanto, *demons* em posições análogas são descritos: após a análise do verbo, um *demon* é acionado agora, que utiliza as informações semânticas de 'espada', recuperadas na fase de análise léxica, para restringir o número de possíveis combinações argumentais do verbo 'ferir'. Desse modo, somente são consideradas as combinações argumentais que apresentam o primeiro argumento com o conjunto de traços [-anim; +con; serve para cortar] ou com uma representação parcial desse conjunto. São possíveis,

assim, as seguintes combinações de traços, retidas para verificação pelo *demon* em questão[16]: a) [-anim; +con/ ação-processo/+ anim]; b) [-anim; +con/ processo]. Em seguida, o segundo *demon*, ao fim da sentença, é acionado, como no exemplo anterior.

Em relação à análise lexical, agora a primeira ocorrência de 'fundo', que corresponde à categoria de *adjetivo*, embora sintaticamente coerente para qualquer uma das duas regras inicialmente disponíveis no revisor, agora é rechaçada, pois não mais se aplicar à valência do verbo 'ferir': os traços semânticos de 'fundo', como adjetivo, são [-anim; +abs; atributo]. Assim, a regra [Suj + VI + PredSuj] deixa de ser validada semanticamente. A próxima regra ativada, i.e., [Suj + VI + ADV [opcional]], permite preencher os traços argumentais do verbo, devido aos traços semânticos de 'fundo', que são, agora, [inanimado; abstrato; modo]. Vale notar que a inexistência de uma combinação explícita, de traços [-anim; +con/ processo/ inanimado; abstrato; modo], não inviabiliza, desta vez, a aceitação dessa regra sintática, pois, como foi enfatizado, a presença do advérbio é prevista como *opcional.* As informações semânticas do léxico, nesse caso, guiam o revisor para a aceitação da regra, promovendo a desambigüização do item lexical 'fundo'.

Várias considerações sobre a proposta de especificação dos traços semânticos dos itens lexicais visando o aprimoramento do ReGra com base nesse protótipo são discutidas a seguir.

## 6. A integração entre sintaxe e semântica no ReGra: dificuldades encontradas

Conforme vimos, a decomposição do sentido lexical em traços semânticos, para a revisão gramatical, remete a problemas complexos, lingüísticos ou computacionais, exigindo uma investigação criteriosa para a escolha de um modelo gramatical que dê conta da tarefa, conforme a complexidade apresentada. No caso em foco no Projeto TraSem, a questão é mais restrita, já que temos um aplicativo em funcionamento que impõe severas restrições sobre as abordagens lingüístico-computacionais. Devido a isso, procuramos eleger uma configuração do protótipo que permitisse abordar (ou manter) as questões mais cruciais apontadas na Seção 4, quais sejam: a) preservar o modelo determinístico do ReGra (visando a decibilidade); b) melhorar sua capacidade gerativa, para contemplar tanto os casos previstos na língua portuguesa quanto os casos agramaticais que não estão sendo adequadamente tratados, no momento; c) garantir que as soluções propostas não atinjam custos proibitivos, mantendo sua complexidade controlada; d) controlar, ainda, o custo de aprimoramento do revisor, em especial, no que diz respeito ao projeto e desenvolvimento dos módulos lexical e sintático-semântico.

Em relação à pesquisa lingüística, um dos grandes desafios da semântica lexical está na elaboração de categorias ontológicas que é, sabidamente, importante para o desenvolvimento de aplicativos de PLN e, para o ReGra, em particular. Como uma fonte de informação altamente refinada, mais rica e estruturada, do conteúdo lexical, poderia levar ao aprimoramento desejado da ferramenta. Muito embora essa iniciativa pudesse contribuir para a melhora de sua capacidade gerativa, o custo de seu desenvolvimento provou ser alto demais para o atual estágio do projeto, conforme já relatamos, pois qualquer iniciativa dessa natureza requer um trabalho profundo sobre a consistência dos dados, a especificação dos processos correspondentes e a análise de desempenho do ReGra. De outro modo, a ontologização não será capaz de fornecer generalizações relevantes para a eliminação de falsos erros ou falsos acertos na revisão gramatical. Entretanto, será possível, no futuro, voltar a essa questão, se considerarmos que vários recursos lingüísticos adicionais, atualmente em desenvolvimento no NILC, poderão ser reutilizados visando a representação ontológica necessária ao Projeto TraSem. Em especial, teremos a Base de Dados Lexicais do NILC (Greghi, 2001), assim como o *thesaurus* da língua portuguesa, ambos propostos como parte deste mesmo Projeto FINEP

[16] Ver abreviaturas no Apêndice 1 e ilustrações semelhantes na Tabela 5.

(Nunes et al., 1999). Além desses, há também o dicionário Português-UNL utilizado no Projeto UNL-Brasil (Dias-da-Silva et al., 1998), de cunho semântico, embora bilíngüe.

Com relação à tarefa de inclusão de traços semânticos no léxico do ReGra, encontramos vários obstáculos. A irregularidade das descrições semânticas do dicionário de Borba, p.ex., acaba por prejudicar a especificação necessária para a incorporação de traços semânticos aos itens lexicais, assim como da estrutura argumental dos verbos ao processamento computacional. Como única fonte da estrutura argumental dos verbos para o português contemporâneo, as descrições dicionarizadas ocorrem, na maioria das vezes, sob a forma de casos (agente, causativo, etc.); raras vezes sob a forma da valência sintática (predicativo, etc), que poderia ser igualmente relevante para tratar casos para os quais a resolução semântica constitui um passo mais complexo do que o necessário. Devido a essa irregularidade e à inexistência de outras fontes lingüísticas, a descrição semântica dos argumentos verbais resulta altamente *ad-hoc*, ficando sob responsabilidade da interpretação de especialistas lingüistas, o que leva à impossibilidade de se assegurar uma especificação criteriosa. Além disso, o esforço despendido nessa tarefa acaba por completar o rol de obstáculos para o desenvolvimento prático da ferramenta.

A resolução computacional segundo a proposta de representação do modelo de estruturas argumentais dos verbos implica, por sua vez, outros problemas de natureza computacional, tais como a necessidade de se garantir a consistência de representação de um grande volume de informações semânticas (tanto para o léxico, quanto para o *parser*) e a queda de eficiência do revisor. Ambos resultam em pior desempenho do revisor, já que a introdução do processamento semântico implica a) a necessidade de manipulação de um volume maior de dados; b) o reconhecimento, determinístico sempre que possível, das regras inadequadas para o contexto sob análise e, portanto, c) o aumento considerável do tempo de resposta do revisor. Situação mais problemática, em relação à piora de desempenho do ReGra, ocorre quando as verificações semânticas acarretam um ônus de processamento muitas vezes inútil e, portanto, injustificável. Este é o caso, particularmente, das sentenças não-ambíguas, de processamento determinístico, para as quais o ReGra já desempenha satisfatoriamente.

Visando minimizar os problemas de desempenho dessa natureza, podemos, inicialmente, considerar algumas técnicas emergenciais ou alternativas. Por exemplo, admitindo o processamento semântico determinístico, como o processamento sintático, ou admitindo sua interrupção ou desconsideração mediante a especificação de um intervalo de tempo limitante, para a produção incondicional de um diagnóstico. Neste caso, uma das seguintes razões seria atribuída ao processamento[17]: a inexistência de correspondência adequada entre sintaxe e semântica, no contexto da sentença sob análise (devido ao determinismo) ou a persistência do comportamento atual do ReGra, implicando a produção de mensagens de revisão baseadas somente nas informações léxico-sintáticas e, portanto, mantendo os diagnósticos inadequados atuais, na pior das hipóteses (devido à ultrapassagem do tempo estipulado para a obtenção de um diagnóstico).

Qualquer uma dessas estratégias provou implicar um alto custo-benefício no estágio de prototipagem, para ser adotada irrestritamente para a revisão gramatical do ReGra. Contudo, não se descarta a continuidade das pesquisas para se encontrar modelos lingüístico-computacionais que dêem conta, de modo cada vez mais satisfatório, de casos como os analisados neste trabalho. Perspectivas dessa natureza são discutidas a seguir.

[17] Lembrando que parte-se de dados não confiáveis, para os quais nem sempre é possível determinar com exatidão o problema em foco, mesmo quando há informações semânticas que auxiliem o processo de revisão.

## 7. Perspectivas de continuidade do projeto

A metodologia de especificação de traços semânticos mínimos e dos procedimentos correspondentes para a integração entre sintaxe e semântica (Seções 4-6) forneceu subsídios valiosos para a prototipagem. Entretanto, a proposta de resolução do problema deixou patente as proporções complexas de especificação e implementação, devido aos seguintes fatores: a) trata-se de um problema de difícil resolução computacional, senão também humana; b) envolve questões altamente problemáticas da Lingüística Computacional, uma vez que admitem a busca qualitativa da interpretação automática de uma língua natural particular em um ambiente de domínio totalmente aberto; c) estando já em operação comercial, disponível para usuários com qualquer nível de conhecimento lingüístico e com qualquer nível de expectativa, o sistema computacional apresenta dificuldades adicionais ante qualquer alteração que se faça necessária: mesmo operando com algumas falhas, seu aprimoramento acarretará um ônus inicial, tanto de produção da nova versão comercial, quanto de adaptação dos usuários.

Além disso, a proposta do modelo do protótipo, que privilegia a estruturação argumental do verbo, não é, necessariamente, a mais adequada para o aprimoramento da ferramenta, no que diz respeito aos falsos erros e falsos acertos. Basta considerar, por exemplo, como perspectiva alternativa, a diversificação da análise categorial realizada: o problema em foco parece ser muito mais de natureza contextual do que de natureza lexical ou de dependência verbal, conforme sugerimos nesse trabalho. Neste caso, uma proposta alternativa está em investigar mais profundamente os contextos de ocorrência das intervenções indevidas do ReGra, mediante a análise do inter-relacionamento entre diversas categorias lexicais, assim como de suas possíveis posições na sentença. Por exemplo, advérbios e adjetivos, em determinados contextos, são classificados diferentemente, como em ‘*Ela andou rápida e distraidamente*.’ e ‘*Ela andou rápida.’* A primeira sentença, diferentemente da segunda, é gramatical pelo simples fato de *rápida* ocupar, nesse contexto, a categoria adverbial, no circunstante *rápida e distraidamente*. Essa conclusão é impossível de se obter automaticamente na segunda sentença, que seria gramatical somente se houvesse um advérbio explícito.

A similaridade desse exemplo com *‘A espada feriu fundo.’* leva à conclusão de que a especificação da estrutura argumental verbal, conforme sugerida na seção anterior, por si só não seria suficiente para resolver a ambigüidade. Em ambos os casos, a questão principal é se o sujeito da sentença acaba por determinar a categoria de seu correspondente, adjetivo ou advérbio, dependendo de seu contexto de ocorrência. Desse modo, o problema central deixa de ser o verbo em si, passando a ser o inter-relacionamento entre os componentes sentenciais que ele determina.

Casos como esses podem implicar a ocorrência de padrões de problemas cuja resolução automática não necessariamente dependa de uma especificação lingüística criteriosa, mas, antes, da inclusão de regras pontuais para o tratamento de cada padrão. Resta saber se é possível, para o tratamento da língua portuguesa, estabelecer tais padrões de modo claro e não ambíguo, sem introduzir inconsistências do ponto de vista da gramaticalidade e do processamento automático correspondente.

As questões fundamentais que devem continuar norteando as decisões de projeto, portanto, devem ser: a) se o sistema pode produzir soluções parciais; b) se o sistema pode produzir soluções úteis, mesmo que parciais.

As dificuldades descritas neste relatório demonstram que, no momento, ainda não conseguimos respondê-las adequadamente. Entretanto, podemos apontar, pelos estudos iniciais (Seções 2 e 3), alguns princípios que cercam o processo de significação e que são, sem dúvida, relevantes para o tratamento da ambigüidade categorial. Dentre eles, destacamos a decomposição do significado em unidades de sentido – traços semânticos – dentro da chamada “análise componencial”; a possibilidade de formalizar as informações do extrato semântico, por meio de uma gramática estruturada sobre os elementos do significado, como é a proposta de

Kaplan and Bresnan (1982); a emergência de uma estrutura semântica a partir da estrutura sintática, através da qual alguns elementos sintáticos se tornam essenciais para a obtenção do significado de uma sentença, como é o verbo na Teoria dos Casos de Fillmore (1963).

Apontamos, ainda, a importância das evidências empíricas obtidas durante o desenvolvimento deste projeto. Particularmente, ressaltamos a) as ocorrências significativas de falsos erros e omissões, cuja análise pode levar a um melhor entendimento sobre os casos que realmente se devem à ausência de tratamento semântico; b) a pertinência das propostas teóricas sobre semântica lexical, para a resolução automática da ambigüidade categorial; assim como c) a pertinência dos modelos gramaticais descritos na literatura, para os propósitos de revisão gramatical.

A preocupação em incorporar ao ReGra um tratamento lingüisticamente mais refinado surgiu da preocupação em construir um sistema motivado pelas descrições dos fenômenos da língua, visando maior robustez durante o processamento de dados não confiáveis. Entretanto, a partir deste relato, concluímos que:

3. É impossível considerar uma proposta metodológica particular de aprimoramento do ReGra, pois ele consiste em uma ferramenta bastante específica, que já conta hoje com um léxico contendo, atualmente, somente informações morfológicas e sintáticas e um *parser* cujo processamento é realizado concomitantemente com a revisão gramatical;
4. As abordagens investigadas revelam-se incompletas para representar os casos e objetivos particulares do ReGra e parecem deixar escapar aspectos da significação que, no que concerne à ambigüidade, podem tornar o ReGra um sistema bastante oneroso. Especificar, p.ex., apenas os “marcadores” e os “distinguidores”, segundo a proposta de Katz e Fodor, não resolveria o problema da ambigüidade presente em palavras como o ‘se’;
5. Faz-se necessário abordar outras perspectivas teóricas, mais próximas do contexto operacional. Por exemplo, seguindo propostas da Psicolingüística, que procuram refletir a aproximação entre o processamento mental da linguagem e o processamento computacional. Neste caso, a proposta de andamento deste projeto recai em alguns modelos estritamente empíricos, sobre os quais pouco temos a elaborar no momento.

Essas conclusões, seguidas aos problemas e alternativas metodológicas apontados nas seções anteriores, resultam nos seguintes requisitos, para o aprimoramento do ReGra:

a) A representação do conteúdo lexical deve envolver não apenas a classificação gramatical mas a consideração de componentes necessários ao equacionamento das ambigüidades lexicais. Entre esses componentes, sobressaem-se: a estrutura de constituintes lexicais (radicais, afixos e vogais e consoantes de ligação), necessários para a desambigüização morfológica, e a estrutura temática, necessária para o equacionamento da ambigüidade sintática.
b) A par da representação de conteúdo declarativo referido em (a), devem estar disponíveis, entre os mecanismos de desambigüização, estratégias de deslizamento semântico (para a recuperação das metáforas, metonímias, etc.) e de instanciação anafórica (para a recuperação dos referentes textuais) das palavras da língua.
c) O processamento da sentença deve envolver algum paralelismo entre léxico e gramática de forma a ser possível interromper ou redirecionar as consultas lexicais em função do conhecimento gramatical disponível ou a optar entre análises sintáticas concorrentes em função da combinatória de informações lexicais disponibilizadas pelo dicionário.

Buscando identificar tais requisitos, passamos a investigar padrões lingüísticos ocorrentes no corpus do NILC, com especial ênfase sobre as diversas combinações categoriais para um

mesmo item lexical, tais como substantivo/adjetivo/advérbio, substantivo/verbo, advérbio/adjetivo ou verbo/adjetivo. Essa nova abordagem para o Projeto TraSem é relatada em (Pinheiro et al., 2001).

## Apêndice 1: Abreviaturas lingüísticas

**A**
A: argumento
abs: abstrato
aç: ação
adj: adjetivo
adv: advérbio
ag: agente
anim: animado
aux: auxiliar

**C**
Ca: causativo
circ: circunstante
con: concreto

**D**
dest: destinatário
dd: discurso direto

**E**
EP: esquema profundo
ES: esquema superficial
Ex: experimentador

**H**
hum: humano

**I**
I: instrumento

**N**
N: nome

**O**
O: oração
Ob: objetivo
Oconj: oração conjuncional
Oinf: oração infinitiva

**P**
P: predicado/ predicador
Pa: paciente
proc: processo
pron: pronome
prep: preposição

**S**
Sadv: sintagma adverbial
SN: sintagma nominal
Sprep: sintagma preposicional
Suj: sujeito

**V**
V: verbo

### Referências bibliográficas

Abrahão, P. R. C. e Lima, V. L. S. (1996). *Um Estudo Preliminar de Metodologias de Análise Semântica da Linguagem Natural.* II Encontro para o Processamento Computacional de Português Escrito e Falado. Centro Federal de Educação Tecnológica do Paraná , Paraná. Outubro.

Almeida, G.M.B. (1995). *Neologismos na informática: natureza e deriva de um vocabulário.* Dissertação de Mestrado, 218p. Faculdade de Ciências e Letras, Universidade Estadual Paulista. Araraquara.

Atkins, B.T.S. and Zampolli, A. (1994). *Computational Approaches to the Lexicon.* Oxford Press, New York.

Bateman, J.A., Kasper, R.T., Moore, J.D., and Whitney, R.A. (1990). A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model. Release Notes. ISI/USC, USA.

Beesley, K.R. and Grefenstette, G. (1996). *Finite-State Analysis of Written Portuguese.* II Encontro para o Processamento Computacional de Português Escrito e Falado. Centro Federal de Educação Tecnológica do Paraná. Paraná, Outubro.

Bresnan, J. (1982). *The Mental Representation of Grammatical Relations.* MIT Press, Cambridge, Mass.

Borba, Francisco da Silva (1990). *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil.* Editora da Universidade Estadual Paulista, São Paulo.

Borba, Francisco da Silva (1991). *Introdução aos Estudos Lingüísticos.* Editora Martins Fontes, Campinas, São Paulo.

Borba, Francisco da Silva (1996). *Uma Gramática de Valência para o Português.* Ática, São Paulo.

Bunt, Harry and Tomita, Masaru (eds.) (1996). Recent Advances in Parsing Technology. Kluwer Academic Publishers, The Netherlands.

Chomsky, N. (1957). *Syntactic Structures.* Mouton, Haia.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge MA.

Chomsky, N. (1972). Remarks on nominalization. In _____ . *Studies on semantics in generative grammar*, pp. 62-119. Mouton, Haia.

Church, K. W. and Mercer R. L. (1993). *Introduction to the Special Issue on Computational Linguistics Using Large Corpora.* Association for Computational Linguistics.

Dias-da-Silva, B.C.; Sossolote, C.; Zavaglia, C.; Montilha, G.; Rino, L.H.M.; Nunes, M.G.V.; Oliveira Jr., O.N.; Aluísio, S.M. (1998). The Design of a Brazilian-Portuguese Machine Tractable Dictionary for an Interlingua Sentence Generator. III Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'98). Porto Alegre - RS. Novembro.

Dubois, J. et al. (1988). *Dicionário de lingüística*. Cultrix, São Paulo.

Fillmore, C. (1968). Lexical Entries for Verbs. *Foundations of Language*, 4, pp. 373-393.

Gazdar, G. and Pullum, G. (1982). *Generalized Phrase Structure Grammar: a theoretical synopsis*. Indiana University Linguistic Club.

Greghi, J.G. (2001). *Uma Base de Dados Lexicais para o Português do Brasil*. Monografia de Qualificação ao Mestrado. Departamento de Computação, ICMC-USP. Março, São Carlos.

Haddock, J.N., Klein, E., and Morrill, G. (1987*). Unification Categorial Grammar, Unification Grammar and Parsing.* University of Edinburgh.

Hirsh-Pasek, K., Reeves, L. M & Golinkoff, R. (1993). Words and meaning: from primitives to complex organization. In J.B. Gleason & N.B. Ratner (org.), Psycholinguistics, pp. 174-180. Harcourt Brace Jovanovich College Publishers, Fort Worth.

Jackendoff, R. (1983). *Semantics and cognition*. MIT Press, Cambridge, MA.

Jackendoff, R. (1990). *Semantic Structures*. MIT Press, Cambridge, MA.

Joshi, A.K. & Shabes, Y. (1992). Tree-adjoining grammars and lexicalized grammars. In *Tree Automata and LGS*. Elsevier Science, Amsterdam.

Kaplan, R. and Bresnan, J. (1982). Lexical-Functional Grammar: a Formal System for Grammatical Representation. In Bresnan, J. (ed.), *The Mental Representation of Grammatical Relations*, pp. 173-281. MIT Press, Cambridge MA.

Katz, J.J. and Fodor, J.A. (1963). The Structure of a Semantic Theory. *Language*, 39, pp. 170-210.

Kay, M. (1984). Functional Unification Grammar: a formalism for machine translation. In *Proceedings of the 10th International Conference on Computational Linguistics.* Stanford University, California.

Lucchesi C. L. and Kowaltowski T. (1993). *Applications of Finite Automata Representing Large Vocabularies*. John Wiley & Sons Ltda.

Lyons, J. (1977). *A Semântica I.* Presença, Lisboa.

Marcus, M. P. (1980). *A Theory of Syntactic Recognition for Natural Language*. The MIT Press. Cambridge, Massachusetts and London, England.

Marrafa, P. (1996). *Representação das Formas Predicativas Verbais do Português numa Base de Conhecimento Lexical.* II Encontro para o Processamento Computacional de Português Escrito e Falado. Centro Federal de Educação Tecnológica do Paraná , Paraná. Outubro.

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), pp. 287-307. Cambridge University Press.

Martins, R.T.; Rino, L.H.M.; Montilha, G. e Nunes, M.G.V. (1999). *Dos modelos de resolução da ambigüidade categorial: O problema do SE*. IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR'99). Universidade de Évora, Portugal. Setembro.

Nunes, M.G.V. (coord.) (1999). Revisor Gramatical e Ferramentas de Auxílio à Escrita. Relatório Parcial - Fevereiro 1999, Processo RC: 3.1.3-0012/98, Convênio: 8.8.98.0591.00. NILC/ICMC-USP.

Pacheco, H. C. F.; Dillinger, M. e Carvalho, M. L. B. (1996). *Uma Nova Abordagem para a Análise Sintática do Português.* II Encontro para o Processamento Computacional de Português Escrito e Falado. Centro Federal de Educação Tecnológica do Paraná , Paraná. Outubro.

Perrault, C.R. (1985). On the Mathematical Properties of Linguistic Theories. In Grosz et al. (eds.), *Readings in Natural Language Processing*. Morgan Kaufmann Publishers, Los Altos, California.

Pinheiro, G.M.; Martins, R.T.; Rino, L.H.M.; Di Felippo, A.; Fillié, V.M.; Hasegawa, R. (2001). *Projeto TraSem: A investigação empírica sobre o problema da ambigüidade categorial.* Tech. Rep. NILC-TR-01-2. São Carlos, Março.

Pollard, C. and Sag, I.A. (1994). *Head-driven Phrase Structure Grammar*. Center for the Study of Language and Information (CSLI). Lecture Notes. Stanford University Press and University of Chicago Press.

Pustejovsky, J. (1995). *The Generative Lexicon: A Theory of Computational Lexical Semantics.* MIT Press, Cambridge, MA.

Pustejovsky, J. and Boguraev, B. (1996). *Lexical Semantics. The Problem of Polysemy*. Clarendon Press, Oxford.

Saussure, Ferdinand de (1916[1975]). Curso de Lingüística Geral. Cultrix, São Paulo.

Silva, J. L. T. e Lima, V. L. S. (1996). *Algumas Considerações sobre Resolução da Ambigüidade Léxica no Processamento da Linguagem Natural*. II Encontro para o Processamento Computacional de Português Escrito e Falado. Centro Federal de Educação Tecnológica do Paraná , Paraná. Outubro.

Tanenbaum, A. S. (1995). *Sistemas Operacionais Modernos*. Prentice-Hall do Brasil Ltda. Rio de Janeiro – RJ.

Viegas, E. and Raskin, V. (1998). *Computational semantic lexicon acquisition. Methodology and Guidelines*. Computing Research Laboratory, New Mexico State University. Las Cruces, USA. Abril.

Woods, W.A. (1970). Transition Network Grammars for Natural Language Analysis. *CACM* 13, pp. 591 – 606.