UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Projeto TraSem:
A Investigação Empírica sobre o Problema da Ambigüidade Categorial

Gisele Montilha Pinheiro
Lucia Helena Machado Rino
Ronaldo Teixeira Martins
Ariani Di Felippo
Vanessa Matsuda Fillié
Ricardo Hasegawa
Maria das Graças Volpe Nunes

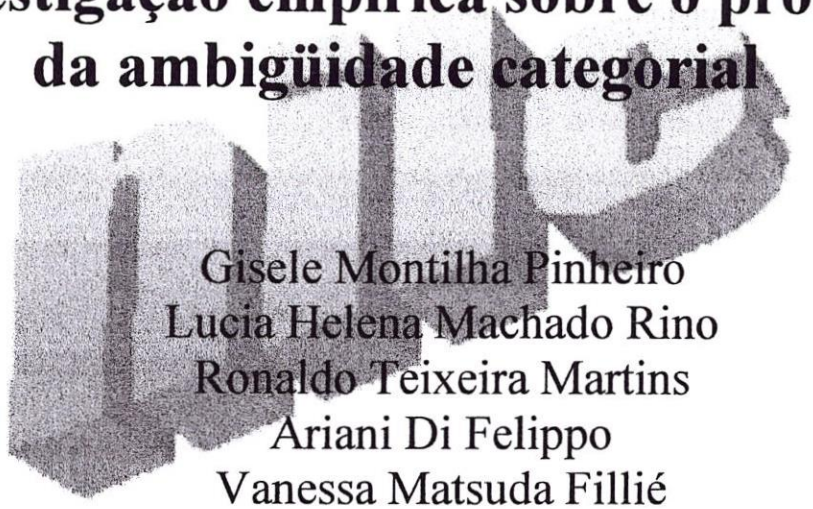
Nº 140

RELATÓRIOS TÉCNICOS

São Carlos - SP

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

Projeto TraSem:
A investigação empírica sobre o problema da ambigüidade categorial



Gisele Montilha Pinheiro
Lucia Helena Machado Rino
Ronaldo Teixeira Martins
Ariani Di Felippo
Vanessa Matsuda Fillié
Ricardo Hasegawa
Maria das Graças Volpe Nunes

NILC-TR-01-2

Abril, 2001

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

Resumo

Este relatório dá seqüência ao relatório técnico NILC-TR-01-1 (Rino et al., 2001), o qual, a partir da exploração teórica dos problemas de especificação semântica para melhorar o desempenho do ReGra, aponta a necessidade de se analisar mais profundamente suas ocorrências de diagnósticos inadequados. A partir do estudo teórico, sugerimos uma metodologia empírica, fundamentada na busca de subsídios estatísticos sobre o corpus do NILC. É esta tarefa, de exploração dos dados lexicais, que descrevemos aqui. Mais particularmente, concentramo-nos, primeiramente, no refinamento das especificações do léxico do ReGra, observando que muitos dos problemas de diagnóstico remetiam a inadequações já existentes no próprio léxico. A partir desse refinamento, a proposta é proceder à análise lingüística do corpus do NILC, etiquetado, com o objetivo de determinar padrões sintáticos que permitam tratar os casos de ambigüidade categorial.

1. Histórico

As pesquisas em torno do tipo e representação das informações semânticas no sistema do ReGra já experimentaram diversas metodologias de trabalho. A investigação em nível semântico justifica a existência do Projeto TraSem, ocupado em subsidiar o ReGra e outros aplicativos desenvolvidos no NILC com o conhecimento semântico suficiente para dirimir os problemas enfrentados no processamento automático do português. A determinação desse objetivo se deveu a diversas etapas de análise sobre as inadequações do ReGra, relatadas em documentos anteriores[1] que merecem ser lidos, para que se tenha uma idéia mais clara sobre o foco deste trabalho.

Apesar das discussões teóricas importantes que tivemos no Projeto TraSem, que remetem ao problema de representação semântica da língua natural, de um modo geral, os experimentos realizados por meio da prototipagem de um módulo semântico do ReGra evidenciaram a impossibilidade – pelo menos no contexto atual – de contemplarmos a modelagem profunda do conhecimento lingüístico necessário ao aprimoramento do ReGra (Rino et al., 2001). Por essa razão, no último semestre de 2000 a equipe empenhou-se em outra perspectiva metodológica, procurando estreitar as hipóteses sobre os fenômenos lingüísticos aos fatos que emergem da base lexical do ReGra. Este relatório, pois, objetiva apresentar essa proposta e apontar os resultados parciais obtidos nesse último período de trabalho no TraSem.

O ponto de partida deste trabalho foi a verificação da existência de casos de ambigüidade categorial que não podem ser facilmente equacionados pelo recurso à freqüência de ocorrência. Ainda que muitas das formas homógrafas da língua portuguesa possam ser resolvidas pela análise estatística de corpora, com a conseqüente identificação de possibilidades de ocorrência e padrões de distribuição, muitas outras formas dependem da representação do contexto semântico da ocorrência, sob o risco de exigirem o estabelecimento de regras de dependência (de regência e de concordância) que, no caso, não seriam pertinentes. É o que se verifica, por exemplo, com a sentença "A espada feriu fundo", que envolve atualmente [falso] erro da ferramenta. Dada a ambigüidade categorial da forma 'fundo', que pode ser adjetivo ou advérbio, e dada a priorização, pela ferramenta, da acepção do adjetivo sobre o advérbio, exige-se uma concordância (entre 'fundo' e 'espada') que, definitivamente, viola as regras de boa-formação sintática das sentenças da língua portuguesa. Este, como inúmeros outros problemas, envolveria estratégias de representação do conteúdo lexical e proposicional que escapam à metodologia atualmente adotada pelo ReGra, pelo que se fez necessária a exploração de alternativas para o problema.

Uma primeira alternativa de resolução de problemas dessa natureza residiu na consideração da hipótese teórica da valência verbal (cf. Borba, 1996), segundo é relatado em (Rino et al., 2001) e apresentado resumidamente neste relatório. Essa hipótese, utilizada na simulação da modelagem lingüística, é a base do protótipo de processamento semântico do ReGra, o qual contempla somente algumas sentenças-problema, buscando sintetizar outros de seus tantos problemas verificáveis. Para a resolução da ambigüidade da sentença acima, por exemplo, postulou-se a necessidade de uma representação mais rica da informação lexical referente ao verbo 'ferir' por sua estrutura argumental,

[1] Relatório Parcial FINEP, Fevereiro/2000 (Processo RC:3.1.3-0012/98, Convênio: 8.8.98.0591.00); Martins et al., 1999; Rino et al., 2001.

combinando os seguintes traços semânticos na posição de agente ou na posição de complemento, a partir do leque de possibilidades sugerido por Borba (1990)[2]:

animado
animado-humano
concreto-inanimado
abstrato-inanimado

Com base nessas possibilidades, na fase de prototipagem procuramos circunscrever a sentença em questão, na expectativa de abstrair uma sistemática essencialmente semântica da estrutura argumental do verbo. Nesse sentido, então, a sentença acima apresentava a seguinte configuração:

[sentença]	*A espada*	*feriu*	*fundo.*
[estrutura sintática]	agente	verbo	complemento
[especificação semântica]	concreto-inanimado	ação-processo	∅

Outros casos de mesma natureza foram estudados para o protótipo, que procurou circunscrever as sentenças-problema na expectativa de abstrair uma sistemática essencialmente semântica para o tratamento computacional da estrutura argumental do verbo. O exame de esquemas semelhantes ao ilustrado fez emergir a insuficiência da especificação semântica nos padrões ditados pelo dicionário, já que, no argumento de complemento, nenhum traço caracterizava especificamente a informação do termo 'fundo' expresso na sentença em questão. Por outro lado, o estudo apontava para um dado interessante acerca da ambigüidade categorial desse termo: em estado de léxico, a palavra admite, ao mesmo tempo, as categorias substantivo (como em "O fundo da panela era branco"), adjetivo (como em "O lago da fazenda é bastante fundo") e advérbio (como a nossa sentença "A espada feriu fundo"). Diante da inexistência de uma ontologia que identificasse a particularidade dos significados distintos de cada uma das acepções do verbo 'ferir' na base lexical do ReGra e, além disso, a par da limitação da especificação semântica para a resolução da ambigüidade instaurada numa ocorrência em que ao verbo segue-se um complemento adverbial do tipo 'fundo', notamos, então, que a diferenciação categorial poderia constituir uma alternativa para a investigação do problema lexical de pares ambíguos. Seguramente, a especificação semântica desempenhava algum papel no processamento lingüístico dessa sentença, porém, o caminho a ser perseguido, no sentido de representar cada item lexical ambíguo segundo a teoria e o modelo simulado, oferecia a desvantagem de ser um esforço humano extraordinário sem o retorno desejado em qualquer prazo que lhe fosse estipulado. Em vista disso, decidimos suspender a aplicação do modelo de valências verbais aos itens lexicais das sentenças-problema em favor de outra direção da pesquisa.

A nova proposta, justificada pela evidência da ambigüidade categorial extraída da sentença em foco, sugeria uma pesquisa ao léxico do revisor, procurando, nas estatísticas, o par categorial que rotulava o maior número de entradas lexicais do sistema. A investigação revelou, então, a seguinte configuração para a ambigüidade categorial[3]:

[2] Segundo Borba, há os seguintes traços para o verbo em questão: animado, animado-humano, concreto-inanimado e abstrato-inanimado.

[3] É interessante ressaltar que, nesta tabela, o número de entradas não corresponde aos verbetes classificados tão como substantivos e verbos, mas a todos os verbetes que, entre todas as suas categorias,

Tabela 1: Números da ambigüidade no léxico total do ReGra

Ambigüidade entre	No. de entradas
Substantivo e Verbo	7028
Substantivo e Adjetivo	2627
Substantivo e Advérbio	49
Substantivo e Conjunção	10
Verbo e Adjetivo	7944
Verbo e Advérbio	49
Verbo e Conjunção	7
Adjetivo e Advérbio	78
Adjetivo e Conjunção	4
Advérbio e Conjunção	22
TOTAL DE ENTRADAS	**17818**

Segundo esses dados, a investigação passou a contemplar o exame do primeiro tipo de relação de ambigüidade, i.e., do par categorial substantivo-verbo, a fim de observar, nos próprios itens lexicais assim rotulados, alguma propriedade lingüística de interesse para a resolução da ambigüidade, que permitisse uma representação genérica e sistemática para toda uma classe morfológica do português. Por exemplo, uma propriedade que descrevesse o comportamento semântico dos substantivos e não dos verbos para um mesmo item lexical categorizado duplamente como substantivo e verbo no léxico do revisor. Essa proposta metodológica, assim como as tarefas decorrentes no Projeto TraSem, são descritas a seguir.

2. Diretrizes e desenvolvimento

A identificação da relação de ambigüidade categorial entre substantivo e verbo no léxico, conforme os procedimentos descritos acima, fez-nos perceber, de imediato, que muitas informações ligadas aos itens lexicais em questão estavam dispostas ali equivocadamente. Tal observação motivou uma iniciativa de correção (ou de "limpeza", como a designamos corriqueiramente) da base lexical antes de partirmos para o trabalho de inserção de informações visando à desambigüização, propriamente dita.

A tarefa de correção foi executada por lingüistas e tomou como diretrizes os seguintes procedimentos:

a) Excluir a categoria 'substantivo' das formas não abonadas como tal no dicionário de referência[4].
b) Excluir a categoria 'substantivo' das formas que, de acordo com o dicionário de referência: b.1) são nomes de objetos ou fatos de um domínio muito específico; b.2) são nomes em desuso ou nomes do português arcaico; b.3) representam gírias de uso pouco difundido; e b.4) representam regionalismos bastante restritos.
c) Excluir a categoria 'substantivo' das formas que não sustentam um significado próprio enquanto substantivo, mas que funcionam como tal em circunstâncias

trazem a marca de substantivo e verbo. Ou seja, o total de entradas pode estar superestimado, se considerarmos que há palavras que são subst+adj+verb e que estão aparecendo três vezes nesta lista: como subs+adj, como adj+verb ou como subst+verb.

[4] *Dicionário Aurélio Eletrônico século XXI*. Nova Fronteira, Versão 3.0, Novembro de 1999.

sintáticas de elipse do substantivo. Além disso, inserir a categoria 'adjetivo', quando o verbete dessas palavras não o previr.

d) Excluir a categoria 'verbo' das formas de uso muito raro ou de domínio bastante específico de acordo com o dicionário de referência.

e) Corrigir o léxico, para casos em que as categorias lexicais estiverem especificadas erroneamente.

f) Inverter a ordem de prioridade entre as categorias dispostas em cada verbete da base lexical, quando a ordem prevista não obedecer ao que, intuitivamente, seria o uso mais freqüente.

No caso (a), a exclusão se deve ao fato de que diversas palavras da base lexical foram geradas automaticamente, de maneira que a flexão de número e gênero incluiu no léxico, de um lado, formas não existentes na língua e, de outro, formas que funcionam como substantivo (ou verbo, ou adjetivo, etc.) somente para um gênero particular, seja ele masculino ou feminino. Alguns exemplos de ocorrências indevidas, de ambas as naturezas, são dados abaixo (gênero não abonado entre parênteses):

alunada (F)	fito (M)	pecada (F)	rodado (M)	tolda (F)
apostolada (F)	gemida (F)	povoada (F)	seriada (F)	tornada (F)
arruada (F)	gramada (F)	quadrada (F)	significada (F)	torrado (M)
bispada (F)	grita (F)	queimado (M)	soída (F)	tratada (F)
cabeceada (F)	ido (M)	reinada (F)	subido (M)	vestida (F)
comunicada (F)	noviciada (F)	resfriada (F)	tecida (F)	
conglomerada (F)	objetiva (F)	resultada (F)	tinida (F)	

Formas não abonadas no dicionário

Notou-se que em todos esses casos do exemplário a acepção de substantivo, no gênero que assumem, não existe na língua, mas há um ou mais significados relacionados à forma do substantivo em gênero contrário. Por exemplo: não existe na língua o substantivo 'pecada', mas sim 'pecado'; da mesma forma, o substantivo 'quadrada', mas sim 'quadrado'; o substantivo 'tecida', mas sim 'tecido'.

O segundo tipo de exclusão – caso (b) – deveu-se à verificação de que esses tipos de ocorrência são recorrentes no léxico e respondem por um número considerável de ambigüidade categorial envolvendo os substantivos. A idéia é que, a fim de atenuar essa ambigüidade, retiremos da base lexical o atributo de substantivo das palavras que, afora a acepção bastante restrita (de regionalismo ou de termo em desuso, por exemplo), não possuem outro significado a elas relacionado. São exemplos desses casos (são dados entre parênteses as iniciais para o gênero do substantivo – (a): fem; (o): masc – e a justificativa para a exclusão):

(a) fura (provincianismo português)	(o) pago (brasileirismo e termo pouco usado)
(o) inço (regionalismo do campo da botânica)	(a) perca (domínio específico da zoologia)
(o) integrando (domínio da análise matemática)	(o) provará (domínio jurídico)
(o) li (termos relacionados à cultura chinesa: tipo de moeda, medida itinerária, etc.)	(a) pule (domínio específico do turfe)
(o) liberando (domínio jurídico)	(a) purga (domínio específico da terapêutica)
(a) manja (lusitanismo)	(a) taxe (domínio específico da medicina)
(o) nego (brasileirismo, popular; gíria)	(o) torce (brasileirismo)
(a) olha (domínio específico da culinária)	(a) trombo (domínio específico da patologia)
(a) ouça (termo pouco usado; var.: 'oiça')	(a) cipoada (brasileirismo, BA)
(o) bobinado(domínio específico da eng. elétrica)	(a) fubecada (brasileirismo, gíria)

Ocorrências inexpressivas no uso geral do português

No caso (c), percebeu-se que as formas lingüísticas que representam o particípio regular de verbos quase sempre também são usadas na função de adjetivo (ex.: 'limpo'), mas não possuem significado próprio como substantivo (ex.: 'abandonado'). A correção do léxico para tais verbetes contribui para a desambigüização categorial da base léxica, mas, mais importante que isso, isenta o léxico de armazenar a informação sobre a elipse do substantivo, que é satisfatoriamente tratada pelo *parser*. Por exemplo, a seguinte regra é aplicada com sucesso, na maioria das vezes: *entenda a palavra seguinte a um artigo como um substantivo ou um adjetivo que antecede a um substantivo*. Assim, 'refogado' em "O refogado da sua mãe é espetacular!" é compreendido pelo revisor como um adjetivo substantivado. A decisão por esse procedimento de exclusão se justifica, ainda, pela maior facilidade em substantivar um adjetivo do que adjetivar um substantivo. Alguns casos de adjetivos dessa natureza são ilustrados abaixo:

abandonado	encarregado	internado	recheado
advogado	exibido	maldito	refugiado
acamado	fiado	manufaturado	sacramentado
acidentado	finado	narrado	safado
beneficiado	flagelado	platinado	segurado
candidato	frustrado	protegido	sorteado
casado	gelado	rachado	tabulado
degenerado	granulado	raptado	torturado
emburrado	homenageado	recenseado	viciado

Ocorrências adjetivadas

A quarta estratégia de exclusão – caso (d) – ao contrário dos substantivos, acarreta uma alteração maior na base lexical, já que não é possível excluir algumas formas do verbo, mantendo-se sua canônica. Assim, esse procedimento foi aplicado em casos muito extremos, a fim de que o léxico não fosse penalizado por conta da desambigüização nesse nível de organização de dados. A título de ilustração, seguem-se alguns casos de exclusão completa da acepção de verbo no léxico, ou seja, casos em que todas as 53 formas flexionais dos verbos do português foram retiradas da base (respectivas canônicas entre parênteses):

iodo (iodar)	jibóia (jiboiar)
intervalo (intervalar)	medicamento (medicamentar)

Ocorrências verbais incomuns

A quinta estratégia, de correção do léxico – caso (e) – foi adotada devido à observação de que algumas classificações não condiziam com as propostas do dicionário de referência, no que diz respeito ao gênero e número de alguns substantivos. Ocorrências desse tipo incluem, por exemplo, 'abas', 'aparas', 'cavalarias', 'curetas', 'gama', entre outras, que estavam classificadas no léxico como sendo s-2G-pl (palavras de dois gêneros), enquanto o dicionário indicava a categorização S-Fem-PL. Variações de gênero para um mesmo substantivo ainda eram indicadas explicitamente por Fem e Masc, em vez de 2G, categoria originalmente assumida no léxico. São exemplos de variações desse tipo 'caças', 'curas', 'gramas', entre outras. Outros tipos de reclassificação lexical foram observados, ainda, para casos que não necessariamente conflitavam com o dicionário de referência, quer em relação ao gênero, quer em relação ao número dos substantivos. Exemplos desse tipo de correção são ilustrados abaixo:

Verbete	Categorização anterior	Correção
abas	s-**2G**-pl-pres-tu	s-**fem**-pl-pres-tu
caças	S-**2G**-PL; PRES-TU	S-**fem**-PL; S-**M**-PL;PRES-TU
raja	subst-**masc**-sing, imper-afirm-tu, pres-ele	subst-**fem**-sing, imper-afirm-tu, pres-ele
asperges	subst-masc-**sing**	subst-masc-**2N**
barris	subst-masc-**sing**, pres-vós	subst-masc-**plural**, pres-vós
labores	subst-masc-**sing**, pres-subj-tu	subst-masc-**plural**, pres-subj-tu

Correção de algumas formas lexicais substantivadas

O sexto procedimento, de inversão das entradas lexicais – caso (f) – diferentemente dos anteriores, que foram baseados no dicionário de referência, foi norteado pela intuição de falantes do português e pelo conhecimento técnico dos lingüistas envolvidos nessa tarefa. A idéia é reduzir o esforço do processamento automático pela hierarquização das categorias morfológicas das palavras no léxico, já que não é possível, nesses casos, eliminar uma ou outra categoria a elas relacionada. Por exemplo: é possível intuir que a palavra 'vestido' é mais comum como substantivo do que particípio, como previa o léxico. Esse é um caso, pois, de inversão na classificação do verbete de 'vestido'. Exemplos de verbetes cuja posição no léxico é devida a esse tipo de análise categorial são apresentados abaixo (as possíveis categorias de cada verbete são indicadas entre parênteses, na ordem hierárquica estabelecida; a ocorrência da categoria 'v' indica a forma verbal no particípio):

chiado (s/v)	ida (s/v)	legado (s/v/adj)	mataria (v/s)
cacetada (s/v)	inibido (adj/v)	levado (adj/v)	moderado (adj/v)
franzida (adj/v)	intrometido (adj/v)	levita (v/s)	narrado (adj/v)
fraseado (s/adj/v)	jorra (v/s)	linguajar (s/v)	ninhada (s/v)
gemido (s/v)	laçaria (v/s)	listra (s/v)	oposto (adj/s/v)
gramado (s/v/adj)	laminado (s/adj/v)	macacada (s/v)	ouvido (s/v/adj)
granulado (adj/v)	lancha (s/v)	macarronada (s/v)	quadrado (s/adj/v)
habilitado (adj/v)	latido (s/v)	machucado (s/adj/v)	vestido (s/v)

Verbetes organizados hierarquicamente segundo sua multiplicidade categorial

Para a execução dessa tarefa de limpeza, foi selecionado somente um dicionário de referência, o *Dicionário Aurélio Eletrônico*. Muitas vezes, porém, a informação lexicográfica disponível não correspondia à impressão lingüística de falantes, o que nos levou a equilibrar as sugestões dicionarizadas com a intuição.

A consideração intuitiva de que nos valemos, por sua vez, procurou respeitar os objetivos da ferramenta de revisão em detrimento da liberdade lingüística que caracteriza

o emprego de algumas formas ou mesmo a criação de palavras não abonadas pelo dicionário. No entanto, as decisões assim fundamentadas não alcançaram grande parte dos dados, porque decidir por um ou outro extremo na linha da ambigüidade, sem o apoio do contexto de ocorrência, é um desafio extremamente complexo, senão inútil. Como resultado, a ambigüidade categorial permaneceu em boa parte dos casos analisados e, em outros tantos, foi substituída por outro tipo de ambigüidade (daquela entre substantivo e verbo para aquela entre adjetivo e verbo, por exemplo). Esse fato era esperado, já que o aprimoramento do léxico, por si só, não deveria impedir a representação da ambigüidade inerente da língua natural em foco. O problema delimitado no Projeto TraSem, portanto, permaneceu centralizado na busca de mecanismos que pudessem auxiliar o revisor a reconhecer a ambigüidade categorial e, eventualmente, tratá-la em algum nível lingüístico.

Os seis procedimentos acima refletem os tipos de imprecisão existentes na base lexical, os quais procuramos sanar para dar prosseguimento ao Projeto TraSem. Entretanto, a resolução lexical, nesse nível, virá a contribuir também a outros projetos de igual relevância para a língua portuguesa. Embora este trabalho tenha sido dedicado a um par específico de categorias substantivo e verbo – ele deixa entrever que os problemas encontrados no léxico podem ser também encontrados nos verbetes das demais categorias morfológicas às quais cada entrada lexical está relacionada. Assim, procedimento semelhante para a análise da base lexical como um todo pode vir a conferir um mapeamento diferente da ambigüidade categorial para outras classes de palavras no léxico do ReGra, ficando como sugestão para desdobramentos futuros da tarefa de limpeza.

A Tabela 2 mostra as alterações lexicais realizadas nessa etapa de correção do léxico, especialmente em relação às ambigüidades entre substantivo e verbo e aos casos de exclusão, inclusão ou inversão de categorias (as correções são em número muito pequeno e, portanto, resolvemos não sintetizá-las nessa tabela). A porcentagem apresentada relaciona-se ao número de ambigüidades encontradas para esse par categorial apontado na Tabela 1 (7.028 verbetes). O sinal negativo da porcentagem indica redução de casos ambíguos.

De acordo com os números apresentados na Tabela 2, o léxico, constituído por 1.518.193 verbetes, passou a apresentar o quadro de ambigüidade categorial da Tabela 3 (a porcentagem aqui diz respeito aos números apontados na Tabela 1).

Tabela 2: Números relativos à tarefa de limpeza do léxico

Tipo de alteração	No. de entradas	%
Exclusão de substantivos	2.630	-37,42
Exclusão de adjetivos	12	-0,17
Exclusão de verbos	7	-0,1
Exclusão de outras categorias	-------	-------
TOTAL DE EXCLUSÕES	**2.649**	**-37,69**
Inserção de substantivos	5	0,07
Inserção de adjetivos	1.533	21,81
Inserção de verbos	2	0,03
Inserção de outras categorias	4	0,06
TOTAL DE INSERÇÕES	**1.544**	**21,97**
Inversão de categorias	2.604	37,05

Tabela 3: Configuração do léxico ao final desta fase de limpeza

Ambigüidade entre	No. de entradas	%
Substantivo e Verbo	4.399	-38
Substantivo e Adjetivo	1.696	-35
Substantivo e Advérbio	46	-6
Substantivo e Conjunção	10	0
Verbo e Adjetivo	9.479	+19
Verbo e Advérbio	49	0
Verbo e Conjunção	7	0
Adjetivo e Advérbio	79	+1
Adjetivo e Conjunção	4	0
Advérbio e Conjunção	22	0
TOTAL DE ENTRADAS	**232.574**	

Retomando, agora, o objetivo inicial, de refinamento da qualidade das informações lexicais já existentes visando a desambigüização, partimos para uma nova tarefa de análise, contemplando o par categorial substantivo-verbo, que é nosso caso de estudo escolhido. Essa tarefa é relatada na próxima seção.

3. A identificação dos tipos combinatoriais de substantivo-verbo

Nesta etapa do trabalho, foi feita uma nova busca no léxico, a partir de sua nova configuração (Tabela 3), para reunir e catalogar todas as formas ambíguas entre substantivos e verbos. Foram criados, então, 107 arquivos, cada qual rotulando um tipo de combinatória de pares ambíguos substantivo-verbo, cujo conteúdo é descrito na Tabela 4. Vale notar que o total de entradas indicado difere do que foi apontado na Tabela 3, com relação à ambigüidade substantivo-verbo. Trata-se de uma alteração movida pelo acréscimo de relações, tais como substantivo aumentativo, diminutivo, 2G (de dois gêneros), etc., anteriormente ignorada no rastreamento.

Tabela 4: Combinatórias categoriais de substantivo-verbo no léxico

Tipo de combinatória	No. de entradas	Palavra(s)-exemplos
1. Subst. Masc. Sing. e Verbo no Part. Fem. Sing.	01	*vista*
2. Subst. Masc. Sing. e Verbo no Part. Masc. Sing.	172	*cuidado / cunhado / passado*
3. Subst. Masc. Plural e Verbo no Part. Fem. Plural	01	*vistas*
4. Subst. Masc. Plural e Verbo no Part. Masc. Plural	180	*compostos / gastos / tratados*
5. Subst. Fem. Sing. e Verbo no Part. Fem. Sing.	125	*chamada / descida*
6. Subst. Fem. Sing. e Verbo no Part. Fem. Sing./ 2ª pessoa Imp-Afirm./ 3ª pessoa Pres. Ind.	04	*junta*
7. Subst. Fem. Plural e Verbo no 2ª pessoa Pres. Ind./ Part. Fem. Plural	03	*sujeitas*
8. Subst. Fem. Plural e Verbo no Part. Fem. Plural / 2ª pessoa Pres. Ind	01	*juntas*
9. Subst. Fem. Plural e Verbo no Part. Fem. Plural	130	*calçadas / despedidas*
10. Subst. Masc. Sing. e Verbo na 1ª pessoa Pres. Ind. / Part. Masc. Sing.	06	*seguro*
11. Subst. Masc. Sing. e Verbo na 3ª pessoa do Fut. Pres. Ind.	04	*abará / cambará*
12. Subst. Masc. Sing. e Verbo na 3ª pessoa pl. Fut. Pres. Ind.	05	*azarão/ porão*
13. Subst. Masc. Sing. e Verbo na 3ª pessoa Fut. Subj. / 1ª pessoa Fut. Subj. / 3ª pessoa Inf-pess / 1ª pessoa Inf-pess	47	*abafar/balar*
14. Subst. Masc. Sing. e Verbo na 3ª pessoa Fut. Subj. / 3ª pessoa Inf-pess	01	*entardecer*
15. Subst. Masc. Sing. e 1ª pessoa Fut. Subj. / 3ª pessoa Fut. Subj. / 1ª pessoa Inf-pess / 3ª pessoa Inf-pess	27	*azar/ placar*
16. Subst. Masc. Sing. e Verbo no Gerúndio	29	*formando/ graduando*
17. Subst. Masc. Sing. e Verbo na 3ª pessoa Imp-Afirm. / 3ª pessoa Pres. Subj. / 1ª pessoa Pres. Subj.	239	*alarme / desgaste / retoque*
18. Subst. Masc. Sing. e Verbo na 3ª pessoa Imp-Afirm. / 3ª pessoa Pres. Subj. / 1ª pessoa Pres. Subj. / 2G Part. Sing.	01	*aceite*
19. Subst. Masc. Sing. e Verbo na 2ª pessoa Imp-Afirm. / 3ª pessoa Pres. Ind.	36	*chega / fecha*
20. Subst. Masc. Sing. e Verbo na 2ª pessoa pl. Imp-Afirm.	02	*papai /estai*
21. Subst. Masc. Sing. e Verbo na 3ª pessoa do Inf-pess	01	*prazer*
22. Subst. Masc. Sing. e Verbo na 3ª pessoa do Inf-pess / 1ª pessoa Inf-pess	08	*dizer / saber*
23. Subst. Masc. Sing. e Verbo na 1ª pessoa do Inf-pess / 3ª pessoa Inf-pess	02	*querer / ver*
24. Subst. Masc. Sing. e Verbo na 3ª pessoa do Pres. Ind.	01	*é*
25. Subst. Masc. Sing. e Verbo na 3ª pessoa do Pres. Ind. / 2ª pessoa Imp-Afirm	01	*vê*
26. Subst. Masc. Sing. e Verbo na 1ª pessoa do Pres. Ind.	1.048	*abono / desmaio / grito*
27. Subst. Masc. Sing. e Verbo na 2ª pessoa do Pres. Ind.	01	*asperges*
28. Subst. Masc. Sing. e Verbo na 2ª pessoa pl. do Pres. Ind.	01	*barris*
29. Subst. Masc. Sing. e Verbo na 1ª pessoa do Pres. Subj. / 3ª pessoa Pres. Subj. / 3ª pessoa Imp-Afirm.	04	*carpete / toque*
30. Subst. Masc. Sing. e Verbo na 1ª pessoa do Pres. Subj. / 3ª pessoa Pres. Subj. / 2ª pessoa Imp-Afirm.	01	*vista*
31. Subst. Masc. Sing. e Verbo na 2ª pessoa do Pres. Subj.	02	*labores*
32. Subst. Masc. Sing. e Verbo na 1ª pessoa do Pret. Imp. Subj. / 1ª pessoa do Pret. Imp. Subj.	01	*impasse*
33. Subst. Masc. Sing. e Verbo na 3ª pessoa Pret. M-Q-Perf. / 1ª pessoa Pret. M-Q-Perf.	01	*fora*
34. Subst. Masc. Sing. e Verbo na 3ª pessoa Pret. Perf. Ind.	02	*leu*
35. Subst. Masc. Sing. e Verbo na 1ª pessoa Pret. Perf. Ind.	03	*pus*
36. Subst. Masc. Sing. e Verbo na 2ª pessoa Pret. Perf. Ind.	03	*guindaste*
37. Subst. Masc. Sing. Aum. e Verbo na 3ª pessoa pl. Fut. Pres. Ind.	01	*casarão*
38. Subst. Masc. Sing. Dim. e Verbo na 1ª pessoa Pres. Ind.	03	*passarinho*

39. Subst. Masc. Plural e Verbo na 2ª pessoa Fut. Pres. Ind.	04	*tangarás*
40. Subst. Masc. Plural e Verbo na 1ª pessoa Fut. Subj. / 3ª pessoa Fut. Subj. / 1ª pessoa Inf-pess. / 3ª pessoa Inf-pess. / 2ª pessoa Fut. Subj. / 2ª pessoa Inf-pess.	01	*soluçares*
41. Subst. Masc. Plural e Verbo na 2ª pessoa Fut. Subj. / 2ª pessoa Inf-pess.	76	*exemplares / olhares / pilares*
42. Subst. Masc. Plural e Verbo na 2ª pessoa Fut. Subj./ 2ª pessoa pl. Inf-pess.	01	*alardes*
43. Subst. Masc. Plural e Verbo na 1ª pessoa pl. Imp-Afirm. / 1ª pessoa pl. Pres. Subj. / 1ª pessoa pl. Pret. Perf. Ind.	01	*demos*
44. Subst. Masc. Plural e Verbo na 1ª pessoa pl. Inf-pess.	01	*termos*
45. Subst. Masc. Plural e Verbo na 2ª pessoa Inf-pess.	10	*poderes / seres*
46. Subst. Masc. Plural e Verbo na 1ª pessoa pl. Pres. Ind.	01	*pomos*
47. Subst. Masc. Plural e Verbo na 2ª pessoa Pres. Ind.	42	*cobres / programas*
48. Subst. Masc. Plural e Verbo na 2ª pessoa pl. Pres. Ind.	54	*casais / terminais*
49. Subst. Masc. Plural e Verbo na 2ª pessoa Pres. Subj.	272	*arames / interesses / professores*
50. Subst. Masc. Plural e Verbo na 2ª pessoa Pres. Subj. / 2G Part. Plural	01	*aceites*
51. Subst. Masc. Plural e Verbo na 2ª pessoa pl. Pres. Subj.	06	*metais*
52. Subst. Masc. Plural e Verbo na 2ª pessoa Pret. Imp. Subj.	01	*impasses*
53. Subst. Masc. Plural e Verbo na 2ª pessoa Pret. M-Q-Perf.	01	*foras*
54. Subst. Masc. Plural e Verbo na 2ª pessoa pl. Pret. Perf. Ind.	03	*lestes*
55. Subst. Masc. 2N e Verbo na 2ª pessoa Pres. Ind.	01	*cais*
56. Subst. Masc. 2N e Verbo na 2ª pessoa Pres. Subj.	01	*pires*
57. Subst. Fem. Sing. e Verbo na 3ª pessoa Fut. Pres. Ind.	01	*ceará*
58. Subst. Fem. Sing. e Verbo na 3ª pessoa Fut. Pret. Ind. / 1ª pessoa Fut. Pret. Ind.	83	*bateria / confeitaria / correria*
59. Subst. Fem. Sing. e Verbo na 1ª pessoa Fut. Pret. Ind. / 3ª pessoa Fut. Pret. Ind.	5	*drogaria*
60. Subst. Fem. Sing. e Verbo na 1ª pessoa Fut. Subj. / 3ª pessoa Fut. Subj. / 1ª pessoa Inf-pess. / 3ª pessoa Inf-pess.	01	*colher*
61. Subst. Fem. Sing. e Verbo na 3ª pessoa Imp-Afirm. / 3ª pessoa Pres. Subj. / 1ª pessoa Pres. Subj.	55	*equipe / peça*
62. Subst. Fem. Sing. e Verbo na 2ª pessoa Imp-Afirm.	01	*busca*
63. Subst. Fem. Sing. e Verbo na 2ª pessoa Imp-Afirm. / 3ª pessoa Pres. Ind.	647	*ajuda / ama / reforma*
64. Subst. Fem. Sing. e Verbo na 2ª pessoa pl. Imp-Afirm.	01	*sede*
65. Subst. Fem. Sing. e Verbo na 3ª pessoa Pres. Ind.	02	*busca*
66. Subst. Fem. Sing. e Verbo na 3ª pessoa Pres. Ind. / 2ª pessoa Imp-Afirm.	32	*janta / tabela*
67. Subst. Fem. Sing. e Verbo na 3ª pessoa pl. Pres. Ind.	01	*vagem*
68. Subst. Fem. Sing. e Verbo na 2ª pessoa Pres. Ind.	01	*esporas*
69. Subst. Fem. Sing. e Verbo na 3ª pessoa Pres. Subj.	01	*lata*
70. Subst. Fem. Sing. e Verbo na 1ª pessoa Pres. Subj. / 3ª pessoa Pres. Subj. / 3ª pessoa Imp-Afirm.	02	*volva*
71. Subst. Fem. Sing. e Verbo na 3ª pessoa Pret. Imp. Ind. / 1ª pessoa Pret. Imp. Ind.	12	*era / garantia*
72. Subst. Fem. Sing. e Verbo na 1ª pessoa Pret. Imp. Ind. / 3ª pessoa Pret. Imp. Ind.	01	*via*
73. Subst. Fem. Sing. e Verbo na 3ª pessoa Pret. M-Q-Perf. / 1ª pessoa Pret. M-Q-Perf.	07	*mentira*
74. Subst. Fem. Sing. e Verbo na 1ª pessoa Pret. M-Q-Perf. / 3ª pessoa Pret. M-Q-Perf.	01	*colhera*
75. Subst. Fem. Sing. e Verbo na 1ª pessoa Pret. Perf. Ind.	01	*devi*
76. Subst. Fem. Sing. Aum. e Verbo na 2ª pessoa Imp-Afirm. / 3ª pessoa Pres. Ind.	01	*beijoca*

77. Subst. Fem. Sing. Dim. e Verbo na 2ª pessoa Imp-Afirm. / 3ª pessoa Pres. Ind.	03	*caminha*
78. Subst. Fem. Sing. Dim. e Verbo na 3ª pessoa Pres. Ind. / 2ª pessoa Imp-Afirm.	02	*parcela*
79. Subst. Fem. Sing. Dim. e Verbo na 3ª pessoa Pret. Imp. Ind. / 1ª pessoa Pret. Imp. Ind.	02	*continha*
80. Subst. Fem. Plural e Verbo na 2ª pessoa Fut. Pres. Ind.	01	*cearás*
81. Subst. Fem. Plural e Verbo na 2ª pessoa Fut. Pret. Ind.	87	*matarias / selarias*
82. Subst. Fem. Plural e Verbo na 2ª pessoa Fut. Subj. / 2ª pessoa Inf-pess.	01	*colheres*
83. Subst. Fem. Plural e Verbo na 1ª pessoa Pres. Ind.	01	*fitas*
84. Subst. Fem. Plural e Verbo na 2ª pessoa Pres. Ind.	682	*casas / mentes / perguntas*
85. Subst. Fem. Plural e Verbo na 2ª pessoa pl. Pres. Ind.	02	*centrais*
86. Subst. Fem. Plural e Verbo na 2ª pessoa Pres. Subj.	61	*cores / tardes*
87. Subst. Fem. Plural e Verbo na 2ª pessoa Pres. Subj. / 2ª pessoa Pret. Imp. Ind.	01	*rias*
88. Subst. Fem. Plural e Verbo na 2ª pessoa Pret. Imp. Ind.	13	*vinhas*
89. Subst. Fem. Plural e Verbo na 2ª pessoa Pret. M-Q-Perf.	08	*trairas*
90. Subst. Fem. Plural Aum. e Verbo na 2ª pessoa Pres. Ind.	01	*beijocas*
91. Subst. Fem. Plural Dim. e Verbo na 2ª pessoa Pres. Ind.	04	*patinhas*
92. Subst. Fem. Plural Dim. e Verbo na 2ª pessoa Pret. Imp. Ind.	03	*mantinhas*
93. Subst. 2G e 2N e Verbo na 2ª pessoa Pres. Ind.	01	*choramingas*
94. Subst. 2G Plural e Verbo na 2ª pessoa Fut. Pret. Ind.	01	*cavalarias*
95. Subst. 2G Plural e Verbo na 2ª pessoa Fut. Subj. / 2ª pessoa Inf-pess.	01	*parlamentares*
96. Subst. 2G Plural e Verbo na 2ª pessoa Pres. Ind.	62	*guias / reservas*
97. Subst. 2G Plural e Verbo na 2ª pessoa pl. Pres. Ind.	03	*morais*
98. Subst. 2G Plural e Verbo na 2ª pessoa Pres. Subj.	07	*testes*
99. Subst. 2G Plural e Verbo na 2ª pessoa Pret. Imp. Ind.	01	*vigias*
100. Subst. 2G Sing. e Verbo na 1ª pessoa Fut. Pret. Ind. / 3ª pessoa Fut. Pret. Ind.	01	*cavalaria*
101. Subst. 2G Sing. e Verbo na 1ª pessoa Fut. Subj. / 3ª pessoa Fut. Subj. Ind. / 1ª pessoa Inf-pess. / 3ª pessoa Inf-pess.	01	*parlamentar*
102. Subst. 2G Sing. e Verbo na 3ª pessoa Imp-Afirm. / 3ª pessoa Pres. Subj. / 1ª pessoa Pres. Subj.	06	*corte*
103. Subst. 2G Sing. e Verbo na 2ª pessoa Imp-Afirm. / 3ª pessoa Pres. Ind.	62	*chama / serra*
104. Subst. 2G Sing. e Verbo na 3ª pessoa Pres. Ind. / 2ª pessoa Imp-Afirm.	03	*samba*
105. Subst. 2G Sing. e Verbo na 1ª pessoa Pres. Ind.	02	*banco*
106. Subst. 2G Sing. e Verbo na 3ª pessoa Pret. Imp. Ind. / 1ª pessoa Pret. Imp. Ind.	01	*vigia*
107. Subst. 2G Sing. e Verbo na 1ª pessoa Pret. M-Q-Perf. / 3ª pessoa Pret. M-Q-Perf.	01	*vira*
TOTAL	**4.456**	

Para cada combinatória apresentada na Tabela 4, a Tabela 5 apresenta os tipos de alteração efetuados em cada arquivo (são apresentados somente os arquivos que sofreram alguma alteração durante a análise, de inclusão ou exclusão de categorias lexicais). Nessa tabela, os símbolos + e - representam, respectivamente, a inclusão e exclusão de categorias. Essas, por sua vez, são representadas por suas iniciais: S – substantivo; ADJ – adjetivo; V – verbo; ADV – advérbio, etc.

Tabela 5: Tipos de alteração de cada arquivo de combinatória

Arquivo	Tipo de alteração
2	-S:2; +ADJ:2
4	-S:2; +ADJ:3
5	-S:14; +ADJ:2; -V:2
9	-S:14;ADJ:2; -V:2; +S:1
11	-S:4
13	-S:13
15	-S:4
16	-S:24
17	-S:18
19	-S:6
20	-S:1
26	-S:44; ADJ:6; +ADV:1;-V:1
31	-S:1
38	-Verb:1
39	-S:2
40	-S:1
41	-S:49
45	-S:3
47	-S:7
48	-S:5
49	-S:17
57	-S:1
58	-S:8
61	-S:4
63	-S:35
70	-S:2
71	-S:3
73	-S:2
74	-S:1
81	-S:7
84	-S:33;+ADJ:1
86	-S:8
88	-S:1
96	-S:4; +S:20
97	+S:1
98	-S:1; +S:2
100	+S:1
102	+S:3
103	-S:2;+S:18
105	+S:1
106	+S:1

Esse conjunto de arquivos, dividido entre a equipe de lingüistas, sofreu dois tipos de análise: a primeira, bastante semelhante à tarefa de limpeza descrita anteriormente, procurou refinar a base lexical ao a) retirar dela as informações irrelevantes, b) corrigir atributos morfológicos e c) inserir dados antes ignorados. A segunda análise, por sua vez, acompanhou uma diretriz particularmente voltada à reclassificação categorial, em favor da desambigüização. Vejamos, em detalhe, cada um dos procedimentos adotados para esses tipos de análise.

3.1. Refinamento da base lexical

Ainda que tivéssemos investido duramente na correção das informações lexicais (Seção 2), a identificação e manipulação dos dados correspondentes aos 107 tipos combinatoriais de ambigüidade entre substantivo-verbo revelou a persistência de algumas imprecisões (sobretudo morfológicas), a ocorrência, ainda, de dados irrelevantes e a necessidade de consideração de informações antes ignoradas. Mostrou-se essencial, portanto, avaliar e corrigir cada uma das entradas lexicais que compunham os 107 arquivos elencados na Tabela 4. Particularmente, nessa etapa do trabalho dedicamo-nos a informações de outra ordem, diferentemente daquelas relatadas na seção anterior: não estão mais em foco dados incorretos, mas sim dados lexicográficos, propriamente ditos, que respondem pelo domínio e situação de uso das formas lingüísticas.

3.1.1. Refinamento por exclusão de casos

Foram definidos, assim, alguns critérios para nortear a padronização das exclusões dos dados lexicais em análise, conforme os seguintes procedimentos:

a) Excluir a categoria 'substantivo' das formas variantes em grau (diminutivo e aumentativo), as quais não representam valor próprio de significado, dependente de suas canônicas.
b) Excluir a categoria 'substantivo' das formas assim rotuladas no léxico, mas que são, na verdade, apenas verbos no infinitivo.
c) Excluir a categoria 'substantivo' quando, a partir do dicionário de referência, ficar constatado que se trata de uma variante semântica de outra forma mais consagrada na língua.
d) Excluir a categoria 'substantivo' daquelas formas que, de acordo com o dicionário de referência: d.1) são nomes de objetos ou fatos de um domínio muito específico; d.2) são nomes em desuso ou do português arcaico; d.3) representam gírias de uso pouco difundido; e d.4) representam regionalismos bastante restritos.

O caso (a) diz respeito à variação em grau, que, apesar de já investigada anteriormente no NILC, não foi alvo de implementação sistemática até o momento. Há alguns poucos substantivos no léxico do ReGra que apresentam o traço de variação em grau, cuja relevância é questionada, em termos de sua representatividade e conseqüente utilidade para o ReGra. Na ausência de um processamento morfológico que tratasse desse fenômeno, essas formas foram inseridas diretamente no léxico, constituindo-se entradas individuais de palavras, distintas de sua canônica apenas pelo traço 'aumentativo' ou 'diminutivo'. Essa inserção, contudo, não atingiu uma escala representativa da língua. Além disso, é bem possível que apenas algumas dezenas de substantivos se incluam nesse perfil, demonstrando que a sua existência na base lexical pouco importa ao processamento lingüístico do revisor. A par dessas considerações, o critério de exclusão das formas variantes em grau que não possuíssem significado próprio foi adotado, posto que, além de nada contribuírem para o bom desempenho da ferramenta, são formas que aumentam a estatística de ambigüidade categorial entre substantivo e verbo. No entanto, foram mantidas as palavras que, ao incorporarem as desinências de diminutivo e aumentativo, produziram significados independentes e fechados. Exemplos de ambos os casos seguem abaixo:

aninho	atinha	continha	mantinha

Representações em grau, inexpressivas no léxico

passarinho	patinho	casarão

Representações em grau, mantidas no léxico

O caso (b), agora, ocupa-se de admitir apenas a categoria verbo para o item lexical, deixando por conta do próprio *parser* decidir pela variação categorial, quando necessário. Algumas formas lexicais sabidamente infinitivas impessoais também são substantivos bastante expressivos do português ordinário. Contudo, essa característica não é um comportamento genérico e sistemático do infinitivo verbal. Ao contrário, a grande maioria dos verbos da nossa língua não apresentam um significado específico correspondendo a um substantivo. Por exemplo: a palavra 'olhar' é um infinitivo impessoal ligado à canônica 'olhar' e também um substantivo da qual se origina a forma flexional 'olhares'. Diferentemente, a palavra 'querer' é infinitivo impessoal relacionado à canônica 'querer', mas não sustenta um significado próprio na acepção de substantivo. Dessa forma, não pode, inclusive, dar origem à flexão 'quereres'. É certo que muitos dicionários costumam abonar, nesses casos, a acepção de substantivo, apontando, nos verbetes, expressões do tipo "é um ato de..." ou fazendo o consulente circular no próprio dicionário já que esses substantivos são, quase sempre, variantes de outras formas, normalmente muito mais usuais na língua. Tal imprecisão não deve ser reproduzida no nosso léxico, motivo pelo qual adotamos o critério de exclusão absoluta da categoria 'substantivo' daqueles verbos que não apresentarem motivação intuitiva ou respaldo razoável do dicionário de referência no que diz respeito à sua natureza substantival. A seguir, ilustramos alguns casos inseridos nesse perfil:

coaxar	pôr	rosnar	soluçar	ver
pensar	proceder	saudar	trovejar	vozear
piscar	quebrar	sentir	vagar	

Verbos de substantivação infreqüente

É importante ressaltar, ainda, que a exemplo do que acontece com relação aos adjetivos que usualmente são substantivados em decorrência da elipse do substantivo ('o abandonado', por exemplo), os casos de substantivação do infinitivo impessoal também são satisfatoriamente tratados por regras específicas do *parser* do revisor. Nesse sentido, a exclusão da categoria substantivo nesses casos não será problema para o processamento de sentenças como "O sentir é diferente do pensar", em que 'sentir' e 'pensar' serão categorizados, pelo contexto, como substantivos masculinos singulares, muito embora não figurem no léxico como tal.

Falamos acima sobre um procedimento peculiar dos dicionários brasileiros que leva o consulente a uma busca incansável pela obra, porque a palavra investigada é uma variante dialetal, regional ou de uso arcaico de uma forma reconhecidamente mais usual da língua. Seguindo raciocínio similar, o caso (c) se ocupa de fazer com que esse tipo de padrão lexicográfico seja inibido no léxico do nosso sistema, já que ele não trata de um repositório do vocabulário diacrônico do português, mas sim de uma base que armazena as formas válidas e genericamente empregadas do português contemporâneo. Por conta disso, foi critério adotado, nessa etapa de refinamento, a exclusão de formas que, na acepção de substantivo, apontam para um movimento semântico circular. Por exemplo: a

palavra 'sofisticaria', como substantivo, é variação pouco usada de 'sofisticação', claramente mais empregada na língua. Os casos a seguir ilustram ainda mais esse procedimento de exclusão:

(a) escapula (var. de 'escapatória')	(a) calca (var. de 'calcadura')
(a) escapa (var. de 'escapada')	(o) arribe (var. de 'arribação)
(a) recolha (var. de 'recolhimento')	(o) empaste (var. de 'empastamento')
(a) sovinaria (var. de 'sovinice')	(o) barral (var. de 'barreira')
(a) arranca (var. de 'arrancamento')	(a) escuma (var. de 'espuma')
(a) abaderna (var. de 'baderna')	(a) folhada (var. de 'folhagem')
(a) cabeceada (var. de 'cabeçada')	(a) arruado (var. de 'arruamento')

Variantes pouco representativas da língua

O procedimento sugerido em (d) já foi considerado na tarefa anterior de correção do léxico (Seção 2, caso (b)), mas, diante da persistência de algumas palavras que se inserem nesse perfil semântico, nessa etapa de refinamento da base lexical continuamos a adotar o critério de exclusão dos substantivos restritos ao domínio, à situação de uso e a regionalismos. A título de ilustração, apresentamos abaixo alguns exemplos de palavras retiradas do léxico por conta desse critério:

(a) arcaria (domínio específico da Arquitetura)	(o) espeque (brasileirismo empregado no NE)
(a) bicharia (termo burlesco e chulo. Além disso, forma variante pouco usada de 'bicharada')	(o) quiete (termo especificamente poético)
(a) luxaria (brasileirismo e forma popular)	(o) abre (gíria pouco usada referente à 'cachaça')
(a) cave (galicismo em desuso no Brasil)	(o) apostema (domínio específico da medicina)
(a) amura (domínio específico da construção naval)	(o) deva (domínio específico da religião)
(a) apaga (domínio específico da marinha)	(a) estruma (domínio específico da endocrinologia)
(a) cruza (brasileirismo empregado no sul)	(a) parangona (lusitanismo referente às artes gráficas)
(a) enclausura (termo pouco usado; var.: clausura')	(a) cavala (brasileirismo e termo folclórico carioca)
(o) aderece (termo em desuso. Forma atual: 'adereço')	

Ocorrências inexpressivas no uso geral do português

Outros

tipos de refinamento lexical por exclusão foram ainda observados. Por exemplo, algumas formas que não constavam no dicionário de referência foram excluídas, tais como a) a canônica do verbo 'caldeiro' (pres-eu), porque não há as palavras 'caldeirar' nem 'calderar' e b) algumas flexões relacionadas a verbos defectivos, sendo exemplos desse caso:

Verbete	Categorização anterior	Categoria excluída
começo	subst-masc-sing (1), pres-eu (começar) (2), pres-eu (comedir) (3)	3
rugir	fut-subj-ele (1), fut-subj-eu (2), inf-pess-ele (3), inf-pess-eu (4), subst-masc-sing (5)	2 & 4
ruge	subst-masc-sing (1), imper-afirm-tu (2), pres-ele (3)	2

Exclusão de algumas flexões verbais

3.1.2. Refinamento por inclusão de casos

Neste caso, não houve um procedimento sistemático para a padronização de inclusões de informações lexicais, pois somente alguns casos isolados foram observados. Por exemplo, foram incluídas com prioridade baixa, pela sua freqüência do uso, algumas categorias de adjetivo, advérbio e adjetivo 2G. É o caso das palavras 'afeto', 'alegro' e 'viúvo', às quais tais categorias foram respectivamente associadas.

Ao final da etapa de refinamento da base lexical, passamos a ter os seguintes dados de alteração no léxico, relativos somente à ambigüidade entre as categorias substantivo-verbo, i.e., às 4.456 entradas apresentadas na Tabela 4.

Tabela 6: Limpeza dos arquivos relativos aos tipos combinatoriais do léxico

Tipo de alteração	No. de entradas	%
Exclusão de substantivos	343	7,697
Exclusão de adjetivos	--------	--------
Exclusão de verbos	05	0,112
Exclusão de outras categorias	--------	--------
Exclusão completa de verbete	01	0,022
TOTAL DE EXCLUSÕES	**349**	**7,831**
Inserção de substantivos	48	1,077
Inserção de adjetivos	16	0,359
Inserção de verbos	--------	--------
Inserção de outras categorias	01	0,022
TOTAL DE INSERÇÕES	**65**	**1,458**

Os números finais, incluindo a correção inicial descrita na seção anterior e esse refinamento, que contou com exclusões e inserções de palavras e/ou categorias morfológicas, podem ser contrastados com os números iniciais do arquivo substantivo-verbo, objeto de análise dessa proposta metodológica de desambigüização. Nesse sentido, os dados puramente estatísticos apontam para o seguinte quadro de alteração da base lexical do sistema:

Tabela 7: Síntese do refinamento do léxico

Tipo de dados	Pré-análise	Pós-análise	Redução da ambigüidade
No. de entradas (subst-verbo)	7.028	4.161	2.867

A etapa seguinte ao refinamento, conforme apontamos no início dessa seção, foi a análise do conteúdo dos 107 arquivos selecionados para o estudo da desambigüização. É o que passaremos a relatar a seguir.

3.2. Análise categorial para a reclassificação do léxico

O esforço da desambigüização lexical, conforme a especificidade da nossa tarefa com respeito ao léxico do ReGra, levou-nos a definir algumas etapas distintas de tratamento da ambigüidade. De um lado, a experiência com alguns resultados satisfatórios relacionados ao estudo da freqüência de uso das palavras e, de outro lado, a influência conhecida dos fatores sintáticos e semânticos na especificação de um determinado comportamento lexical conduziram-nos à busca de uma forma de

desambigüização que refletisse esses três aspectos lingüísticos: freqüência, contexto sintático e informação semântica. Relatamos, aqui, somente o trabalho concluído sobre a freqüência de uso. A abordagem analítica envolvendo os demais aspectos constitui trabalho em progresso no projeto e, por essa razão, ainda não temos resultados conclusivos.

Supondo que seja possível priorizar certas acepções de formas ambíguas quando essas sugerirem à intuição de falante e de especialista da língua uma maior taxa de ocorrência, a informação sobre a freqüência de uso pode representar uma redução dos casos de ambigüidade lexical que importam ao revisor. Por exemplo: a ambigüidade envolvendo as 2as. pessoas (singular e plural) dos verbos e qualquer outra categoria morfológica (substantivo, adjetivo, etc.), pode ser tratada pela freqüência ao definirmos que o processamento lingüístico do português deve prescindir da acepção do verbo em favor da outra categoria que a forma requerida acumula na classificação lexical. Assim, se tomarmos a palavra 'portais' (2ª pessoa do plural do verbo 'portar' e substantivo masculino plural ligado à canônica 'portal'), diremos que a intuição nos leva a decidir pela prioridade do substantivo em relação ao verbo, porque as 2as. pessoas verbais são de uso muito raro no português atual.

É certo que um estudo exaustivo da freqüência seria mais adequado a uma iniciativa como essa. Contudo, também é sabido que esse tipo de informação não determina o fim da ambigüidade lexical, além de somente ser possível aplicá-la a um número pouco expressivo do vocabulário do português. Sendo assim, determinamos que o estudo da freqüência, para esse trabalho de desambigüização entre os substantivos e verbos, deveria recair sobre alguns casos específicos, muito especialmente aqueles que a sugestão intuitiva pudesse decidir. Na consecução dessa diretriz, portanto, propusemos a inversão da ordem da classificação lexical segundo o critério da freqüência nos casos que apresentamos a seguir:

a) Quando a ambigüidade envolver as 2as. pessoas verbais.
b) Quando a ambigüidade envolver um tempo ou modo verbal de emprego menos comum que outros presentes na classificação da palavra no léxico.
c) Quando a ambigüidade envolver um substantivo, cujo significado aponta: c.1) para um domínio muito particular (embora não seja específico de uma área de conhecimento); c.2) para gíria ou coloquialismo ou termo muito popular; ou c.3) quando o substantivo, pelo seu significado, for notadamente menos comum que a forma verbal.

No caso (a), convertemos as pessoas verbais em questão para a última posição, na ordem de prioridade de especificação lexical. Curiosamente, esses casos não foram encontrados nos arquivos (cf. Tabela 4) envolvendo o substantivo masculino plural, único material em que apenas duas categorias determinavam a ambigüidade (no caso, o substantivo e alguma forma verbal). Ali, as 2as. pessoas verbais já ocupavam a última posição na classificação morfológica das palavras e, portanto, a inversão não foi aplicada. Em relação aos demais arquivos, embora fosse alterada a ordem de prioridade das formas ambíguas, resultando na última posição para tais categorias, vale notar que a ambigüidade entre o substantivo e o verbo persiste, para outras pessoas verbais. Isto quer dizer que nossa tarefa somente distanciou as 2as. pessoas do quadro geral de ambigüidade categorial representada no léxico. No entanto, as resoluções possíveis de ocorrências ambíguas para tais classes morfológicas, agora, tendem a ser mais eficientes,

devido a esse distanciamento. Casos dessa natureza são apresentados a seguir (os números entre parênteses indicam as categorias morfológicas sob análise):

Verbete	Categorização anterior	Inversão realizada
ajuda	imp-afirm-tu (1), pres-ind-ele (2), subst-fem-sing (3)	231
anestesia	imp-afirm-tu (1), pres-ind-ele (2), subst-fem-sing (3)	321
casa	subst-fem-sing (1), imp-afirm-tu (2), pres-ind-ele (3)	132
fita	subst-fem-sing (1), imp-afirm-tu (2), pres-ind-ele (3), adj-fem-sing (4)	1342
posta	imp-afirm-tu (1), pres-ind-ele (2), adj-fem-sing (3), subst-fem-sing (4), part-fem-sing (5)	35214
revista	subst-fem-sing (1), imp-afirm-ele-rever (2), pres-subj-ele (3), pres-subj-eu (4), imp-afirm-tu-revistar (5), pres-ind-ele (6), part-fem-sing (7)	1763425

Inversão da classificação lexical para 2as. pessoas verbais

Similarmente a várias etapas anteriores deste trabalho, assumimos, no caso (b), que o processamento lingüístico do revisor pode se tornar mais ágil se puder ser estabelecida a prioridade das formas verbais segundo sua prioridade (presumida) de processamento mental, de compreensão das sentenças, ou segundo sua prioridade de uso. Por exemplo, é lícito dizer que o imperativo, no português, é menos comum que o presente do indicativo, razão pela qual a acepção correspondente ao presente deve ser privilegiada, durante o processamento. No entanto, muitos eram os casos, no léxico, de palavras ambíguas cujos verbetes apresentavam, como primeira categorização verbal, o modo imperativo. A inversão da ordem de classificação, nesses casos, seguiu mais uma vez a hipótese de que é possível reduzir a incidência de erros ou o tempo de processamento durante a resolução da ambigüidade. Exemplos de palavras relativas a este caso são dados a seguir:

Verbete	Categorização anterior	Inversão realizada
corra	imp-afirm-ele (1), pres-subj-ele (2), pres-subj-eu (3), subst-fem-sing (4)	2314
diagrama	imp-afirm-tu (1), pres.ele (2), subst-masc-sing (3)	321
esgrima	imp-afirm-ele (1), pres-subj-ele (2), pres-subj-eu (3), subst-fem-sing (4)	4123
objetivo	adj-masc-sing (1), subst-masc-sing (2), pres-eu (3)	213

Inversão da classificação lexical para tempos ou modos verbais menos usuais

Já a inversão da classificação no caso (c) busca refletir o uso mais consagrado de algumas formas do verbo em detrimento das formas substantivais correspondentes, pouco usuais. Diferentemente dos procedimentos anteriores, desta vez não se trata de excluir a categoria substantivo, mas de afastar a prioridade com que essa acepção é acionada pelo processador. Tal medida se baseia na impossibilidade de se decidir, a priori, pela exclusão de palavras cujas acepções podem ser tão infreqüentes que mereçam já não figurar no léxico. Assim, assumimos, neste caso, que quando a ambigüidade envolver somente categorias verbais ou substantivas, as mais freqüentes serão privilegiadas; quando envolver o par substantivo-verbo, a prioridade será dada ao verbo. Seguem alguns exemplos desses casos:

Verbete	Categorização anterior	Inversão realizada
cantaria	subst-fem-sing (1), fut-pret-ele (2), fut-pret-eu (3)	231
contrastaria	subst-fem-sing (1), fut-pret-ele (2), fut-pret-eu (3)	231

pene	subst-masc-sing (1), imper-afirm-ele (2), pres-subj-ele (3), pres-subj-eu (4)	4321

Reclassificação lexical para substantivos e verbos de significação particular

Conforme assinalamos, a análise dos arquivos segundo o critério da freqüência, para o par substantivo-verbo, procurou minimizar o problema da ambigüidade categorial durante a revisão gramatical, invertendo a prioridade de classificação morfológica das palavras da língua. Com base nos procedimentos descritos, é possível notar que nosso trabalho, nesta etapa, acabou recaindo no problema de classificação categorial, em vez de recair sobre o problema de resolução categorial para a desambigüização, propriamente dita. É certo que, apesar de se tratarem de alterações que podem valer a qualquer objetivo de processamento lingüístico, por serem de natureza geral – de uso ou de interpretação – da língua, elas foram motivadas pela necessidade de minimizar o esforço de processamento frente ao fenômeno da ambigüidade lexical. Logo, esse exame empírico ofereceu-nos a vantagem de tomarmos ciência e procurarmos minimizar os problemas de especificação lexical e, ainda, reiterarmos a já apontada necessidade de, oportunamente, expandirmos esse trabalho para outros pares categoriais que apresentem ambigüidade.

A aplicação do estudo da freqüência, nos padrões que definimos para essa tarefa em particular, resultou no quadro de alterações que apresentamos a seguir, cuja porcentagem é relacionada ao total de verbetes definidos na Tabela 4 (4.456 entradas):

Tabela 8: Síntese do refinamento do léxico (por reclassificação de substantivo-verbo)

Tipo de alteração	No. de entradas	%
Inversão de categorias entre duas formas	175	3,92
Inversão de categorias entre mais de duas formas	368	8,25

4. Perspectivas futuras do Projeto TraSem{XE "Perspectivas futuras do Projeto TraSem"}

A alteração da base de dados lexical, levada a efeito para a correção dos problemas de dicionarização verificados, reduziu efetivamente o número de escolhas equivocadas da ferramenta, em relação à categoria gramatical das formas homógrafas. No entanto, é forçoso considerar que os problemas semânticos relatados no início deste trabalho não foram ainda abordados. Para a sua consideração, e para evitar que nos detenhamos sobre pseudo-problemas de classificação, além da correção das entradas lexicais já empreendidas para palavras com dupla classificação de substantivo e verbo, está em curso o estudo do contexto sintático de ocorrência de algumas de suas formas ambíguas, selecionadas com base em uma nova análise de ocorrência no corpus do NILC.

Primeiramente, estamos procedendo à verificação do contexto de ocorrência das palavras de interesse pela distribuição sintática dos componentes à sua direita e esquerda, visando à formulação de regras de desambigüização sintática que venham a eliminar boa parte dos problemas de intervenções indevidas ou de não-intervenções indevidas atualmente cometidos pelo ReGra. No entanto, desconfiamos que haverá, ainda assim, problemas residuais cuja solução não encontrará respaldo no contexto sintático, devido à sua natureza essencialmente semântica. Esses serão os casos de interesse real neste projeto de pesquisa, que merecerão atenção especial pela necessidade de representação do conteúdo semântico. Muito provavelmente, então, recairemos no problema inicial, de

especificar os traços semânticos dos itens lexicais com base em fundamentos ontológicos (cf. apontados em Rino et al., 2001). No entanto, vale notar que, agora, dada a metodologia estratificada em três etapas – de freqüência de uso, de verificação sintática e, finalmente, de verificação semântica – nosso problema de representação e manipulação semântica foi minimizado, podendo concentrar-se naqueles casos em que esse tipo de resolução é realmente imprescindível.

Referências bibliográficas

Borba, F. da Silva (1990), *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil*. Editora da UNESP, 2ª ed. São Paulo.

Borba, F. da Silva (1996). *Uma gramática de valências para o português*. Ática. São Paulo.

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.

Rino, L.H.M.; Martins, R.T.; Marchi, A.R.; Kuhn, D.C.S.; Pinheiro, G.M; Pardo, T.A.S.; Felippo, A. (2001). *Projeto TraSem: A investigação teórica sobre o problema da ambigüidade categorial*. Tech. Rep. NILC-TR-01-1. São Carlos, Março.