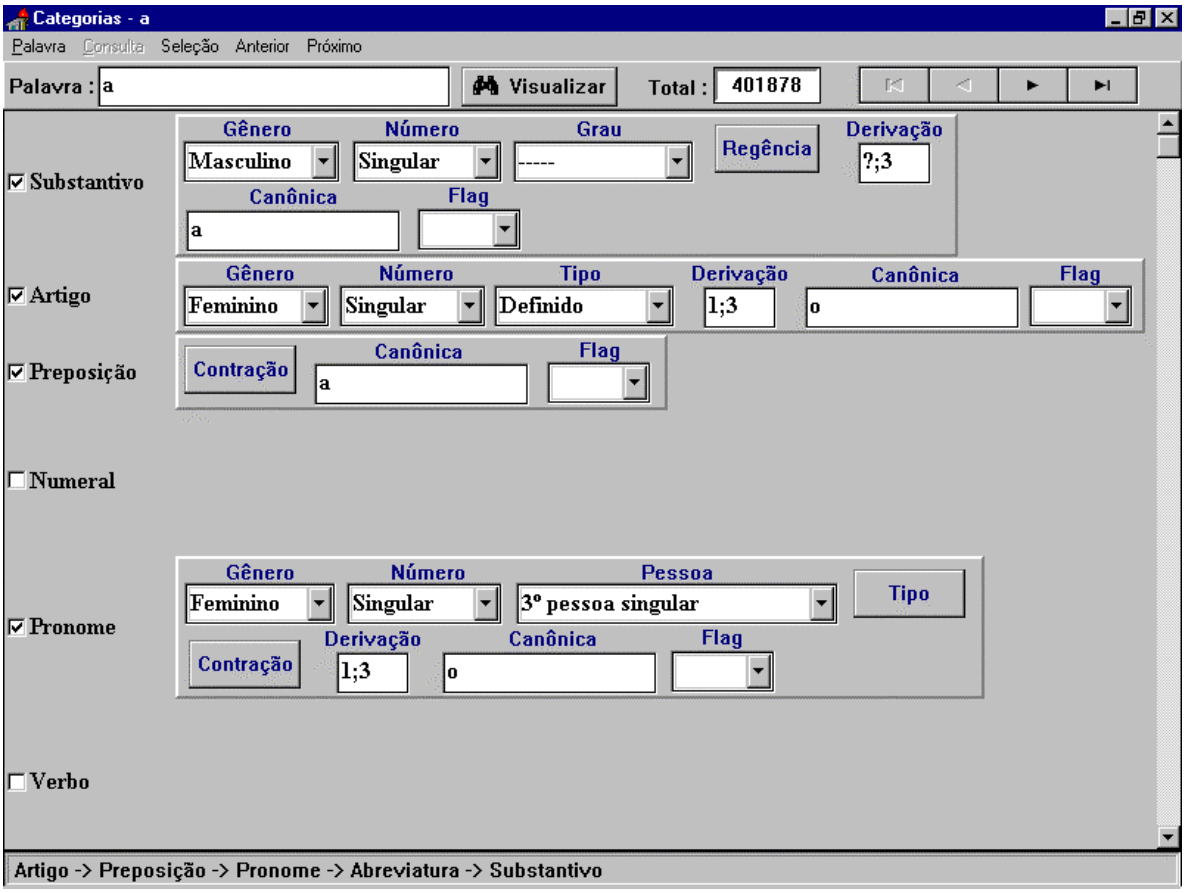

Figura 6.1: Janela do gerenciador apresentando a palavra ***a*** com suas classes.

Agradecimentos

Os autores gostariam de agradecer a colaboração de toda a equipe do convênio USP-Itautec/Philco, especialmente a Ricardo Hasegawa, Marcelo Turine, Marcelo Couto, Ronaldo T. Martins, Osvaldo N. Oliveira Jr., Claudete Moreno, Maria Cristina F. Oliveira, Gisele Montilha e Teresa Martins, pela ajuda efetiva em inúmeras oportunidades dessa jornada. Este trabalho também é fruto do trabalho árduo de cada um deles. Os autores também agradecem o suporte financeiro da Itautec-Philco e do CNPq.

Referências

[Bechara 92] Bechara, E. *Moderna gramática portuguesa*. São Paulo: Nacional,1992.

[Cunha 85] Cunha, C. *Gramática da língua portuguesa*. Rio de Janeiro: Nova Fronteira, 1985.

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

[Kowaltowski et al. 1995a] Kowaltowski, T.; Lucchesi, C.L.; Stolfi, J. "Minimization of Binary Automata". *Journal of the Brazilian Computing Society*, 3(1), 36-42, 1995.

[Kowaltowski et al. 1995b] Kowaltowski, T.; Lucchesi, C.L.; Stolfi, J. "Application of finite automata in debugging natural language vocabularies". *Journal of the Brazilian Computing Society*, 3(1), 05-11, 1995.

[Lima 72] Lima, C. H. da R. *Gramática normativa da língua portuguesa*. Rio de Janeiro: José Olympio, 1972.

[Nunes et al. 1996] Nunes, M.G.V.; Vieira, F.M.C.; Zavaglia, C.; Sossolote, C.R.C.; Hernandez, J. "A Construção de um Léxico para a Língua Portuguesa do Brasil para Suporte à Correção Automática de Textos". *Relatório Técnico do ICMSC-USP*, nro.42, 36p., Setembro 1996.

forma de abonar a inserção de itens não-dicionarizados, utilizou-se a informação dos itens faltantes para a ampliação e o enriquecimento da base lexical.

Um outro papel importante do corpus é o de possibilitar pesquisas que visem uma hierarquização morfológica mais substanciada, a partir do cálculo de freqüência das ocorrências no corpus. Estatísticas sobre o corpus que visem o preenchimento de informações semânticas no Léxico também estão previstas como tarefas num futuro próximo. Com o intuito de tornar o corpus uma base de apoio à pesquisa ainda mais efetiva, recentemente foi adquirido um conjunto de ferramentas de manipulação de corpora lingüísticos — *the IMS corpus query tools*, da Universidade de Stuttgart — que inclui diversas facilidades para acesso, pesquisa e estatística em corpora. Uma possibilidade bastante interessante da ferramenta é a de manipular corpus anotado (*tagged corpus)*, seja morfológica, sintática ou semanticamente. Com o uso do nosso Léxico, já somos capazes de anotar o corpus com informações morfológicas; com o uso de nosso revisor, temos uma boa perspectiva de, em pouco tempo, conseguir o mesmo com informações sintáticas. A partir de um corpus anotado, de tamanho tão expressivo como o nosso, acreditamos poder progredir muito em nossas pesquisas envolvendo o uso da língua escrita portuguesa correntemente empregada no Brasil, em tarefas automáticas que incluem a revisão gramatical, a tradução e a geração de textos.

6. O Sistema Gerenciador do Léxico

O sistema gerenciador do Léxico está dividido em duas partes funcionais: armazenamento das palavras e interface de acesso.

A primeira parte é responsável pela organização dos dados pertencentes ao Léxico, ou seja, as palavras devem estar armazenadas de forma segura e de fácil acesso. Para isso, utilizou-se o gerenciador de base de dados Paradox 5.0 (Windows) no tratamento da inserção, remoção e atualização das palavras. Recentemente, a base de dados foi transportada para o Gerenciador Oracle, a fim de buscar uma maior segurança no armazenamento das informações existentes no Léxico. Atualmente, o Léxico conta com cerca de 1.500.000 entradas que, juntamente com seus atributos, ocupam cerca de 88MB de arquivo texto. Quando compactado, já no formato adequado para uso dos corretores, este arquivo passa a ocupar apenas 1 MB. Estatísticas apontam para a presença de cerca de 1.400.000 formas verbais, sendo que mais de um milhão delas são constituídas de ênclises e mesóclises, 46.000 substantivos, 47.400 adjetivos, 4.000 nomes próprios, 1.750 advérbios, para citar apenas as classes maiores.

A interface de acesso é ilustrada pela Figura 6.1. Através desta e outras janelas, o usuário pode: visualizar uma palavra já existente no léxico; inserir, remover e atualizar os diversos atributos uma palavra; consultar uma determinada classe de palavras; consultar por prefixo ou sufixo; hierarquizar das classes de uma palavra. Funções de busca de palavras, através do uso de máscaras, também estão disponíveis.

O usuário conta com menus de opções em várias situações, facilitando o preenchimento dos campos necessários, como é o caso da apresentação de listas de preposições para regência verbal e nominal, predicação e tempo verbais e pronomes.

7. Conclusões

Este trabalho descreve o projeto de construção de um léxico que dá suporte a sistemas revisores ortográfico e gramatical de textos em Português do Brasil. A versão aqui descrita consumiu trabalho integral de três lingüistas, dois informatas, além de vários colaboradores eventuais, por um período de 11 meses. Este projeto continua em desenvolvimento, sendo que nossa meta é construir um léxico o mais abrangente possível em termos de quantidade de palavras e de suas categorizações. Além disso, almejamos inserir informações semânticas às entradas lexicais, com o objetivo de dar suporte, num futuro próximo, a sistemas de interpretação e de tradução.

classe, maior divergência entre os gramáticos quanto à categorização a ser observada. A tipologia adotada foi a seguinte: Afirmação, Negação, Dúvida, Intensidade, Circunstância de Lugar, Circunstância de Tempo, Circunstância de Modo, Interrogativo de Lugar, Interrogativo de Causa, Interrogativo de Tempo, Interrogativo de Modo. Para esta classe, o registro da canônica e do grau constituiu uma informação necessária.

É importante salientar que, das palavras constantes do Léxico, os advérbios terminados em *-mente* foram aqueles gerados automaticamente, através da seguinte instrução: identificados os adjetivos, acresça-se *-mente* à forma feminina dos itens cuja flexão se marca através de sufixos. Ex: atent*o* (adj.), atent*a* (adj.) **/** atenta*mente***.** Para os categorizados como 2g., a formação do advérbio seria feita apenas pelo acréscimo de *-mente*. Houve, contudo, a necessidade de proceder a certas correções, visando a acomodação gráfica. Assim, na geração automática, derivou-se, de maneira inapropriada, por exemplo, a forma comu*mm*ente a partir do adj. 2g. (Ex: vocábulo *comum* / expressão *comum*). Esta verificação, todavia, foi a das mais simples. Como a derivação automática desses advérbios foi feita, inicialmente, a partir de um léxico pluricategorizado, ou seja, a partir de um léxico em que as palavras estavam distribuídas por mais de uma classe de palavras, ao aplicar-se a regra já descrita, gerou-se formas inadmissíveis. *Abacaxi* é uma delas, já que este item encontrava-se classificado no Léxico como substantivo e como adjetivo. *Abacaximente* constitui uma forma inaceitável que, em hipótese alguma, poderia constar do Léxico que serve, igualmente, de base ao corretor ortográfico. Nesta primeira fase, foi necessário proceder a uma exaustiva verificação manual. Em um momento posterior, quando a filosofia do projeto mudou radicalmente, dada a opção pela seleção entre substantivos e adjetivos, de forma a evitar-se a pluricategorização, a geração automática mostrou-se mais eficiente. Assim, pôde-se, à medida que se inseria novos itens lexicais, proceder a uma derivação mais segura de advérbios em *-mente* a partir de bases adjetivas. É bem verdade que há restrições a esta derivação a partir de adjetivos, diagnosticadas pelo recurso à intuição.

5. O Papel do Corpus

À medida em que se constatava a ausência de palavras de alta freqüência na Língua (a base inicial foi o conjunto de verbetes do Dicionário Eletrônico da Melhoramentos), procurava-se inseri-las no Léxico. Antes, contudo, procedia-se à consulta aos dicionários disponíveis, a fim de "legitimar" a inserção dos itens faltantes. No entanto, constatou-se, através da consulta, que palavras comuns para usuários de nível médio não se encontravam dicionarizadas. Ex: *acessar, mentalização*.

Tornou-se necessário, então, definir parâmetros para a ampliação do Léxico em construção. Se, por um lado, são os dicionários que legitimam o uso, por outro, é impossível desconsiderar as ocorrências de uso de um dado item lexical, a partir da intuição do especialista enquanto falante da língua. Esta dupla questão nos levou a utilizar o banco de textos existente como parâmetro para a ampliação de nossa base lexical. Coube, então, aos informatas manipular os textos do banco de forma a quantificar em ordem decrescente os vocábulos de maior ocorrência (20 foi o índice de freqüência definido). Feita a quantificação, o próximo passo consistiu na comparação dos itens quantificados com aqueles constantes da base lexical. Os itens faltantes no Léxico e presentes no banco de textos alimentado durante todo o projeto foram, então, inseridos. Assim, o banco de textos, criado com o objetivo de possibilitar, de um lado, o levantamento dos erros mais freqüentes dos usuários da língua — usuários dos quais se espera um conhecimento de nível médio definido a partir de sua escolaridade, o segundo grau completo — e de outro, a execução de testes com a finalidade de avaliar o grau de otimização das regras implementadas com vistas à correção automática de textos escritos não-literários do Português Contemporâneo, passou a ter uma nova função: licenciar a inserção de palavras até o momento não-dicionarizadas. Ao mesmo tempo em que se buscou o banco de textos como

"nenhures", "quem" foram classificados como 2g e 2n*; "nada", "tudo", "alguém/ninguém", "outrem"* foram classificados como Masculino / singular.

Aos pronomes possessivos foi mantida a classificação que relaciona o pronome com as pessoas do discurso, mesmo que a sua concordância verbal dentro de uma frase seja feita em 3ª pessoa ora do singular ora do plural. E assim temos *"meu"* classificado como 1ª pessoa do singular, masculino, singular; *"tuas"* como 2ª pessoa do singular, feminino, plural e assim por diante.

Os pronomes pessoais retos não apresentaram nenhum tipo de problema uma vez que eles designam diretamente as pessoas do discurso. Quanto ao gênero, os únicos pronomes pessoais retos marcados de tal flexão são *"ele"* e *"ela". "Eu", "tu", "nós", "vós"* não trazem marcas de gênero e intuitivamente se decidiu classificá-los como 2g, já que podem se referir tanto a uma pessoa do sexo masculino quanto do feminino. Quanto ao número, os pronomes *"eu", "tu", "ele", "ela"* foram classificados como singulares e *"eles", "elas", "nós", "vós"* como plurais.

Os pronomes pessoais oblíquos *"me", "mi", "comigo", "te", "ti", "contigo", "nos", "conosco", "vos", "convosco", "lhe", "lhes",* quanto ao gênero, foram classificados como 2g. *"Se", "si"* e *"consigo"* são pronomes oblíquos tanto da terceira pessoa do singular quanto do plural e por isso foram classificados na opção 3ª pessoa singular e plural, além de 2g.

Os pronomes de tratamento são de terceira pessoa quando assunto do discurso, mas são de segunda pessoa quando se referem à pessoa. A classificação realizada tratou-os como pronomes pessoais, o que de fato o são, mesmo sabendo que eles se comportam gramaticalmente como 3ª pessoa na concordância verbal de uma frase, e assim, temos por ex.:
"Você" = 2g / Sing. / 2ª p. sing.; *"Vocês"* = 2g / Pl. / 2ª p. pl.; *"Senhor"* = M / Sing. / 2ª p. sing.; *"Senhores"* = M / Pl. / 2ª p. pl.; *"Senhora"* = F / Sing. / 2ª p. sing.; *"Senhoras"* = F / Pl. / 2ª p. pl.

4.6 Conjunções

Dado que as palavras foram tipificadas a partir das classes de palavras propostas pelas Gramáticas Tradicionais, procedeu-se a um estudo exaustivo junto às Gramáticas Normativas do Português, com a finalidade tanto de fazer um levantamento exaustivo sobre as formas conjuncionais, como no sentido de chegar a uma categorização e subcategorização congruentes entre os gramáticos. [Bechara 92], [Cunha 85] e [Lima 72] constituíram obras de referência para a pesquisa, fato que não dispensou a consulta a outros manuais pedagógicos. Para esta classe também estava previsto o registro da canônica. A classificação e subclassificação das conjunções do Português incluem as conjunções coordenadas (aditiva, adversativa, alternativa, conclusiva, explicativa) e subordinadas (integrante, causal, comparativa, concessiva, condicional, consecutiva, final, temporal, proporcional, conformativa).

Tal como já foi citado em relação a outras classes de palavras, o problema da homografia/homonímia colocou-se agora no interior de uma única classe, revelando assim a dificuldade de desambigüizar, tanto no domínio do Léxico quanto das regras gramaticais, a função sintática dos itens homógrafos. Tomando-se como exemplo a partícula *que*, pode-se observar, através das possibilidades de classificação, a dificuldade de se estabelecer uma hierarquização que revelasse a sua freqüência em termos de ocorrência. *Que* encontra-se classificada como uma conjunção *coordenada aditiva, explicativa*, mas também como *subordinada causal, concessiva, final, temporal, integrante*. O papel do corpus no estabelecimento da freqüência para as classes cujos itens são finitos passa a ser indispensável, não prescindindo, contudo, de análise especializada, já que a forma por si só e contexto sintático, os quais podem ser identificados automaticamente, não permitem estabelecer o "uso" do *que,* por exemplo.

4.7 Advérbios

No tratamento dos advérbios, os procedimentos metodológicos foram os mesmos utilizados em relação a outras classes de palavras, a saber, estudo exaustivo para o levantamento das formas seguido de sua categorização. Observe-se que há, em relação a esta

presente do indicativo, presente do subjuntivo, pretérito imperfeito, pretérito imperfeito do subjuntivo, pretérito mais-que-perfeito, pretérito perfeito e *pessoas possíveis*: eu, tu, ele, nós, vós, eles), colocação pronominal (nula, ênclise, mesóclise), regência do verbo (*preposições*: a, ante, após, até, com, contra, de, desde, em, entre, para, perante, por, sem, sob, sobre, trás), seguidas da forma canônica. Para os verbos no particípio passado, são necessárias as informações de gênero (masculino, feminino, 2g) e número (singular, plural e 2n). Ex.: Fazer=<V.[INT.PRONOM.TD.TI][INF-PESS.ELE.EU.]N.[a.de.em.por.][fazer]>

Com o objetivo de evitar a pluricategorização, realizou-se a união das categorias particípio passado e adjetivo, prevalecendo a forma verbal, pois a adjetivação a partir de particípios é facilmente recuperada pelas regras gramaticais.

As preposições que são regidas pelos verbos foram inseridas manualmente via interface do Léxico, bem como a informação de pronominal na predicação dos verbos pronominais.

Verbos que possuem conjugação irregular e não puderam ser gerados automaticamente, foram inseridos manualmente.

4.4 Numerais

Os numerais possuem as informações morfológicas de gênero, número, tipo (cardinal, ordinal, multiplicativo, fracionário), seguidas de forma canônica. Por exemplo:

dez=<NUM.2G.PL.CAR.[dez]>

A classificação dos numerais gerou inúmeras pesquisas em gramáticas bem como a troca de opinião e conhecimento sobre o assunto entre os lingüistas.

Os gramáticos da língua portuguesa são unânimes ao afirmar que todos os numerais cardinais são invariáveis quanto ao gênero e ao número, salvo o *"um"* (fem.: *uma*), o *"dois"* (fem.: *duas*) e as centenas acima de *"cem" (quatrocentos/quatrocentas, quinhentos/quinhentas).* No entanto, a categoria "invariável", que constava do Léxico deixou de existir por questões de incompatibilidade com a sintaxe do Léxico prevista pela equipe parceira da UNICAMP, responsável pela compactação do mesmo. Assim, os numerais cardinais classificados como invariáveis passaram a ocupar as informações "2G"(tanto masculino quanto feminino) e "plural".

Os numerais cardinais quando substantivados se pluralizam normalmente e, portanto, no Léxico eles são categorizados também como substantivos no singular e no plural. Ex.:
Quatro=<S.M.SI.[].[quatro]>; Quatros=<S.M.PL.[].[quatro]>

Os numerais ordinais variam em gênero e número e estão classificados nas suas diversas formas: *primeiro, primeiros, primeira, primeiras*.

Os numerais fracionários concordam com os cardinais que indicam o número das partes e, portanto, a flexão de gênero e número ocorre do mesmo modo que para os cardinais, e as suas formas flexionadas estão inseridas no Léxico. Ex: *Um terço das mulheres está presente; obviamente, duas terças partes estão ausentes.*

Os numerais multiplicativos que funcionam também como adjetivos encontram-se flexionados em gênero e número no Léxico, além de possuirem a informação da categoria numeral, usando-se a hierarquização para classificá-los na interface (no caso, a categoria adjetivo vem em 1º lugar). É o caso dos numerais *"triplo", "duplo".* Já as formas multiplicativas *"dúplice", "tríplice"* variam apenas em número, enquanto que para a informação de gênero, optou-se por 2g. Essa classe gramatical foi inserida manualmente via interface do Léxico.

4.5 Pronomes

Pronomes possuem as informações morfológicas de gênero, número, tipo (reto, oblíquo átono, oblíquo tônico, possessivo, demonstrativo, indefinido, interrogativo, relativo, reflexivo, tratamento), pessoa gramatical (1ª sing., 2ª sing., 3ª sing., 1ª pl., 2ª pl., 3ª pl., 3ª sing. e pl.), seguidas de contração e canônica. Ex.: tua=<PRON.F.SI.[POSS.]2S.C.[][teu]>

Ao contrário dos critérios descritos acima, por questões implementacionais, os pronomes *"isto", "isso", "aquilo", "cada"* foram classificados como 2g / singular*; "mais", "menos", "disto/disso", "nisto/nisso", "daquilo", "naquilo", "que", "algo", "onde", "algures", "alhures",*

pronominais, denotadores expressivos e expressões e locuções latinas foram tratados em listas à parte.

4.1 Substantivos

Os substantivos possuem as informações morfológicas de gênero, número, grau e regência nominal acompanhadas da canônica da palavra. Por ex.: menino=<SM. SI. ?. []. [menino]>

Há muitos casos no Léxico em que um substantivo coincide com uma forma verbal, como é o caso, por exemplo, do lexema "casa" que está classificado no dicionário da seguinte forma:
casa=<SF.SI.?.[].[casa]>
-----= <V.[TD.TI.INT. PRONOM.][PRES.IND.ELE.IMP.AFIR.TU.]N.[a,com] [casar]>

As duas categorias gramaticais para o lexema são relevantes na língua portuguesa e não podem deixar de existir. Assim, a única solução encontrada para casos como esse foi a hierarquia entre a categoria substantivo e a do verbo. O método utilizado foi o intuitivo diante das limitações de tempo do qual dispúnhamos e mesmo da escassez de referências bibliográficas sobre o assunto. Dessa forma, o lexema "casa" encontra-se hierarquizado no Léxico primeiramente como substantivo e em segundo lugar como verbo.

No caso da coincidência entre as categorias substantivo e adjetivo, prevaleceu apenas a de adjetivo, uma vez que a sua flexão, tanto de gênero quanto de número, é completa na maioria dos casos. Assim, para qualquer substantivo que fosse também adjetivo optou-se por classificá-lo apenas como adjetivo, já que podemos obter a substantivação de adjetivos a partir do determinante. É o caso por exemplo do vocábulo *gordo*. Pelas regras gramaticais, o adjetivo da frase "*O homem gordo saiu*" transforma-se em substantivo se acrescido do artigo definido, como em: "*O gordo que aqui estava acabou de sair*" e, portanto, é recuperada a categoria substantivo do lexema.

Palavras como *gelo, abóbora, limão, laranja, groselha*, entre outras, segundo a lista original de palavras, estavam categorizadas tanto como substantivos masculinos ou femininos e como adjetivos 2g (os chamados uniformes, porque acompanham substantivos de ambos os gêneros) e 2n (aqueles cuja identificação do número se faz pelo substantivo) quando se referiam à cor de alguma coisa. Ao classificá-las ou como adjetivos ou como substantivos, nova questão surgiu: se a opção fosse pelo adjetivo, qualquer frase do tipo: "*As abóboras são gostosas*" seria dada como errada, porque com a informação 2n, o revisor só aceitaria " abóbora" ou "as abóbora", no momento da substantivação do adjetivo. A solução encontrada foi classificá-las, então, como substantivos.

4.2 Adjetivos

Adjetivos possuem as informações morfológicas de gênero, número, grau e regência nominal acompanhadas da forma canônica. Ex.: magríssimo=<ADJ.M.SI.SU.[][magro]>.

Nas formas adjetivas do tipo *"parasito/parasita", "magricelo/magricela"*, o problema surgiu quanto às canônicas e não quanto às categorias. As duas formas foram classificadas como adjetivas sendo elas próprias as suas canônicas, pois, se fosse feita a opção pela forma *"parasito",* que seria a mais coerente gramaticalmente, uma vez que a sua flexão é completa, perder-se-ia a forma *"o parasita"* que é corretamente usada e empregada.

Os adjetivos que possuíam regência nominal diferente da regência verbal dos particípios passados foram reinseridos, uma vez que todo adjetivo que fosse também particípio passado havia sido unido nesta última categoria e, portanto, retirado do Léxico.

4.3 Verbos

Os verbos possuem as seguintes informações morfológicas: predicação (intransitivo, transitivo direto, transitivo indireto, bitransitivo, verbo de ligação, verbo auxiliar, pronominal), forma nominal (nula, gerúndio, particípio, particípio e gerúndio), tempo (*tempos possíveis*: futuro do presente, futuro do pretérito, futuro do subjuntivo, imperativo afirmativo, infinitivo pessoal,

Na geração de lexemas a partir de bases nominais (substantivo/adjetivo), quinze padrões morfológicos foram arrolados com vistas à flexão nominal de gênero. Para a geração de lexemas nominais flexionados em número, foram criadas 14 regras morfológicas baseadas em normas gramaticais elaboradas a partir de um estudo realizado nas principais gramáticas da língua portuguesa.

Algumas situações interessantes se apresentaram como conseqüência da geração automática. É o caso das palavras terminadas em *-idade, -ismo, -logia, -ez*, que não são usadas no plural segundo o dicionário [Biderman 92]. A geração automática de número criou todas as formas de plural para palavras com essas terminações, uma vez que não pusemos nenhuma restrição à máquina quanto a não realizá-las, mesmo porque, das pesquisas realizadas, encontramos apenas restrições quanto ao uso do plural e não quanto à existência do mesmo. Sabemos que frases do tipo "As surdezes das pessoas atrapalham a sua compreensão" não são praticáveis quanto ao seu uso, porém elas são passíveis de existência num contexto poético ou metafórico. E assim temos, no Léxico, plurais como: "sinceridades", "biologias", "comunismos" mesmo tendo presente que não são usados.

Para a geração automática das formas verbais, foram criadas regras computacionais que permitiram a criação de lexemas dentro dos padrões regulares da flexão verbal, ou seja, a geração de formas verbais só é possível para verbos que possuam conjugações regulares dentro da língua portuguesa. Através da indicação da terminação verbal, do tempo a que se conjuga, da pessoa verbal utilizada e do seu sufixo verbal correspondente, foram criadas regras de derivação para as três conjugações. A partir do conjunto de regras, as formas foram geradas obedecendo a seguinte norma: radical + sufixo, acompanhado da sua correspondente classificação de tempo e pessoa, sendo que radical é "ar", para as formas canônicas terminadas em -ar; "er", para as formas canônicas terminadas em -er; "ir", para as formas canônicas terminadas em -ir.

Outro processo de geração automática incidiu, por exemplo, sobre as formas do futuro do presente e do futuro do pretérito do indicativo. Foram geradas, a partir da canônica dos verbos transitivos diretos, transitivos indiretos e bitransitivos, todas as formas mesoclíticas da Língua Portuguesa. Todas as regras que regem essas gerações automáticas podem ser encontradas, em detalhe, em [Nunes et al. 1996].

4. A classificação das palavras

Todos os lexemas que fazem parte do Léxico possuem informações gramaticais de forma a atender as exigências morfológicas e sintáticas de cada palavra-entrada quando requisitada pelas regras gramaticais. Para cada classe gramatical do lexema foram inseridas categorias específicas de acordo com a sua própria natureza e que serão descritas abaixo.

Na implementação automática das informações gramaticais dos lexemas, visando uma categorização que fosse a menos complexa possível, objetivando, assim, amenizar a ambigüidade gramatical, os homônimos tornaram-se um difícil problema a ser resolvido, bem como a pluricategorização de uma palavra. Com isso, alguns critérios para a classificação dos lexemas foram criados, com o intuito de suavizar os problemas das regras gramaticais do revisor.

A hierarquização, ou seja, indicação hierárquica quanto à maior freqüência de uso para os homógrafos, e a união das classes gramaticais em uma única categoria para a pluricategorização foram alguns dos critérios utilizados na implementação das informações gramaticais dos lexemas. São descritas abaixo as medidas tomadas para as classes gramaticais mais relevantes. Deixamos de mencionar neste artigo, por questões de espaço, as categorias dos Artigos, Prefixos e Preposições. Siglas, acrônimos, abreviaturas e nomes próprios foram inseridos após levantamento exaustivo no corpus e posterior classificação manual. É importante notar que a presença desses itens, além de servir ao corretor ortográfico (uma vez que sua ausência implica em aviso do corretor), evita que o analisador sintático deixe de prosseguir sua tarefa devido à impossibilidade de classificar um item lexical. Finalmente, vale ressaltar que, pelas características da estrutura adotada, locuções adverbiais, prepositivas, conjuntivas,

meninos, por exemplo, estão todos ligados à forma canônica *menino* e, conseqüentemente, todos ligados entre si.

2) A complementação mais importante do conjunto original foi no sentido de se gerar, de forma automática, as flexões de gênero, número e grau das classes de substantivos, adjetivos, artigos e particípios, e conjugação de verbos regulares. Para tanto, estudos foram feitos no sentido de se criar regras de flexões bastante abrangentes e sujeitas a um número mínimo de erros. Aplicadas essas regras, de forma automática, os conjuntos flexionados obtidos foram conferidos manualmente pela equipe de lingüistas, de modo a eliminar os erros. Toda a verificação/inclusão/alteração/eliminação é feita via interface especialmente planejada para tal, conforme descrito na seção 6. É importante, nesse ponto, justificar a opção por gerar as formas flexionadas, derivadas e as conjugações, em contraposição à utilização de um esquema de representação de radicais e suas expansões legais. O convênio da Itautec com a UNICAMP possibilitou o uso de tecnologia desenvolvida pela equipe daquela universidade que prevê o uso de autômatos finitos como base para a construção de verificador e aconselhador ortográfico bastante eficientes [Kowaltowski & Lucchesi 1993]. Além disso, a tecnologia de minimização de autômatos, baseada no fato de que vocabulários de línguas naturais são bastante esparsos (no sentido de que a maioria das transações levam a um estado de rejeição), possibilita compactar a lista de palavras de forma que, por exemplo, um arquivo de 2.1 MB, contendo palavras em Português, seja convertido, sem perda de informações, em um arquivo com menos de 68 KB [Kowaltowski et al. 1995a,1995b]. Finalmente, a descompactação ocorre de maneira bastante eficiente, não prejudicando seu uso pelo corretor. Dessa forma, não nos pareceu compensador trabalhar com uma estrutura de dados mais complexa e que, além de também necessitar de compactação, iria demandar maior tempo de acesso às entradas.

3) Em paralelo à geração de lexemas, ocorreu o processo de classificação gramatical das entradas do Léxico. Essa tarefa foi, sem dúvida, das mais complexas, tanto pelo volume de trabalho envolvido, quanto pelas inúmeras questões teóricas que se colocavam no decorrer do trabalho. Essas questões estão discutidas nas seções seguintes.

4) Infelizmente, o trabalho de verificação de abrangência e corretude do Léxico não pode ser feito em paralelo à sua construção, devido à ausência das funções de acesso ao Léxico devidamente compactado, que nos foram fornecidas posteriormente pela equipe da UNICAMP. A esta altura, o Léxico já era bastante volumoso e toda modificação sistemática acabava por colocar em risco a estabilidade das informações, além de demandar um trabalho que consumiu muito tempo. No entanto, esse tipo de trabalho já era previsto. Quando de posse das funções de acesso, pode-se testar a abrangência do Léxico tanto contra o corpus quanto contra outros léxicos eletrônicos disponíveis. Testes com o próprio revisor gramatical, por outro lado, indicavam a qualidade das informações sintáticas das entradas. Dessa forma, milhares de novas palavras foram inseridas e/ou (re)classificadas.

3. A Geração de Lexemas

O Dicionário Eletrônico da Melhoramentos (1987), tal como se encontrava disponível inicialmente, apresentava os itens dicionarizados em sua forma não-marcada, ou seja, sem as marcas de flexão de gênero, número e grau para os nomes, e sem a flexão modo-temporal e número-pessoal para os verbos. Lema constitui, assim, formas não-marcadas. No entanto, como a correção automática é feita a partir do reconhecimento das formas morfológicas derivadas a partir de um lema, foi necessário proceder-se à geração de lexemas, que são as formas flexionadas em gênero e número, no caso dos nomes, e em modo-tempo e número-pessoa, no caso dos verbos.

2. Metodologia Adotada

Cada entrada do Léxico é constituída de uma palavra ou, no máximo, palavras compostas hifenizadas. Essa característica faz do Léxico uma lista fechada de palavras e impossibilitada de aceitar lemas e de conseqüência, seus lexemas, que não estejam ali listados. Se por um lado esse procedimento é positivo, por oferecer segurança lingüística ao usuário quando encontra-se em dúvida sobre a existência de certas formações de palavras, por outro lado apresenta o aspecto negativo da limitação vocabular.

Léxicos computacionais que tenham a característica de representar suas entradas na forma de aglutinação de morfemas oferecem a vantagem de que potencialmente infinitos novos vocábulos podem ser criados automaticamente e, portanto, aceitos, já que é comum existirem palavras que são aceitas gramaticalmente e reconhecidas pelos seus falantes, mas que muitas vezes não se encontram dicionarizadas ainda. É o caso de várias palavras terminadas em "-idade" e em "-or" por exemplo, tais como "praticidade" e "desentupidor". Essa vantagem pode transformar-se, porém, em um aspecto negativo se as novas formações não forem aceitas gramaticalmente. Assim, enquanto a lista de palavras rejeitaria o vocábulo "imexível", por exemplo, o aglutinador de morfemas o aceitaria, já que tal formação teoricamente é possível dentro da língua portuguesa. Finalmente, é pertinente dizer que o conjunto inicial de verbetes foi extraído de um dicionário eletrônico (Melhoramentos), o que certamente economizou tempo e esforços.

A decisão de se dicionarizar palavras implica também em que expressões que se queira considerar como *tokens* únicos, para efeito de análise sintática, como as locuções, devem ser manipuladas num contexto extra-léxico. Neste caso, a saída é indicar, de forma *ad hoc* no Léxico, os prováveis componentes de expressões que, por sua vez, devem estar disponíveis na forma de listas (caso das locuções) ou mesmo inferidos via programas (caso dos nomes próprios, em que se consideram ocorrências consecutivas de nomes próprios como um único nome próprio).

Partindo-se de um conjunto de aproximadamente 120 mil palavras do Dicionário Melhoramentos da Língua Portuguesa (1987), o trabalho de expansão consistiu em gerar: a) as conjugações dos verbos, b) as flexões de gênero, c) as flexões de número, d) as flexões de grau. Essas tarefas foram todas feitas de forma automática, a partir de algoritmos formulados pelos lingüistas. Para a maioria dessas tarefas foi necessária uma revisão não-automática cuidadosa para a detecção de malformação de palavras.

Testes do revisor ortográfico empregando o Léxico (parcial) construído a partir do conjunto de verbetes inicial mostraram um desempenho insuficiente. Por isso, adicionalmente às formas previstas, um grande trabalho de verificação de formas faltantes foi feito utilizando-se um corpus que conta, hoje, com aproximadamente 27 milhões de palavras. O corpus contém textos de livros científicos, literários e jornalísticos, com predominância para este último tipo, por questão de disponibilidade. Não houve preocupação em garantir representatividade do corpus quanto às diferentes tipologias de texto do Português do Brasil, mas sim reunir um banco de textos para testes. Através desse trabalho com o corpus, o Léxico foi expandido e atualmente conta com cerca de 1.200.000 lexemas gerados a partir de aproximadamente 100.000 lemas.

Em resumo, a metodologia utilizada contou com os seguintes passos:

1) Em paralelo a uma primeira verificação do conjunto original de palavras, tanto em relação à sua abrangência, quanto à sua exatidão nas conjugações de verbos, teve início a especificação das informações sintáticas que deveriam fazer parte de cada palavra do Léxico. Isto foi feito levando-se em conta as especificidades de cada categoria gramatical. Esta especificação sofreu algumas alterações durante todo o processo e sua versão final pode ser encontrada em [Nunes et al. 1996]. É importante ressaltar o papel da forma canônica, presente em toda entrada, que tem a função de ligar toda palavra à forma básica que lhe deu origem. Com isso, possibilita-se recuperar as várias flexões e derivações de uma mesma forma básica. Assim, *menina, meninas,*

A CONSTRUÇÃO DE UM LÉXICO PARA O PORTUGUÊS DO BRASIL: LIÇÕES APRENDIDAS E PERSPECTIVAS*

Maria das Graças V. Nunes# Fabiano M. Costa Vieira

Cláudia Zavaglia% Cássia R. C. Sossolote% Josélia Hernandez%

Departamento de Ciências da Computação e Estatística
Instituto de Ciências Matemáticas de São Carlos
Universidade de São Paulo
{mdgvnune|fabiano|claudia|cassia|joselia}@icmsc.sc.usp.br

Resumo

Neste trabalho são relatados alguns passos relevantes da construção de um léxico para o Português do Brasil, que dá suporte ao revisor gramatical desenvolvido sob coordenação do Núcleo Interdepartamental de Lingüística Computacional (NILC), no âmbito do convênio entre o ICMSC-USP e a Itautec-Philco S.A, em vigor desde 1993. Este Léxico também dá suporte ao corretor ortográfico desenvolvido por equipe da UNICAMP, também em convênio com a Itautec-Philco. Este artigo apresenta a metodologia usada, os principais problemas encontrados e as soluções adotadas e características do sistema computacional gerenciador.

1. Introdução

Este artigo relata a construção de um léxico para o Português do Brasil, que serve de suporte aos revisores ortográfico e gramatical passíveis de serem acoplados ao processador de textos Word for Windows da Microsoft, e que são comercializados pela Itautec-Philco S.A. Estes sistemas computacionais foram desenvolvidos através de convênios entre aquela empresa e a UNICAMP (ortográfico), e a USP-São Carlos (gramatical). Entende-se por léxico, neste contexto, um conjunto bastante abrangente de verbetes de nossa língua, compreendendo uma série de informações sobre cada um deles, de modo a tornar possível as correções ortográfica e gramatical de textos escritos em Português do Brasil. Além disso, essas informações devem suportar algumas funções que podem ser úteis para o usuário de um processador de textos, tais como: conjugação de verbos, busca de palavras por máscaras, determinação de categorias gramaticais de palavras, etc. Tais funções fazem parte do sistema que engloba o corretor ortográfico, desenvolvido pela equipe da UNICAMP.

Os requisitos de um léxico que dê suporte aos dois tipos de revisores — ortográfico e gramatical — são, aparentemente, conflitantes. Para o revisor ortográfico, o léxico deve ser o mais abrangente possível, contendo inclusive nomes próprios, siglas, abreviaturas, etc. Já o revisor gramatical busca, em geral, através de uma análise sintática automática e da detecção de certos padrões, identificar desvios lingüísticos que fujam ao padrão da norma culta da língua portuguesa em textos escritos não-literários. Portanto, as palavras do léxico precisam ser categorizadas quanto a sua morfologia, o que dificulta a manipulação de grandes massas de dados requeridas pela abrangência do revisor ortográfico. A compilação de um conjunto de palavras para a formação de um léxico é conceitualmente simples, apesar do enorme volume de trabalho envolvido. Além disso, o que em princípio parecia um trabalho mecânico, ainda que exaustivo, acabou mostrando facetas interessantes com perspectivas de uma nova gama de pesquisas em lexicografia, como comentado nas conclusões desse trabalho.

* Este projeto foi financiado, através de convênio com a USP, pela Itautec-Philco.
# Pesquisadora parcialmente financiada pelo CNPq, proc. nro. 301365-91.1
% Lingüistas pesquisadoras do convênio USP-Itautec-Philco.