

Instituto de Ciências Matemáticas de São Carlos

ISSN - 0103-2569

**A CONSTRUÇÃO DE UM LÉXICO DA LÍNGUA
PORTUGUESA DO BRASIL PARA SUPORTE À
CORREÇÃO AUTOMÁTICA DE TEXTOS**

**MARIA DAS GRAÇAS V. NUNES
FABIANO M. COSTA VIEIRA
CLÁUDIA ZAVAGLIA
CÁSSIA R.C. SOSSOLOTE
JOSÉLIA HERNANDEZ**

Nº 42

RELATÓRIOS TÉCNICOS DO ICMSC

São Carlos
Ago./1996

SYSNO	906300
DATA	/ /
ICMC - SBAB	

**A Construção de um
Léxico da Língua Portuguesa do Brasil
para suporte à
Correção Automática de Textos***

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
mdgynune@icmsc.sc.usp.br*

Resumo

Neste trabalho são relatados os passos 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 relatório apresenta a metodologia usada, os principais problemas encontrados e as soluções adotadas e características do sistema computacional gerenciador.

Abstract

This work reports on the construction of a lexicon for Brazilian Portuguese which supports the grammar checker developed at ICMSC-USP in conjunction with Itautec-Philco S.A. This lexicon also supports a spelling checker. This paper presents the adopted methodology and some problems involved in the different phases of this task.

1. Introdução

O NILC - Núcleo Interdepartamental de Lingüística Computacional - é um grupo formado por docentes do Instituto de Ciências Matemáticas de São Carlos e do Instituto de Física de São Carlos, da Universidade de São Paulo, campus de São Carlos, que vem, desde 1991, atuando em pesquisas na área de processamento de linguagem natural. Este grupo

* 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.

iniciou, há três anos, um projeto patrocinado pela Itautec-Philco, visando a elaboração de um corretor gramatical para o português do Brasil. Ao longo deste período, o projeto contou com a supervisão de 4 a 6 professores da universidade, e com equipes de desenvolvimento distintas, em geral envolvendo cerca de 10 pessoas, entre cientistas da computação e linguistas. Uma versão comercial do Revisor desenvolvido, que opera acoplado ao processador de textos Redator, da Itautec, foi lançada em Julho de 1995 [Nunes 96]. Uma segunda versão, mais abrangente e baseada em léxico desenvolvido para esse fim, para ser instalada em processadores Word/Windows, está sendo lançada no mercado em agosto/96. A ferramenta conta ainda com um revisor de erros mecânicos, ou seja, erros típicos de digitação, não detectáveis por corretores ortográficos, e um analisador de legibilidade. O índice de legibilidade, que é uma adaptação do índice Flesch amplamente utilizado para o Inglês, classifica o texto em quatro faixas de dificuldade de leitura, servindo de apoio à verificação da adequação do texto ao leitor a que este se destina [Martins 96]. Essas faixas correspondem aos diferentes níveis de escolaridade: 1a. a 4a. série (muito fácil); 5a. a 8a. série (fácil); 2o. grau (difícil); e texto acadêmico em áreas específicas (muito difícil).

Neste trabalho é relatada a construção de um léxico para o português do Brasil, para dar suporte aos revisores ortográfico e gramatical da Itautec-Philco S.A, desenvolvidos através de convênios com a UNICAMP (ortográfico) e com 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 já presentes no Redator, atualmente suportadas pelo dicionário do Proverb, 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 para os 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, através de 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. De fato, a compilação do presente léxico tomou praticamente um ano de trabalho de três linguistas e dois informatas, dedicando-se respectivamente 30 e 20 horas semanais, atuando no ICMSC-USP de São Carlos. Foram necessários, também, o trabalho de digitadores e cooperação eventual de colaboradores, principalmente com vistas à adequação do léxico para o sistema de correção gramatical. 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 relatório está organizado da seguinte forma: na seção 2 apresentamos a metodologia usada na construção do léxico; a seção 3 descreve como foram gerados automaticamente as várias formas de lexemas a partir de lemas; as seções 4 e 5 apresentam problemas interessantes e as soluções adotadas durante as tarefas de classificação sintática de nomes e verbos; a seção 6 discute o papel relevante do corpus utilizado neste trabalho; a seção 7 apresenta características do sistema computacional gerenciador do léxico e a seção 8 apresenta as conclusões e aponta para perspectivas de trabalhos futuros. Os Apêndices A e B apresentam, respectivamente, as estruturas dos vários tipos de entrada do léxico e a sintaxe definida para sua estrutura inteira. Finalmente, o Apêndice C mostra as regras de conjugação verbal que foram usadas para a geração automática das formas verbais do Léxico.

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 um aspecto negativo que é, comumente, o de todos os dicionários de língua, ou seja, a limitação vocabular.

Estudos e pesquisas, nos EUA e Inglaterra, por exemplo, estão sendo desenvolvidos com o objetivo de se criarem léxicos computacionais por aglutinação de morfemas, o que seria vantajoso se pensarmos que infinitos novos vocábulos poderiam ser criados automaticamente e aceitos dentro de um contexto escrito, 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”, em “-or” por exemplo, tais como “desentupidor” e “praticidade”. Essa vantagem pode transformar-se facilmente, porém, em um aspecto negativo quando tratamos de revisão gramatical computadorizada se as novas formações não forem aceitas gramaticalmente. E, 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. Adicionalmente, o conjunto inicial de verbetes foi extraído de um dicionário eletrônico, 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). Tais tratamentos são feitos pelo módulo de entrada dos revisores, chamado de Analisador Léxico.

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 categorial gramatical. Esta especificação sofreu algumas alterações durante o processo e sua versão final é apresentada no Apêndice A. A sintaxe definida para a estrutura interna do léxico é apresentada no Apêndice B. É importante ressaltar o papel da forma canônica, presente em toda entrada, que tem o papel de ligar toda palavra à sua forma dita canônica, isto é, a forma que lhe deu origem. Com isso, possibilita-se recuperar as várias flexões e derivações de uma mesma forma básica. Assim, *menina*, *meninas*, *meninos*, por exemplo, estão todos ligados à forma canônica *menino* e, conseqüentemente, todos ligados entre si.
- 2) Alguns problemas foram detectados no conjunto original, como por exemplo, a falta de conjugação de alguns verbos. Sanados esses problemas, passou-se à fase de complementação dos verbetes, ou seja, visando-se a abrangência do conjunto.
- 3) A complementação mais importante foi no sentido de se gerar, de forma automática, as flexões de gênero, número e grau das classes de substantivos, adjetivos, artigos e participípios, 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 7. 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 Itaotec com a UNICAMP também previa o uso do compactador desenvolvido pela equipe daquela universidade, que gera um código compactado tal que o arquivo texto contendo cerca de 1.200.000 entradas produz um arquivo, já em outro formato, compactado de 1MB. Além disso, a descompactação ocorre de maneira bastante eficiente, não prejudicando seu uso pelo corretor. Dessa forma, não pareceu-nos compensador trabalhar com uma estrutura de dados mais complexa que, além de também ser necessária sua compactação, demandasse maior tempo de acesso às entradas.

4) Em paralelo com a 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.

5) Infelizmente, o trabalho de verificação de abrangência e corretude do léxico não pode ser feito em paralelo com 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 paralelo 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 faz-se a partir do reconhecimento das formas morfológicas derivadas a partir de um lema, foi necessário proceder-se à geração de **lexemas**, formas flexionáveis em **gênero e número**, no caso dos nomes, e em **modo-tempo, número-pessoa**, no caso dos verbos.

Na geração de lexemas a partir de bases nominais (substantivo/adjetivo), os padrões morfológicos arrolados com vistas à flexão nominal de gênero foram 15 (Fig. 3.1).

Regras Morfológicas para a Derivação da Flexão Nominal em Gênero
1º padrão: Para todas as palavras terminadas em -o, troca-se o -o por -a. Ex: menino > menina; garoto > garota.
2º padrão: Para todas as palavras terminadas em -e, troca-se o -e por -a. Ex: mestre > mestra; monge > monja.
3º padrão: Para todas as palavras terminadas em consoante, acresce-se a vogal -a. Ex: doutor > doutora; professor > professora.
4º padrão: Para todas as palavras terminadas em -ês, troca-se o -ês por -esa. Ex: português > portuguesa; inglês > inglesa.
5º padrão: Para todas as palavras terminadas em -iz, acresce-se a vogal -a. Ex: juiz > juíza
6º padrão: Para todas as palavras terminadas em -e, acresce-se -ssa. Ex: abade > abadessa; visconde > viscondessa.
7º padrão: Para todas as palavras terminadas em -e, acresce-se -sa. Ex: duque > duquesa; príncipe > princesa.
8º padrão: Para todas as palavras terminadas em -a, troca-se o -a por -isa. Ex: poeta > poetisa; sacerdote > sacerdotisa.
9º padrão: Para todas as palavras terminadas em -l, acresce-se -esa. Ex: cônsul > consulesa.
10º padrão: Para todas as palavras terminadas em -l, acresce-se -essa.
11º padrão: Para as palavras terminadas em -ão, elimina-se o -o. Ex: anão > anã; irmão > irmã.
12º padrão: Para as palavras terminadas em -ão, troca-se o -ão por -ona. Ex: respondão > respondona; comilão > comilona.
13º padrão: Para as palavras terminadas em -ão, troca-se o -ão por -oa. Ex: leão > leoa; leitão > leitoa.
14º padrão: Para as palavras terminadas em -ão, troca-se -ão por -ana. Ex: sultão > sultana.
15º padrão: Para as palavras terminadas em -ão, elimina-se o -ão e acresce-se o -a. Ex: ladrão > ladra.

Figura 3.1 Regras de Flexão Nominal de Gênero

Para a geração de lexemas nominais flexionados em número, foram criadas 14 regras morfológicas baseadas em normas gramaticais que foram elaboradas a partir de um estudo realizado nas principais gramáticas da língua portuguesa (Fig. 3.2).

REGRA 1 Para todas as palavras terminadas em -ão , troca-se o -ão por -ões . Ex.: botão > botões
REGRA 2 Para todas as palavras terminadas em -ão , troca-se o -ão por -ães . Ex.: guardião > guardiães
REGRA 3 Todas as palavras terminadas em vogal são acrescidas de -s . Ex.: monte > montes
REGRA 4 Todas as palavras terminadas em -n são acrescidas de -s (o acento anterior desaparece). Ex.: pólen > polens
REGRA 5 Para todas as palavras terminadas em -em , -im , -om , -um elimina-se o -m e acrescenta-se -ns . Ex.: pudim > pudins
REGRA 6 Todas as palavras terminadas em -r , -z e -m são acrescidas de -es . Ex.: mulher > mulheres
REGRA 7 Para todas as palavras terminadas em -al , -el e -ul , elimina-se o -l e acrescenta-se -is . Ex.: pombal > pombais
REGRA 8 Para as palavras oxítonas terminadas em -il , elimina-se o -l e acrescenta-se -s . Ex.: barril > barris
REGRA 9 Para todas as palavras oxítonas acentuadas e seguidas de -s , troca-se o acento pela letra sem acento e acrescenta-se -es . Ex.: ananás > ananases
REGRA 10 Para palavras paroxítonas terminadas em -il , elimina-se o -il e acrescenta-se -eis . Ex.: projétil > projéteis
REGRA 11 Todos os participios passados são acrescidos de -s . Ex.: utilizado > utilizados
REGRA 12 Para palavras terminadas em -ol , eliminar o -ol e inserir -óis . Ex.: caracol > caracóis
REGRA 13 Para palavras oxítonas acentuadas terminadas em -ês , eliminar o acento e acrescentar -es . Ex.: polonês > poloneses
REGRA 14 Para palavras oxítonas terminadas em -el , trocar o -el por -éis . Ex.: farnel > farnéis

Figura 3.2: Regras de Flexão Nominal de Número

Para as formas geradas a partir dessas regras gramaticais realizou-se uma checagem manual das mesmas para que pudesse ser constatada a pertinência da criação lexical.

Palavras terminadas em **-idade**, **-ismo**, **-logia**, **-ez** não seriam 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 atrapalha 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 constantes do Apêndice C.

A partir destas tabelas, as formas foram geradas obedecendo a seguinte norma: radical + sufixo, acompanhado da sua correspondente classificação de tempo e pessoa, sendo que radical = - **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, conforme ilustra a Figura 3.3, onde:

VTI = verbo transitivo indireto

VTDI = verbo transitivo direto e indireto (bitransitivo)

OI = me, te, lhe, nos, vos, lhes

TEMPO	PRED	PESSOA	FORMA PRIMITIVA	FORMA MESOCLÍTICA	EXEMPLO
FUTPRES	VTI/VTDI	EU	RADICAL + EI	RADICAL + OI + EI	dar-lhe-ei
FUTPRES	VTI/VTDI	TU	RADICAL + ÁS	RADICAL + OI + ÁS	dar-lhe-ás
FUTPRES	VTI/VTDI	ELE	RADICAL + À	RADICAL + OI + À	dar-lhe-á
FUTPRES	VTI/VTDI	NÓS	RADICAL + EMOS	RADICAL + OI + EMOS	dar-lhe-emos
FUTPRES	VTI/VTDI	VÓS	RADICAL + EIS	RADICAL + OI + EIS	dar-lhe-eis
FUTPRES	VTI/VTDI	ELES	RADICAL + ÃO	RADICAL + OI + ÃO	dar-lhe-ão
FUTPRET	VTI/VTDI	EU	RADICAL + IA	RADICAL + OI + IA	dar-lhe-ia
FUTPRET	VTI/VTDI	TU	RADICAL + IAS	RADICAL + OI + IAS	dar-lhe-ias
FUTPRET	VTI/VTDI	ELE	RADICAL + IA	RADICAL + OI + IA	dar-lhe-ia
FUTPRET	VTI/VTDI	NÓS	RADICAL + ÍAMOS	RADICAL + OI + ÍAMOS	dar-lhe-íamos
FUTPRET	VTI/VTDI	VÓS	RADICAL + ÍEIS	RADICAL + OI + ÍEIS	dar-lhe-íeis
FUTPRET	VTI/VTDI	ELES	RADICAL + IAM	RADICAL + OI + IAM	dar-lhe-iam
FUTPRES	VTI/VTDI	ELE	RADICAL + Á	RADICAL + 'SE' + Á	dar-se-á
FUTPRES	VTI/VTDI	ELES	RADICAL + ÃO	RADICAL + 'SE' + ÃO	dar-se-ão
FUTPRET	VTI/VTDI	ELE	RADICAL + IA	RADICAL + 'SE' + IA	dar-se-ia
FUTPRET	VTI/VTDI	ELES	RADICAL + IAM	RADICAL + 'SE' + IAM	dar-se-iam

Figura 3.3 Formas do Futuro do Presente e do Futuro do Pretérito do Indicativo

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 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 frequência 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. Serão descritas abaixo, dada a menção do problema, as medidas tomadas para cada classe gramatical.

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 para tal empreendimento foi o intuitivo diante das limitações de tempo do qual dispúnhamos e mesmo da escassez de referências bibliográficas sobre o assunto. De consequência, o lexema “casa” encontra-se hierarquizado no Léxico primeiramente como substantivo e em segundo lugar como verbo.

As categorias substantivo e adjetivo foram unidas em uma somente e prevaleceu na escolha o 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 fez-se a união 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 e assim temos: “*O gordo que aqui estava acabou de sair*” e, portanto, recuperada a categoria substantivo do lexema.

Palavras como *gelo*, *abóbora*, *limão*, *laranja*, *groselha*, entre outras, segundo o banco de dados, 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 Artigos

Os artigos possuem as informações morfológicas de gênero, número, tipo (definido ou indefinido) e a forma canônica. Por ex.: os=<ART.M.PL.DE.[o].

Foram inseridas no Léxico as formas flexionadas de gênero e número a partir dos lemas “*o*” e “*um*” e assim temos:

- Artigos definidos: *o, os, a, as*
- Artigos Indefinidos: *um, uns, uma, umas*.

Por ser uma classe fechada e constando de um número pequeno de integrantes, as formas foram inseridas manualmente, diretamente na interface do programa de acesso ao Léxico.

4.3 Preposições

As preposições possuem uma informação chamada “contração” na qual identificamos se a preposição é simples ou combinada, seguida de forma canônica. Ex.: nas=<PREP.C.[em.as][no].

A canônica da preposição é a forma não marcada que permite recuperar as informações de gênero e número das formas combinadas e é a própria preposição para as simples. Temos:

- Preposições simples ou essenciais: *a, ante, após, até, com, contra, de, desde, em, entre, para, perante, por, sem, sob, sobre, trás*;
- Preposições combinadas: são as preposições essenciais combinadas com os artigos definidos/indefinidos e/ou com os pronomes: *dos, deste, nela, disso, da*, etc.;
- Preposições acidentais são certas palavras que pertencem a outras classes gramaticais, mas que funcionam como preposições: *afora, fora, conforme, consoante, durante, exceto, mediante, menos, não obstante, salvo, salvante, segundo, senão, tirante, visto*, etc.

As preposições simples ou combinadas foram inseridas manualmente via interface.

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 os temos 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 possuírem 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].

Para iniciarmos a inserção dos pronomes na interface do Léxico, usou-se o critério de que “os pronomes relacionam o ser com os participantes de um ato de comunicação, as pessoas (falante/ouvinte) e o assunto do discurso”. Assim, a 1ª pessoa refere-se ao falante: *eu, nós* (pronomes pessoais), se pertence ou interessa a ele: *meu* (pronomes possessivos) e se está perto dele: *este, esta, isto: aqui* (pronomes demonstrativos). A 2ª pessoa é o ouvinte:

tu, você; se pertence ou interessa a ele: *teu, seu*(= *de você*) e se está perto dele: *esse, essa, isso: aí*. A 3ª pessoa substitui o nome do ser de que se fala (= assunto): *ele*, se pertence ou interessa a ele: *seu* (= dele). Se não está perto do falante nem do ouvinte: *aquele, aquela, aquilo: ali ou lá* e se é um ser de que se fala sem apresentar a seu respeito uma idéia precisa: *alguém* (pronome indefinido): *algures, alhures* [Luft 94].

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”, “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 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, presente do indicativo, presente do subjuntivo, pretérito

imperfeito, pretérito imperfeito do subjuntivo, pretérito mais-que-perfeito, pretérito perfeito e *peças possíveis*: eu, tu, ele, nós, vós, eles), **colocação pronominal** (nula, ênclise, mesôclise), **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 na 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.

4.7 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.

A categoria do adjetivo prevaleceu, através da hierarquização, sobre a do substantivo para os lexemas que possuíam as duas classificações, para se evitar a pluricategorização do Léxico.

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.

4.8 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. Eis a classificação e subclassificação das conjunções do português:

1. **Conjunções Coordenadas:**

- 1.1. Aditiva;
- 1.2. Adversativa;
- 1.3. Alternativa;
- 1.4. Conclusiva;
- 1.5. Explicativa.

2. **Conjunções Subordinadas:**

- 2.1. Integrante;
- 2.2. Causal;
- 2.3. Comparativa;
- 2.4. Concessiva;
- 2.5. Condicional;
- 2.6. Consecutiva;
- 2.7. Final;
- 2.8. Temporal;
- 2.9. Proporcional;
- 2.10. 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 de suas possibilidades classificatórias, 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. O estabelecimento de qualquer taxonomia ou a distribuição dos itens lexicais e gramaticais por tipologias já aceitas pressupõe sempre a análise dos itens enquanto **forma e conteúdo**, fato que permite estabelecer valores e funções lexicais e gramaticais das palavras do português.

4.9 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 classe, maior divergência entre os gramáticos quanto à categorização a ser observada. A tipologia adotada no momento da inserção dos itens adverbiais em nossa base de dados 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: atento (adj.), atenta (adj.) / **atentamente**. Para os categorizados como **2g.**, a formação do advérbio dar-se-ia 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 **comummente** 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.

4.10 Prefixos

Na definição da inserção dos prefixos, o critério utilizado para tanto foi a presença ou não do hífen no processo de formação de palavras compostas. Novamente, coube a consulta às Gramáticas Normativas do Português, da qual resultou o seguinte elenco de prefixos, os quais foram inseridos, posteriormente, no Léxico: *ab, ad, ante, além, aquém, anti, arqui, auto, bem, circum, contra, extra, ex, infra, inter, intra, hiper, mal, neo, ob, pan, pós, pré, pró, pseudo, recém, sem, semi, sobre, sob, sota, soto, sub, supra, ultra, super, vice*.

4.11 Locuções Adverbiais, Prepositivas, Conjuntivas, Pronominais, Denotadores Expressivos e Expressões e Locuções Latinas.

Para a elaboração da lista acima, recorreremos às gramáticas normativas da língua portuguesa¹. Em muitos instantes os gramáticos não são unânimes em seus postulados, e é

¹ Cf. Referências e Bibliografia.

justamente em momentos como esses que o respaldo da lingüística se faz necessário. Vejamos alguns exemplos desses problemas:

- 1º) A gramática de Napoleão de Almeida apresentou uma enorme lista de locuções adverbiais não usadas comumente, que passou por uma análise dos lingüistas.
- 2º) Dentro das locuções pronominais, as gramáticas somente trazem as locuções pronominais indefinidas;
- 3º) O *Novo manual da redação*, da Folha de São Paulo, apresenta algumas classificações de locuções muito diferentes das outras gramáticas em geral;
- 4º) A *Moderna gramática brasileira*, de [Luft 94], apresenta algumas conjunções que, em posição inicial na frase, passam a ter o valor de denotadores expressivos.

No término do levantamento das locuções, chegamos a resultados que podem ser melhor observados na Figura 4.1.

Locuções	Total
LOCUÇÕES ADVERBIAIS mais freqüentes	491
LOCUÇÕES ADVERBIAIS menos freqüentes	120
LOCUÇÕES ADVERBIAIS E ADVÉRBIOS LATINOS	57
DENOTADORES EXPRESSIVOS	38
LOCUÇÕES CONJUNCIONAIS	114
LOCUÇÕES PREPOSITIVAS	201
LOCUÇÕES PRONOMINAIS	88
TOTAL GERAL	1109

Figura 4.1: Tabela de Ocorrências das Locuções

4.12 Siglas, Acrônimos e Assemelhados

Após o levantamento das locuções acima mencionadas, fizemos uma primeira relação das siglas, dos acrônimos e dos assemelhados mais usuais na Língua Portuguesa, tomando como parâmetro o *Novo Manual da Redação*, da Folha de São Paulo, para também incluir no Léxico. Nessa pesquisa tivemos 398 ocorrências.

4.13 Listagem de Nomes Próprios

Dando seqüência ao arrolamento de palavras para constar do Léxico, passamos a inventariar os nomes próprios. Para tanto, nos dispusemos das gramáticas, dos dicionários e do *corpus* deste projeto.

Como nas locuções, observamos posturas divergentes entre as gramáticas quanto ao uso de determinadas palavras como substantivo próprio ou comum, ao tomarmos como base o fato de os nomes próprios serem escritos com maiúsculas. Observemos as mais recorrentes:

1ª) Para Sacconi [Sacconi 92], acidentes geográficos (baía de Guanabara, oceano Atlântico, rio Amazonas, etc.), e, em nomes que designam cargos ou funções (presidente, governador, prefeito, bispo, ministro, rei, etc.), usa-se a inicial minúscula;

2ª) Para Almeida [Almeida 92], nas locuções: O rio Amazonas, A rua da Glória, O mar Vermelho, o primeiro nome se escreve com minúscula, quando é suscetível de várias especificações e vem depois de iniciado o período;

3ª) Para *O Manual de Redação do Estadão*, em nomes de acidentes geográficos e sua denominação (Baía de Guanabara, Oceano Atlântico, Rio Amazonas, etc.), usa-se a inicial maiúscula, apesar de o *Manual* advertir que a norma oficial reza pela minúscula, e, em nomes que designam cargos ou funções (presidente, governador, prefeito, ministro, rei, etc.), usa-se a minúscula, ainda apesar de o *Manual* advertir que a norma oficial determina maiúscula para os nomes que designam altos cargos;

4ª) Também para o *Manual*, nos nomes de entidades e personagens do folclore (saci, mula-sem-cabeça, caipora, curupira, etc.), usa-se a minúscula;

5ª) Para Bechara [Bechara 92], em nomes de vias e lugares públicos (Avenida de Rio Branco, Rua do Ouvidor, Praia do Flamengo, etc.), e, em nomes que designam altos cargos, dignidade ou postos (Presidente da República, Governador do Estado, Prefeito, Bispo, Ministro da Educação, etc.), usa-se a inicial maiúscula.

Os dicionários de Aulete [Caldas 64] e de Houaiss [Koogan/Houaiss 94] foram de grande importância para o levantamento dos nomes próprios, bem como a varredura pelo *corpus*, o que definiu os nomes mais correntes nos textos escritos da atualidade.

Foram realizadas várias listas de nomes próprios, e a que foi implementada somou aproximadamente 4690 ocorrências.

4.14 Abreviaturas

Passada a categorização dos nomes próprios, passamos a categorizar as abreviaturas já incluindo a sua classificação na sintaxe computacional criada para representá-la no Léxico. As abreviaturas foram tomadas principalmente dos dicionários de nossa língua. Observemos alguns exemplos: apto=<ABREV.M.SI.[apto]0.>; caps=<ABREV.M.PL.[caps]0.>; dª=<ABREV.F.SI.[da]0.>; dd=<ABREV.2G.SI.[dd]0.>; dras=<ABREV.F.PL.[dras]0.>; drs=<ABREV.M.PL.[drs]0.>.

4.15 Palavras com Apóstrofos

Relacionamos, também, todas as palavras com apóstrofos encontradas nos dicionários, por exemplo: *barriga-d'água*, *caixa-d'óculos*, *d'alembertiano*, *dor-d'olhos*, *estrela-d'alva*, *galinha-d'angola*, *mestre-d'armas*, *pé-d'água*.

4.16 Estrangeirismos

Para que fossem formalizadas regras que assinalassem o uso indevido ou não de algumas expressões ou verbetes estrangeiros, foram mapeados, através dos dicionários Aurélio [Ferreira 86], Melhoramentos, e Houaiss [Houaiss 94], os anglicismos, galicismos, germanismos, italianismos, helenismos, latinismos, ou seja, os estrangeirismos da língua. Abaixo, deixamos alguns exemplos dos, aproximadamente, 1427 estrangeirismos levantados: *à côté* (francês), loc. adv. - de lado, acessoriamente; *Ablaut* (alemão), s.m. - apofonia, s.f.; *affrettando* (italiano), adv. - apressando; *apartheid* (holandês), s.f. - separação, s.f.; *apeiron* (grego) - s.m. - infinito, s.m.; *baby-sitter* (inglês), s.2g. - babá, s.f.; *dacha* (russo), s.f. - casa de campo, s.f.; *kabuki* (japonês), s.m. - cabúqui, s.m.; *kibutz* (hebraico), s.m. - colcós, s.m.; *momentum* (latim), s.m. - momento, s.m.; *yin-yang* (chinês), s.m.

5. O Papel do Corpus

Na construção do Léxico com vistas à implementação do Revisor Gramatical, o Dicionário Eletrônico da Melhoramentos (1987) constituiu o “acervo lexical” inicial do qual se partiu. No entanto, pôde-se observar, desde o início, que o fato deste dicionário estar entre os de nível médio não garantia a sua representatividade, já que a classificação em “médio”, “inframédio”, etc. parece estabelecer-se mais em função do número de verbetes do que em função da frequência de uso dos itens dicionarizados.

Assim, na medida em que se constatava a ausência de palavras de alta frequência na Língua, procurava-se inseri-las no Léxico. Antes, contudo, procedia-se à consulta aos dicionários disponíveis, a saber, Aurélio Eletrônico, Aulete (1956), Biderman (1992), Koogan/Houaiss (1994), 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*.

Era 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. 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. Ex: *música*, *sem-terra*, *criminalística*. Convém dizer que 20 foi o índice de frequência definido para a inserção das formas faltantes. Assim, o banco de textos, criado com o objetivo de possibilitar, de um lado, o levantamento dos erros mais frequentes 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 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 frequência 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.

7. 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 está sendo transportada para o Gerenciador Oracle, a fim de buscar uma maior segurança no armazenamento das informações existentes no Léxico.

Como descrito anteriormente, cada palavra possui dois campos: a própria palavra e suas classes. Essas classes obedecem uma estrutura previamente definida (cf. Apêndice A).

Exemplo:

Palavra: *abafa*

Categorias: <V.[INT.PRONOM.TD.][IMPER-

AFIRM.TU.PRES.ELE.]N.[[*abafar*]/0.#S.M.SI.N.[/?].3.[*abafa*]/0.>

Portanto, para armazenar o Léxico foi necessária a criação de apenas uma tabela contendo dois campos: *Palavra* e *Classes*. Como forma de facilitar o acesso às palavras, criou-se um índice de palavras (Campo Palavra).

A Tabela 8.1 ilustra como essa indexação é organizada pelo gerenciador:

Palavra	Classes
...	...
desfiladas	<V./INT.TD./PARTIC.F.PL./N./[[desfilar]0.>
desfiladeiro	<S.M.SI.N./[?].3.[desfiladeiro]0.>
...	...

Tabela 8.1: Indexação das entradas do Léxico

A segunda parte do sistema é composta de uma interface de acesso aos dados (Léxico) armazenados pelo gerenciador. A seguir estão as funções presentes no programa:

1. Visualização de uma palavra já existente no Léxico.
2. Inserção, Remoção e Atualização de uma palavra.
3. Consulta a uma determinada classe de palavras.
4. Consulta por prefixo ou sufixo.
5. Hierarquização das classes de uma palavra.

A visualização das classes pertencentes à palavra é ilustrada pela Figura 8.2. Para buscar uma determinada palavra, basta digitá-la no campo *Palavra* e pressionar o botão *Visualizar*.

Figura 8.2: Janela do gerenciador apresentando a palavra *a* com suas classes.

As funções de Inserção, Remoção e Atualização podem ser acessadas através do menu *Palavra*.

Quando o usuário tiver a necessidade de trabalhar somente com as palavras pertencentes a uma determinada classe (Por exemplo: Preposições), deve-se utilizar a função 3.

Analogamente, a função 4 permite que o usuário selecione as palavras por prefixo ou sufixo, por exemplo:

- Digitando *pro** no campo *Palavra* serão selecionadas todas as palavras que iniciam com o prefixo *pro*.
- Digitando **mente* no campo *Palavra* serão selecionadas todas as palavras que terminam com o sufixo *mente*.

As setas localizadas na região superior à direita da janela principal possibilitam ao usuário navegar através do dicionário.

Como uma palavra pode receber mais de uma classificação, conforme comentado nas seções anteriores, é necessário que haja uma hierarquização das classes pertencentes à palavra. O sistema gerenciador facilita a tarefa do linguista-usuário, apresentando todas as opções morfológicas possíveis, de modo que ele componha, hierarquicamente, essa classificação. Uma vez estabelecida pelo usuário, esta hierarquização aparece no rodapé da janela principal (Fig. 8.2). Quando o usuário tiver a necessidade de alterar a ordem das classes, a palavra deverá ser novamente gravada (Menu *Palavra|Gravar Palavra*) com a hierarquização desejada, conforme ilustra a Figura 8.3.

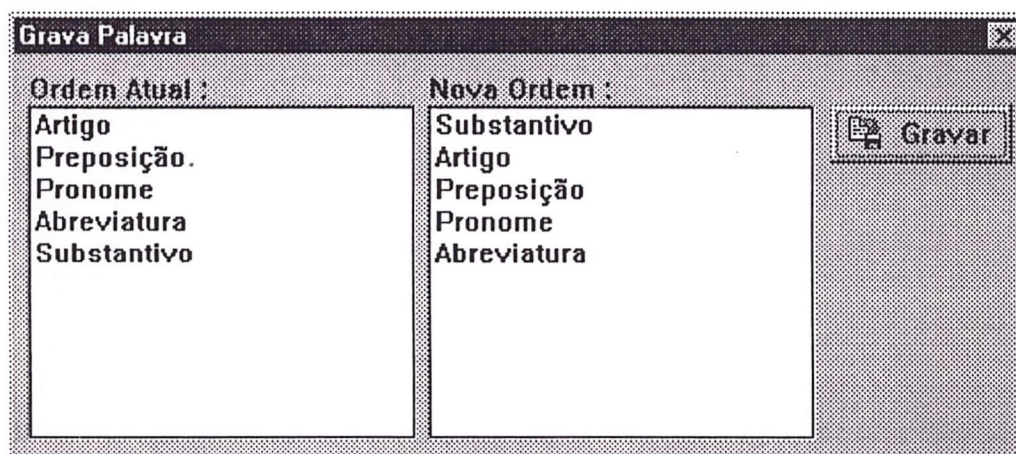


Figura 8.3: Janela para efetuar a hierarquização e gravação de uma palavra.

Informações pertencentes à determinada classe podem ser facilmente inseridas, como é o caso da regência verbal, regência nominal, predicação verbal, tempo verbal, contração, tipo do pronome etc. (Fig. 8.4).

Categorias - a

Palavra: Seleção Anterior Próximo

Palavra: Visualizar Total: 402371

☒ Substantivo	Gênero: Canônica:	Número: Flag:	Grau:	Regência:	Derivação:
☒ Artigo	Gênero: Canônica:	Número: Flag:	Tipo:	Derivação:	Canônica: Flag:
☒ Preposição	Contração: Canônica:	Flag:			
☒ Numeral	Gênero: Canônica:	Número: Flag:	Tipo:	Derivação:	
☒ Pronome	Gênero: Canônica:	Número: Flag:	Pessoa: Tipo:	Derivação: Canônica: Flag:	
☒ Verbo	Gênero: Colocação Pronominal:	Número: Regência:	Forma Nominal: Canônica: Flag:	Predicação:	Tempo:

Artigo -> Preposição -> Pronome -> Abreviatura -> Substantivo

Figura 8.4: Classes de palavras com suas respectivas categorias.

Adicionalmente, para cada uma das categorias que requeiram subclassificação ou mesmo a escolha de um valor dentro de um conjunto finito e conhecido, o sistema apresenta janelas de modo a facilitar a entrada de dados, conforme ilustram as Figuras 8.5 e 8.6.

Regência do Substantivo

Outra Preposição:

Preposições Seleccionadas:

Preposições Possíveis:

- a
- ante
- após
- até
- com
- contra
- de
- desde
- em
- entre
- para
- perante

Figura 8.5: Janela para composição de Regências Verbal e Nominal.

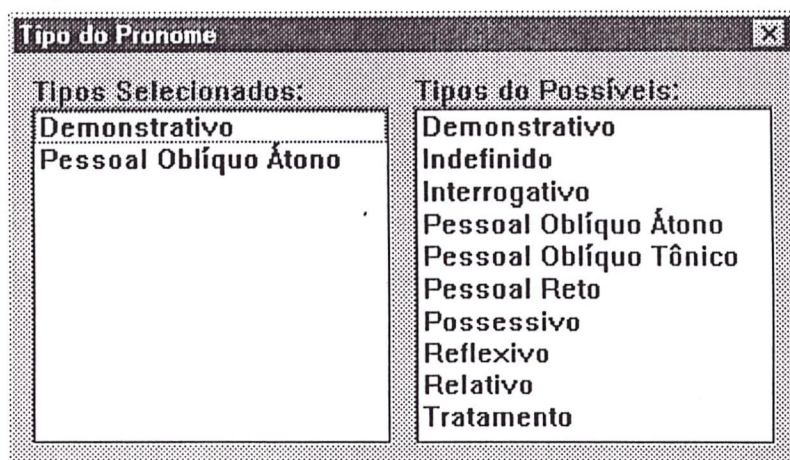


Figura 8.6: Janela para escolha de tipo do Pronome.

8. Conclusões e Perspectivas Futuras

Este relatório descreveu 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.

Computacionalmente, foram três as atividades principais para a construção e manutenção do léxico, sendo que em todas o grande volume de dados foi sempre o fator crucial. A primeira diz respeito ao gerenciamento da base de dados, com a adoção de procedimentos rigorosos de segurança para a manutenção do léxico. As outras duas atividades consistiram no desenvolvimento de programas de software para geração automática de novas palavras, e manipulação do corpus.

A experiência adquirida na construção do léxico mostrou a necessidade de se dispor de um corpus representativo da língua em uso. Com a possibilidade de empregar ferramentas de software para lidar com estas grandes massas de dados, abrem-se novas perspectivas de pesquisas em lexicografia, e mesmo construção de novos tipos de dicionários. Para isso, muito contribuirá a geração automática de novas palavras (por exemplo: advérbios terminados em "mente", adjetivos com sufixo "ável", palavras justapostas como em "interdiscurso"), que obedecem as regras de formação de palavras para

o português. É óbvio que a verificação, por um especialista, das palavras geradas é essencial, como já ocorreu na compilação do nosso léxico. Além disso, levando-se em conta a frequência de uso, como já o fez a Profa. Tereza Bidermann em seu dicionário, é possível criar dicionários adequados para um dado grupo de usuários, incluindo palavras de uso freqüente geradas automaticamente e que em geral não constam dos dicionários impressos, e evitando-se palavras que jamais são empregadas por aquele grupo-alvo. Outra possibilidade é a criação de dicionários técnicos. Uma busca no corpus pode não só identificar os termos técnicos mais freqüentes, mas também apontar estrangeirismos que se incorporam à língua por falta de similar em português e mesmo termos técnicos "aportuguesados", muitas vezes sem o cuidado de obedecer as regras de formação de palavras em português.

Finalmente, porém não menos importante, vale ressaltar a experiência em se trabalhar com uma equipe interdisciplinar de lingüistas e informatas. A inexistência de cursos de formação, em nível de graduação, em Lingüística Computacional faz de experiências como esta uma oportunidade ímpar de se formar recursos humanos, atraindo-os para trabalhos de pesquisa e desenvolvimento nesta área.

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 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 e Bibliografia

- [Almeida 92] Almeida, N. M. de. *Gramática metódica da língua portuguesa*. 38ª ed. São Paulo: Saraiva, 1992.
- [Basílio 89] Basílio, M. *Teoria Lexical*. São Paulo: Ática, 1989.
- [Bechara 92] Bechara, E. *Moderna gramática portuguesa*. São Paulo: Nacional, 1992.
- [Biderman 92] Biderman, M. T. C. *Dicionário contemporâneo de português*. Petrópolis: Vozes, 1992.
- [Borba 91] Borba, F. da S. *Dicionário gramatical de verbos*. 2ª ed. São Paulo: Ed. UNESP, 1991.
- [Caldas 64] Caldas, A. *Dicionário contemporâneo da língua portuguesa*. 2ª ed. Rio de Janeiro: Delta, 1964.
- [Câmara 91] Câmara Jr., J. M. *Problemas de Lingüística Descritiva*. 14ª ed. Petrópolis: Vozes, 1991.
- [Câmara 92] -----, *Estrutura da Língua Portuguesa*. 21ª ed. Petrópolis: Vozes, 1992.
- [Carone 94] Carone, F. de B. *Morfossintaxe*. 4ª ed. São Paulo: Ática, 1994.
- [Cunha 85] Cunha, C. *Gramática da língua portuguesa*. Rio de Janeiro: Nova Fronteira, 1985.

- [D'Andrade 93] D'Andrade, E. *Dicionário inverso do português*. Lisboa: Edições Cosmos, 1993.
- Dicionário Eletrônico da Melhoramentos, 1987.
- [Faraco 94] Faraco, C. & Moura, F. M. de. *Gramática*. 13ª ed. São Paulo: Ática, 1994.
- [Fernandes 93a] Fernandes, F. *Dicionário de regimes de substantivos e adjetivos*. 22ª ed. São Paulo: Globo, 1993.
- [Fernandes 93b] -----, *Dicionário de verbos e regimes*. 39ª ed. São Paulo: Globo, 1993.
- [Ferreira 86] Ferreira, A. B. de H. *Novo dicionário do Aurélio*. 2ª ed. Rio de Janeiro: Nova Fronteira, 1986.
- [Hauy 87] Hauy, A. B. *Da Necessidade de uma Gramática-Padrão da Língua Portuguesa*. 3ª Ed. São Paulo: Editora Ática, 1987.
- [Kehdi 90] Kehdi, V. *Morfemas do Português*. São Paulo: Ática, 1990.
- [Kehdi 92] -----, *Formação de Palavras em Português*. São Paulo: Ática, 1992.
- [Koch 94] Koch, I. G. V. & Silva, M. C. P. de S. e. *Linguística Aplicada ao Português: Morfologia*. 7ª ed. São Paulo: Cortez, 1994.
- [Koch 95] -----, *Linguística Aplicada ao Português: Sintaxe*. 6ª ed. São Paulo: Cortez, 1995.
- [Koogan 94] Koogan/Houaiss. *Enciclopédia e dicionário ilustrado*. Rio de Janeiro: Delta, 1994.
- [Kury 93] Kury, A. da G. *Novas Lições de Análise Sintática*. São Paulo: Ática, 1993.
- [Lima 72] Lima, C. H. da R. *Gramática normativa da língua portuguesa*. Rio de Janeiro: José Olympio, 1972.
- [Luft 93] Luft, C. P. *Dicionário prático de regência verbal*. São Paulo: Ática, 1993.
- [Luft 94] Luft, C. P.. *Moderna gramática brasileira*. 12ª ed. São Paulo: Globo, 1994.
- Manual de redação da Folha*. São Paulo: Folha de S. Paulo, 1992.
- [Martins 96] Martins, T.B.F.; Ghiraldello, C.M.; Nunes, M.G.V.; Oliveira Jr., O.N. Readability Formulas Applied to Textbooks in Brazilian Portuguese. *Notas do ICMSC-USP, Série Computação*, nro. 28, 1996, 11p.
- [Nicola 94a] Nicola, J. De & Infante, U. *Gramática essencial*. 8ª ed. São Paulo: Scipione, 1994.
- [Nicola 94b] -----, *Gramática contemporânea da língua portuguesa*. 13ª ed. São Paulo: Scipione, 1994.
- [Nunes 96] Nunes, M.G.V. et alli. Style and Grammar Checkers for Brazilian Portuguese. *Notas do ICMSC-USP, Série Computação*, nro. 25, 1996.
- O Estado de S. Paulo. *Manual de redação e estilo*. São Paulo: O
- [Perini 89] Perini, M. A. *Sintaxe Portuguesa: Metodologia e Funções*. São Paulo: Ática, 1989.
- [Ryan 93] Ryan, M. A. *Conjugação dos Verbos em Português*. 9ª ed. São Paulo: Ática, 1993.
- [Sacconi 89] Sacconi, L. A. *Gramática essencial da língua portuguesa*. 9ª ed. São Paulo: Atual, 1989.
- [Sandmann 89] Sandmann, A. J. *Formação de palavras no português brasileiro contemporâneo*. Curitiba: Ícone, 1989.
- [Sandmann 93] Sandmann, A. J. *Morfologia Geral*. 2ª ed. São Paulo: Contexto, 1993.
- [Savioli 94] Savioli, F. P. *Gramática em 44 lições*. 24ª ed. São Paulo: Ática, 1994.
- [Terra 94a] Terra, E & Nicola, J. de *Gramática de hoje*. São Paulo: Scipione, 1994.

[Terra 94b] Terra, E. & Nicola, J. de. Verbos. *Guia Prático de Emprego e Conjugação*. São Paulo: Scipione, 1994.

[Tufano 90] Tufano, D. *Estudos de língua portuguesa: gramática*. 2ª ed. São Paulo: Moderna, 1990.

APÊNDICE A: ESTRUTURAS DAS ENTRADAS DO LÉXICO

As categorias (classes) básicas do verbete são:

1- Substantivo

Gênero: M/F/2G

Número: SI/PL/2N/INV

Grau: A/D/N

Regência do Substantivo: ($p_1, \dots, p_n$) uma lista (podendo ser nula) de preposições.

Regra Derivação Gênero: Identificador da regra (número inteiro ou ?).

Regra Derivação Número: Identificador da regra. (número inteiro ou ?)

Forma canônica: Uma palavra entre colchetes

Exemplo:

bóia=<S.F.SI.?.[?].3.[bóia]0.>

2- Adjetivo

Gênero: M/F/2G

Número: SI/PL/2N

Grau: A/D/SU/N

Regência do Adjetivo: ($p_1, \dots, p_n$) uma lista (podendo ser nula) de preposições.

Regra Derivação Gênero: Identificador da regra.

Regra Derivação Número: Identificador da regra.

Forma canônica: Uma palavra

Exemplo:

seniores=<ADJ.2G.PL.N.[?].6.[sênior]0.>

afabilíssimo=<ADJ.M.SI.SU.[?].?.[afável]0.>

3- Artigo

Gênero: M/F

Número: SI/PL

Tipo: DE/I

Regra Derivação Gênero: Identificador da regra.

Regra Derivação Número: Identificador da regra.

Forma canônica: Uma palavra

Exemplos:

os=<ART.M.PL.DE.1.3.[o]0.>

uns=<ART.M.PL.I.3.5.[um]0.>

4- Preposição

Contração: símbolo C seguido de uma lista de palavras (par preposição, palavra); ou nula.

Forma canônica: Uma palavra

Exemplos:

desse=<PREP.C.[de.esse.][dc]0.>

ao=<PREP.C.[a.o.][a]0.>

5- Conjunção

Tipo: (COORD, SUBORD)

Complemento Coordenativa: (ADIT, ADVE, ALTER, CONCL, EXPL)

Complemento Subordinativa: (INTEG, CAUS, COMP, CONC, COND, CONS, FIN, TEMP, PROPOR, CONFOR)

Forma canônica: Uma palavra

Exemplos:

e=<CONJ.[COORD.[ADIT.ADVE.]][e]0.>

que=<CONJ.[COORD.[ADIT.ALTER.]SUBORD.[INTEG.CAUS.COMP.FIN.CONC.]] [que]0.>

6- Numeral

Gênero: M/F/2G/INV

Número: SI/PL/2N/INV

Tipo: CAR/ORD/MUL/FRA/FRA-ORD

Regra Derivação Gênero: Identificador da regra.

Regra Derivação Número: Identificador da regra.

Forma canônica: Uma palavra

Exemplos:

quarto=<NUM.M.SI.FRA-ORD.1.3.[quarto]0.>

duplo=<NUM.M.SI.MUL.1.2.[duplo]0.>

7- Pronome

Gênero: M/F/2G/INV

Número: SI/PL/2N/INV

Tipo: (RET., OBL-AT, OBL-TO, POSS, DEM, INDE, INTER, REL, REFL, TRAT)

Pessoa: 1S/ 2S/ 3S/ 1P/ 2P/ 3P/3SP/N

Regra Derivação Gênero: Identificador da regra.

Regra Derivação Número: Identificador da regra.

Contração: a letra C seguida de uma lista da forma (preposição, palavra), ou nula

Forma canônica: Uma palavra

Exemplos:

desse=<PRON.M.SI.[DEM.]2S.2.3.C.[de.esse.][esse]0.>

os=<PRON.M.PL.[DEM.OBL-AT.]3P.1.3.C.][o]0.>

tuas=<PRON.F.PL.[POSS.]2S.N.3.C.][teu]0.>

uns=<PRON.M.PL.[INDE.]N.3.5.C.][um]0.>

8- Nomes Próprios

Gênero: M/F/2G

Número: SI/PL/2N

Forma canônica: Uma palavra

Exemplos:

darci=<NOM.2G.SI.[darci]1.>

9- Verbo

Predicação do Verbo: (INT, TD, TI, BI, LIG, AUX, PRONOM) - (só no caso de infinitivo).

Formas Nominais: INF-PESS/GERUN/PARTIC

(no caso de *Particípio*, são necessárias as informações de gênero - M/F/2G e número - SI/PL/2N)

(no caso de *Infinitivo Pessoal*, o verbo possui Pessoa - EU, TU, ELE, NÓS, VÓS ELES)

Par(Tempo,Pessoa):

Tempo: (PRES, PRET-IMPERF, PRET-PERF, PRET-M-Q-P, FUT-PRES, FUT-PRET, PRES-SUBJ, PRET-IMPERF-SUBJ, FUT-SUBJ, IMP-AFIRM)

Pessoa: (EU, TU, ELE, NÓS, VÓS, ELES)

Colocação Pronominal: (N, ENC, MESOC)

Regência do Verbo: (p₁, ..., p_n) uma lista (podendo ser nula) de preposições (só no caso de infinitivo).

Forma canônica: Uma palavra

Obs: A primeira e a terceira pessoa do Infinitivo Pessoal coincidem com o Infinitivo Impessoal. Assim, não colocamos a forma nominal Infinitivo Impessoal como complemento do verbo.

Exemplos:

deram=<V.[INT.TD.TI.BI.][PRET-PERF.ELES.PRET-M-Q-P.ELES.]N.[][dar]0.>

fizesse=<V.[INT.TD.TI.BI.][PRET-IMPERF-SUBJ.EU.PRET-IMPERF-SUBJ.ELE.]N.[][fazer]0.>

10- Advérbio

Tipo: CIRC-LUG / CIRC-TEMP / CIRC-MOD / NEG / DUV / INT / AFIR / INT-LUG / INT-TEMP / INT-MOD / INT-CAUS

Grau: A/D/SU/N

Forma canônica: Uma palavra

Exemplos:

antes=<ADV.[CIR-TEMP]N.[antes]0.>

11.- Prefixos

Forma canônica: Uma palavra

Exemplos:

super=<PREF.[super]0.>

12- Siglas

Forma canônica: Uma palavra

Exemplos:

onu=<SIGL.[onu]0.>

13- Abreviaturas

Forma canônica: Uma palavra

Exemplos: exma=<ABREV.[exmo]2.>

APÊNDICE B: NOTAÇÃO BNF DAS ENTRADAS DO LÉXICO

Abreviaturas utilizadas na sintaxe:

1S. = primeira pessoa do singular	INT. = intransitivo
2S. = segunda pessoa do singular	INTE. = intensidade
3S. = terceira pessoa do singular	INTEG. = integrante
1P. = primeira pessoa do plural	INTER. = interrogativo
2P. = segunda pessoa do plural	INT-CAUS. = interrogativo de causa
3P. = terceira pessoa do plural	INT-LUG. = interrogativo de lugar
3SP. = terceira pessoa do singular e terceira pessoa do plural	INT-MOD. = interrogativo de modo
A. = aumentativo	INT-TEMP. = interrogativo de tempo
ABREV. = abreviatura	INV. = invariável
ADIT. = aditiva	LIG. = ligação
ADJ. = adjetivo	M. = masculino
ADVE. = adversativa	MESOC. = mesóclise
AFIR. = afirmação	MUL. = multiplicativo
ALTER. = alternativa	N. = nulo
ART. = artigo	NEG. = negação
AUX. = auxiliar	NOM. = nome próprio
BL. = bitransitivo	NUM. = numeral
C. = contração	OBL-AT. = pessoal oblíquo átono
CAR. = cardinal	OBL-TO. = pessoal oblíquo tônico
CAUS. = causal	ORD. = ordinal
CIR-LUG. = circunstância lugar	PARTIC. = participio
CIR-TEMP. = circunstância tempo	PL. = plural
CIR-MOD. = circunstância modo	POSS. = possessivo
COMP. = comparativa	PREF. = prefixo
CONC. = concessiva	PREP. = preposição
CONCL. = conclusiva	PRES. = presente
COND. = condicional	PRES-SUBJ. = presente do subjuntivo
CONFOR. = conformativa	PRET-IMPERF. = pretérito imperfeito
CONJ. = conjunção	PRET-IMPERF-SUBJ. = pretérito imperfeito do subjuntivo
CONS. = consecutiva	PRET-M-Q-P. = pretérito mais que perfeito
COORD. = coordenativa	PRET-PERF. = pretérito perfeito
D. = diminutivo	PRON. = pronome
DE. = definido	PRONOM. = pronominal
DEM. = demonstrativo	PROPOR. = proporcional
DUV. = dúvida	RET. = pessoal reto
ENC. = ênclise	REL. = relativo
EXPL. = explicativa	REFL. = reflexivo
F. = feminino	S. = substantivo
FIN. = final	SI. = singular
FRA. = fracionário	SIGL. = sigla
FRA-ORD. = fracionário e ordinal	SU. = superlativo
FUT-PRES. = futuro do presente	SUBORD. = subordinativa
FUT-PRET. = futuro do pretérito	TD. = transitivo direto
FUT-SUBJ. = futuro do subjuntivo	TEMP. = temporal
GERUN. = gerúndio	TI. = transitivo indireto
I. = artigo indefinido	TRAT. = tratamento
IMP-AFIRM. = imperativo afirmativo	V. = verbo
INDE. = pronome indefinido	2G. = dois gênero
INF-PESS. = infinitivo pessoal	2N. = dois número

::= Significa "é definido como".
 | Significa "ou".
 {} Significa que o conteúdo pode ser repetido nenhuma ou várias vezes.
 <> Significa que o conteúdo é uma construção sintática BNF.

<verbo> ::= <palavra> =< <classificação> { # <classificação> } >
 <abreviatura> ::= **ABREV.**
 <adjetivo> ::= **ADJ.** <gênero-2> <número-2> <grau-2> <regência> <regra-de-derivação>
 <advérbio> ::= **ADV.** <tipo-do-advérbio> <grau-2>
 <artigo> ::= **ART.** <gênero-1> <número-1> <tipo-do-artigo> <regra-de-derivação>
 <canônica> ::= [<palavra>]
 <caracter> ::= a..z | á | é | í | ó | ú | à | â | ê | ô | ã | õ | ç | ü | - | ' | ^ | °
 <classe> ::= <substantivo> | <artigo> | <preposição> | <numeral> | <pronome> | <verbo> |
 <adjetivo> | <conjunção> | <advérbio> | <nome-próprio> | <prefixo> | <sigla> | <abreviatura>
 <classificação> ::= <classe> <canônica> <flag-composição>
 <colocação-pronominal> ::= **N.** | **ENC.** | **MESOC.**
 <conjugação> ::= [<tempo> < Pessoa-1> | <forma-nominal>
 { <tempo> < Pessoa-1> | <forma-nominal> }]
 <conjunção> ::= **CONJ.** <tipo-da-conjunção>
 <conjunções-coordenadas> ::= [<tipos-de-conj-coord> { <tipos-de-conj-coord> }]
 <conjunções-subordinadas> ::= [<tipos-de-conj-subord> { <tipos-de-conj-subord> }]
 <contração> ::= **C.** <lista-de-palavras>
 <dígito> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
 <flag-composição> ::= <dígito>
 <forma-nominal> ::= **GERUN.** | **PARTIC.** <gênero-2> <número-2>
 <gênero-1> ::= **M.** | **F.**
 <gênero-2> ::= **M.** | **F.** | **2G.**
 <gênero-3> ::= **M.** | **F.** | **2G.** | **INV.**
 <grau-1> ::= **N.** | **A.** | **D.**
 <grau-2> ::= **N.** | **A.** | **D.** | **SU.**
 <lista-de-palavras> ::= [<palavra> . { <palavra> . }]
 <nome-próprio> ::= **NOM.** <gênero-2> <número-2>
 <numeral> ::= **NUM.** <gênero-3> <número-3> <tipo-do-numeral> <regra-de-derivação>
 <número-1> ::= **SI.** | **PL.**
 <número-2> ::= **SI.** | **PL.** | **2N.**
 <número-3> ::= **SI.** | **PL.** | **2N.** | **INV.**
 <palavra> ::= <caracter> { <caracter> }
 < Pessoa-1> ::= **EU.** | **TU.** | **ELE.** | **NOS.** | **VOS.** | **ELES.**
 < Pessoa-2> ::= **1S.** | **2S.** | **3S.** | **1P.** | **2P.** | **3P.** | **3SP.** | **N.**
 <predicação> ::= [<tipo-da-predicação> { <tipo-da-predicação> }]
 <prefixo> ::= **PREF.**
 <preposição> ::= **PREP.** <contração>

 <pronome> ::= **PRON.** <gênero-3> <número-3> <tipo-do-pronome> < Pessoa-2>
 <regra-de-derivação> <contração>
 <regência> ::= <lista-de-palavras>
 <regra-de-derivação> ::= <dígito> | ?
 <sigla> ::= **SIGL.**
 <substantivo> ::= **S.** <gênero-2> <número-3> <grau-1> <regência> <regra-de-derivação>
 <tempo> ::= **PRES.** | **PRET-IMPERF.** | **PRET-PERF.** | **PRET-M-Q-P.** | **FUT-PRES.** |
FUT-PRET. | **PRES-SUBJ.** | **PRET-IMPERF-SUBJ.** | **FUT-SUBJ.** |
IMPER-AFIRM. | **INF-PESS.**
 <tipo-da-predicação> ::= **INT.** | **TD.** | **TI.** | **BI.** | **LIG.** | **AUX.** | **PRONOM.**

<tipo-do-advérbio> ::= **CIR-LUG.** | **CIR-TEMP.** | **CIR-MOD.** | **NEG.** | **DUV.** | **INT.** |
 AFIR. | **INT-LUG.** | **INT-TEMP.** | **INT-MOD.** | **INT-CAUS.**
 <tipo-do-artigo> ::= **DE.** | **I.**
 <tipo-da-conjunção> ::= [**COORD.** <conjunções-coordenadas> |
 SUBORD. <conjunções-subordinadas>
 { **COORD.** <conjunções-coordenadas> |
 SUBORD. <conjunções-subordinadas> }]
 <tipo-do-numeral> ::= **ORD.** | **MUL.** | **FRA.** | **FRA-ORD.** | **CAR.**
 <tipo-do-pronome> ::= **TRAT.** | **RET.** | **OBL-AT.** | **OBL-TO.** | **POSS.** | **DEM.** | **INDE.** |
 INTE. | **REL.** | **REFL.**
 <tipos-de-conj-coord> ::= **ADIT.** | **ADVE.** | **ALTER.** | **CONCL.** | **EXPL.**
 <tipos-de-conj-subord> ::= **INTE.** | **CAUS.** | **COMP.** | **CONC.** | **COND.** | **CONS.** | **FIN.**
 TEMP. | **PROPOR.** | **CONFOR.**
 <verbo> ::= **V.** <predicação> <conjugação> <colocação-pronominal> <regência>

APÊNDICE C: REGRAS DE CONJUGAÇÃO VERBAL

TABELA 1 - PRIMEIRA CONJUGAÇÃO

<u>TEMPO-PESSOAL</u>	<u>SUFIXOS</u>
PRES. IND. EU	O
PRES. IND. TU	AS
PRES. IND. ELE	A
PRES. IND. NOS	AMOS
PRES. IND. VOS	AIS
PRES. IND. ELES	AM
IMPERF. EU	AVA
IMPERF. TU	AVAS
IMPERF. ELE	AVA
IMPERF. NOS	ÁVAMOS
IMPERF. VOS	ÁVEIS
IMPERF. ELES	AVAM
PRET. PERF. EU	EI
PRET. PERF. TU	ASTE
PRET. PERF. ELE	OU
PRET. PERF. NOS	AMOS
PRET. PERF. VOS	ASTES
PRET. PERF. ELES	ARAM
PRET. M.Q.P. EU	ARA
PRET. M.Q.P. TU	ARAS
PRET. M.Q.P. ELE	ARA
PRET. M.Q.P. NOS	ÁRAMOS
PRET. M.Q.P. VOS	ÁREIS
PRET. M.Q.P. ELES	ARAM
FUT. PRES. EU	AREI
FUT. PRES. TU	ARÁS
FUT. PRES. ELE	ARÁ
FUT. PRES. NOS	AREMOS
FUT. PRES. VOS	AREIS
FUT. PRES. ELES	ARÃO
FUT. PRET. EU	ARIA
FUT. PRET. TU	ARIAS
FUT. PRET. ELE	ARIA
FUT. PRET. NOS	ARÍAMOS
FUT. PRET. VOS	ARÍEIS
FUT. PRET. ELES	ARIAM
PRES. SUBJ. EU	E
PRES. SUBJ. TU	ES
PRES. SUBJ. ELE	E
PRES. SUBJ. NOS	EMOS
PRES. SUBJ. VOS	EIS
PRES. SUBJ. ELES	EM
IMPERF. SUBJ. EU	ASSE
IMPERF. SUBJ. TU	ASSES
IMPERF. SUBJ. ELE	ASSE
IMPERF. SUBJ. NOS	ÁSSEMOS
IMPERF. SUBJ. VOS	ÁSSEIS
IMPERF. SUBJ. ELES	ASSEM

FUT. SUBJ. EU	AR
FUT. SUBJ. TU	ARES
FUT. SUBJ. ELE	AR
FUT. SUBJ. NOS	ARMOS
FUT. SUBJ. VOS	ARDES
FUT. SUBJ. ELES	AREM
IMP. AFIRM. TU	A
IMP. AFIRM. ELE	E
IMP. AFIRM. NOS	EMOS
IMP. AFIRM. VOS	AI
IMP. AFIRM. ELES	EM
INF. PES. EU	AR
INF. PES. TU	ARES
INF. PES. ELE	AR
INF. PES. NOS	ARMOS
INF. PES. VOS	ARDES
INF. PES. ELES	AREM
GER.	ANDO
PART.	ADO

TABELA 2 - SEGUNDA CONJUGAÇÃO

<u>TEMPO-PESSOAL</u>	<u>SUFIXOS</u>
PRES. IND. EU	O
PRES. IND. TU	ES
PRES. IND. ELE	E
PRES. IND. NOS	EMOS
PRES. IND. VOS	EIS
PRES. IND. ELES	EM
IMPERF. EU	IA
IMPERF. TU	IAS
IMPERF. ELE	IA
IMPERF. NOS	ÍAMOS
IMPERF. VOS	ÍEIS
IMPERF. ELES	IAM
PRET. PERF. EU	I
PRET. PERF. TU	ESTE
PRET. PERF. ELE	EU
PRET. PERF. NOS	EMOS
PRET. PERF. VOS	ESTES
PRET. PERF. ELES	ERAM
PRET. M.Q.P. EU	ERA
PRET. M.Q.P. TU	ERAS
PRET. M.Q.P. ELE	ERA
PRET. M.Q.P. NOS	ÊRAMOS
PRET. M.Q.P. VOS	ÊREIS
PRET. M.Q.P. ELES	ERAM
FUT. PRES. EU	EREI
FUT. PRES. TU	ERÁS
FUT. PRES. ELE	ERÁ
FUT. PRES. NOS	EREMOS
FUT. PRES. VOS	EREIS
FUT. PRES. ELES	ERÃO

FUT. PRET. EU	ERIA
FUT. PRET. TU	ERIAS
FUT. PRET. ELE	ERIA
FUT. PRET. NOS	ERÍAMOS
FUT. PRET. VOS	ERÍEIS
FUT. PRET. ELES	ERIAM
PRES. SUBJ. EU	A
PRES. SUBJ. TU	AS
PRES. SUBJ. ELE	A
PRES. SUBJ. NOS	AMOS
PRES. SUBJ. VOS	AIS
PRES. SUBJ. ELES	AM
IMPERF. SUBJ. EU	ESSE
IMPERF. SUBJ. TU	ESSES
IMPERF. SUBJ. ELE	ESSE
IMPERF. SUBJ. NOS	ÊSSEMOS
IMPERF. SUBJ. VOS	ÊSSEIS
IMPERF. SUBJ. ELES	ESSEM
FUT. SUBJ. EU	ER
FUT. SUBJ. TU	ERES
FUT. SUBJ. ELE	ER
FUT. SUBJ. NOS	ERMOS
FUT. SUBJ. VOS	ERDES
FUT. SUBJ. ELES	EREM
IMP. AFIRM. TU	E
IMP. AFIRM. ELE	A
IMP. AFIRM. NOS	AMOS
IMP. AFIRM. VOS	EI
IMP. AFIRM. ELES	AM
INF. PES. EU	ER
INF. PES. TU	ERES
INF. PES. ELE	ER
INF. PES. NOS	ERMOS
INF. PES. VOS	ERDES
INF. PES. ELES	EREM
GER.	ENDO
PART.	IDO

TABELA 3 - TERCEIRA CONJUGAÇÃO

<u>TEMPO-PESSOAL</u>	<u>SUFIXOS</u>
PRES. IND. EU	O
PRES. IND. TU	ES
PRES. IND. ELE	E
PRES. IND. NOS	IMOS
PRES. IND. VOS	IS
PRES. IND. ELES	EM
IMPERF. EU	IA
IMPERF. TU	IAS
IMPERF. ELE	IA
IMPERF. NOS	ÍAMOS
IMPERF. VOS	ÍEIS
IMPERF. ELES	IAM

PRET. PERF. EU	I
PRET. PERF. TU	ISTE
PRET. PERF. ELE	IU
PRET. PERF. NOS	IMOS
PRET. PERF. VOS	ISTES
PRET. PERF. ELES	IRAM
PRET. M.Q.P. EU	IRA
PRET. M.Q.P. TU	IRAS
PRET. M.Q.P. ELE	IRA
PRET. M.Q.P. NOS	ÍRAMOS
PRET. M.Q.P. VOS	ÍREIS
PRET. M.Q.P. ELES	IRAM
FUT. PRES. EU	IREI
FUT. PRES. TU	IRÁS
FUT. PRES. ELE	IRÁ
FUT. PRES. NOS	IREMOS
FUT. PRES. VOS	IREIS
FUT. PRES. ELES	IRÃO
FUT. PRET. EU	IRIA
FUT. PRET. TU	IRIAS
FUT. PRET. ELE	IRIA
FUT. PRET. NOS	IRÍAMOS
FUT. PRET. VOS	IRÍEIS
FUT. PRET. ELES	IRIAM
PRES. SUBJ. EU	A
PRES. SUBJ. TU	AS
PRES. SUBJ. ELE	A
PRES. SUBJ. NOS	AMOS
PRES. SUBJ. VOS	AIS
PRES. SUBJ. ELES	AM
IMPERF. SUBJ. EU	ISSE
IMPERF. SUBJ. TU	ISSES
IMPERF. SUBJ. ELE	ISSE
IMPERF. SUBJ. NOS	ÍSSEMOS
IMPERF. SUBJ. VOS	ÍSSEIS
IMPERF. SUBJ. ELES	ISSEM
FUT. SUBJ. EU	IR
FUT. SUBJ. TU	IRES
FUT. SUBJ. ELE	IR
FUT. SUBJ. NOS	IRMOS
FUT. SUBJ. VOS	IRDES
FUT. SUBJ. ELES	IREM
IMP. AFIRM. TU	E
IMP. AFIRM. ELE	A
IMP. AFIRM. NOS	AMOS
IMP. AFIRM. VOS	I
IMP. AFIRM. ELES	AM
INF. PES. EU	IR
INF. PES. TU	IRES
INF. PES. ELE	IR
INF. PES. NOS	IRMOS
INF. PES. VOS	IRDES
INF. PES. ELES	IREM
GER.	INDO
PART.	IDO