Instituto de Ciências Matemáticas e de Coputação

ISSN - 0103-2569

ETIQUETAÇÃO DE UM CORPUS E CONSTRUÇÃO DE UM ETIQUETADOR DE PORTUGUÊS

Sandra M. Aluísio
Rachel V. Aires

Nº 107

RELATÓRIOS TÉCNICOS DO ICMC

São Carlos
Mar./2000

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

ETIQUETAÇÃO DE UM CORPUS E CONSTRUÇÃO DE UM ETIQUETADOR DE PORTUGUÊS

Sandra M. Aluísio
Rachel V. Aires

NILC-TR-00-2

Março, 2000

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

Etiquetação de um Corpus e Construção de um Etiquetador de Português

Resumo

Neste trabalho, relatamos as atividades desenvolvidas no ano de 1999 relativas ao subprojeto "Etiquetação de um Corpus e Construção de um Etiquetador de Português" inserido no Projeto "Revisor Gramatical e Ferramentas de Auxílio à Escrita", aprovado no âmbito do programa PADCT-III, CDT, fase 1, em Julho de 1998.

1. Introdução

Este subprojeto possui duas tarefas relacionadas. Uma delas visa à construção de um etiquetador do tipo Part-Of-Speech (POS) Tagger para o português brasileiro. Uma das principais fontes de problemas para a etapa de análise sintática, ou parsing, de uma língua é a pluricategorização gramatical de suas palavras. No ReGra (Martins et al, 1998), várias intervenções inadequadas (ou falsos erros) decorrem da associação indevida de uma categoria gramatical a uma palavra, que geram representações (árvores) sintáticas incorretas. Um etiquetador do tipo POS tem como meta a associação de atributos morfossintáticos a cada palavra de uma sentença, para minimizar as possibilidades estruturais e, portanto, excluir regras gramaticais que antes se mostravam ambíguas por falta de informações suficientes para decidir por uma estruturação sintática adequada. Porém, realizar a primeira etapa da análise sintática não é a única utilidade de um etiquetador do tipo POS. Os processadores de corpus, por exemplo, são ferramentas de muita utilidade, uma vez que são capazes de realizar importantes estatísticas sobre a língua escrita, subsidiando pesquisas e desenvolvimento de processadores automáticos de língua natural. Porém, para extrair o máximo de informações de um corpus através dessas ferramentas, é necessário fornecer, como entrada, um corpus já etiquetado com marcas do tipo POS. Desta forma, etiquetar a seção de textos classificados como corrigidos[1] (32 milhões de palavras) do corpus do NILC, com a ajuda de um etiquetador que apresente uma precisão comparável aos melhores etiquetadores para o inglês constitui a segunda tarefa deste subprojeto. A Seção 2 contextualiza este subprojeto. As Seções 3 a 5 reportam os trabalhos realizados durante 1999: (1) elaboração de várias versões do NILC Tagset, hoje estável, e contando do 37 etiquetas morfossintáticas e 16 de pontuação; (2)

[1] A porção corrigida do corpus está acessível para consulta, através do *IMS corpus query tools*, da Universidade de Stuttgart, no repositório de recursos computacionais para a língua portuguesa: http://cgi.portugues.mct.pt/acesso/.

marcação de um corpus inicial de treinamento, através da utilização de um software para marcação semi-automática do corpus e de um etiquetador simbólico para o português; (3) estudo e transporte de etiquetadores independentes de língua, disponíveis na WWW, para o português/NILC tagset. A Seção 6 indica as próximas etapas deste subprojeto.

2. Contextualização

A etiquetação morfossintática é uma tarefa bastante conhecida em PNL. Ela consiste em se atribuir uma categoria gramatical (e talvez outros atributos referentes a cada categoria como, por exemplo, gênero e número de substantivos) à cada palavra de um texto, escolhida dentre todas as listadas no léxico. A anotação de um texto com etiquetas morfossintáticas é útil para várias manipulações subseqüentes do texto, pois fornece: a) uma forma de abstração das palavras do texto quando desejamos processar todas as palavras que pertençam a certa categoria de alguma forma específica, por exemplo, levantar todos os nomes próprios do texto; b) um grau de desambigüização que pode ser benéfico para os níveis subseqüentes de análise como o *parsing* (Zavrel and Daelemans, 1998).

O processo de etiquetação pode ser dividido em duas fases: a de treinamento, que corresponde à aquisição do modelo lingüístico, geralmente a partir de um corpus anotado, e a de teste, quando o algoritmo de etiquetação é aplicado a uma parte reservada do corpus para avaliação do aprendizado. A divisão típica de um corpus em treinamento e teste é 90 a 80% para treinamento e 10 a 20% para teste. Uma vez que a precisão alcançada esteja de acordo com o valor esperado, o algoritmo de etiquetação pode ser utilizado para etiquetar os textos da língua em que foi treinado. Caso a precisão esteja aquém da esperada, pode-se utilizar um corpus de treinamento maior, sendo que, para etiquetadores estatísticos o tamanho típico deste corpus está entre 1 milhão e 2 milhões de palavras, enquanto que, etiquetadores neurais permitem um corpus de treinamento menor — entre 100.000 e 500.000 palavras.

Existem vários sistemas transportáveis e disponíveis na WWW que geram etiquetadores para qualquer língua a partir de corpus anotado previamente. Entretanto, a etiquetação morfológica não possui uma solução ótima, embora os melhores etiquetadores alcancem uma precisão de 97%[2]. Os etiquetadores desenvolvidos podem ser classificados em dois grupos principais de acordo com o tipo de conhecimento que eles usam para representar o modelo lingüístico: simbólicos (regras, base de casos, árvores de decisão sem probabilidade) e probabilísticos (modelos de Markov, baseados em entropia, árvores de decisão estatísticas,

---

[2] Etiquetadores humanos são consistentes entre si em 98%, dando assim um limite superior à performance esperada de etiquetadores computacionais.

redes neurais). Abordagens híbridas são também consideradas já que alguns etiquetadores empregam conhecimento probabilístico e simbólico. A precisão reportada por muitos etiquetadores estatísticos está entre 96% e 97%, enquanto que os simbólicos, baseados em Constraint Grammars[3], ultrapassam 99%. Considerando esses dados, pode-se pensar que o problema de etiquetação morfológica está resolvido desde que esta precisão é perfeitamente aceitável para muitos sistemas de PLN. Por que então estudar e investigar neste subprojeto a construção de um etiquetador para o português com uma precisão melhor ainda que os disponíveis para o inglês? O que significa um acréscimo de 0.3% na precisão?

Orações em texto têm, em média, cerca de 30 palavras. Se nós admitirmos uma taxa de erro de 3 a 4%, então, em média, cada oração conteria um erro. Desde que a etiquetação morfossintática é uma tarefa importante para a interpretação em PLN, começar com um erro em cada oração pode ser muito desvantajoso, especialmente se considerarmos que a propagação de tais erros pode crescer não linearmente. Para superar a barreira da taxa de erro de 3-4%, o projeto de etiquetação de textos em português do NILC combinará vários etiquetadores disponíveis na WWW, juntamente com o etiquetador simbólico derivado do ReGra[4], seguindo a abordagem de vários outros trabalhos (Brill and Wu, 1998; Zavrel and Daelemans, 1998).

3. Versões do NILC Tagset

O primeiro NILC tagset (conjunto de etiquetas) foi criado de acordo com as recomendações constantes em (EAGLES, 1996) para se incluir atributos e valores para as 14 categorias principais do tagset inicial:

1. S (substantivo)	6. PP (preposição)	11. E (residual-palavras estrangeiras) SE (residual-símbolos especiais) FM (residual-fórmulas matemáticas)
2. V (verbo)	7. C (conjunção)	
3. AJ (adjetivo)	8. N (numeral)	
4. PR (pronome)	9. AR (artigo)	12. símbolo (pontuação)
5. AV (advérbio)	10. I (interjeição)	13. SI (siglas)
		14. AB (abreviaturas)

[3] Uma regra de uma Constraint Grammar coloca o problema da ambigüidade em foco pela especificação das leituras impossíveis (que devem ser descartadas) ou requeridas (e assim devem ser escolhidas) no contexto de uma sentença (Bick, 1996).

[4] Este etiquetador combina as informações do léxico com as regras de desambigüização usadas pelo *parser* do revisor (Martins et al., 1998)..

Neste caso, o tag de uma palavra é formado pelos valores, em seqüência, de cada atributo relativo a cada categoria do tagset. Note que, separando cada valor usa-se um ponto. Por exemplo, os atributos e valores da categoria Substantivo são:

Tipo	Comum (C)	Próprio (P)		
Gênero	Masculino (M)	Feminino (F)	2G (2G)	
Número	Singular (S)	Plural (P)	2N (2N)	Invariável (I)
Grau	Aumentativo (A)	Diminutivo (D)	Neutro (N)	

Exemplos: mesa /S.C.F.S.N — carrinhos /S.C.M.P.D — personagem /S.C.2G.S.N

Quatro outras versões do tagset foram desenvolvidas tentando-se sanar problemas de eficiência de treinamento de um dos algoritmos de etiquetagem utilizados e também para cobrir novos casos encontrados no corpus. No primeiro caso, o tagset foi reduzido, pois não havia necessidade do aprendizado de muitos dos atributos das categorias principais. Uma vez que o etiquetador consegue descobrir a categoria, o texto pode ser pós-processado com a ajuda do léxico do NILC, preenchendo-se os valores para os atributos restantes. Para o segundo, a classe das palavras residuais recebeu nova subcategorização incorporando REGIONALISMOS e COLOQUIALISMOS presentes em muitos textos literários; as categorias de locuções adverbial, conjuncional, prepositiva e pronominal foram criadas sob a influencia do projeto temático Fapesp para a etiquetagem do português clássico baseada em corpus (http://www.ime.usp.br/~tycho), e também foram os nomes próprios compostos etiquetados igualmente à forma seguida em tal projeto. Siglas e abreviaturas desapareceram como categorias principais sendo renomeadas com a categoria de nomes. A classificação dos verbos e pronomes foi bastante alterada. Resolvemos privilegiar o atributo predicação para os verbos, tornando este atributo uma categoria principal — uma tentativa inédita na construção de tagsets que geralmente privilegiam as formas nominais.

O NILC tagset, encontra-se estável, apresentando 37 etiquetas morfossintáticas e 16 de pontuação. O número de etiquetas é uma das preocupações no projeto de etiquetadores, pois ele afeta o desempenho do algoritmo de aprendizado de certos etiquetadores encontrados no WWW. Por exemplo, o etiquetador de Brill (Brill, 1995) tem uma complexidade de tempo da ordem de O ($t^2 * (3\ t^2 + 7t) * n$), onde t é o número de etiquetas e n o tamanho do corpus de treinamento. De acordo com Brill, com um tagset de 36 etiquetas, o aprendizado de regras contextuais levaria 1 dia em uma Sun Sparc-10 com um corpus de 500,000 palavras. Assim, para um tagset de 176 tags, o tempo de treinamento é proibitivo, durando 543 dias, ou seja, 1 ano e meio. Como pretendemos portar o etiquetador de Brill para o NILC tagset, o número de etiquetas atual é adequado para tal empreitada. O NILC tagset é apresentado abaixo.

Adjetivo	ADJ
Advérbio	ADV
Artigo	ART
Número Cardinal	NC
Número Ordinal	ORD
Outros Números	NO
Substantivo Comum	N
Nome Próprio	NP
Conj. Coordenativa	CONJCOORD
Conj. Subordinativa	CONJSUB
Pronome Demonstrativo	PD
Pronome Indefinido	PIND
Pronome Oblíquo Átono	PPOA
Pronome Pessoal Caso Reto	PPR
Pronome Possessivo	PPS
Pronome Relativo	PR
Pronome Oblíquo Tônico	PPOT
Pronome Interrogativo	PINT
Pronome Apassivador	PAPASS
Pronome de Realce	PREAL
Pronome Tratamento	PTRA
Preposição	PREP
Verbo Auxiliar	VAUX
Verbo de Ligação	VLIG
Verbo Intransitivo	VINT
Verbo Transitivo Direto	VTD
Verbo Transitivo Indireto	VTI
Verbo Bitransitivo	VBI
Interjeição	I
Locução Adverbial	LADV
Locução Conjuncional	LCONJ
Locução Prepositiva	LPREP
Locução Pronominal	LP
Palavra Denotativa	PDEN
Locução Denotativa	LDEN
Palavras ou Símbolos Residuais	RES
Dúvida	DUV
;	;
.	.
:	:
-	-
(	(
!	!
?	?
...	...
)	)
"	"
[	[
]	]

{	{
}	}
'	'
,	,

OBSERVAÇÕES:

1) RESIDUAL

É atribuída a palavras que estão fora das classes tradicionais, embora ocorram com freqüência. Exemplo: palavras estrangeiras, fórmulas matemáticas (quando aparecem sem espaço em branco, como em X/21, 2+3=5), símbolos especiais, como @, R$, $, %, #, +, -, =, ^, ~, <, >, /, \

É dividida em 4 subcategorias:

(a) PALAVRAS ESTRANGEIRAS (E)

Não se trata de estrangeirismos (ex. flat, shopping, abajour, pizza), ou seja, não são palavras já incorporadas ao Português.
Ex. home
Love

(b) SÍMBOLOS ESPECIAIS (SE)

São os símbolos com exceção daqueles da classe 14. Pontuação:
@, R$, $, %, #, +, -, =, ^, ~, <, >, /, \ e todos os demais que podem ser inseridos dos alfabetos especiais (Symbol, etc.)

(c) FÓRMULAS MATEMÁTICAS (FM)

São expressões, sem espaço entre os símbolos, que não constituem palavras do português.
Ex. f(x)
x+y (observe que x + y – com espaço – seria classificado como x/S +/SE y/S
35/4
10%

(d) REGIONALISMOS e COLOQUIALISMOS

'magina, t'esconjuro, p'ra, inté.

2) Em conjunto com as etiquetas, utilizamos três operadores:

- Dois operadores para tratamento de contrações, mesóclises e ênclises:

+: para contrações e ênclises. Exemplo: do = PREP+ART; dar-lhe = V??+PPOA (por exemplo VTD+PPOA).

!: para mesóclises. Exemplo: dar-lhe-ei = V??!PPOA (por exemplo VTD!PPOA).

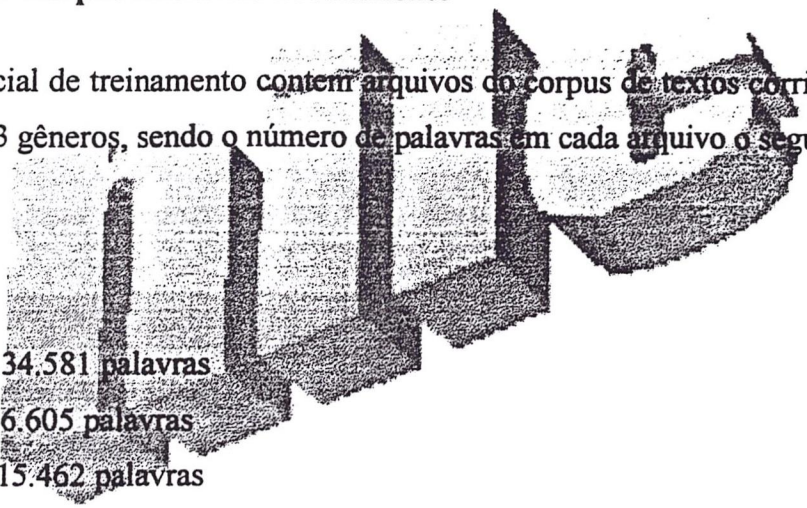
- O operador "-" para separar as tags NP, LADV, LPREP, LCONJ, LP, LDEN de seus índices (por exemplo: NP-31, LPREP-22, LCONJ-41)

3) Os índices são utilizados para marcar nomes próprios compostos e locuções

- Nomes próprios compostos são marcados com um índice NP-ij, onde i é o tamanho do NP e j a posição da palavra no NP, como no exemplo: Rio_NP-31 de_NP-32 Janeiro_NP-33
- Locuções são marcados com um índice L??-ij, onde i é o tamanho da locução e j a posição da palavra na locução, como no exemplo: a_LADV-21 fundo_LADV-22

4. Marcação do Corpus Inicial de Treinamento

O corpus inicial de treinamento contem arquivos do corpus de textos corrigidos do NILC, selecionados de 3 gêneros, sendo o número de palavras em cada arquivo o seguinte:



- Jornalisticos

ie96ma01.tag --- 34.581 palavras
ie96ma15.tag --- 6.605 palavras
rura95ag.tag --- 15.462 palavras
--- 56.648 palavras

- Didáticos

9biol3.tag --- 16.257 palavras

- Literários

litbras.tag --- 32.054 palavras

totalizando: 104.959 palavras marcadas.

Para auxiliar a marcação manual do corpus de treinamento para vários etiquetadores foi desenvolvido um sistema gráfico em Borland Delphi 4.0. Foi também incorporada uma biblioteca desenvolvida no NILC para tornar disponível ao usuário do sistema de marcação todas as opções de classes gramaticais possíveis para a palavra sendo marcada, obtidas do Léxico de português do Brasil. O usuário deve, então, escolher uma opção adequada ao

contexto para as palavras que possuem mais de uma classe gramatical. O sistema também possui um *help* online com as etiquetas do NILC tagset para ser utilizado quando a palavra do texto não pertence ao léxico. A interface do software consiste basicamente de duas janelas.

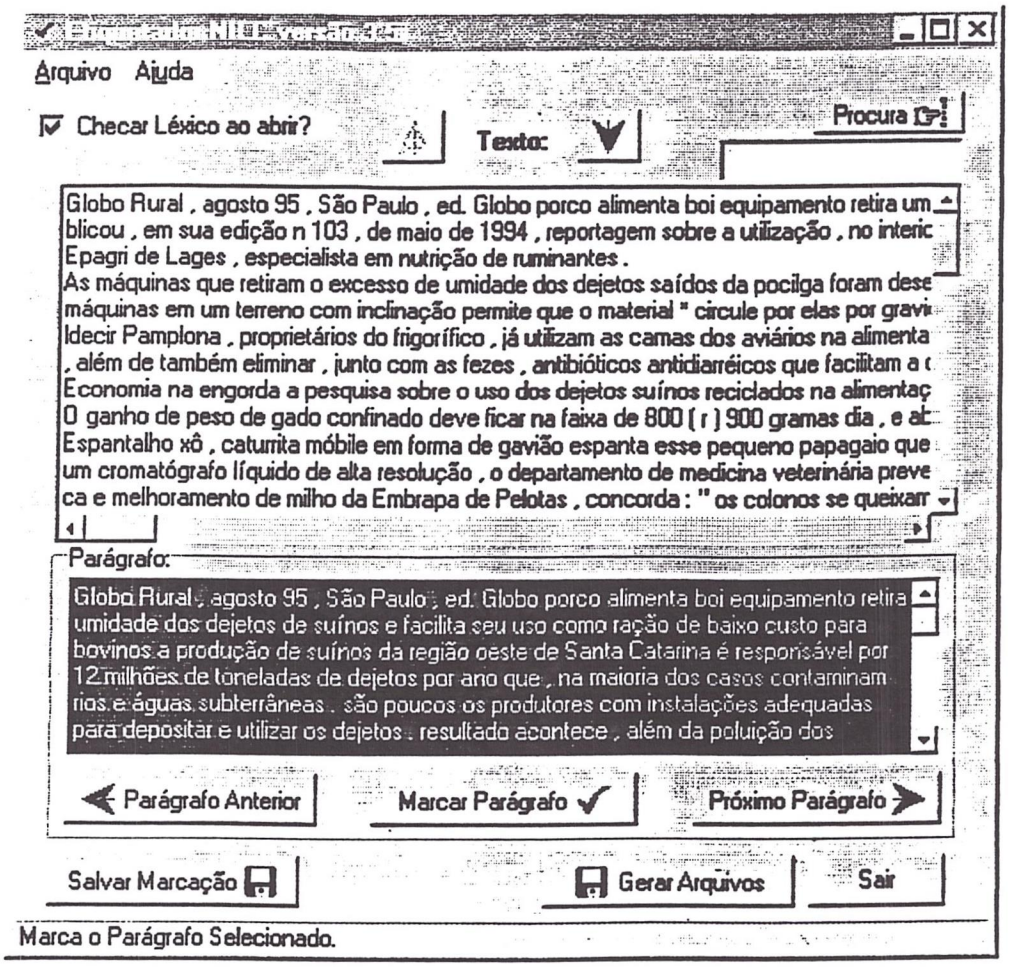


Figura 2: Janela Principal do Sistema de Marcação.

A primeira, como ilustrado na Figura 2, tem por objetivo principal abrir um arquivo de texto parcialmente marcado, ou a ser marcado por inteiro. Isso é feito selecionando-se a opção Abrir Texto do menu "Arquivo". Após a seleção do arquivo, a interface selecionará automaticamente a primeira sentença e a colocará na caixa com rótulo "Sentença". O usuário também poderá escolher por qual sentença ele pretende começar a marcação fazendo uso dos botões "Sentença Anterior" e "Próxima Sentença", que atualizam o conteúdo da caixa de rótulo "Sentença". Também, para maior facilidade do usuário foi inserido na interface um botão para procura de cadeias de caràcteres, permitindo assim que este reinicie a marcação de textos mais longos a partir de onde havia parado. Para isto basta que este se lembre de qualquer parte do texto para qual deseja avançar, digitá-la na caixa de textos localizada acima da caixa de texto principal (como mostra a Figura 2) e, ao apertar o botão rotulado "Procura", a caixa que indica o parágrafo atual avançará até encontrar o parágrafo que contém o texto

procurado ou até o último parágrafo no caso da cadeia procurada não exista no texto aberto entre o parágrafo atual e o final do texto.

Quando a sentença desejada estiver selecionada, o usuário deverá fazer uso do botão "Marcar Sentença" para prosseguir com a marcação. Após a marcação, o botão "Salvar Marcação" deverá ser usado para salvar para o disco as marcações executadas e o botão "Sair" para encerar o programa. Na última versão do sistema, as cores da caixa de texto foram mudadas para azul com fonte amarela, para melhor visualização prolongada. Após o botão Marcar Sentença ser selecionado, a janela de marcação mostrada na Figura 3 se abrirá.

Para a marcação do corpus inicial também contamos com uma adaptação do etiquetador simbólico baseado em Constraint Grammars desenvolvido por Eckhard Bick[5] (Bick, 1996) para trabalhar com as etiquetas do NILC tagset. Este etiquetador possui uma precisão variando de 97.6 a 99.7% quando trabalha com seu tagset original que é um pouco diferenciado do NILC tagset principalmente quanto a pronomes e verbos. Entretanto, estas duas classes recebem informação do analisador sintático para suprir a transitividade dos verbos, por exemplo. Desta forma, sempre que havia dúvida de marcação do corpus inicial o corpus marcado pelo etiquetador simbólico foi pesquisado servindo como fonte de referência dado sua alta precisão.

[5] http://visl.hum.ou.dk/tagwebsite/visl/visltop.html

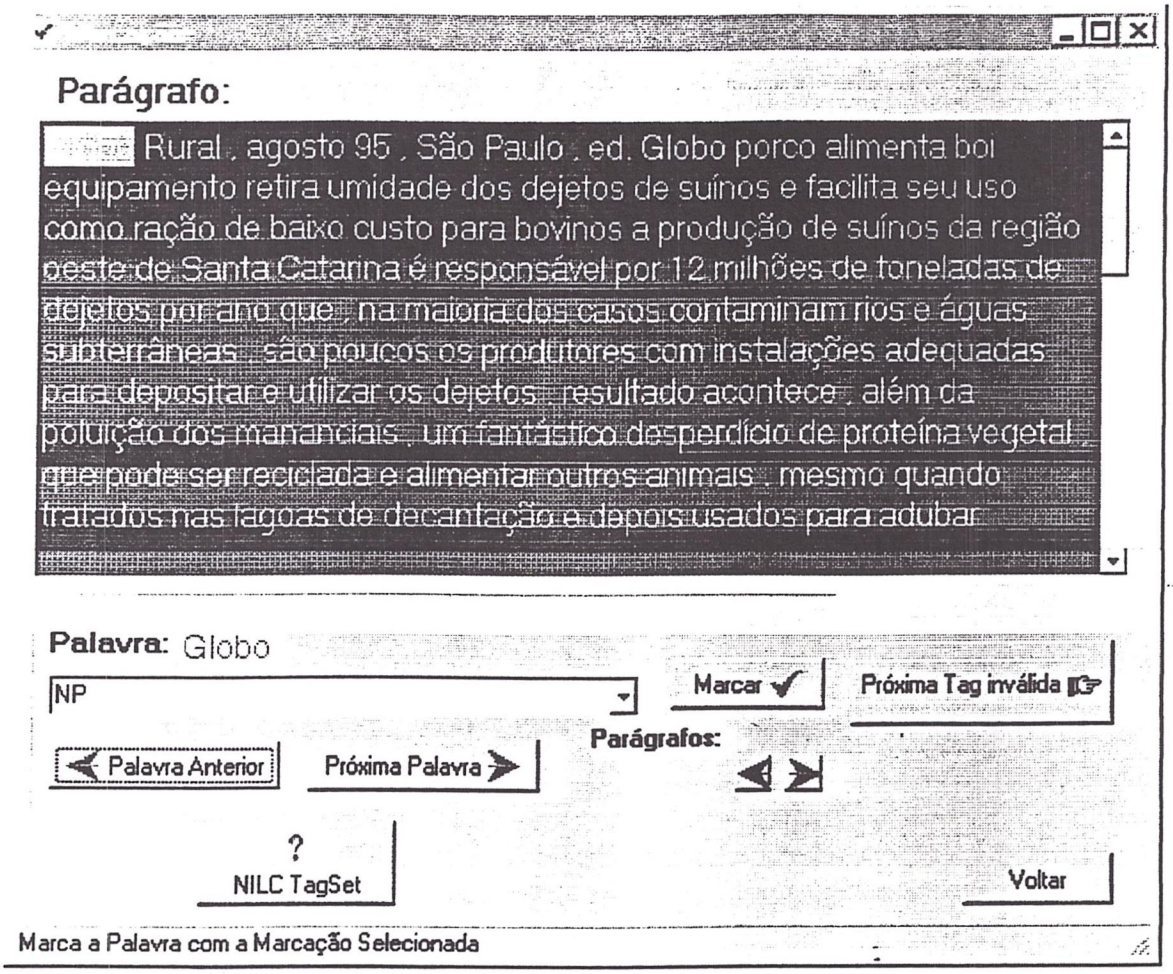


Figura 3: Janela de Marcação.

5. Etiquetadores Avaliados e Utilizados

A tarefa relacionada com a construção de um etiquetador para o português do Brasil prevê o porte de três etiquetadores disponíveis na WWW: TreeTagger (Schmid, 1995), MXPOST (Ratnaparkhi, 1996), e o baseado em transformações dirigidas por erros (TBE) Brill (Brill, 1995). Todos estes três etiquetadores foram estudados e avaliados e um experimento com o MXPOST e o corpus marcado de 104.959 palavras foi realizado.

Este corpus inicial palavras seria suficiente para treinar um etiquetador baseado em redes neurais, porém o TreeTagger, o TBE, e o MXPOST exigem um corpus etiquetado de um milhão de palavras para termos uma precisão melhor. Este corpus deverá ser obtido incrementalmente a partir do uso dos etiquetadores testados, sendo posteriormente revisado manualmente.

A implementação desses etiquetadores tem por objetivo verificar a) se existe um etiquetador particularmente adequado para o português do Brasil, b) avaliar o impacto de diferenças lingüísticas entre textos de treinamento e teste e, c) avaliar o desempenho de cada etiquetador com três tipos de texto: didáticos, jornalísticos e literários. Como dito acima

(Seção 2), na busca do melhor etiquetador para o português o projeto de etiquetação de textos em português do NILC também combinará vários etiquetadores disponíveis na WWW, juntamente com o etiquetador simbólico derivado do ReGra, seguindo a abordagem de vários outros trabalhos (Brill and Wu, 1998; Zavrel and Daelemans, 1998).

A avaliação de um etiquetador é baseada na comparação da saída produzida por diferentes etiquetadores no mesmo texto. O conjunto de etiquetas é o mesmo para todos os etiquetadores, portanto a comparação se concentra na precisão dos mesmos. Os seguintes parâmetros são identificados, comparados e tem os resultados interpretados: a) todas as taxas de erro de cada etiquetador (com dicionário aberto e fechado); b) as etiquetas que são mais problemáticas para cada etiquetador podem ser identificadas por um contador de freqüência de erro de marcação e a classificação por etiqueta; c) uma comparação entre os etiquetadores, e pares de etiquetas confundidas, com o objetivo de identificar etiquetas que são confundidas da mesma forma em todos os etiquetadores testados.

Uma avaliação por tipo de texto também foi feita para avaliar como textos de treinamento e de teste com estilos diferentes afetam o resultado e como um certo tipo de texto interfere no comportamento de um determinado etiquetador. Seguem abaixo o procedimento de treinamento do etiquetador MXPOST com dicionário aberto e fechado[6] e o resultado da avaliação de ambas versões.

Procedimento de Treinamento e Teste no MXPOST

O MXPOST roda no sistema operacional unix. Para instalá-lo bastou descompaquitá-lo em um diretório e editar a variável "classpath" dos arquivos do etiquetador com o nome do diretório da instalação. Ele foi treinado para dicionário aberto e fechado. No caso do dicionário aberto o arquivo de treinamento continha 94.444 palavras e o de teste 10.515 palavras (divisão 90 e 10% para treinamento e teste, respectivamente). No treinamento com dicionário fechado, o arquivo de treinamento continha 104.959 palavras. Os arquivos de treinamento devem estar sempre no formato "palavra_tag" e devem ter apenas uma sentença por linha. Para treinar basta usar o comando:

trainmxpost diretórioprojeto arquivotreinamento.

---

[6] Um dicionário aberto é constituído por todas as palavras do *corpus* de treinamento e por suas respectivas classes de ambigüidade. enquanto que um dicionário fechado é constituído por todas as palavras do *corpus* de treinamento e de teste e por suas respectivas classes de ambigüidade.

onde diretórioprojeto é nome dado para o diretório onde ficará o modelo que será treinado e arquivotreinamento é o seu arquivo de dados para o treinamento. Para marcar um texto basta usar o comando:

mxpost diretórioprojeto < arquivoteste.

onde arquivoteste é o arquivo que será marcado. O cálculo de precisão foi realizado usando o comando "diff" do unix:

diff -bi Texto100A.res Teste100.tag

diff -bi Texto100F.res Teste100.tag

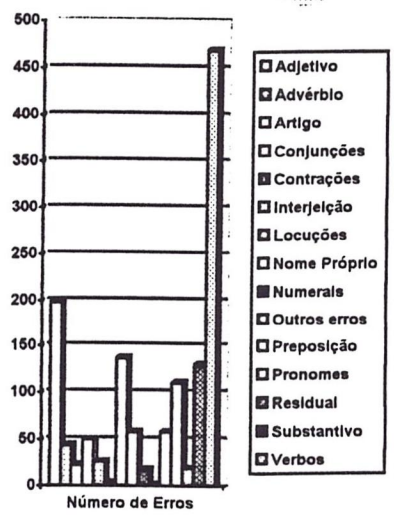
O treinamento do MXPOST para português com um arquivo de 104.959 palavras demorou cerca de 4 dias em uma Silicon Graphics O2 com 64 Mb. A marcação de um arquivo de cerca de 10.000 palavras demora entre 20 e 35 minutos na mesma máquina.

Os testes no dicionário fechado foram feitos com os 3 tipos de texto presentes no corpus: didático, jornalístico e literário. O arquivo de teste literário continha 10.515 palavras, o didático 10.514 palavras e o jornalístico 10.513 palavras.

Avaliação Inicial

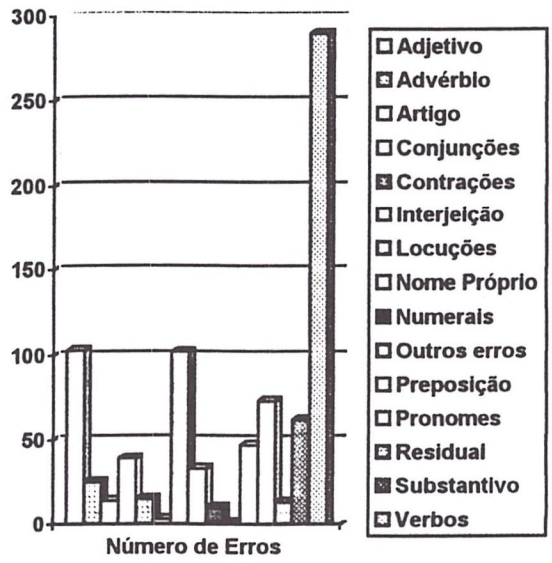
A precisão para dicionário aberto é de **87,2%** (1342 erros), quando testado com o tipo de texto didático. A tabela e gráfico abaixo mostram o desempenho do etiquetador para cada categoria do tagset.

Classes	Número de Erros
Adjetivo	198
Advérbio	44
Artigo	22
Conjunções	50
Contrações	26
PREP+ADJ	*2*
PREP+ART	*13*
PREP+N	*1*
PREP+PD	*6*
PREP+PIND	*1*
PREP+PPOT	*3*
Interjeição	3
Locuções	137
Nome Próprio	58
Numerais	17
Número Cardinal	*9*
Número Ordinal	*5*
Outros Números	*3*
Outros erros	1
Preposição	58
Pronomes	110
Residual	19
Substantivo	130
Verbos	469



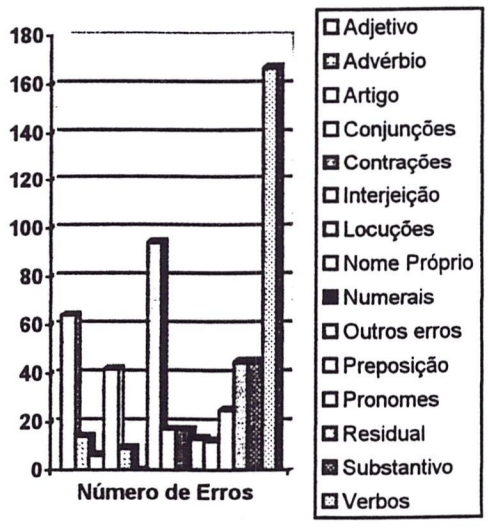
A precisão para dicionário fechado foi de **92,1%** (830 erros), quando testado com o tipo de texto literário. A tabela e gráfico abaixo mostram o desempenho do etiquetador para cada categoria do tagset.

Classes	Número de Erros
Adjetivo	103
Advérbio	25
Artigo	14
Conjunções	39
Contrações	15
PREP+ADJ	*2*
PREP+ART	*6*
PREP+N	*1*
PREP+PD	*3*
PREP+PIND	*1*
PREP+PPOT	*2*
Interjeição	3
Locuções	102
Nome Próprio	33
Numerais	10
Número Cardinal	*3*
Número Ordinal	*4*
Outros Números	*3*
Outros erros	1
Preposição	47
Pronomes	73
Residual	13
Substantivo	62
Verbos	290



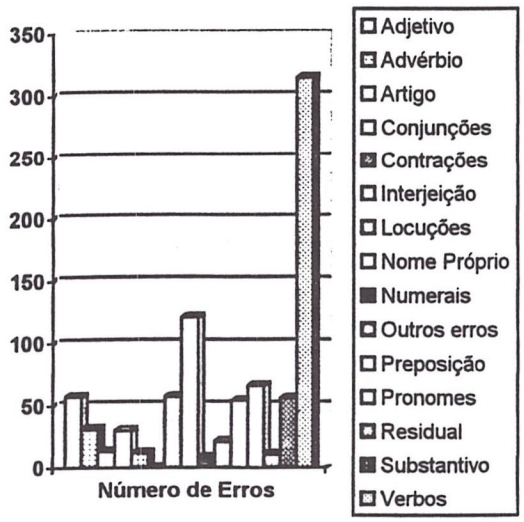
A precisão para dicionário fechado foi de **94,6%** (570 erros), quando testado com o texto didático. A tabela e gráfico abaixo mostram o desempenho do etiquetador para cada categoria do tagset.

Classes	Número de Erros
Adjetivo	64
Advérbio	14
Artigo	6
Conjunções	42
Contrações	9
PREP+ART	*8*
PREP+PPR	*1*
Interjeição	0
Locuções	94
Nome Próprio	17
Numerais	17
Número Cardinal	*2*
Número Ordinal	*1*
Outros Números	*14*
Outros erros	13
Preposição	12
Pronomes	25
Residual	45
Substantivo	45
Verbos	167



A precisão para dicionário fechado foi de **92%** (852 erros), quando utilizado o tipo de texto jornalístico. A tabela e gráfico abaixo mostram o desempenho do etiquetador para cada categoria do tagset.

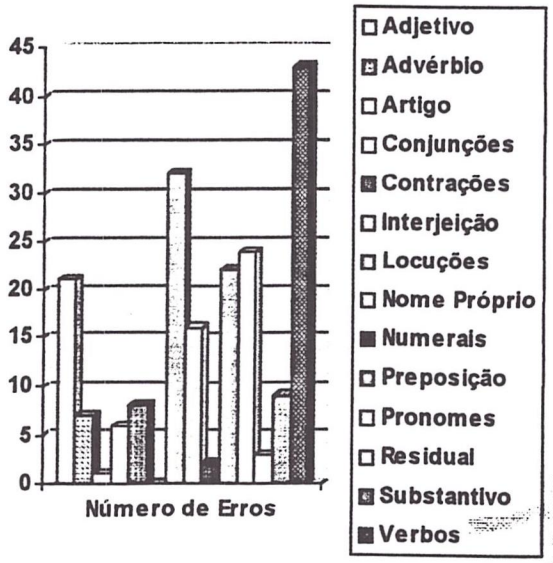
Classes	Número de Erros
Adjetivo	57
Advérbio	31
Artigo	13
Conjunções	30
Contrações	12
PREP+ART	*12*
Interjeição	0
Locuções	57
Nome Próprio	121
Numerais	8
Número Cardinal	*3*
Número Ordinal	*3*
Outros Números	*2*
Outros erros	21
Preposição	54
Pronomes	66
Residual	11
Substantivo	56
Verbos	315



Os testes acima mostram que o desempenho das duas versões do MXPOST (dicionário fechado e aberto) quando treinado com o NILC tagset e textos em português ainda está aquém da desejada. Este resultado era esperado, pois para se obter melhores valores para a precisão precisaríamos ter um corpus com 1 milhão de palavras. Comparando-se os erros do treinamento com dicionários aberto e fechado vimos que eles erram 314 palavras em comum,

mas em apenas 194 destes erros o erro é o mesmo. Um exemplo de erro diferente é um caso onde a classificação correta para a palavra "divertires" era verbo transitivo direto. Neste caso, o erro do treinamento aberto foi marcá-la como verbo bitransitivo e do fechado marcá-la como adjetivo. Quanto às locuções, quando o número de exemplos de uma locução é pequeno não existe aprendizado, mas quando a locução é bastante freqüente o aprendizado acontece (por exemplo, a locução denotativa "é que" é marcada corretamente como "é_LDEN-21 que_LDEN-22"). Existe também o problema de aprendizado de somente uma parte da locução, que deve ser solucionado no treinamento das novas versões, provavelmente com o pós-processamento do texto marcado. Por exemplo, o etiquetador marca "de cima à baixo" como "de_LADV-21 cima_LADV-22 a_ART baixo_ADJ" quando o esperado seria marcar como "de_LADV-41 cima_LADV-42 a_LADV-43 baixo_LADV-44". A tabela e gráfico abaixo mostram a distribuição dos 194 erros comuns entre os etiquetadores para cada categoria do tagset.

Classes	Número de Erros
Adjetivo	21
Advérbio	7
Artigo	1
Conjunções	6
Contrações	8
PREP+ADJ	*2*
PREP+ART	*4*
PREP+N	*0*
PREP+PD	*1*
PREP+PIND	*0*
PREP+PPOT	*1*
Interjeição	0
Locuções	32
Nome Próprio	16
Numerais	2
Número Cardinal	*1*
Número Ordinal	*0*
Outros Números	*1*
Preposição	22
Pronomes	24
Residual	3
Substantivo	9
Verbos	43



6. Próximas Etapas

As próximas etapas quanto à construção do etiquetador são: porte para o português dos dois outros etiquetadores já estudados (TreeTagger e TBE) e combinação dos três etiquetadores estudados e do etiquetador simbólico derivado do ReGra. Quanto à tarefa de marcação do corpus do NILC, esta tarefa será realizada tão logo se obtenha um etiquetador para o português com uma precisão que se compare a dos melhores etiquetadores para o inglês (entre 97 e 98%).

Referências

Bick, E. (1996) Automatic Parsing of Portuguese. Anais do II Encontro para o Processamento Computacional de Português Escrito e Falado (PROPOR'96), pp. 91-100. Curitiba.

Brill, E. & Wu. J. (1998) Classifier Combination for Improved Lexical Disambiguation. ACL98. (http://www.cs.jhu.edu/~brill/acadpubs.html).

Brill, E. Transformation-based error-driven learning of natural language: A case study in part of speech tagging. Computational Linguistics, v. 21, n.4, p. 543-565, 1995. (http://www.cs.jhu.edu/~brill/papers.html)

EAGLES (1996) EAGLES - Expert Advisory Group on Language Engineering Standards. Recommendations for the Morphosyntactic Annotation of Corpora. (http://www.ilc.pi.cnr.it/EAGLES96/browse.html).

Martins, R.T., Hasegawa, R., Nunes, M.G.V., Montilha, G., Oliveira Jr., O.N. (1998b). Linguistic isuues in the develoment of ReGra: a Grammar Checker for Brazilian Portuguese. Natural Language Engineering, 4(4), pp.287-307.

Ratnaparkhi, A. (1996) A maximum entropy part-of-speech tagger. Proceedings of the Conference on Empirical Methods in Natural Language Processsing. University of Pennsylvania.

Schmid, H. (1995). Probabilistic Part-of-Speech Tagging using Decision Trees. (http://www.ims.unituttgart.de/Tools/DecisionTreeTagger.html).

Zavrel, H. & Daelemans, J.(1998). Improving Data Driven Wordclass Tagging by System Combination. Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL'98); cmp-lg/9807013.