

# **Bases de Conhecimento para Pré-Etiquetagem de Itens Lexicais**

Marco Gonzalez  
Maurício C. de Oliveira  
Rodrigo C. Picada  
Vera L. S. de Lima  
PUCRS - Faculdade de Informática  
Av.Ipiranga, 6681 – Prédio 16 - PPGCC  
90619-900 Porto Alegre, Brazil  
**{gonzalez, vera}@inf.pucrs.br**

## **Resumo**

Um sistema de etiquetagem de texto acrescenta informações, através de etiquetas, associadas a cada item lexical do texto analisado. Utilizando bases de conhecimento, nosso sistema executa uma pré-etiquetagem, abrindo caminho para o processo completo a ser realizado no âmbito da etiquetagem de textos. Na etapa que abordamos aqui, alguns itens lexicais podem receber mais de uma etiqueta, configurando, nestes casos, uma ambigüidade. Numa primeira fase, os itens lexicais são analisados a partir de uma base de palavras conhecidas e, em seguida, através de seus sufixos. Numa segunda fase, a etiquetagem ambígua é reduzida com o auxílio de uma base de regras sintáticas.

## **Abstract**

An annotation system inserts tags in a text. These tags add information which are associated with each lexical item of the analyzed text. Using databases of knowledge, our system executes a preliminary tagging which will include initial information for the complete process of text annotation. In the stage that is presented here, it is possible that some lexical items receive more than one tag, and that defines an ambiguous tagging. In a first phase, the lexical items are analyzed using a database of known words and, after, a database of suffixes is used. In a second phase, a database of syntactic rules is applied. This last procedure tries to reduce the ambiguous tags number.

**Palavras-chave:** bases de dados de conhecimento, processamento da linguagem natural, morfologia, regras sintáticas, etiquetagem de texto.

## 1 Introdução

Etiquetadores de texto [1,2,3,5,6] constituem importante ferramenta às etapas de análise de documentos textuais. Acrescentam informações sobre itens lexicais<sup>1</sup> em um texto. Estas informações podem ser morfológicas, sintáticas, semânticas, entre outras, sobre os componentes do texto. Um etiquetador morfológico insere, junto a cada palavra de um texto, uma etiqueta que revela sua categoria morfológica. Neste trabalho, são consideradas as seguintes categorias morfológicas: artigo definido e indefinido, adjetivo, advérbio, conjunção coordenativa e subordinativa, interjeição, numerais cardinal e ordinal, pronomes demonstrativo, indefinido, relativo, pessoal e possessivo, preposição, substantivo, verbo, verbo auxiliar e particípio. São também etiquetadas as pontuações e, separadamente delas, a vírgula.

Utilizamos, na realização deste trabalho, bases de conhecimento, formadas respectivamente por palavras conhecidas (base *PALAVRAS*), sufixos (base *SUFIXOS*) e regras sintáticas (base *REGRAS*), que apoiam um processo de pré-etiquetagem de textos na língua portuguesa.

Note-se que a um pré-etiquetador, na concepção adotada, é permitida a ambigüidade de etiquetas, ou seja, uma palavra pode ser indicada como pertencente a mais de uma categoria morfológica. As bases *PALAVRAS* e *SUFIXOS*, utilizadas na fase inicial da pré-etiquetagem, não são suficientes para eliminar esta ambigüidade, já que uma palavra pode pertencer a mais de uma categoria morfológica se não for considerado o contexto sintático onde a mesma se encaixa. Esta situação fica ampliada se visualizarmos, na palavra, apenas seu sufixo. Da mesma forma, a aplicação de regras sintáticas, consideradas em uma segunda fase da pré-etiquetagem, conforme a abordagem que seguimos, não tem também a pretensão de eliminar totalmente a ambigüidade das etiquetas morfológicas inseridas.

A contribuição destas duas fases iniciais é inserir no texto informações suficientes para a etiquetagem propriamente dita (que poderá ter um objetivo específico ou seguir regras mais rigorosas), levando em consideração, então, outras informações morfológicas e sintáticas. Além disto, embora o trabalho seja voltado à língua portuguesa, sua essência pode ser diretamente aplicada à língua espanhola.

Este artigo se organiza em 5 seções. A seção 2 informa sobre o elemento mórfico enfatizado nesta abordagem e lista as etiquetas e as categorias morfológicas consideradas; a seção 3 apresenta o algoritmo utilizado para a pré-etiquetagem; a seção 4 descreve as três bases de conhecimento utilizadas: *PALAVRAS*, *SUFIXOS* e *REGRAS*; a seção 5 apresenta a avaliação inicial de um protótipo do sistema; e a seção 6 tece algumas considerações sobre o trabalho realizado.

---

<sup>1</sup> Palavras simples ou compostas ou, ainda, pontuações.

## 2 Sufixos e etiquetas

Os constituintes mórficos, ou seja, os componentes de uma palavra são: radical, tema, vogal temática, afixos (sufixos e prefixos), desinências, vogal de ligação e consoante de ligação [4]. Interessam-nos, nesta abordagem, especialmente os sufixos. Com eles pretendemos obter as informações iniciais para definir a categoria morfológica das palavras analisadas. Assim como o radical permite reunir uma família de palavras em torno de um conceito comum, o sufixo, elemento que se acresce ao radical para formar nova palavra, pode dar pistas sobre a categoria morfológica da palavra formada [4,8].

Neste trabalho, as categorias morfológicas consideradas são marcadas através do conjunto de etiquetas apresentado na Tabela 1.

etiqueta	categoria morfológica
_AD	artigo definido
_AI	artigo indefinido
_AJ	adjetivo
_AP	verbo no particípio
_AV	advérbio
_CC	conjunção coordenativa
_CS	conjunção subordinativa
_IN	interjeição
_NC	numeral cardinal
_NO	numeral ordinal
_PD	pronome demonstrativo
_PI	pronome indefinido
_PL	pronome relativo
_PN	pontuação (exceto a vírgula)
_PP	pronome pessoal
_PR	preposição
_PS	pronome possessivo
_SU	substantivo
_VB	verbos (exceto particípio)
_VG	vírgula

Tabela 1. Etiquetas morfológicas

## 3 Procedimento para a pré-etiquetagem

O procedimento, adotado para a pré-etiquetagem, *tokeniza* o texto em itens lexicais, identificando palavras e pontuações. A partir daí, utiliza o seguinte algoritmo:

Para cada item lexical *I* identificado no texto:

*I* é pesquisado na base *PALAVRAS*

Se *I* é identificado então é etiquetado

Senão

*I* é pesquisado na base *SUFIXOS*

Se *I* é identificado então é etiquetado

Para cada item lexical com etiquetagem ambígua *Ea*:

*Ea* é pesquisada na base *REGRAS*

Se alguma etiqueta de *Ea* é identificada então

a regra sintática correspondente é aplicada

## 4 Bases de conhecimento

### 4.1 Palavras conhecidas

Na base *PALAVRAS* são armazenados os itens lexicais mais freqüentes encontrados na língua portuguesa, principalmente as palavras invariáveis, como as preposições, mas também algumas variáveis, como os pronomes e os artigos. Também aqui são armazenadas as pontuações. Os registros são armazenados em ordem alfabética e o formato de cada um, nesta base, é apresentado na Tabela 2. Esta base armazena 743 registros e um trecho dela é apresentado na Figura 1.

campo	tipo de dado	descrição
<i>Item</i>	literal	Item lexical
<i>Etiqueta</i>	literal	conjunto de etiquetas preferenciais <sup>2</sup>

Tabela 2. Configuração da base *PALAVRAS*

<i>Item</i>	<i>Etiqueta</i>
...	
?	_PN
...	
a	_AD_PD_PP_PR
abaixo	_AV
acaso	_AV_SU
...	

Figura 1. Trecho da base *PALAVRAS*

<sup>2</sup> Nas bases *PALAVRAS* e *SUFIXOS*, as etiquetas associadas, respectivamente, às palavras conhecidas e aos sufixos são aquelas preferenciais, ou seja, são descartadas as etiquetas com pequena probabilidade de ocorrência em Português.

No trecho apresentado na Figura 1, o item lexical “?” seria etiquetado com \_PN. A palavra “a” seria etiquetada com \_AD\_PD\_PP\_PR, já que ela pode assumir cada uma destas categorias, conforme é apresentado a seguir, através dos exemplos (1), (2), (3) e (4), onde ocorre respectivamente como artigo definido, pronome demonstrativo, pronome possessivo e preposição.

- (1) a conferência iniciou,
- (2) esta não é a que me referi,
- (3) eu a esperei,
- (4) foi dito a ele.

A palavra “abaixo” receberia a etiqueta \_AV (advérbio), como em (5), sendo desprezada a possibilidade da etiqueta \_IN (interjeição) pela pequena probabilidade de ocorrência, como em (6).

- (5) ele está abaixo do diretor
- (6) abaixo o diretor!.

E, finalmente, conforme o trecho da base *PALAVRAS* apresentado na Figura 1, a palavra “acaso” seria etiquetada com \_AV\_SU, atendendo as ocorrências de advérbio, como em (7), e de substantivo, como em (8).

- (7) se acaso você chegasse,
- (8) aquilo foi um acaso.

## 4.2 Sufixos

Na base *SUFIXOS* é armazenada a maioria dos sufixos encontrados na língua portuguesa. A palavra cujo sufixo não está presente nesta base é considerada desconhecida. Aqui também os registros são armazenados em ordem alfabética. O formato de cada registro na base *SUFIXOS* é apresentado na Tabela 3. Esta base armazena 521 registros e um trecho dela é apresentado na Figura 2.

campo	tipo de dado	descrição
<b>Sufixo</b>	literal	sufixo <sup>3</sup> invertido
<b>Etiqueta</b>	literal	conjunto de etiquetas preferenciais

Tabela 3. Configuração da base *SUFIXOS*

<sup>3</sup> Em alguns casos, o “sufixo” armazenado na base não constitui exatamente um sufixo, mas um conjunto formado por sufixo e por parte (ou totalidade) de outros elementos mórficos. É o caso de “anos” (ver Figura 2), por exemplo, para a palavra “anos” ou para a palavra “panos”. Entretanto, no caso de “americanos”, a terminação “anos” já não engloba o radical.

<b>Sufixo</b>	<b>Etiqueta</b>
...	
sai	_AJ_SU_VB
sahla	_SU_VB
sarie	_SU_AJ_VB
serod	_SU_AJ
sona	_SU_AJ
sue	_SU_AJ_PS
...	

Figura 2. Trecho da base *SUFIXOS*

No caso do trecho apresentado na Figura 2, poderíamos ter, como exemplo, as ocorrências que seguem:

sai (ias): feias (adjetivo ou substantivo), saías (substantivo ou verbo);

sahla (alhas): falhas (substantivo ou verbo);

sarie (eiras): cadeiras (substantivo), arruaceiras (adjetivo ou substantivo), peneiras (substantivo ou verbo);

serod (dores): contadores (adjetivo ou substantivo);

sona (anos): anos (substantivo), americanos (adjetivo ou substantivo); e

sue (eus): ateneus (substantivo ou adjetivo), seus (pronome possessivo)

Os sufixos são armazenados de forma invertida para facilitar a pesquisa à base. A base *SUFIXOS* é implementada como uma árvore binária transformada. Assim, o trecho apresentado na Figura 2 tem origem na árvore apresentada na Figura 3.

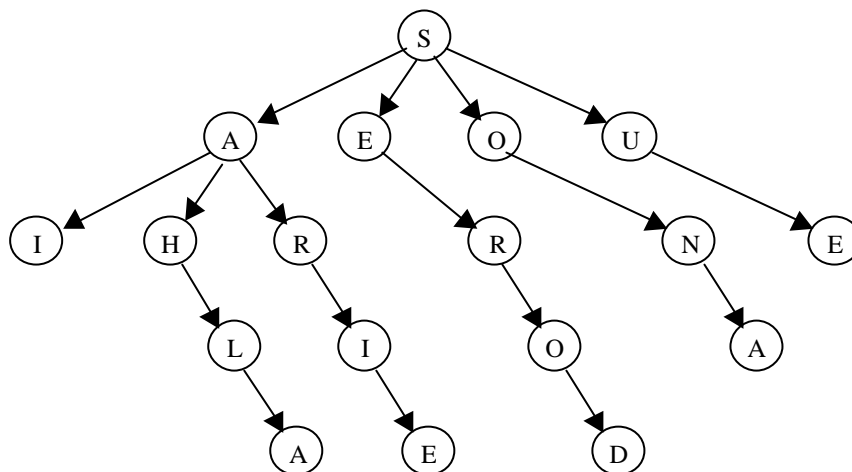


Figura 3. Trecho da árvore de sufixos original

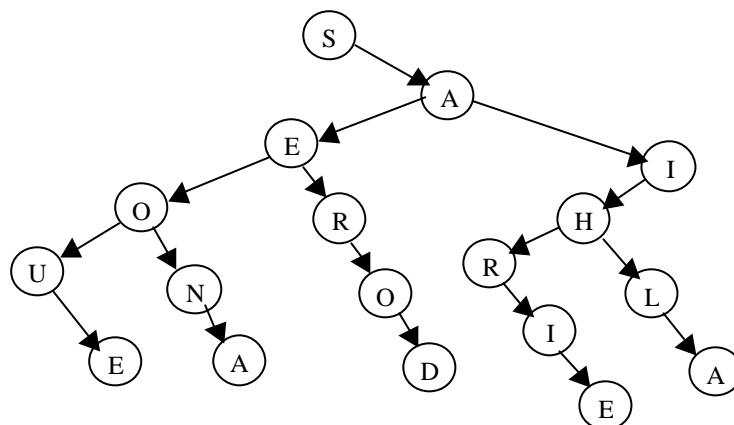


Figura 4. Trecho da árvore de sufixos transformada

Na Figura 4 é apresentada a árvore binária transformada a partir daquela que aparece na Figura 3. Na árvore binária transformada temos, no caso proposto, cada nó-filho à direita e cada nó-irmão à esquerda. Desta forma, de acordo com o trecho incluído na Figura 4, ao receber, por exemplo, a palavra “peneiras”, o sistema inicia a pesquisa pelo “s” e tenta encontrar, à direita, a letra “a”. Encontrando, continua a procurar à direita a letra “i”. Ao encontrar, agora tenta encontrar a letra “r”. Não tendo sucesso ao descer à direita, procura à esquerda até encontrar o “r”. Conseguindo encontrar, busca então novamente à direita e obtém a letra “i” e, posteriormente, nesta mesma direção, a letra “e”. Ao ser encontrado o sufixo (ou a terminação) da palavra pesquisada, a etiqueta (ou conjunto de etiquetas) correspondente é inserida. No caso de “peneiras” teríamos a etiquetagem com \_SU\_AJ\_VB, embora a etiqueta \_AJ pudesse ser descartada para esta palavra. Não o é em virtude da terminação “eiras”.

### 4.3 Regras sintáticas

Na base *REGRAS* são armazenadas regras sintáticas com a finalidade de apoiar a fase final deste processo de pré-etiquetagem. O uso destas regras se baseia no princípio que estabelece que as informações vinculadas a uma palavra são afetadas pelas palavras vizinhas a ela [7]. Assim, as palavras com etiquetagem ambígua são tratadas através de regras que tentam diminuir o número de etiquetas associadas. Esta base armazena 62 registros e um trecho dela é apresentado na Figura 5. O formato de cada registro desta base é apresentado na Tabela 4.

campo	tipo de dado	descrição
<i>EC</i>	literal	etiqueta corrente
<i>PC</i>	literal	palavra corrente
<i>EA</i>	literal	etiqueta anterior
<i>E2A</i>	literal	etiqueta que está duas posições antes da palavra/etiqueta corrente
<i>PA</i>	literal	palavra anterior
<i>P2A</i>	literal	palavra que está duas posições antes da palavra/etiqueta corrente
<i>EP</i>	literal	etiqueta posterior
<i>E2P</i>	literal	etiqueta que está duas posições depois da palavra/etiqueta corrente
<i>PP</i>	literal	palavra posterior
<i>P2P</i>	literal	palavra que está duas posições depois da palavra/etiqueta corrente
<i>ECNova</i>	literal	etiqueta corrente nova

Tabela 4. Configuração da base *REGRAS*

<i>EC</i>	<i>PC</i>	<i>EA</i>	<i>E2A</i>	<i>PA</i>	<i>P2A</i>	<i>EP</i>	<i>E2P</i>	<i>PP</i>	<i>P2P</i>	<i>ECNova</i>
...										
_AD								que		_PD
...										
_VB		_AI								_SU
...										

Figura 5. Trecho da base *REGRAS*

Naturalmente, nem todos os campos são preenchidos em todas as regras sintáticas. Exemplos disto são as duas regras apresentadas na Figura 5 que podem ser interpretadas da seguinte forma:

Se **EC** = "\_AD" e **PP** = "que" então **ECNova** = "\_PD"

e

Se **EC** = "\_VB" e **EA** = "\_AI" então **ECNova** = "\_SU".

Um exemplo de aplicação da primeira destas regras, tendo a palavra "a" como corrente, seria

... a AD PD PP PR que PL saiu VB...

Neste caso, teríamos a alteração da etiquetagem de "a" para \_PD:

... a PD que PL saiu VB...

Um exemplo de aplicação da segunda regra, tendo a palavra "canto" como corrente, seria

... um AI canto SU VB...

Neste caso, a alteração da etiqueta de "canto" seria para \_SU:

... o AI canto SU...



## 5 Avaliação

Um protótipo para o nosso sistema (o ETQ-SUF) foi construído e pode ser encontrado em <http://www.inf.pucrs.br/~gonzalez/etq>. A avaliação deste protótipo, teve como entrada um pequeno texto de 234 palavras e obteve os seguintes resultados:

- palavras desconhecidas = 5,98%
- palavras com etiquetagem ambígua = 59,83%
- palavras com etiquetagem não ambígua = 34,19%

Considerando apenas as palavras com etiquetagem não ambígua, temos:

- etiquetagem incorreta = 1,25%
- etiquetagem correta = 98,75%

Esta é apenas uma avaliação preliminar. Ainda são necessários ajustes e acréscimos na base *SUFIXOS* e, principalmente, na base *REGRAS*. Nesta última há ainda muito o que fazer, tanto em relação à inclusão de novas regras sintáticas quanto à revisão daquelas já inseridas. Assim que a base *REGRAS* estiver mais robusta, será elaborada uma avaliação mais rigorosa.

## 6 Considerações finais

Duas vantagens desta abordagem são encontradas na não exigência de um léxico (a base *PALAVRAS*, por sua cobertura não chega a ser considerada como tal) e por dispensar um corpus de treino<sup>4</sup>. Pode ser observada alguma semelhança do nosso sistema com o SMORPH [6], quando à utilização de terminações dos itens lexicais e, entre outros, com o PALAVRAS [3], quanto ao uso de regras para o tratamento das ambigüidades da etiquetagem. Entretanto, a comparação em termos de resultados do nosso sistema com outros etiquetadores fica prejudicada pela abordagem de pré-etiquetagem que adotamos.

A grande dificuldade de implementação deste sistema de pré-etiquetagem está exatamente na construção de duas de suas bases (*SUFIXOS* e *REGRAS*). Para se obter resultados aceitáveis na etiquetagem, a busca de regras sintáticas e sufixos se constitui em um trabalho de pesquisa exaustivo. Além deste aspecto de abrangência das bases, temos outros:

- (i) a ausência de alguma etiqueta, que deveria estar associada a um sufixo, compromete obviamente os resultados; e
- (ii) uma regra sintática deve ser verdadeira para todos os casos onde a configuração que tem como condição ocorre, e isto não é trivial quanto à pesquisa dessas ocorrências.

---

<sup>4</sup> Corpus de treino corresponde a um conjunto de textos utilizados com o objetivo de fazer com que o sistema reconheça determinados padrões, servindo como aprendizagem para a tarefa, no caso, de etiquetagem.

Em razão disto, a construção destas bases constitui fundamentalmente a contribuição que este trabalho pode dar ao processo de etiquetagem. A troca das mesmas (incluindo a base *PALAVRAS*) permite tratar outros idiomas além do Português, sem alteração do algoritmo do sistema.

## **Referências bibliográficas**

- [1] AIRES, Rachel Virgínia Xavier. Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil. Dissertação, Instituto de Ciências Matemáticas de São Carlos - USP, São Carlos, 2000.
- [2] ALVES, Carlos Daniel Chacur. Etiquetagem do Português Clássico Baseada em Corpus. Dissertação, Instituto de Matemática e Estatística da Universidade de São Paulo – USP, São Paulo, SP, 1999.
- [3] BICK, Eckhard. The Parsing System Palavras, Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.
- [4] CEGALLA, Domingos Paschoal. **Novíssima Gramática da Língua Portuguesa**. São Paulo, SP: Editora Nacional, 1998. 587p.
- [5] FINGER, Marcelo. Técnicas de Otimização da Precisão Empregadas no Etiquetador Tycho Brahe. V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR00), ICMC/USP, São Paulo, 2000.
- [6] HAGÈGE, Caroline. SMORPH: um analisador/gerador morfológico para o Português. Workshop sobre taggers para o português. Lisboa, Portugal: Instituto de Lingüística Teórica e Computacional, 1997.
- [7] JURAFSKY, D.; MARTIN, J. **Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. New Jersey, USA: Prentice Hall, 2000. 934 p.
- [8] SACCONI, Luiz Antonio. **Nossa Gramática – Teoria e Prática**. São Paulo, SP: Atual Editora. 1999. 576p.