# Directivas para identificação e classificação morfológica na colecção dourada do HAREM

*Nuno Cardoso, Diana Santos e Rui Vilela*
Última versão: 29 de Março de 2006

Neste documento, apresentamos as directivas usadas na etiquetagem da colecção dourada do HAREM e, consequentemente, qual o comportamento esperado pelos sistemas que nele participem.

Começamos por descrever o formato do que consideramos um texto anotado com entidades mencionadas (EMs), e qual a definição operacional da classificação morfológica destas. Depois indicamos quais os critérios usados na anotação morfológica da colecção dourada.

Noutro texto (Cardoso & Santos, 2006) será indicada a metodologia seguida na classificação semântica.

## 1 Regras Gerais

Cada EM é rotulada por uma etiqueta de abertura e uma etiqueta de fecho, semelhante às etiquetas usadas em XML. Na etiqueta de abertura, coloca-se a categoria atribuída, um tipo, e em alguns casos a classificação morfológica. Na etiqueta de fecho, coloca-se somente a categoria (que faz parte da classificação semântica). Veja-se um exemplo de uma EM etiquetada:

```
<PESSOA TIPO="GRUPO" MORF="M,P">Beatles</PESSOA>
```

Os nomes das categorias e dos tipos não devem incluir caracteres com acentos e/ou cedilhas, e devem estar em maiúsculas. Ou seja, deverá ser usado `<ORGANIZACAO>` em vez de `<ORGANIZAÇÃO>` ou `<Organizacao>`.

Os valores dos atributos `TIPO` e `MORF` devem ser rodeados por aspas, e o atributo `MORF`, se existir, deve seguir sempre o atributo `TIPO`.

Não deve haver nenhum espaço imediatamente a seguir à etiqueta de abertura e antes da etiqueta de fecho.

```
Certo  : O <PESSOA TIPO="INDIVIDUAL" MORF="M,S">João</PESSOA> é um professor.
Errado : O<PESSOA TIPO="INDIVIDUAL" MORF="M,S"> João</PESSOA> é um professor.
Errado : O <PESSOA TIPO="INDIVIDUAL" MORF="M,S">João </PESSOA>é um professor.
Errado : O <PESSOA TIPO="INDIVIDUAL" MORF=" M, S ">João</PESSOA>é um professor.
```

Se a EM contém espaços, esses devem manter-se inalterados.

```
Certo  : O <PESSOA TIPO="INDIVIDUAL" MORF="M,S">João Mendes</PESSOA> é um
professor.
Errado : O <PESSOA TIPO="INDIVIDUAL" MORF="M,S">JoãoMendes</PESSOA> é um
professor.
```

As aspas, parênteses, plicas ou travessões não são para incluir na etiqueta, se englobarem a EM como um todo (ver caso 1). No entanto, são para incluir, caso apenas se apliquem a partes da EM (caso 2) ou façam parte integrante da mesma .

Caso 1:

```
Certo  : A "<OBRA TIPO="ARTE" MORF="F,S">Mona Lisa</OBRA>"
Errado : A <OBRA TIPO="ARTE" MORF="F,S">"Mona Lisa"</OBRA>
```

Caso 2:

```
Certo  : O "<PESSOA TIPO="INDIVIDUAL" MORF="M,S">Mike "Iron" Tyson</PESSOA>
Certo  : <PESSOA TIPO="INDIVIDUAL" MORF="M,S">John (Jack) Reagan</PESSOA>
Certo  : Os resultados foram semelhantes aos produzidos por Diana Santos e
colegas <OBRA TIPO="PUBLICACAO">(Santos et al, 2005)</OBRA>.
```

## 1.1 Recursividade das etiquetas

Não se deve usar etiquetas dentro de etiquetas, como nos exemplos (errados) seguintes:

```
Errado : <PESSOA TIPO="GRUPO" MORF="M,P"><ORGANIZACAO TIPO="SUB"
MORF="M,P">Bombeiros</ORGANIZACAO></PESSOA>
Errado : <ORGANIZACAO TIPO="INSTITUICAO" MORF="M,S">Departamento de <ABSTRACCAO
TIPO="DISCIPLINA" MORF="F,S">Informatica</ABSTRACCAO> do IST</ORGANIZACAO>
```

## 1.2 Vagueza na classificação semântica

No caso de haver dúvidas entre várias categorias ou tipos, deve utilizar-se o operador '|'. Por exemplo, em A*judem os Bombeiros!*, se se considerar que não existe razão para preferir uma das duas seguintes classificações para *Bombeiros*, nomeadamente `<PESSOA TIPO="GRUPO">` e `<ORGANIZACAO TIPO="INSTITUICAO">`, devem-se colocar ambas:

```
Ajudem os <PESSOA|ORGANIZACAO TIPO="GRUPO|INSTITUICAO"
MORF="M,P">Bombeiros</PESSOA|ORGANIZACAO>!
```

Podem ser especificados mais do que uma categoria ou tipo, ou seja, `<A|B|C|...>`, ou ainda um maior número de |, são aceites.

Caso a dúvida seja entre tipos, deve-se repetir a categoria. Por exemplo, se se estiver em dúvida sobre qual o tipo de organização (`EMPRESA` ou `INSTITUICAO`?) na frase *O ISR trata dessa papelada*, deve-se repetir a categoria `ORGANIZACAO` tantas vezes quantos os tipos indicados:

```
O <ORGANIZACAO|ORGANIZACAO TIPO="EMPRESA|INSTITUICAO"
MORF="M,S">ISR</ORGANIZACAO|ORGANIZACAO> trata dessa papelada.
```

Seja como for, haverá apenas um atributo MORF.

## 1.3 Vagueza na identificação

Se houver dúvidas (ou análises alternativas) de qual a identificação da(s) EM(s) que deverá ser considerada correcta, as várias alternativas são marcadas entre as etiquetas `<ALT>` e `</ALT>`, que delimitam e juntam as várias alternativas, que são separadas pelo caracter '|'. O exemplo abaixo mostra a etiquetagem a usar, quando não se consegue decidir por uma única identificação:

```
O <ALT><PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Governo de Cavaco Silva</PESSOA> |
Governo de <PESSOA TIPO="INDIVIDUAL" MORF="M,S">Cavaco Silva</PESSOA></ALT>
```

Neste caso, cada alternativa, e cada EM, terá o seu atributo MORF.

## 1.4 Critérios de identificação de uma EM

Uma EM deve conter pelo menos uma letra em maiúsculas, e/ou algarismos.

```
Certo  : <TEMPO TIPO="DATA">Agosto</TEMPO>
Errado : <TEMPO TIPO="DATA">ontem de manhã</TEMPO>
```

A única excepção a esta regra abrange os nomes dos meses, que devem ser considerados EMs ou parte de EMs, mesmo se grafados com minúscula. Esta excepção deve-se ao facto de haver grafia maiúscula em Portugal e minúscula no Brasil.

```
Certo  : <TEMPO TIPO="DATA" MORF="M,S">agosto de 2001</TEMPO>
```

Existe também um conjunto de palavras relativas a certos domínios que também são excepções a esta regra, descritas em Cardoso & Santos (2006). Mais pormenores quanto a este assunto podem ser lidos no documento agora citado.

## 1.5 Relação entre a classificação e a identificação

Embora a classificação deva ter em conta o significado da EM no texto, a identificação (ou seja a sua delimitação) deve restringir-se às regras das maiúsculas enunciadas acima. Ou seja, apenas a parte associada ao nome próprio deve ser identificada, embora classificada, se for caso disso, a entidade maior em que se enquadra. Vejam-se os seguintes exemplos:

```
Certo  : a filha de <PESSOA TIPO="INDIVIDUAL" MORF="M,S">Giuteyte</PESSOA>
Certo  : o tratado de <ACONTECIMENTO TIPO="EVENTO"
MORF="M,S">Tordesilhas</ACONTECIMENTO>
```

Isso também se aplica aos casos em que no texto um fragmento ou parte da EM é compreendida como relatando anaforicamente a uma entidade não expressa na sua totalidade. Por exemplo, na frase *A Revolução de 1930 foi sangrenta, e a de 1932 ainda mais*, deve marcar-se *1932* como `<ACONTECIMENTO TIPO="EFEMERIDE" MORF="F,S">` e não como `<TEMPO TIPO="DATA">`.

Nos casos em que houve claramente um engano na grafia, escolhemos (e note-se que isto é uma excepção às regras enunciadas acima) corrigir mentalmente a grafia (maiúscula /minúscula) de forma a poder classificar correctamente, tanto a nível de identificação como a nível de classificação, semântica e morfológica. Além disso, estamos a pensar em marcar estes casos, na colecção dourada, com uma classificação `META="ERRO"`.

```
Certo  : O grupo terrorista <PESSOA TIPO="GRUPOMEMBRO" MORF="M,S"
META="erro">Setembro negro</PESSOA>...
```

Outras excepções, mais sistematicamente apresentadas, são as seguintes:

Para poder distinguir mais facilmente os casos de classes de objectos cujo nome inclui um nome próprio (geralmente de uma pessoa), adicionámos a seguinte regra de identificação para a categoria `COISA`: a preposição anterior também deve fazer parte da EM em *constante de Planck*, *bola de Berlim* ou *porcelana de Limoges*. Nesse caso, o atributo MORF refere-se à constante, à bola e à porcelana, devendo portanto ser morfologicamente classificada como "F,S".

Por outro lado, consideramos que as EMs de categoria `VALOR` e dos tipos `QUANTIDADE` ou `MOEDA` devem incluir a unidade, independentemente de esta ser grafada em maiúscula ou minúscula.

Finalmente, no caso de doenças, formas de tratamento e certo tipo de acontecimentos, consideramos aceitáveis um conjunto finito de nomes comuns precedendo a própria EM, cuja lista exaustiva se encontra mais uma vez em Cardoso & Santos (2006).

# 2 Classificação morfológica

Considerámos como passíveis de ser classificadas morfologicamente (isto é, EMs que devem ter o atributo MORF):

- As categorias PESSOA, ORGANIZACAO, COISA, ABSTRACCAO, ACONTECIMENTO, OBRA, e VARIADO na sua totalidade;
- Na categoria LOCAL, os tipos ADMINISTRATIVO e GEOGRAFICO;
- Na categoria TEMPO, o tipo CICLICO.

As seguintes EM não têm atributo MORF:

- A categoria VALOR na sua totalidade;
- Na categoria LOCAL, os tipos CORREIO;
- Na categoria TEMPO, o tipo HORA.

E finalmente, nos seguintes casos as EMs podem ou não ter o atributo MORF:

- Na categoria LOCAL, o tipo VIRTUAL;
- Na categoria TEMPO, os tipos DATA e PERIODO.

Uma série de exemplos de aplicação são apresentados posteriormente para clarificar em que situações ocorrem estas excepções.

**Género (morfológico)**

Consideramos que o género de uma EM pode ter três valores:
M – EM com género masculino;
F – EM com género feminino;
? - Para os casos em que o género é indefinido.

**Número**

Consideramos que o número de uma EM pode ter três valores:
S – EM no singular;
P – EM no plural;
? - Para os casos em que o número é indefinido.

## 2.1 Exemplos de não atribuição de MORF na categoria LOCAL

Em alguns casos particulares da sub-categoria VIRTUAL, o atributo MORF foi omitido, devido ao facto de não ser possível avaliar morfologicamente números de telefone.

```
Certo: <LOCAL TIPO="VIRTUAL">(48) 281 9595</LOCAL>
```

Os casos que possuam a etiqueta MORF são, pelo contrário, geralmente casos em que a entidade é de outro tipo básico, mas é empregue no contexto na acepção de LOCAL.

```
Certo: Como capturar da <LOCAL TIPO="VIRTUAL" MORF="F,S">Internet</LOCAL> os
endereços
Certo: uma ordem do governo local publicada na "<LOCAL TIPO="VIRTUAL"
MORF="F,S">Gazeta de Macau</LOCAL>" ordenava
Certo: E só depois da publicação no '<LOCAL TIPO="VIRTUAL" MORF="M,S">Diário da
República</LOCAL>' é que tomámos conhecimento do traçado.
```

## 2.2 Exemplos de não atribuição de MORF na categoria TEMPO

Nos tipos PERIODO e DATA há casos distintos em que são aplicados o atributo MORF.
As datas especificadas em termos de anos ou de dias não possuem nunca a etiqueta MORF.

```
Certo: Este ano de <TEMPO TIPO="PERIODO">1982</TEMPO> deve
Certo: <TEMPO TIPO="PERIODO">1914-1918</TEMPO>
Certo: ia ser a <TEMPO TIPO="DATA">17 de Dezembro</TEMPO> porque saiu
Certo: Em <TEMPO|TEMPO TIPO="DATA|PERIODO">91</TEMPO>, foram angariados
```

As classificações que possuem atributo MORF são meses, séculos, e períodos históricos.

```
Certo: Cinema para o mês de <TEMPO TIPO="PERIODO" MORF="M,S">Maio</TEMPO>
Certo: Mas já vem do <TEMPO TIPO="DATA" MORF="M,S">século XVI</TEMPO> o feriado
Certo: os povoadores cristãos da <TEMPO|ACONTECIMENTO TIPO="PERIODO|EFEMERIDE"
MORF="F,S">Reconquista</TEMPO|ACONTECIMENTO>.
Certo: Nesta <TEMPO TIPO="PERIODO" MORF="F,S">Primavera</TEMPO>, encontrei  me
com membros da
Certo: está agora previsto para <TEMPO TIPO="DATA" MORF="M,S">Outubro</TEMPO> ou
<TEMPO TIPO="DATA" MORF="M,S">Novembro</TEMPO>
```

# 3 Regras de atribuição de classificação morfológica

Considera-se o contexto e o texto adjacente para determinar o género e o número de uma dada EM, que à partida pode não ter género ou número definido.

Quando nem esse contexto nem o conhecimento lexical dos anotadores permite atribuir valores definidos, usa-se o valor ?, não especificado.

Exemplos:

```
Certo  : O <PESSOA TIPO="INDIVIDUAL" MORF="M,S">João</PESSOA> é um professor.
Certo  : A <PESSOA TIPO="INDIVIDUAL" MORF="F,S">João</PESSOA> não veio.
Certo  : O apelido <ABSTRACCAO TIPO="NOME" MORF="?,S">João</ABSTRACCAO> é muito
raro.
```

Ou seja, o nome *João* tem diferentes interpretações da sua classificação morfológica, consoante o contexto em que se encontra inserido.

## 3.1 Exemplos na categoria LOCAL

Algumas localidades administrativas são precedidas por artigo, determinando assim o género e número da entidade que designam (*o Porto*, *a Madeira*, *o Brasil*, *a Guarda*, *o Minho*, *o Rio Grande do Sul*, *os Estados Unidos*). Contudo, muitas outras não levam artigo e torna-se mais difícil de atribuir uma classificação morfológica.

Pareceu-nos em alguns casos haver consenso, tal como para *Portugal* (M,S), *Lisboa* (F,S), *Bragança* (F,S), *Brasília* (F,S), *Nova Iorque* (F,S) e *Colónia* (F,S), mas noutros casos apenas pudemos usar ? no género, tal como em *Chaves*, *São Paulo* (estado ou cidade), *Castelo Branco*, *Braga* ou *Madrid*, excepto quando tal é especificado no contexto.

```
Certo: <LOCAL TIPO=ADMINISTRATIVO MORF="F,S">Leiria</LOCAL> é linda.
Certo: do concelho de <LOCAL TIPO="ADMINISTRATIVO" MORF="?,S">Aregos</LOCAL>
Certo: todo o noroeste(de <LOCAL TIPO="ADMINISTRATIVO"
MORF="?,S">Resende</LOCAL> ao
Certo: em <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">Portugal</LOCAL> seria
Certo: aqui em <LOCAL TIPO="ADMINISTRATIVO" MORF="M,S">São Paulo</LOCAL>
Certo: em <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Nova Iorque</LOCAL> e saímos
Certo: polícia de <LOCAL TIPO="ADMINISTRATIVO" MORF="F,S">Colónia</LOCAL> foram
suspensos
```

## 3.2 Exemplos na categoria ORGANIZACAO

Geralmente o número e género de uma organização são definidos pelo número e género da primeira palavra do nome, *Charcutaria Brasil* (F,S), *Armazéns do Chiado* (M,P) *Banco X* (M,S) ou *Caixa Y* (F,S), enquanto empresas internacionais têm geralmente associado o género feminino: A *Coca-Cola*, a *Benetton*, a *IBM*, a *Microsoft*, a *Sun*, a *Lotus*, a *Ferrari*, etc.

```
Certo: junto do <ORGANIZACAO TIPO="EMPRESA" MORF="M,S">Banco Sotto
Mayor</ORGANIZACAO>
Certo: Uma acção da <ORGANIZACAO TIPO="EMPRESA" MORF="F,S">Cartier</ORGANIZACAO>
```

```
Certo: A acção da <ORGANIZACAO TIPO="EMPRESA" MORF="F,S">Portugal
Telecom</ORGANIZACAO> resultou
Certo: Esta página tem o apoio da <ORGANIZACAO TIPO="EMPRESA"
MORF="F,S">IP</ORGANIZACAO>
```

### 3.3 Exemplos na categoria PESSOA

No caso de GRUPOMEMBRO, ou seja, grupos de pessoas, o número é geralmente plural, e o género depende do sexo dos membros. As *Doce*, os *ABBA*, os *Xutos e Pontapés*, os *Beatles*, as *Spice Girls*, os *GNR*...

```
Certo: os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Stones</PESSOA>
Certo: e antes dos <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">R.E.M.</PESSOA>
Certo:<PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Peruanos</PESSOA> com diamantes
falsos
Certo: depois os <PESSOA TIPO="GRUPOMEMBRO" MORF="M,P">Mouros</PESSOA> que lhe
deram o nome
Certo: dez minutos o <PESSOA TIPO="GRUPOMEMBRO" MORF="M,S">Bastia</PESSOA>
assegurou a presença na final
```

### 3.4 Exemplos na categoria ACONTECIMENTO

No caso de EVENTO, os acontecimentos desportivos que tenham duas equipas, o número é singular, e o género é masculino, visto que correspondem a um jogo.

```
Certo: seguintes jogos: <ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Penafiel-Rio
Ave</ACONTECIMENTO>
Certo: e o <ACONTECIMENTO TIPO="EVENTO" MORF="M,S">Nacional-
Académica</ACONTECIMENTO>
```

### 3.5 Exemplos na categoria ABSTRACCAO

No caso do tipo DISCIPLINA, a maior parte das EMs que se refiram a disciplinas na área da educação tem género feminino, o número pode variar consoante o primeiro átomo.

```
Certo: e <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">Filosofia</ABSTRACCAO> em
todas as universidades
Certo: <ABSTRACCAO TIPO= DISCIPLINA  MORF= F,S >Ciência da
Informação</ABSTRACCAO>
Certo: futuros professores de <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,S">Educação
Física</ABSTRACCAO>
Certo: As <ABSTRACCAO TIPO="DISCIPLINA" MORF="F,P">TI</ABSTRACCAO> são uma
ferramenta
```

Já em relação a desportos, o género é em geral masculino, embora haja alguns que, por serem originários de palavras portuguesas femininas, mantêm o género, tal como *Vela* ou *Luta livre*.

```
Certo: Página do time de <ABSTRACCAO TIPO="DISCIPLINA"
MORF="M,S">Handebol</ABSTRACCAO>
```

## Referências

[Cardoso & Santos 2006] Nuno Cardoso & Diana Santos. “Directivas para identificação e classificação semântica na colecção dourada do HAREM”. Março de 2006.

# Índice