Apêndice A

Segundo HAREM: Directivas de anotação

Diana Santos, Paula Carvalho, Cláudia Freitas e Hugo Gonçalo Oliveira

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice A, p. 277–286.

Nota das editoras: Este apêndice reproduz a versão 4.1 das directivas do HAREM clássico, que foi actualizada pela última vez no dia 12 de Março de 2008. Incluímos também o elenco de categorias e a descrição da sintaxe, que se encontravam em páginas distintas na rede, nas secções A.4 e A.5, respectivamente, do presente documento. Incluímos igualmente na secção A.6 a lista de minúsculas disponibilizada e actualizada pela última vez no dia 7 de Abril de 2008.

Este texto descreve a tarefa objecto do Segundo HAREM, concentrando-se nas modificações em relação ao Primeiro, já bem documentado em Cardoso e Santos (2007).

Como seria de esperar, o Segundo HAREM vai ser mais abrangente que o anterior, não só ao corrigir e melhorar algumas arestas em relação ao Primeiro (muitas delas já discutidas no livro (Santos e Cardoso, 2007a)), mas por incluir duas novas tarefas/pistas, nomeadamente a normalização de expressões temporais (apêndice B) e a detecção de relações semânticas entre EM, o ReRelEM (apêndice C).

De forma a compatibilizar todas estas alterações num único formato, tornámos a sintaxe mais flexível, combinando numa mesma caracterização de saída a identificação de (i) apenas categorias, (ii) categorias e tipos, e (iii) categorias, tipos e subtipos, sendo todas estas classificações opcionais.

Todas as EM começam com <EM ID="xxx"> e acabam com </EM>. O único atributo obrigatório é o ID; que, para facilidade de processamento, restringimos a uma combinação de apenas letras não acentuadas (maiúsculas ou minúsculas), algarismos, e os caracteres "-" e "_". Veja a secção A.5 para mais pormenores.

Note-se também que, visto que CATEGs, TIPOs e SUBTIPOs são opcionais, passa a haver uma maior clarificação no significado de OUTRO, que não significará ignorância, visto que esta será marcada pela **falta** de valor desse atributo. OUTRO indica assim explicitamente uma classificação distinta do elenco sugerido (seja a nível das CATEGorias, dos TIPOs ou dos SUBTIPOs).

Da mesma forma, considera-se opcional a identificação da relação entre duas EM (COREL) e o tipo de relação (TIPOREL), assim como os vários atributos associados a uma análise mais fina de expressões temporais.

A.1 Motivação para as presentes directivas

Embora a nova organização tenha naturalmente algumas opiniões divergentes em relação à anterior (em particular em relação à elegância de algumas distinções), tentámos, excepto nos casos mais problemáticos, manter aquilo que já tinha sido feito na edição anterior, para poupar trabalho aos antigos participantes e garantir alguma continuidade.

Descrevemos, em seguida, as modificações que as diferentes categorias sofreram. Excepto em relação aos subtipos, todas essas modificações se encontram reflectidas na nova versão das colecções douradas do Primeiro HAREM.

A.2 Questões de delimitação

Mudámos ligeiramente a definição operacional de EM, de três formas.

A.2.1 Desaparecimento de entidades complexas

No Primeiro HAREM, tínhamos algumas categorias que poderiam ser designadas como entidades complexas, ou semi-estruturadas, cuja identificação – embora extremamente relevante num contexto de extracção de informação – era difícil de conceber como REM, como era o caso de

- moradas (anterior `CATEGORIA LOCAL` e `TIPO CORREIO`)
- referências bibliográficas (anterior `CATEGORIA OBRA` e `TIPO PUBLICACAO`)

que pensamos agora fazer mais sentido analisar em termos dos mais pequenos constituintes, aliás em termos semelhantes ao que já tinha sido feito para outro tipo de "entidades complexas", como, por exemplo

- informações sobre direitos de autor (copyright notices) em páginas Web

A.2.2 Tratamento mais convencional de expressões com várias palavras

Além disso, e visto que a sugestão de ter EM iniciadas por "de" não foi considerada satisfatória, passámos a considerar que

- algumas das expressões que haviam sido classificadas como EM não o eram de facto [Por exemplo, "de Belém" em *pastéis de Belém*.]
- outras deveriam continuar a ser identificadas como constituintes de uma EM maior, a qual compreenderia todos os termos (eventualmente grafados com inicial minúscula) que designam a classe ou o objecto que essa EM representa. [Ex: "gaiola de Faraday" e não apenas "de Faraday"]

Resumindo, o critério formal da obrigatoriedade de maiúscula na identificação de EM mantém-se (ou seja, "médio oriente" não é considerado EM), excepto para o `TEMPO`, em que as regras são diferentes.

Contudo, quando outras expressões que fazem claramente parte da EM se encontram grafadas em minúsculas devem ser igualmente identificadas, pois a incorrecta identificação das EM põe em causa a sua própria classificação.

- correcto: [ministro da Administração Interna] — `PESSOA/CARGO`
- incorrecto: ministro da [Administração Interna] — `DISCIPLINA/ABSTRACCAO`
- correcto: [relógio de Sol] — `COISA/CLASSE`
- incorrecto: relógios de [Sol] — `FISICO/PLANETA`

Note-se contudo que isto é apenas válido para casos em que se pode defender que estamos em presença de uma expressão com várias palavras, e não é para ser generalizado à detecção de sintagmas nominais. Assim, em *a casa do João* apenas *João* como pessoa deve ser marcado.

A.2.3 Introdução de intervalos de valores como EM

Também inspirados pela proposta de reclassificação das entidades temporais, decidimos que intervalos de valores, assim como a especificação mais fina desses valores passava a fazer parte integrante da EM de VALOR.

Por exemplo, veja-se a frase

> Ele saltou <EM ID="" CATEG="VALOR" TIPO="QUANTIDADE" SUBTIPO="n">**entre 7 a 10 metros**</EM>a sua fuga.

em que uma EM substitui as duas que seriam esperadas no Primeiro HAREM.

A.3 Mudanças por categoria

Passamos agora a fazer um apanhado das mudanças nas categorias.

A.3.1 VALOR

Mantemos a classificação anterior, com os tipos CLASSIFICACAO, MOEDA e QUANTIDADE.

A única diferença é que intervalos de valores, como *entre 3 e 4%* ou *de 5 a 10 kg*, passam a ser uma única EM, assim como as EM também incluem as preposições ou quantificadores relacionados com outras formas de descrever uma quantidade, tal como *cerca de 200 gramas*, *menos de 10%* ou *aproximadamente 15 euros*.

A.3.2 VARIADO

Deixa de haver a categoria VARIADO, passando a haver também a categoria OUTRO, com a mesma interpretação do OUTRO nos tipos ou subtipos.

A.3.3 PESSOA

Foi adicionado mais um tipo, o de POVO, para cobrir casos como *Não há música como a do Brasil*, *A House Music conquistou Inglaterra, Holanda, Alemanha e Ibiza* ou *Lisboa ficou horrorizada com essa notícia*.

Além disso, não sofreu modificações, excepto na lista de formas de tratamento, que foi actualizada, entre outras coisas ao tentar incluir-se mais sistematicamente as usadas no Brasil.

A.3.4 ORGANIZACAO

Passou-se SUB para SUBTIPO do tipo de organização em questão, ou seja, passam a ser possíveis os casos

- CATEG="ORGANIZACAO" TIPO="INSTITUICAO" SUBTIPO="SUB"
- CATEG="ORGANIZACAO" TIPO="EMPRESA" SUBTIPO="SUB"
- CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO" SUBTIPO="SUB"

O subtipo SUB não será contudo alvo de análise, anotação ou comparação no Segundo HAREM, mas foi mantido nas CD do Primeiro HAREM por uma questão de consistência.

A.3.5 LOCAL

Como indicado acima, deixámos de considerar o tipo CORREIO como uma EM, preferindo a marcação separada de ruas, estados e países dentro de moradas.

Além disso a informação marcada como LOCAL ALARGADO no Primeiro HAREM passou a ser considerada como informação adicional em relação aos tipos ADMINISTRATIVO ou GEOGRAFICO (agora rebaptizados de HUMANO ou FISICO). Assim, as EM anteriormente marcadas como LOCAL tipo ALARGADO passam a ter um SUBTIPO.

Passa pois a existir apenas uma tripartição da categoria LOCAL em FISICO, HUMANO e VIRTUAL, em que FISICO substitui o anterior termo GEOGRAFICO, e HUMANO o anterior termo ADMINISTRATIVO.

Além da categoria TEMPO, esta foi a única categoria onde os participantes desejaram uma classificação mais fina de subtipos.

Esta lista é o resultado da discussão pelos participantes envolvidos (mencionados na secção dos agradecimentos), a qual não pretende de forma alguma ser uma descrição exaustiva de todos os tipos conceptuais de lugares em português, mas apenas a soma das várias sensibilidades, experiências e opiniões da organização e dos já mencionados participantes.

Para locais de tipo HUMANO (note-se que os nomes são indicativos, não exaustivos)

- PAIS: inclui países, principados, e uniões de países, como é, por exemplo, o caso da União Europeia
- DIVISAO: inclui agregados populacionais como metrópoles, cidades, aldeias, vilas ou freguesias, assim como outras divisões administrativas tais como estados (Brasil), concelhos, distritos, províncias (Portugal), continentes, ou bairros fiscais
- REGIAO: localização cultural ou tradicional, sem valor administrativo, tal como a Baixa, o Grande Porto, o Médio-Oriente, o Terceiro Mundo ou o Nordeste (brasileiro)
- CONSTRUCAO: inclui todo o tipo de construções, desde edifícios, aglomerados de edifícios ou zonas específicas de um edifício (por exemplo, sala, galeria, jardim ou piscina), a pontes, barragens, portos, etc.
- RUA: inclui todo o tipo de arruamentos, como como ruas, avenidas, estradas, travessas, praças, pracetas, becos, largos, etc.
- OUTRO

Para locais de tipo FISICO

- AGUACURSO: inclui rios, ribeiros, riachos, afluentes, quedas de água, etc.
- AGUAMASSA: inclui lagos, mares, oceanos, golfos, estreitos, canais, bacias, barragens, etc.
- RELEVO: inclui montanhas, cordilheiras, montes, serras, planícies, planaltos, vales, etc.
- PLANETA: inclui todos os corpos celestes
- ILHA: inclui ilhas e arquipélagos

- REGIAO: designa uma região geográfica/natural, tal como o Bósforo, ou os Bálcãs, a Meseta Ibérica, a região do Amazonas, o Deserto do Sahara, ou os continentes vistos como região da geografia física
- OUTRO

Quanto a locais de tipo VIRTUAL, que indica localização abstracta, propomos os seguintes SUBTIPOS

- COMSOCIAL: inclui todos os meios de comunicação social, como jornais, televisão, rádio
- SITIO: inclui todos os locais virtuais no sentido electrónico: Web, WAP, ftp etc.
- OBRA: referência a uma obra impressa
- OUTRO

ilustrados respectivamente pelos seguintes exemplos:

- Essa afirmação saiu no *Diário de Notícias* ontem.
- Vai ao *Público on-line* ou à *Linguateca* e vê os anúncios que lá estão.
- No último *Harry Potter* vem a explicação da morte do Dumbledore.

Note-se que se a comunicação social é explicitamente na internet, então é SITIO. De resto, se nada for dito sobre isso (ou seja, não estiver a indicação *online* ou outra forma de o indicar, por exemplo através de um URL), então considera-se COMSOC.

Note-se também que, além de deixar de considerar URL como EM, também deixámos fora números de telefone e de fax.

A.3.6 ACONTECIMENTO

Não sofreu alteração, ou seja, mantém inalterados os tipos EFEMERIDE, EVENTO e ORGANIZADO.

A.3.7 OBRA

Em relação às EM da categoria OBRA, reduzimos os tipos de OBRA aos seguintes:

- REPRODUZIDA, da qual há muitas cópias/exemplares
- ARTE, que significa peça única
- PLANO, que se distingue das outras OBRAs pelo seu carácter contingente e circunstancial (note-se que estava anteriormente em ABSTRACCAO)

Na mesma linha que retirámos a categoria de CORREIO dos LOCAL, deixámos de entrar em conta com PUBLICACAO, que deixa de ser considerada uma EM de todo.

A.3.8 ABSTRACCAO

Esta categoria foi consideravelmente simplificada, retendo apenas os tipos DISCIPLINA, ESTADO, IDEIA e NOME. Por um lado, foram retirados desta categoria os tipos MARCA (convertido para categoria COISA de tipo CLASSE ou IDEIA) e PLANO (passado para categoria OBRA tipo PLANO). Por outro lado, DISCIPLINA, ESCOLA e OBRA foram todas juntas em DISCIPLINA. Muito resumidamente, então:

- DISCIPLINA: passou a referir quer uma disciplina ou área, quer uma escola literária, científica, artística, religiosa ou ideológica, ou mesmo um estilo musical.
- ESTADO representa, como anteriormente, sobretudo doenças.
- IDEIA é a mais abstracta das abstracções
- NOME, como anteriormente, representa um objecto linguístico e não a entidade que designa.

A.3.9 COISA

Esta categoria foi a que sofreu mais alterações, e por isso decidimos reescrever completamente as directivas no que lhe diz respeito.

Em primeiro lugar e como já mencionado, mudámos os critérios de identificação, de forma a que casos em que apenas por questões de convenção se grafam com letra maiúscula deixem de ser abrangidas pela noção de EM: ou seja, *pastéis de Belém*, *flauta de Bisel*, visto que estão em variação livre com *pastéis de feijão* ou *guitarra acústica*, e apenas se grafam em maiúscula por os seus nomes derivarem de locais ou pessoas.

Noutros casos sobretudo de terminologia científica, mantivemos a classificação de COISA tipo CLASSE mas identificando o conceito todo, ou seja, as EM passam a ser *constante de Planck* e *aparelho de Golgi*.

Basicamente a principal questão associada à categoria COISA é que debaixo desta designação estão "coisas" ontologicamente muito diferentes mas que a linguagem natural e em particular o português não distingue formalmente, como classe/membro, classe/subclasse e exemplo/classe. Para tentar produzir sobretudo critérios mais claros de anotação, sem querer forçar distinções que não estão lá (ou que os anotadores humanos têm dificuldade em fazer), redefinimos o seguinte elenco de tipos de COISA objecto de classificação no Segundo HAREM:

- OBJECTO que tem um nome individualizado, e que inclui desde animais (vivos, individuais) a planetas, passando por meios de locomoção tal como barcos ou foguetões e ursos de peluche.
- SUBSTANCIA nome de uma substância que, por ser massiva, não permite em geral distinções entre individuos ou espécies. É no entanto concreta e por isso não pode ser classificada como abstracção
- CLASSE passa pois a ter apenas as classes que são designadas por nomes próprios, tais como marcas ou modelos, assim como raças de animais ou programas de computador

- MEMBROCLASSE designa elementos que não têm nome individual mas que são designados pelo nome da classe a que pertencem, tal como *Ford* ou *iPod* em *o meu Ford* ou *o Ipod dela*, ou mesmo *Coca Cola* em *Quem me roubou a minha Coca Cola?* ou *Fox Terrier* em *Viste o meu Fox Terrier?*

A.4 Elenco de categorias do Segundo HAREM

Na tabela A.1 encontra-se o elenco de categorias, tipo e subtipos do HAREM clássico. Entre parênteses encontra-se: i) à frente das categorias o número de tipos; ii) à frente dos tipos, o número de subtipos.

A.5 Segundo HAREM: sintaxe

Uma EM é identificada pela etiqueta <EM> com atributos e terminada por </EM>.

Por exemplo,

<EM ID="xxx" CATEG="A" TIPO="B" SUBTIPO="C" COREL="corel" TIPOREL="tiporel">**Qualquer Coisa**</EM>

Os atributos possíveis

- têm de aparecer em maiúsculas;
- só podem ser ID, CATEG, TIPO, SUBTIPO, COREL, TIPOREL, TEMPO_REF, SENTIDO, VAL_NORM, VAL_DELTA, COMENT;
- o seu valor tem de ser incluído entre aspas, a seguir ao sinal de igual.

O único atributo obrigatório é o ID, que tem de ser uma combinação de apenas letras não acentuadas (maiúsculas ou minúsculas), algarismos, e os caracteres "-" e "_". A cada EM corresponde um ID único.

Os valores dos atributos COREL e TIPOREL estão descritos nas directivas do ReRelEM (Santos et al., 2008b).

Os valores dos atributos TEMPO_REF, SENTIDO, VAL_NORM e VAL_DELTA estão descritos nas directivas do TEMPO (Hagège et al., 2008).

Se várias possibilidades de identificar uma expressão correspondem a segmentações diferentes, usa-se <ALT>, separando as várias alternativas pelo caracter |.

Por exemplo, <ALT> alt1 | alt2 | alt3 </ALT>, em que alt1, alt2, alt3 são texto eventualmente marcado com <EM>. Para cada alternativa alt1, alt2, alt3 deve corresponder um ID diferente.

Para a tarefa de classificação, para todas as EM excepto as do TEMPO, uma EM no máximo terá a forma

<EM ID="xxx" CATEG="A" TIPO="B" SUBTIPO="C">**Entidade**</EM>.

Os valores possíveis para CATEG, TIPO e SUBTIPO:

- podem ser omitidos

- o TIPO só pode ser especificado se a CATEG também o for, e tem de pertencer a essa categoria
- o SUBTIPO só pode ser especificado se o TIPO também o for, e tem de pertencer a esse tipo
- o SUBTIPO só está definido para os tipos FISICO, HUMANO, VIRTUAL da CATEG LOCAL e para os TIPOS TEMPO_CALEND da CATEG TEMPO
- podem ser simples (veja-se a tabela na secção A.4), ou complexos
- valores complexos (correspondendo a vagueza) criam-se através da concatenação de vários valores através do caracter |.
- se um dado valor é omitido, usa-se o vazio
- a ordem dos valores complexos tem de ser idêntica nos três atributos, ou seja a ordem dos tipos tem de ser igual à ordem das categorias a que correspondem
- é necessário repetir a categoria se se quiser especificar alternativas entre tipos dessa mesma categoria
- é necessário repetir o tipo se se quiser especificar alternativas entre subtipos desse mesmo tipo

É possível incluir o que se quiser dentro do atributo COMENT, excepto caracteres especiais do XML como & < > ou aspas.

A.6 Lista de minúsculas

Nesta página, encontra-se a lista de palavras ou expressões em minúsculas que devem fazer parte da EM, no âmbito do Segundo HAREM. Relembramos que as regras para a identificação e classificação das EMs temporais se encontram separadamente definidas em Hagège et al. (2008).

Esta lista não é exaustiva, nem pretende descrever a língua portuguesa, tendo sido criada apenas com o objectivo de fornecer a todos os participantes no HAREM os mesmos critérios de identificação, neste caso, por extenso.

Note-se que as (sequências de) palavras listadas em seguida apenas devem ser tidas em consideração se surgirem imediatamente acompanhadas por uma outra palavra ou expressão iniciadas por maiúscula, as quais podem ser eventualmente antecedidas da preposição *de* e respectivas contracções.

Por exemplo:

<EM ID="Ex1" CATEG="PESSOA" TIPO="INDIVIDUAL">**presidente Lula**</EM>

presidente italiano <EM ID="Ex2" CATEG="PESSOA" TIPO="INDIVIDUAL">**Romano Prodi**</EM>

<EM ID="Ex3" CATEG="PESSOA" TIPO="CARGO">**duque de Bragança**</EM>

Os elementos das listas encontram-se organizados alfabeticamente por CATEG/TIPO (isto é, podem surgir no âmbito de uma EM classificada com os atributos a seguir especificados):
PESSOA/CARGO ou PESSOA/GRUPOCARGO ou PESSOA/INDIVIDUAL ou PESSOA/GRUPOIND

alta-comissária, altas-comissárias, alto-comissário, altos-comissários; bispo, bispos; chanceler, chanceleres; chefe, chefes; condessa, condessas, conde, condes; cônsul, cônsules, consulesa, consulesas; czar, czares, czarina, czarinas; dire(c)tor, dire(c)tora, dire(c)toras, dire(c)tores; dire(c)tor-geral, dire(c)tora-geral, dire(c)toras-gerais, dire(c)tores-gerais; duque, duques, duquesa, duquesas; embaixador, embaixadores, embaixatriz, embaixatrizes; infanta, infantas, infante, infantes; governador, governadora, governadoras, governadores; líder, líderes; ministra, ministras, ministro, ministros; padre, padres; patrão, patroa, patroas, patrões; porta-voz, porta-vozes; presidente, presidentes; primeira-ministra, primeiras-ministras, primeiro-ministro, primeiros-ministros; princesa, princesas, príncipe, príncipes; rabi(no), rabi(no)s; rainha, rainhas, rei, reis; reitor, reitora, reitoras, reitores; secretária de Estado, secretárias de Estado, secretário de Estado, secretários de Estado; secretária-geral, secretárias-gerais, secretário-geral, secretários-gerais; sultão, sultãos, sultões; visconde, viscondes, viscondessa, viscondessas

As palavras listadas podem ser, eventualmente, precedidas de "ex-", "vice", "co" ou "sub".

PESSOA/INDIVIDUAL ou PESSOA GRUPOIND

arquite(c)ta, arquite(c)tas, arquite(c)to, arquite(c)tos, avó, avós, avô, avôs; bispo, bispos; dom, dona, donas; doutor, doutora, doutoras, doutores; engenheira, engenheiras, engenheiro, engenheiros; irmã, irmão, irmãos, irmãs; madre, madres; mestre, mestres; padre, padres; professor, professora, professoras, professores; rabi(no), rabi(no)s; senhor, senhora, senhoras, senhores; seu; sir, sô; tia, tias, tio, tios; vovó, vovós, vovô, vovôs.

Todas as combinações de *senhor* seguido de cargo ou de título, tal como *senhor ministro* ou *senhor padre*, são também aceites. Estes casos podem ser antecedidos por *excelentíssimo* (ou respectiva abreviatura).

Aceitam-se igualmente as abreviaturas convencionalmente associadas aos elementos destas listas, sempre que estas existam.

ABSTRACCAO/ESTADO

doença; mal; sindroma; síndroma; síndrome

COISA/SUBSTANCIA

vitamina, vitaminas.

As palavras compreendidas nas listas são consideradas da mesma forma pela avaliação do Segundo HAREM quer estejam em maiúsculas ou minúsculas.

Agradecimentos

Agradecemos ao Marcirio Chaves, Nuno Cardoso, Caroline Hagège, Nuno Mamede, Bruno Martins e Mário Silva os comentários, sugestões e dúvidas formulados na discussão dos subtipos de LOCAL, cujo resultado final é, contudo, da nossa responsabilidade. Agradecemos também ao Nuno Cardoso e à Cristina Mota a correcção de muitos problemas em versões anteriores.

Tabela A.1: Elenco de categorias do Segundo HAREM

Categorias	Tipos	Subtipos
ABSTRACCAO (5)	DISCIPLINA	
	ESTADO	
	IDEIA	
	NOME	
	OUTRO	
ACONTECIMENTO (4)	EFEMERIDE	
	EVENTO	
	ORGANIZADO	
	OUTRO	
COISA (5)	CLASSE	
	MEMBROCLASSE	
	OBJECTO	
	SUBSTANCIA	
	OUTRO	
LOCAL (4)	FISICO (7)	ILHA, AGUACURSO, PLANETA, REGIAO, RELEVO, AGUAMASSA, OUTRO
	HUMANO (6)	RUA, PAIS, DIVISAO, REGIAO, CONSTRUCAO, OUTRO
	VIRTUAL (4)	COMSOCIAL, SITIO, OBRA, OUTRO
	OUTRO	
OBRA (4)	ARTE	
	PLANO	
	REPRODUZIDA	
	OUTRO	
ORGANIZACAO (4)	ADMINISTRACAO	
	EMPRESA	
	INSTITUICAO	
	OUTRO	
PESSOA (8)	CARGO	
	GRUPOCARGO	
	GRUPOIND	
	GRUPOMEMBRO	
	INDIVIDUAL	
	MEMBRO	
	POVO	
	OUTRO	
TEMPO (5)	DURACAO	
	FREQUENCIA	
	GENERICO	
	TEMPO_CALEND (4)	HORA, INTERVALO, DATA, OUTRO
OUTRO		
VALOR (4)	CLASSIFICACAO	
	MOEDA	
	QUANTIDADE	
	OUTRO	
OUTRO (1)		