O passar do TEMPO no HAREM

Cristina Mota
Linguateca (FCCN)
cmota@ist.utl.pt

Paula Carvalho
XLDB (FCUL)
pcc@di.fc.ul.pt

Resumo

O presente artigo apresenta um estudo contrastivo entre as propostas de análise do tempo na primeira e segunda edições do HAREM. Discutimos, entre outros aspectos, as principais vantagens e inconvenientes de uma e outra, tendo em linha de conta os princípios teórico-metodológicos subjacentes ao modelo geral do HAREM. A discussão é feita com base nas expressões temporais compreendidas nas colecções douradas do Primeiro e do Segundo HAREM, que reanalisámos de acordo com, respectivamente, as directivas do TEMPO adoptadas no Segundo e no Primeiro HAREM. Em forma de um balanço final, apresentamos algumas sugestões de melhoria no tratamento do tempo num próximo HAREM.

1 Introdução

A extracção automática de expressões temporais consiste na identificação e classificação de expressões que ajudam a localizar temporalmente os eventos descritos num texto. A primeira avaliação conjunta que se dedicou a avaliar sistemas que reconhecem este tipo de expressões foi a MUC (Message Understanding Conference) para textos em inglês. Inicialmente, o reconhecimento de expressões temporais estava integrado no reconhecimento de eventos (uma das informações que os sistemas precisavam de fornecer sobre o evento era o momento da ocorrência), e mais tarde, a partir da MUC-6 (Grishman e Sundheim, 1996), a sexta edição dessa avaliação, passou a ser uma sub-tarefa da nova tarefa de reconhecimento de entidades mencionadas. Essa opção pode ser discutível (veja-se, por exemplo, Hagège, Baptista e Mamede (2010), que argumentam a favor da separação das duas tarefas), mas dado que (i) várias dessas expressões são constituídas por maiúsculas e (ii) a intersecção do conjunto destas expressões e das entidades mencionadas não é vazio, ter as tarefas em conjunto não é completamente indefensável, uma vez que para poder reconhecer umas é preciso saber excluir as outras.

Assim, para a língua portuguesa, a Linguateca, tendo como uma das linhas directoras a promoção de avaliações conjuntas em diversas áreas do processamento computacional do português (Santos e Rocha, 2003), organizou o HAREM.

O HAREM é, então, uma avaliação conjunta de sistemas de reconhecimento de entidades mencionadas em português. A primeira iniciativa desta avalição, o Primeiro HAREM, consistiu em dois eventos: um que decorreu entre 2004 e 2005, e outro, designado Mini-HAREM, que decorreu no ano de 2006 e que teve por objectivo medir o progresso dos sistemas participantes no evento anterior[1]. A segunda edição desta avaliação, o Segundo HAREM, teve início em Setembro de 2007, culminando com o Encontro do Segundo HAREM, um ano depois.

O Segundo HAREM, no entanto, foi uma avaliação mais abrangente do que a anterior, que não só corrigiu e aperfeiçoou algumas arestas em relação ao Primeiro, como incluiu duas novas pistas, concretamente o reconhecimento e normalização de expressões temporais (Hagège, Baptista e Mamede, 2008b; Hagège, Baptista e Mamede, 2008a) e a detecção de relações semânticas entre entidades mencionadas (EM), o ReRelEM (Freitas et al., 2008b; Freitas et al., 2008a). A avaliação do reconhecimento de entidades mencionadas excluindo as da categoria TEMPO, que têm no Segundo HAREM uma pista exclusivamente a elas dedicada, foi então designada HAREM clássico[2].

[1] Apesar de as directivas terem sofrido de um evento para o outro pequenas alterações em algumas categorias (consulte-se Santos e Cardoso (2007) para mais informação sobre os dois eventos), iremos tratar indistintamente os dois eventos do Primeiro HAREM, a não ser nos casos específicos em que importa distingui-los, uma vez que a categoria TEMPO, sobre a qual este artigo se debruça, não foi afectada por alterações entre os dois eventos do Primeiro HAREM.

[2] Na verdade, a avaliação feita aos sistemas no HAREM clássico incluiu também as entidades com a categoria TEMPO sem ter em conta os atributos de normalização, mas neste artigo quando nos referimos ao HAREM clássico é para designar apenas a proposta de avaliação das restan-

Se em relação ao HAREM clássico e ao ReRelEM foi mantida a filosofia subjacente ao Primeiro HAREM, nomeadamente o modelo semântico (Santos, 2007) e modelo geral de avaliação (Santos, Cardoso e Seco, 2006), o mesmo não aconteceu com a pista do TEMPO, cuja proposta assentou numa abordagem diferente da da organização do HAREM ao problema do reconhecimento de entidades temporais. Saliente-se, a este respeito, que (i) embora o HAREM se tenha inspirado inicialmente na MUC, estas duas avaliações seguiram abordagens diferentes, como é discutido em detalhe em Seco (2007), e (ii) a proposta de avaliação do TEMPO no Segundo HAREM foi inspirada na TimeML (consulte-se, por exemplo, Pustejovsky et al. (2003)), uma avaliação dedicada exclusivamente ao reconhecimento de entidades temporais, eventos e relações entre ambos.

Neste artigo, discutimos então a proposta de reconhecimento e normalização das expressões temporais no âmbito do Segundo HAREM, procurando colocá-la em confronto com a proposta de classificação destas expressões, implementada na primeira edição desta avaliação conjunta. Começaremos por dar uma panorâmica da avaliação do TEMPO nas duas edições do HAREM; depois, faremos uma análise crítica contrastiva entre as duas propostas, destacando primeiro os aspectos que consideramos positivos na nova proposta e, em seguida, quais as características importantes do Primeiro HAREM que se perderam. Ilustraremos igualmente o impacto da aplicação das novas directivas no reconhecimento das expressões temporais anotadas no Primeiro HAREM, bem como a situação inversa de aplicação das directivas do Primeiro HAREM à colecção dourada do Segundo HAREM. Em forma de um balanço final, sumarizamos, na secção 6, algumas sugestões de melhorias no tratamento do TEMPO num próximo HAREM.

2 *Panorâmica sobre o tempo no HAREM*

No Primeiro HAREM, a proposta de análise de entidades temporais encontrava-se integrada com a das restantes entidades e totalmente a cargo da Linguateca. Porém, no Segundo HAREM, a proposta de reconhecimento de entidades temporais, juntamente com a nova tarefa de normalização dessas expressões, constituiu uma pista independente, proposta por um dos grupos participantes (Hagège, Baptista e Mamede, 2008b). Os proponentes da nova proposta pretendiam completar, enriquecer e alargar a definição desta categoria tal como proposta no Primeiro HAREM (cf. (Cardoso e Santos, 2007)), a uma noção mais geral de expressão temporal (Hagège, Baptista e Mamede, 2008b). Convém, no entanto, salientar que a implementação (criação da colecção dourada e desenvolvimento dos programas de avaliação) foi levada a cabo pela Linguateca. Tal como discutido em Mota et al. (2008a), não ter sido um mesmo grupo responsável pela proposta e implementação da avaliação do TEMPO levantou algumas dificuldades, que não serão aqui discutidas.

Uma das principais diferenças entre a prosposta de tratamento do TEMPO no Primeiro e no Segundo HAREM diz respeito aos critérios de identificação utilizados no reconhecimento de EM, o que leva a que o próprio conceito de entidade possa diferir (e difere, efectivamente na maioria dos casos) nessas propostas.

Assim, de modo a não deixar de fora expressões como, por exemplo, *ontem*, *depois de amanhã*, *há muitos anos atrás*, no Segundo HAREM as entidades temporais não ficaram limitadas a ter de conter maiúsculas[3] ou números. Como consequência natural, o total de entidades anotadas como TEMPO no Segundo HAREM cresceu substancialmente. A Tabela 1[4] mostra que o número de entidades com essa categoria aumentou cerca de 50% (806 entidades temporais no Primeiro HAREM e 1200 no Segundo HAREM), embora o total de entidades no Segundo HAREM até tenha decrescido.

Além disso, a classificação também sofreu alterações. Como se pode observar na Tabela 2, as categorias DATA e HORA foram consideradas tipos da nova categoria TEMPO_CALEND, deixou de se considerar as categorias PERIODO e CICLICO, e passaram a existir as categorias DURACAO, FREQUENCIA e GENERICO (na secção seguinte iremos detalhar melhor as noções que estas classificações tentam cobrir).

Ilustramos, com os exemplos (1) e (2), o formato das anotações para a categoria TEMPO no Primeiro e Segundo HAREM, respectivamente. Estes exemplos servem igualmente para ilustrar que no Segundo HAREM, as expressões temporais são reconhecidas e classificadas essencialmente com base num conjunto de critérios linguísticos frequentemente utilizados para iden-

tes categorias.

[3]Notamos, no entanto, que, por exemplo, a anotação dos meses do ano em português do Brasil, que são escritos em minúsculas, estava prevista nas directivas do Primeiro HAREM

[4]Estes valores foram calculados a partir das colecções douradas do Primeiro HAREM com as entidades anotadas no formato do Segundo HAREM, as quais foram disponibilizadas como material de treino no Segundo HAREM.

	CDPH	(CDPH-pe; CDPH-mh)	CDSH
Total de entidades	8734	(5065; 3669)	7845
Total de entidades tempo	806	(441; 365)	1200
Percentagem de entidades tempo	9,23%	(8,71%; 9,95%)	15,30%

Tabela 1: Total de entidades classificadas como TEMPO na colecção dourada do primeiro evento do Primeiro HAREM (CDPH-pe), na CD do Mini-HAREM do Primeiro HAREM (CDPH-mh) e na CD do Segundo HAREM (CDSH); CDPH reune as duas colecções do Primeiro HAREM.

	TIPO	SUBTIPO	Exemplo
PH	DATA	-	*Nasci em Angola, a **2 de Junho de 1968**.*
	HORA	-	*Chegamos a Presidente Figueiredo às **19h30***
	PERIODO	-	*actividades ligadas à campanha eleitoral de **Setembro***
	CICLICO	-	*quando vinha o padre benzer na **Páscoa***
SH	TEMPO_CALEND	DATA	*rocha ornamental utilizada **desde antes de 7000 a.C.***
		HORA	*o impacto dos destroços teria ocorrido **às 18h06***
	DURACAO	-	*tocando um vinil **por mais de um minuto***
	FREQUENCIA	-	*vinha **três vezes por semana**, dar-me lição de alemão*
	GENERICO	-	*percorreu **todo o século vinte***

Tabela 2: Classificação do TEMPO no Primeiro HAREM (PH) e no Segundo HAREM (SH).

tificação de constituintes sintácticos, em particular, grupos nominais com valor temporal (eventualmente antecedidos de preposição), e, portanto, os limites das entidades temporais foram alargados para incluir a preposição e outros modificadores que com elas formam o complemento adverbial de tempo. Veja-se que, por exemplo, a anotação do Primeiro HAREM delimita **1880** enquanto a do Segundo HAREM delimita **A partir de 1880**.

(1) A partir de <TEMPO TIPO="DATA≫ **1880** </TEMPO> ensinou psicologia e filosofia em Harvard, universidade que abandonou em <TEMPO TIPO="DATA≫**1907**</TEMPO>, proferindo conferências nas universidades de Columbia e Oxford. Morreu em Chocorua, New Hampshire, a <TEMPO TIPO="DATA≫**26 de Agosto de 1910**</TEMPO>.

(2) <EM ID="12" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+1880----T----E-- LMP≫**A partir de 1880**</EM> ensinou psicologia e filosofia em Harvard, universidade que abandonou <EM ID="14" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+1907----T----E-- LM-≫**em 1907**</EM>, proferindo conferências nas universidades de Columbia e Oxford. Morreu em Chocorua, New Hampshire, <EM ID="19" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+19100826T----E-- LM-≫**a 26 de Agosto de 1910**</EM>.

3 O tempo no bom caminho

No Primeiro HAREM, um dos critérios de base para a identificação de uma dada palavra ou expressão como potencial entidade mencionada prendia-se com o uso de maiúsculas ou algarismos. Neste contexto, apesar de as expressões abaixo assinaladas serem semanticamente equivalentes, apenas a expressão ilustrada em (3) seria reconhecida, dado que a ilustrada em (4) não obedece a nenhum dos requisitos formais explicitados naquelas directivas.

(3) Essa lei entrou em vigor a **1 de Janeiro de 2008**

(4) Essa lei entrou em vigor no **primeiro dia do ano de dois mil e oito**

Este critério formal exclui, pois, à partida, um número significativo de expressões temporais que seria igualmente importante reconhecer no âmbito da extracção de informação temporal de um dado texto. Neste sentido, parece inquestionável que a adopção das novas directivas do TEMPO, que não impõem tal requisito, permite um reconhecimento mais alargado (e, na nossa opinião, completamente justificado) das expressões temporais. De acordo com as directivas do Segundo HAREM, as duas expressões exemplificadas em (5) e (6) seriam, portanto, adequadamente reconhecidas e classificadas de forma idêntica:

(5) a 1 de Janeiro de 2008 [TEMPO TEMPO_CALEND DATA]

(6) no primeiro dia do ano de dois mil e oito [TEMPO TEMPO_CALEND DATA]

Além da melhoria em termos de cobertura, as novas directivas possibilitaram também, em alguns casos, uma melhor definição das próprias expressões temporais. Referimo-nos, em particular, ao alargamento da extensão da noção de entidade aos seus eventuais quantificadores e modificadores. Por exemplo, de acordo com o Primeiro HAREM, apenas a data (*2009*) deveria ser reconhecida como EM na construção ilustrada em (7).

(7) O Tratado foi ratificado antes de **2009**.

Se procurássemos estabelecer/compreender a relação que existe entre esssa data e a entidade a que a mesma se refere[5], estaríamos a representar/inferir uma relação errada, uma vez que o tratado em questão foi ratificado não em 2009, mas num ano anterior àquele (uma informação veiculada por todo o complemento adverbial, isto é, pela locução prepositiva *antes de* e a data, propriamente dita).

No Segundo HAREM, a EM temporal (data e respectivo modificador/localizador temporal) passaria, neste caso em concreto, a ser integralmente reconhecida.

Com o alargamento do reconhecimento a toda a expressão temporal, a proposta do tempo levou também à reclassificação de certas EM que de outra forma seriam classificadas como VALOR QUANTIDADE. Considere-se o seguinte exemplo:

(8) Isso aconteceu há **3 anos** [`VALOR QUANTIDADE`]

No âmbito do primeiro HAREM, apenas a expressão *3 anos* seria identificada como EM, recebendo a classificação de VALOR QUANTIDADE (a qual não capta a referência temporal de toda a expressão). No âmbito da nova proposta, todo o complemento temporal (*há 3 anos*) seria reconhecido como EM, ao qual seria atribuído a subclassificação de `TEMPO_CALEND`.

(9) Isso aconteceu **há 3 anos** [`TEMPO TEMPO_CALEND DATA`]

Um outro aspecto que consideramos positivo no âmbito da nova especificação das EM temporais diz respeito ao tratamento conferido às expressões temporais que representam um intervalo específico de tempo. Considere-se o exemplo ilustrado em (10).

(10) A edição deste ano do Festival de Paredes de Coura decorre **entre 12 e 15 de Agosto**.

[5]De facto, no ReRelEM estava previsto o reconhecimento de relações entre entidades com a categoria TEMPO e entidades com as categorias OBRA, ACONTECIMENTO e PESSOA (ver Tabela 4.4 em Freitas et al. (2008b)

No Primeiro HAREM, seriam reconhecidas duas EM temporais independentes, *12* e *15 de Agosto*, o que levaria ao desmembramento da unidade sintáctico-semântica e consequente perda de informação veiculada por esta expressão. De facto, o significado individual de *12* e *15 de Agosto* não é idêntico ao significado de todo o constituinte, que também engloba, implicitamente, os dias 13 e 14.

Este tipo de expressão passou, pois, a ser tratado como uma única EM (do tipo INTERVALO) no âmbito da nova proposta de avaliação, tendo, aliás, motivado a decisão de considerar nas directivas do HAREM clássico também como entidades (com a categoria VALOR) os intervalos de valores, e respectivos especificadores (como ilustrado em (11) e (12)).

(11) Ele saltou **entre 7 a 10 metros** na sua fuga. [`VALOR QUANTIDADE`]

(12) O bilhete custa **mais de 5 euros** [`VALOR MOEDA`]

Ao nível da classificação, a nova proposta do TEMPO também é mais abrangente do que a anterior, ao considerar dois novos tipos de expressões temporais, os quais se encontram associados à noções de frequência e duração temporal (cf. (13) e (14), respectivamente).

(13) Costuma ir ao cinema **3 vezes por mês** [`TEMPO FREQUENCIA`]

(14) Demorou **3 anos** a escrever o livro [`TEMPO DURACAO`]

A proposta de análise das expressões temporais no Segundo HAREM introduziu a tarefa de normalização de expressões temporais. Embora consideremos que o novo desafio é importante para a análise temporal de um texto, não o iremos discutir por nos parecer uma tarefa que pode ser feita em paralelo ou posteriormente à do reconhecimento das entidades mencionadas, não fazendo, portanto, parte integrante do mesmo.

4 O tempo a regredir

A filosofia do Primeiro HAREM, e do HAREM em geral proposta pela Linguateca, assenta numa abordagem de baixo para cima ao reconhecimento de entidades mencionadas, partindo da análise de textos para determinar quais as entidades de interesse e propor uma classificação das mesmas com base no contexto semântico em que se encontram.

Na nova proposta do TEMPO, a noção de EM temporal é definida essencialmente com base num conjunto de critérios sintácticos e/ou formais, aproximando-se, em muitos casos, mais de uma unidade sintáctica do que lexical.

(15) Ficámos de nos encontrar **às três da tarde**

De facto, como ilustra o exemplo (15), a expressão destacada não corresponde a uma unidade lexical, mas a todo o complemento requerido pelo verbo *encontrar*. Porém, nem sempre assim é. Observa-se que, um número considerável de expressões que cabem na definição de EM temporal não corresponde a uma unidade linguística coesa, nem do ponto de vista lexical nem sintáctico-semântico. Um conjunto de restrições descritas ou ilustradas nas directivas leva, por exemplo, a que se considere que qualquer preposição que introduza um dado núcleo temporal faça parte integrante da expressão temporal, mesmo nos casos em que essa preposição nada tem a ver com o complemento temporal, servindo apenas de elemento de ligação entre esse complemento e o predicador (verbal, nominal ou adjectival) que a seleccionou, como acontece em (17), por oposição a (16) (e ao exemplo (15), ilustrado antes) .

(16) **À noite**, vou ao cinema

(17) Isso remonta **aos anos 80**

Os proponentes justificaram esta opção argumentando que os sistemas teriam dificuldade em distinguir os dois casos. Sendo uma simplificação, defenderam naturalmente a separação dos dois casos numa futura avaliação. Consideramos, no entanto, que este critério da simplificação leva a que se tenha uma colecção dourada com incorrecções do ponto de vista sintático e semântico, embora esteja conforme às directivas, o que não deveria acontecer.

No que respeita aos modificadores do núcleo de tempo (adjectivos, orações relativas ou sintagmas preposicionais), adoptou-se uma abordagem exactamente contrária à anteriormente mencionada. De facto, neste caso, a instrução geral das directivas é para não incluir os modificadores, mesmo que sejam modificadores obrigatórios do núcleo temporal (i.e., mesmo que sejam requeridos por esse nome), como exemplificado em (18). Um tratamento diferente é, no entanto, conferido aos modificadores adjectivais iniciados por maíusculas, como ilustrado em (19).

(18) Os poetas **do período** barroco

(19) Os poetas **do período Barroco**

Muito embora os restantes modificadores não possam, como referimos antes, ser identificados como fazendo parte da EM temporal, eles constituem um requisito para a identificação de certas expressões como entidades mencionadas. Observem-se, a título ilustrativo, os seguintes exemplos:

(20) **Nessa altura**, ele não estava consciente disso

(21) **Na altura do nascimento do filho**, ele não estava consciente disso

(22) **Na altura**, ele não estava consciente disso

De acordo com as novas directivas, apenas as expressões assinaladas em (20) e (21) devem ser consideradas EM, uma vez que o núcleo nominal *altura* é, respectivamente, especificado por um demonstrativo ou acompanhado de um modificador, neste caso, de tipo preposicional (*do nascimento do filho*), que, como dissemos, não fará, no entanto, parte da EM. No entanto, apesar de (20) e (22) serem construções quasi-equivalentes, a expressão temporal presente nesta última construção não deverá, de acordo com as directivas, ser reconhecida como EM, uma vez que não obedece a nenhuma das restrições anteriormente explicitadas.

A falta de sistematicidade no tratamento destes casos é, pois, algo que, na nossa perspectiva, deverá ser revisto numa futura edição de avaliação destas expressões temporais.

As questões anteriormente apontadas colocam-se fundamentalmente ao nível da identifição das EM. No que respeita à classificação propriamente dita das EM temporais, consideramos que, embora tenha havido uma maior especificação em relação à classificação de algumas EM, perderam-se também algumas distinções contempladas já no âmbito do Primeiro HAREM, que consideramos importantes. É, por exemplo, o caso da noção de PERIODO (anteriormente entendido como um período de tempo contínuo, não repetido), ilustrada nos exemplos (23) e (24) a seguir apresentados .

(23) Vou a Londres no próximo **Inverno** [`TEMPO PERIODO`]

(24) Nos **anos 80**, surgiram centenas de novas bandas musicais. [`TEMPO PERIODO`]

Actualmente, excepto nos casos em que esse período de tempo aparece expresso no texto através de dois limites temporais explícitos (que indicam, respectivamente, o início e o fim desse período), como acontece em (10), as restantes referências a períodos de tempo deixaram de ser classificadas como tal no âmbito desta avaliação. Os exemplos acima seriam, de acordo com as directivas do Segundo HAREM, anotados especificamente como datas.

(25) Vou a Londres **no próximo Inverno** [`TEMPO TEMPO_CALEND DATA`]

(26) **Nos anos 80**, surgiram centenas de novas bandas musicais [TEMPO TEMPO_CALEND DATA]

A solução adoptada, em que se valoriza a forma em detrimento da semântica das expressões, origina incongruência na análise de expressões, formalmente distintas mas semanticamente equivalentes, como, por exemplo, as ilustradas abaixo:

(27) **Entre 2006 e 2008** foram registados centenas de acidentes de viação [TEMPO TEMPO_CALEND INTERVALO]

(28) **Nos dois últimos anos** foram registados centenas de acidentes de viação [TEMPO TEMPO_CALEND DATA]

Em (27), como existem duas referências temporais explícitas, a entidade é marcada com o subtipo INTERVALO; a ausência das fronteiras explícitas desse intervalo de tempo, em (28), leva a que a EM em questão seja classificada com o subtipo DATA.

Por outro lado, o inverso também pode acontecer. Com as novas directivas do TEMPO, uma mesma expressão temporal que tenha sentidos diferentes, por se encontrar em contextos diferentes, pode ser anotada do mesmo modo. Por exemplo, a expressão *entre 12 e 15 de Agosto* no exemplo (10) e no exemplo (29) não referencia a mesma entidade temporal.

(29) O exame será realizado **entre 12 e 15 Agosto**.

Repare-se que em (10) o festival decorre em cada um dos dias expresso pelo intervalo, mas em (29) o exame só tem lugar num desses dias. Portanto neste último caso, a expressão está a ser usada para referir uma entidade que é uma data, embora formalmente não o seja.

A noção de tempo cíclico (abrangida pelo tipo CICLICO), que servia no Primeiro HAREM para representar períodos ou datas recorrentes / que se repetem no tempo, como é o caso das EM abaixo, também deixou de existir no novo modelo de classificação temporal.

(30) Costumo viajar na **Páscoa** [TEMPO CICLICO]

(31) A restauração da independência da República comemora-se a **1 de Dezembro** [TEMPO CICLICO]

No novo modelo de classificação, ambas as entidades *na Páscoa* e *a 1 de Dezembro* passariam a ser identificadas como datas. Na verdade, as expressões temporais abrangidas, no Primeiro HAREM, pelos tipos PERIODO ou CICLICO, são, na nova edição de avaliação, geralmente classificadas com o tipo DATA (como anteriormente ilustrado) ou ainda com o tipo GENERICO, como ilustrado em (32).

(32) Vários modelos inspirados no **século 18** [TEMPO PERIODO]

A classificação dessas expressões com os subtipos DATA ou GENERICO depende exclusivamente do critério geral de identificação e classificação das expressões temporais adoptado nas directivas do TEMPO no Segundo HAREM e que consiste na verificação de que a expressão temporal em questão, pode, ou não, responder em contexto à interrogativas "PREP quando?". Essa interrogativa é produtiva, por exemplo, em (30) e (31), o que leva a que as EM aí presentes sejam classificadas como TEMPO_CALEND DATA. Pelo contrário, essa interrogativa parece marginal ou mesmo inaceitável em (32), o que leva a que a EM presente nessa construção seja classificada como GENERICO.

Relativamente ao reconhecimento de expressões temporais com o tipo GENERICO, vale a pena ainda acrescentar que julgamos que muitas vezes estas expressões não representam/mencionam de facto entidades temporais. Compare-se, por exemplo, a classificação de *Natal* no Primeiro e Segundo HAREM na seguinte frase:

(33) Domingos Afonso, na maré do **Natal**, dava a todos os pobres um quilo de bacalhau [ABSTRACCAO ESCOLA]

(34) Domingos Afonso, na maré **do Natal**, dava a todos os pobres um quilo de bacalhau [TEMPO GENERICO]

De facto, *Natal* no contexto em causa, é uma referência ao *espírito natalício* e aos costumes da *época natalícia*, não sendo propriamente um locativo temporal: não é esta expressão que localiza temporalmente o evento referido na frase (a oferta do quilo de bacalhau), mas sim a ocorrência de *a noite de Consoada* no contexto anterior (não ilustrado)[6]. A classificação adequada desta expressão como ABSTRACCAO, em vez de TEMPO, fica mais clara no seguinte exemplo:

(35) Em Abril, Domingos Afonso, na maré do **Natal**, dava a todos os pobres um quilo de bacalhau [ABSTRACCAO ESCOLA]

[6]Destaque-se, aliás, que *Consoada* no Primeiro HAREM foi classificada como CICLICO (uma vez que está a representar várias noites de Consoada, em vez de uma Consoada em particular) e que no Segundo HAREM *a noite de Consoada* foi classificado como GENERICO (por, no contexto, não responder de forma adequada a uma pergunta com a estrutura "PREP quando?").

Finalmente, um aspecto que consideramos muito positivo no modelo de classificação do HAREM (Primeiro HAREM e HAREM clássico), a vagueza na classificação das entidades mencionadas, foi ignorado nas novas directivas do TEMPO[7].

Por exemplo, expressões que, num dado contexto, podem ser referências quer a um período (ou intervalo de tempo) quer a uma data, e cuja vagueza era possível de representar no antigo modelo de classificação, passaram a ter uma única classificação no âmbito da nova proposta.

(36) O mês de **Outubro** é marcado pela passagem para o euro de alguns produtos e serviços BPI.[TEMPO|TEMPO PERIODO|DATA]

Veja-se o caso de *Outubro*, ilustrado em (36), que tanto pode ser uma referência a um período, se a passagem ao euro foi sendo feita ao longo do mês, como a uma data, se a passagem ao euro se deu apenas num dia desse mês. No Primeiro HAREM, foi-lhe associados dois tipos (PERIODO e DATA). No Segundo HAREM, pelo contrário, seria necessário optar por anotar este caso (que incluiria *o mês de*), de uma de duas maneiras: (i) como GENERICO, porque que a expressão sintacticamente não corresponde a um complemento adverbial de tempo, e, consequentemente, não é possível formular com base na sintaxe da frase a pergunta "PREP quando" que a teria como resposta; (ii) como TEMPO_CALEND DATA, porque mesmo assim, com base na semântica da frase, é possível formular a pergunta *quando é que se deu a passagem para o euro de alguns produtos e serviços BPI?* e se trata formalmente de uma data.

5 *Troca de directivas entre as duas edições do HAREM*

De forma a tipificar e avaliar as diferenças no tratamento das expressões temporais em cada uma das edições do HAREM, procedemos a dois exercícios distintos, mas complementares, tomando como referência as directivas do TEMPO adoptadas no Primeiro e no Segundo HAREM.

O primeiro, cujos resultados descrevemos em 5.1, consistiu em verificar quais as entidades da CD do primeiro evento do Primeiro HAREM, classificadas como TEMPO de acordo com as directivas do Primeiro HAREM, que sofreriam alterações, tanto ao nível da segmentação como da classificação, se fossem analisadas de acordo com as directivas adoptadas no Segundo HAREM.

[7]Isso talvez seja uma consequência de a análise do contexto ser muito local, o que advém dos critérios de delimitação utilizados serem essencialmente lexico-sintácticos.

O segundo exercício, cujos resultados apresentamos em 5.2, consistiu no processo inverso ao anteriormente descrito. Em concreto, analisámos as entidades classificadas como TEMPO na CD do Segundo HAREM, por forma a verificar quais as alterações que essas EM sofreriam se tivéssemos utilizado, em vez das directivas do Segundo HAREM, as directivas do Primeiro HAREM.

Daqui resultaram novas colecções douradas que serão disponibilizadas no sítio da Linguateca dedicado ao HAREM (`http://www.linguateca.pt/HAREM/`).

As tabelas 3 e 5.2 sumarizam as principais modificações observadas, em cada um dos casos, respectivamente. Para simplificar a tabela usámos apenas o tipo ou subtipo mais específico, não mostrando a categoria nem o tipo (caso exista o subtipo). Relembramos então que:

Na classificação do Primeiro HAREM:

- DATA, HORA, CICLICO e PERIODO são tipos da categoria TEMPO
- EFEMERIDE e ORGANIZADO são tipos da categoria ACONTECIMENTO
- QUANTIDADE é um tipo da categoria VALOR
- PUBLICACAO é um tipo da categoria OBRA
- IDEIA é um tipo da categoria ABSTRACCAO

Na classificação do Segundo HAREM:

- GENERICO, DURACAO e FREQUENCIA são tipos da categoria TEMPO
- DATA, HORA e INTERVALO são subtipos do tipo TEMPO_CALEND, que é um tipo da categoria TEMPO
- EFEMERIDE é um tipo da categoria ACONTECIMENTO

Para efeitos de contabilização nas tabelas, considerámos que:

- a segmentação era diferente de uma edição para outra do HAREM quando de uma entidade se passasse a ter duas ou o contrário;
- a entidade seria alargada se passasse a incluir palavras em minúsculas, excluindo os casos em que a palavra é uma preposição;
- a entidade seria mais curta se deixasse de incluir palavras em minúsculas, excluindo os casos em que a palavra é uma preposição;
- a classificação TEMPO DATA do Primeiro HAREM e a classificação TEMPO TEMPO_CALEND DATA são a mesma classificação;

- a classificação TEMPO HORA do Primeiro HAREM e a classificação TEMPO TEMPO_CALEND HORA são a mesma classificação.

Falaremos dessas experiências com mais pormenor, em seguida.

5.1 Análise das entidades classificadas como TEMPO no Primeiro HAREM de acordo com as directivas do Segundo HAREM

Para efeitos deste exercício, tomámos como ponto de partida a CD usada no primeiro evento do Primeiro HAREM (Mota et al., 2008b; Mota et al., 2008a; Carvalho et al., 2008), a qual é constituída por 129 documentos (distribuídos por oito géneros: web, jornalístico, entrevista, expositivo, correio electrónico, literário, político e técnico), que compreendem, no seu conjunto, um total de 5.065 EM (as quais foram anotadas de acordo com as directivas estabelecidas no âmbito do Primeiro HAREM). Dessas EM, 441 (8.7%) encontram-se associadas à categoria TEMPO.

Ao analisarmos aquele subconjunto de EM[8] com as directivas do TEMPO do Segundo HAREM, verificámos que:

- 19% das EM (81 entidades) não sofreriam quaisquer alterações;
- 77% das EM (334 entidades) passariam a ser segmentadas de forma diferente, nomeadamente devido ou a) à inclusão da preposição que introduz a expressão temporal (223 entidades), como em (37), ou b) ao alargamento da própria noção de expressão temporal (32 entidades), como em (38), ou, finalmente, c) à observação de ambas as situações (79 entidades), como em (39).

  (37) Foi constituída por escritura pública **em Junho de 1992**

  (38) Realiza-se **dia 20 de Maio**

  (39) A Evolução Da Colônia portuguesa na América, **a partir da segunda metade do século XVII**, será profundamente marcada pelo novo rumo

- 18% das EM (79 entidades) passariam a receber uma nova classificação. Mais especificamente:

  – 45 entidades EM anteriormente classificadas como DATA ou PERIODO passariam a ser classificadas como INTERVALO (cf. (40) e (41), respectivamente);

  (40) O 17º congresso mundial em Gestão de Projectos decorrerá **entre os dias 4 a 6 do Junho**

  (41) foi o ditador alemão que comandou a Alemanha nazista na Segunda Guerra Mundial (**1939-1945**)

  – 21 entidades EM anteriormente classificadas como PERIODO ou CICLICO (cf. (42) e (43), respectivamente) passariam a ser classificadas como DATA;

  (42) Eles teriam vendido **em março** acima da média

  (43) Juntamo-nos sempre **pela Páscoa**

  – 8 entidades EM anteriormente classificadas como DATA, PERIODO ou CICLICO (cf. (44), (45) e (46), respectivamente) passariam a ser classificadas como GENERICO;

  (44) e dois meses depois veio **o 24 de Abril**

  (45) Comentários Jornadas da Juventude animam **mês de Abril**

  (46) Ainda me lembro do primeiro texto que eu estive a ler. Era sobre a **Primavera**

  – 2 entidades EM anteriormente classificada como DATA ou PERIODO passaria a ser classificada como DURACAO (ver, respectivamente, exemplos (47) e (48));

  (47) Dada a escassez de tempo e a abrangência da matéria, já se firmou um compromisso de **durante 1997**

  (48) encontraram a capivara no Brasil **durante o século XVI**

  – 1 entidade EM anteriormente classificada como CICLICO passaria a ser classificada como FREQUENCIA:

  (49) Desde aquele ano tem continuado este serviço na Igreja Metodista (**primeiro Domingo de cada ano**))

  – Apenas 2 entidades EM deixariam de ser reconhecidas como TEMPO (cf. (50) e (51))

  (50) os povoadores cristãos da **Reconquista**

  (51) parecia uma mina daquela época do **Velho Oeste**

[8] Para simplificar a análise, excluímos das 441, nove entidades por fazerem parte de uma estrutura `ALT`.

Mesma delimitação e classificação	81
Mesma delimitação, mas muda a classificação:	17
- PERIODO − > DATA	4
- DATA, CICLICO ou PERIODO − > GENERICO	4
- PERIODO − > INTERVALO	7
- DATA|EFEMERIDE − > EFEMERIDE	1
- PERIODO − > IDEIA	1
Segmentação diferente:	19
- Uma entidade PERIODO passa a duas entidades DATA	1
- Uma DATA passa a duas entidades (DATA e LOCAL)	1
- Duas entidades DATA formam INTERVALO	2
- Duas entidades ou mais formam INTERVALO:	31
– introduzido por preposição	24
– introduzido por preposição e incluindo modificadores	7
Entidade seria alargada:	26
- Para incluir modificadores	2
- Para incluir outros elementos (dia, ano, ...):	24
– Com mesma classificação	20
– PERIODO − > DATA	1
– DATA ou PERIODO − > GENERICO	3
Continua a ser TEMPO, mas com preposição inicial:	223
- Com mesma classificação	211
- CICLICO ou PERIODO − > DATA	7
- DATA − > GENERICO	1
- PERIODO − > INTERVALO	2
- DATA ou PERIODO − > DURACAO	2
Continua a ser TEMPO, mas com preposição inicial e outros elementos:	48
- Com mesma classificação	39
- PERIODO − > DATA	8
- PERIODO − > INTERVALO	1

Tabela 3: Análise das entidades classificadas com TEMPO no Primeiro HAREM com as directivas do Segundo HAREM

5.2 Análise das entidades classificadas como TEMPO no Segundo HAREM de acordo com as directivas do Primeiro HAREM

No segundo exercício, tomámos como ponto de partida a CD do Segundo HAREM (Carvalho et al., 2008; Mota et al., 2008b). Este corpo é constituído por 129 documentos distribuídos por 13 géneros: notícia, didáctico, blogue jornalístico, blogue pessoal, perguntas, ensaio, opinião, blogue humorístico, legislativo, promocional, entrevista, texto privado manuscrito e perguntas faq, que compreende um total de 7.836 EM, identificadas de acordo com as directivas em vigor no Segundo HAREM. Dessas EM, 1.195 (15%) encontram-se associadas à categoria TEMPO, quase três vezes mais do que as entidades temporais compreendidas na CD do primeiro evento do Primeiro HAREM.

Ao aplicarmos as directivas do Primeiro HAREM àquele subconjunto de EM, observámos que:

- 41% das entidades temporais (491 EM) deixariam de ser reconhecidas por não conterem nem dígitos nem maiúsculas, o que constituía, como já referimos antes, um requisito essencial no Primeiro HAREM. Isso mostra a importância de reconhecer também expressões em minúsculas, pois de outra forma uma grande parte do conteúdo temporal dos textos é perdida.

  Na sua maioria (62%), essas EM pertencem à nova categoria TEMPO_CALEND (291 são DATA, 11 são HORA e 3 são INTERVALO). Nos restantes casos, as expressões exprimem valores de frequência (70 EM) ou de duração (46 EM), não previstos nas directivas do Primeiro HAREM, ou, em vez disso, encontram-se associadas à categoria GENERICO (65 casos), que constitui igualmente uma novidade nas novas directivas do TEMPO. Tal como é referido nas directivas, estas últimas expressões, muito embora contenham uma unidade de tempo,

Não seria entidade no Primeiro HAREM	491
Mesma delimitação e classificação	53
Mesma delimitação, mas muda a classificação:	25
- GENERICO, DATA ou INTERVALO − > PERIODO	19
- DATA − > CICLICO	1
- DURACAO|QUANTIDADE ou FREQUENCIA − > QUANTIDADE	3
- EFEMERIDE|GENERICO − > EFEMERIDE	2
Segmentação diferente:	69
- Em duas entidades e mantém-se a classificação	1
- INTERVALO passa a duas entidades DATA	36
- INTERVALO passa a duas entidades HORA	2
- DURACAO ou QUANTIDADE|INTERVALO passa a duas entidades QUANTIDADE	2
- DATA faz parte de ORGANIZADO	19
- GENERICO ou DATA faz parte de EFEMERIDE	2
- DATA faz parte de PUBLICACAO	5
- INTERVALO passa a uma entidade PERIODO, e a outra não reconhecida	2
Entidade seria mais curta:	15
- Com mesma classificação	5
- GENERICO − > PERIODO	4
- DATA, DURACAO, GENERICO ou DURACAO|QUANTIDADE − > QUANTIDADE	5
- GENERICO − > EFEMERIDE	1
Continua a ser TEMPO, mas sem preposição inicial e sem outros elementos:	527
- Com mesma classificação	441
- DATA, DURACAO, GENERICO, ou INTERVALO − > PERIODO	81
- DATA ou GENERICO − > CICLICO	5
Deixa de ser TEMPO, sem preposição inicial e sem outros elementos:	15
- DATA, DURACAO, GENERICO, DATA|DURACAO ou DURACAO|QUANTIDADE − > QUANTIDADE	9
- DATA ou GENERICO − > EFEMERIDE	6

Tabela 4: Análise das entidades classificadas com TEMPO no Segundo HAREM com as directivas do Primeiro HAREM

não representam um tempo de calendário específico no contexto em questão. Estes casos encontram-se ilustrados, respectivamente, em (52), (53) e (54). Cinco casos estão ainda marcados como vagos entre DURACAO e GENERICO, DURACAO e DATA (3 deles), e DATA e GENERICO.

(52) **A maior parte das vezes** Mills fala de si mesmo como o originador da mensagem[FREQUENCIA]

(53) **Durante muitos anos** o fotógrafo trabalhou nas agências Sygma e Gamma [DURACAO]

(54) não há nada como **o outono** [GENERICO]

- 48% seriam reconhecidas como entidades temporais, mas seriam segmentadas de forma distinta (577 casos), quer por deixarem de fora da EM a preposição que as introduz (449 casos), quer por deixarem de incluir unidades lexicais - especificadores, modificadores, etc. (em minúsculas), que antes não seriam contempladas (9 casos), quer pela observação de ambas as situações (78 casos). Nos restantes 41 casos, a entidade seria segmentada em duas entidades temporais, deixando de fora outros elementos em mínúsculas (como a preposição inicial, por exemplo). Este último caso acontece sobretudo com entidades associadas à noção de intervalo (38 casos), que seriam fragmentadas e classificadas como duas EM independentes, pertencentes ao tipo DATA ou HORA. Cada um dos casos é ilustrado, respectivamente, em (55), (56), (57) e (58).

(55) O que nos é dito é que **em** [**Janeiro**] tudo vai mudar

(56) foi inaugurada **dia** [**9 de setembro de 2004**]

(57) deverá situar-se nos 407,4 euros (..) **a partir de** [**Janeiro**] **do próximo ano**

(58) Que acontenceu na Argélia **na noite de** [**17**] **para** [**18 de Agosto de 1994**]

De notar, no entanto, que a classificação dessas entidades não seria a mesma em 22,5% dos casos, tal como se ilustra de (59) a (62).

(59) assistimos **no** [**século XVI**] ao fermentar de um enorme debate [DATA -> PERIODO]

(60) evento que acontece sempre **dia** [**06 de janeiro**] [DATA -> CICLICO]

(61) Mas a festa vai continuar **ao longo do ano de** [**2008**] [DURACAO -> PERIODO]

(62) porque só funciona **das** [**08h00**] **às** [**09h00**] [INTERVALO -> HORA]

- Apenas 6% das entidades anotadas no Segundo HAREM como TEMPO (73 entidades) continuariam a ser TEMPO com a mesma delimitação, mas, dessas, 20 passariam a ser classificadas com um tipo diferente, como em (63) e (64). De referir, no entanto, que esta percentagem ascenderia aos 43% se tivermos igualmente em conta as EM que diferem destas simplesmente por causa da introdução da preposição. Isto sugere que sistemas, como o da Priberam, que por opção não incluíram a preposição, poderiam ter tido resultados significativamente melhores no reconhecimento das entidades temporais, uma vez que no Segundo HAREM entidades que não estivessem bem delimitadas não contribuíam para a pontuação final do sistema.

  (63) fecha **2.ª** [DATA -> CICLICO]

  (64) Afirmando esperar um ano de 2008 mais exigente do que **2007** [DATA -> PERIODO]

- Aproximadamente 4,5% das entidades temporais do Segundo HAREM (54 entidades) passariam a ser classificadas com uma categoria que não TEMPO, das quais apenas 5 casos manteriam a mesma delimitação. Na sua maioria, estas entidades fariam parte de uma entidade maior com a categoria ACONTECIMENTO ORGANIZACAO, como se ilustra em (65). Se tivermos em conta apenas as entidades que são reconhecidas de acordo com as directivas de ambas as edições do HAREM, então cerca de 8% não seriam classificadas como TEMPO, o que acentua as diferenças de interpretação do sentido das entidades no Primeiro e no Segundo HAREM.

  (65) O [Tour de França **de 2009**] vai começar no Mónaco [TEMPO -> ACONTECIMENTO ORGANIZADO]

Nos casos em que a entidade existe tanto no Segundo HAREM como no Primeiro HAREM, mesmo que com ajustes na delimitação, verificámos ainda que:

- as entidades com o tipo GENERICO continuariam a ser entidades temporais, mas com o tipo PERIODO (9 entidades) ou CICLICO (4 entidades), ou então passariam à categoria ACONTECIMENTO EFEMERIDE (8 entidades);
- as entidades com o tipo DURACAO continuariam a ser entidades temporais com o tipo PERIODO em 2 casos mas maioritariamente passariam a ser classificadas com VALOR QUANTIDADE (12 entidades);
- a única entidade com valor de frequência a manter-se, seria com a classificação de VALOR QUANTIDADE;
- as datas continuariam a ser datas na maioria dos casos, e passariam a ser classificadas como períodos em 82 casos e como datas cíclicas em 2 casos. Em 34 casos deixariam de ser tempo para passar a ser ou ser integradas em ACONTECIMENTO EFEMERIDE (5 casos), ACONTECIMENTO ORGANIZADO (19 casos), VALOR QUANTIDADE (5 casos) e OBRA PUBLICACAO (5 casos);
- Nenhuma entidade com o tipo HORA seria reclassificada;
- Entidades do tipo INTERVALO passariam a períodos, quando a entidade não é partida em duas (no Primeiro HAREM, datas separadas por '/' ou '-' correspondiam a uma única entidade), ou a datas e horas, quando a entidade é segmentada em duas que representam os limites do intervalo (como acontece, por exemplo, a *entre Novembro e Dezembro*).

5.3 Discussão

Estes dados mostram claramente que a noção de entidade temporal não é idêntica no Primeiro e no Segundo HAREM.

As principais diferenças observadas têm sobretudo a ver com a tarefa de identificação, uma vez que a percentagem de entidades temporais que seriam delimitadas do mesmo modo quer se usem umas directivas ou outras seria baixa (22% no caso da CD do Primeiro HAREM e 11% no caso da CD do Segundo HAREM, se não contarmos as entidades que não seriam identificadas por só conterem minúsculas - se tivermos em conta

também essas, então a percentagem seria ainda mais baixa: 6%).

No entanto, também se observam diferenças na classificação. Especificamente, 18% das entidades do Primeiro HAREM deixariamm de ter a mesma classificação, enquanto no Segundo HAREM 29% das entidades que seriam também reconhecidas no Primeiro HAREM (mesmo que com ajustes na delimitação) teriam uma classificação diferente. Isso é, naturalmente, uma consequência de as categorias serem diferentes em ambas as edições do HAREM, mas também se deve ao facto de os conceitos por elas representados serem diferentes. Por esse motivo, não é possível estabelecer um mapeamento entre as categorias das duas edições. Repare-se que mesmo as entidades DATA num dos HAREM podem não ser DATA no outro.

Ao nível da classificação, um facto interessante é que as directivas do Segundo HAREM tendem a confirmar que as entidades do Primeiro HAREM são entidades temporais, pois só duas entidades é que deixariam de serem reconhecidas como referências temporais. Porém, na situação inversa é mais provável que uma entidade temporal deixe de o ser ao aplicar-se as directivas do Primeiro HAREM, pois 8% das entidades que seriam reconhecidas por ambas as directivas deixariam de ser tempo no Segundo HAREM.

Se, por um lado, se pode argumentar que com o Segundo HAREM, as entidades temporais ficaram delimitadas com maior precisão e abrangendo um maior leque de entidades, por também permitir o reconhecimento de expressões só em minúsculas (relembre-se que 41% das entidades do Segundo HAREM não existiriam no Primeiro, e que 52% passaram a ser mais bem delimitadas), pelo outro, pensamos que o Segundo HAREM ficou a perder em termos da classificação das entidades referidas pelas expressões do texto (consideramos que em cerda de metade dos casos em que houve alteração da classificação, houve perda de significado).

De facto, no Primeiro HAREM, os critérios de reconhecimento de entidades temporais são essencialmente semânticos, obrigando muitas vezes a uma leitura que vai além da frase para poder determinar a classificação. No Segundo HAREM, pelo contrário, basta um contexto muito local para atribuir a classificação, pois os critérios são de natureza essencialmente formal.

É por essa razão que uma expressão que superficialmente é uma data, no Segundo HAREM é em geral classificada como DATA (ou será GENERICO quando não verifica o critério da pergunta-resposta), mas no Primeiro HAREM poderia ser classificada como DATA, PERIODO, ou CICLICO (ou mesmo, estar integrada em entidades não elementares que referenciam entidades não temporais), mostrando que uma mesma expressão pode designar referentes temporais com propriedades distintas. Por exemplo, *Páscoa* em (66) e (67) não designa o mesmo referente temporal. No primeiro caso, **Páscoa** designa um domingo de Páscoa concreto e único (DATA), mas, no segundo caso, designa vários domingos de Páscoa (CICLICO).

(66) Numa pista em bom estado, apesar da ameaça de chuva que pairou na região durante o domingo de **Páscoa**

(67) Juntamo-nos sempre pela **Páscoa**

6 Sugestões para o futuro do tempo no HAREM

De acordo com a análise que fizemos, e de modo a preservar quer o modelo semântico subjacente ao Primeiro HAREM, quer a maior precisão na identificação das entidades temporais do Segundo HAREM, julgamos que em futuras edições do HAREM se deve ter em conta os seguintes aspectos na avaliação do TEMPO:

- não constrangimento das entidades temporais a terem apenas maiúsculas ou números;
- inclusão de todos os modificadores que façam parte da entidade temporal de forma a delimitar todo o complemento adverbial de tempo;
- reconhecimento de intervalos de tempo cujos limites estejam expressos por meio de duas datas;
- reconhecimento de frequências e de durações;
- classificação semântica mais fina, nomeadamente voltar a considerar os tipos PERIODO e CICLICO.

Agradecimentos

Este trabalho foi desenvolvido no âmbito da Linguateca, co-financiada pelo governo português, pela União Europeia (FEDER e FSE), sob o contrato POSC/339/1.3/C/NAC, e também financiada pela UMIC e pela FCCN. O trabalho da segunda autora foi ainda financiado pela Fundação para a Ciência e a Tecnologia através de uma bolsa de pós-doutoramento com a referência SFRH/BPD/45416/2008. A primeira autora agradece ainda o apoio que o grupo Proteus da New York University lhe tem dado en-

quanto desenvolve o seu trabalho para a Linguateca.

Estamos igualmente gratas à Diana Santos pela motivação e revisão de versões anteriores do artigo, bem como ao José Carlos Medeiros e ao Pablo Gamallo pelo seu cuidadoso trabalho de revisão que nos forneceu valiosas sugestões de melhoria, que tentámos ter em conta tanto quanto o tempo permitiu.

A bibliografia foi construída com o apoio do SUPeRB (Cabral, Santos e Costa, 2008).

Referncias

Cabral, Luís Miguel, Diana Santos, e Luís Fernando Costa. 2008. SUPeRB: Building bibliographic resources on the computational processing of Portuguese. Em Daniela Braga, Miguel Sales Dias, e António Teixeira, editores, *Propor 2008 Special Session: Applications of Portuguese Speech and Language Technologies (full proceedings)*, September 10, 2008.

Cardoso, Nuno e Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 211–238, 12 de Novembro, 2007. Documento original publicado no sítio do HAREM a 29 de Março de 2006. Republicado como Relatório técnico DI-FCUL TR-06-18 : Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Novembro de 2006.

Carvalho, Paula, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas, e Cristina Mota. 2008. Segundo HAREM: Modelo geral, novidades e avaliação. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 11–31, 31 de Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Paula Carvalho, e Hugo Gonçalo Oliveira. 2008a. Apêndice C: ReRelEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Freitas, Cláudia, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho, e Cristina Mota. 2008b. Relações semânticas do ReRelEM: além das entidades no Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 77–96, 31 de Dezembro, 2008.

Grishman, Ralph e Beth Sundheim. 1996. Message understanding conference-6: a brief history. Em *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, pp. 466–471, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2008a. Apêndice B: Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 31 de Dezembro, 2008.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2008b. Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 33–54, 31 de Dezembro, 2008.

Hagège, Caroline, Jorge Baptista, e Nuno Mamede. 2010. Caracterização e processamento de expressões temporais em português. *Linguamática*, 2(1):63–76, Abril, 2010.

Mota, Cristina, Paula Carvalho, Cláudia Freitas, Hugo Gonçalo Oliveira, e Dia. 2008a. É tempo de avaliar o TEMPO. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 55–75, 31 de Dezembro, 2008.

Mota, Cristina, Diana Santos, Paula Carvalho, Cláudia Freitas, e Hugo Gonçalo Oliveira. 2008b. Apêndice H: Apresentação detalhada das colecções do Segundo HAREM. Em Cristina Mota e Diana Santos, editores, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*. Linguateca, pp. 355–377, 31 de Dezembro, 2008.

Pustejovsky, James, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, e Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. Em *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, 15-17 de Janeiro, 2003.

Santos, Diana. 2007. O modelo semântico usado no Primeiro HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 43–57, 12 de Novembro, 2007.

Santos, Diana e Nuno Cardoso. 2007. Breve Introdução ao HAREM. Em Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, pp. 1–16, 12 de Novembro, 2007.

Santos, Diana, Nuno Cardoso, e Nuno Seco. 2006. Avaliação no HAREM: Métodos e medidas. Relatório Técnico DI-FCUL TR-06-17, Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Novembro, 2006. `http://www.linguateca.pt/Diana/download/SantosCardosoSecoMedidas2006.pdf`.

Santos, Diana e Paulo Rocha. 2003. AvalON: uma iniciativa de avaliação conjunta para o português. Em Amália Mendes e Tiago Freitas, editores, *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística (APL 2002)*, pp. 693–704, Lisboa, 2-4 de Outubro de 2002, 2003. APL.

Seco, Nuno, 2007. *MUC vs HAREM: a contrastive perspective*, pp. 35–41. Linguateca, 12 de Novembro, 2007.