Capítulo 2

Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro

Jorge Baptista, Caroline Hagège e Nuno Mamede

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Capítulo 2, p. 33–54.

Neste capítulo apresentamos a proposta que elaborámos (Hagège et al., 2008) para a tarefa de reconhecimento, classificação e normalização de expressões temporais (ET) no âmbito da segunda avaliação conjunta de sistemas de reconhecimento entidades mencionadas (EM) do português – o Segundo HAREM. Procurámos, além disso, reflectir sobre a experiência desta avalição conjunta para que, baseados na forma de intervenção dos vários sistemas participantes e nos resultados globais desta pista de avaliação, pudéssemos sugerir futuros desenvolvimentos e novas iniativas de avaliação desta tarefa.

2.1 Introdução

2.1.1 Generalidades

O reconhecimento, classificação e representação das ET não é uma tarefa trivial. Apesar de o conjunto de elementos lexicais (pelo menos em termos de palavras simples) envolvidos ser relativamente extenso, é, ainda assim, suficientemente bem limitado para que se conceba como meta exequível atingir-se uma cobertura lexical próxima da exaustividade. Já o mesmo não se passa com o conjunto de construções em que se podem combinar estes elementos lexicais associados à expressão do tempo, que poderão representar várias centenas de construções diferentes[1], os quais se podem combinar entre si segundo padrões sintácti-co-semânticos que, tanto quanto sabemos, ainda não foram sistematicamente recenseados.

Este tipo de expressão apresenta também a dificuldade suplementar que resulta na diversidade de valores semânticos (interpretação) que podem ser associados aos elementos gramaticais ou formais que introduzem a expressão temporal. Assim, por exemplo, nas ET *no próximo ano* e *em duas semanas*, não é possível fazer depender apenas da presença da preposição *em* a interpretação global de cada uma destas expressões. Pelo contrário, só levando em consideração toda a expressão bem como o preenchimento lexical das várias posições estruturais (preposição, determinante, nome de tempo e eventual modificador) é possível classificá-las de forma adequada, nomeadamente, considerando a primeira ET como uma **data** (exemplo (2.1)) e a segunda como uma **duração** (exemplo (2.2)).

(2.1) O João só vai fazer isso *no próximo ano*.

(2.2) O Pedro concluiu a tarefa *em duas semanas*.

Além disso, e como em muitos outros aspectos da linguagem natural, verifica-se um determinado grau de vagueza na interpretação de muitas ET. Assim, por exemplo, uma ET como *há dois anos* deverá ser interpretada como se referindo ao intervalo de tempo entre 1 de Janeiro e 31 de Dezembro de 2006 ou a uma data exacta nesse ano, mas relativamente ao momento da enunciação (*hoje*)? Repare-se que as línguas têm geralmente mecanismos (quantificadores) que tanto permitem controlar (contrariar?) como reforçar esta dimensão (vagueza) intrínseca do discurso:

(2.3) O João fez isso *há precisamente/aproximadamente/mais de dois anos*.

(2.4) O João fez isso *há imensos/vários/alguns/poucos/uns poucos de anos*.

[1] Como exemplo de uma exploração sistemática de famílias de expressões temporais em português, veja-se, entre outros, Móia (2000) e Baptista (2003).

A indefinição, a que acima nos referimos, poderá eventualmente ser esclarecida pelo contexto comunicativo ou discursivo. Contudo, noutros casos, ela é um mecanismo expressivo da língua, dando origem a formas cuja interpretação não é necessariamente literal, como acontece em situações de hipérbole (como em (2.5)) ou de eufemismo (como em (2.6)).

(2.5) O Pedro fez isso *há séculos/mais de três quinze dias*!

(2.6) Espera só *um minuto* que eu já te faço isso.

Finalmente, salientamos que uma adequada interpretação das ET depende muitas vezes da frase em que se insere. Assim, por exemplo, até uma data como *5 de Dezembro* só pode ser localizada relativamente ao momento da enunciação se se levar em conta o tempo-modo do verbo que a ET modifica:

(2.7) O avião aterrou em Lisboa *no dia 5 de Dezembro*.

(2.8) O avião vai aterrar em Lisboa *no dia 5 de Dezembro*.

Por outro lado, esta expressão tem, nas frases acima, um valor aspectual pontual, resultado da combinatória com um predicado como *aterrar* (avião); se se tratar de outro tipo de predicado, com outro valor aspectual, a modificação que o advérbio exerce parece ser aspectualmente diferente:

(2.9) O Pedro esteve em casa doente *no dia 5 de Dezembro*.

Um caso semelhante, ocorre nas construções temporais com *haver*, que podem ter leituras diferentes consoante o tempo-modo do verbo da frase que modificam: **data** em (2.10) e **duração** em (2.11).

(2.10) O João fez isso *há 5 anos*.

(2.11) O João faz isso *há 5 anos*.

A proposta de avaliação da categoria `TEMPO` apresentada ao Segundo HAREM procurou abordar algumas destas questões, dando particular ênfase ao tratamento da referência e tentando contribuir no sentido da construção de um standard de normalização das ET.

2.1.2 Motivação da proposta

Com a normalização de ET, temos como objectivo final a tarefa, bem mais complexa, de reconhecer as ET presentes no texto para as associar aos eventos e estados de coisas que aquelas modificam, de modo a podermos ordenar parcialmente, segundo uma sequência cronológica, esses mesmos eventos e estados de coisas.

Naturalmente, esta meta constitui um objectivo demasiado ambicioso, em particular no quadro de um evento como o HAREM, cujo foco é o reconhecimento e a classificação de EM. Pretendemos, pois, com a nossa proposta dar um passo naquela direcção, passando pela incontornável tarefa de reconhecimento e classificação de ET, na continuidade do Primeiro HAREM, ao mesmo tempo que fazemos uma primeira abordagem a um dos

grandes problemas levantados por este tipo de expressões, nomeadamente o problema da referência temporal.

A proposta de reconhecimento, classificação e normalização de expressões temporais que fizemos no âmbito do Segundo HAREM (Hagège et al., 2008) encontra a sua principal motivação em trabalhos recentes e num interesse renovado da comunidade do processamento de linguagem natural (PLN) pela problemática do tratamento do tempo, no domínio mais vasto da extracção de informação. Com efeito, é necessário tomar em conta a dimensão temporal veiculada nos textos para levar a cabo de maneira satisfatória diversas tarefas que visam a extracção de informação a partir de textos. Por exemplo, as respostas a perguntas como *Qual é a capital da Alemanha?* *Quem era o vice presidente de Bush?* serão diferentes conforme os momentos da história a que se possam referir e, naturalmente, consoante a data dos textos que estarão acessíveis para poder responder a estas perguntas. Para aplicações de PLN que trabalham com vários documentos como, por exemplo, a sumarização, uma representação adequada da dimensão temporal dos textos deverá permitir relacionar entre si os eventos neles referidos.

Vários indicadores mostram o interesse crescente na área do processamento do tempo: é disto exemplo a primeira avaliação conjunta TempEval[2], em 2007 (Verhagen et al., 2007), que teve lugar no âmbito da conferência Senseval 2007[3]. A Google também oferece na Google Trends[4] a possibilidade de visualizar o resultado de uma pesquisa usando a dimensão temporal. Além do mais, já foram feitas propostas para anotação fina de ET e, para o inglês, existem alguns recursos, tais como os textos anotados com a norma TimeML (Saurí et al., 2006)[5]. Para outras línguas (o francês e o romeno, pelo menos), estão já em desenvolvimento diversos trabalhos nesta área (ver, por exemplo, Battistelli et al. (2008)).

Pareceu-nos importante abordar este problema para o português e a avaliação conjunta do HAREM constituiu uma excelente plataforma para o fazer, embora a nossa proposta ultrapasse o quadro estrito de reconhecimento de entidades mencionadas (REM).

2.1.3 Questões operacionais da proposta

Nesse sentido, na elaboração da proposta, procurámos seguir alguns princípios norteadores que aqui apresentamos sucintamente, embora tenhamos de retomar alguns deles mais adiante:

(i) uma tarefa executável em seis meses de desenvolvimento, a fim de permitir não só a continuidade dos anteriores participantes, dando-lhes tempo de reconverterem os seus sistemas, se necessário, mas também incentivar a participação de novos actores;

(ii) compatibilidade com propostas já existentes, garantido uma continuidade natural com a tarefa da anterior edição do HAREM (Cardoso e Santos, 2007), aproximando-a ou adaptando-a, no entanto, aos standards que se estão a constituir em torno das mais recentes avaliações conjuntas internacionais;

(iii) limitação da dependência entre eventos e ET, procurando minimizar as por vezes complexas interacções entre o tipo de construção e a ET que a modifica;

[2] http://www.timeml.org/tempeval/
[3] http://nlp.cs.swarthmore.edu/semeval/
[4] http://www.google.com/trends
[5] http://www.timeml.org/site/index.html

(iv) independência entre a tarefa de delimitação das ET e o tratamento da subcategorização verbal, o que nos levou a propor a inclusão de certas preposições na EM;

(v) adopção de critérios claros de atomização das ET;

(vi) adesão ao princípio de classificar antes de resolver a referência temporal;

(vii) normalização parcial das ET, isto é, apresentar para um conjunto de situações, suficientemente claras, uma proposta de normalização, deixando para momento posterior o tratamento de outras expressões; do mesmo modo, permitir que uma expressão para a qual está disponível apenas parte da informação necessária à sua adequada normalização seja, ainda assim, normalizada pelo menos parcialmente;

(viii) os agregados temporais[6] não são, por ora, considerados, dada a sua especificidade;

(ix) tentar assegurar o critério de intersubjectividade máxima na anotação, procedendo sempre que possível à listagem e/ou descrição intensional dos elementos lexicais que entram na formação das ET.

2.2 Proposta para o Segundo HAREM

2.2.1 Delimitação das ET

A fim de se poder anotar de maneira unívoca as entidades da categoria TEMPO, convém definir rigorosamente os critérios sintáctica e semanticamente motivados que deverão ser seguidos a fim de se poder delimitar com precisão as fronteiras das entidades a anotar. Neste sentido, a proposta que apresentámos representa uma evolução e modificação importantes relativamente à estratégia adoptada no Primeiro HAREM (Cardoso e Santos, 2007, pp. 223-225).

Assim, nesta proposta, considera-se que deverá ser delimitada entre as balizas <EM ID=... CATEG="TEMPO"> e </EM> a totalidade da expressão temporal, isto é, **incluindo a preposição** que a introduzir, no caso da expressão temporal ser um sintagma preposicional (e.g. *no ano passado*), **ou o determinante** no caso de ser um sintagma nominal (e.g. *todos os dias*).

Por detrás desta opção está a noção de que na maioria das ET, os elementos ditos gramaticais (preposições e determinantes, sobretudo) são não apenas parte integrante destas locuções, apresentando muitas delas um elevado grau de fixidez combinatória interna, como contribuem de modo crucial para a classificação das ET nos diferentes tipos da categoria TEMPO. Naturalmente, este tipo de decisão acarretou, sobretudo por uma questão de coerência mas também de simplicidade, que se incluíssem nas EM certas preposições que não fazem parte da ET propriamente dita mas que são seleccionadas (regidas) por outros elementos lexicais (operadores). Tal sucede, sobretudo, nos casos das ET genéricas, como se pode ver em (2.12).

(2.12) Eu gosto <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">**do Verão**</EM>.

[6] Um agregado temporal é uma expressão complexa que inclui simultaneamente valores de DATA e de FREQUENCIA, como, por exemplo: no primeiro domingo de cada mês.

A preposição *de*, neste caso, enquanto elemento que introduz o complemento de *gostar*, em nada contribui para a interpretação da EM. Considerámos, no entanto, que o tratamento das regências verbais (para usar um termo mais tradicional) deveria constituir um problema distinto, a resolver independemente do reconhecimento das EM.

2.2.2 Delimitação das ET complexas

Decidimos também integrar na EM certos elementos gramaticais, tradicionalmente analisados como advérbios, que entram na formação de ET complexas:

(2.13) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias depois**</EM>.

De facto, este tipo de ET complexa é formado por dois elementos: uma expressão quantificadora do tipo `DURACAO` (*alguns dias*) e o adverbial *depois*. Esta última forma pode introduzir outros constituintes ligando-se-lhes por meio da preposição *de* e, assim, receber diferentes análises consoante seja seguida de uma oração (conjunção subordinativa temporal), como em (2.14), ou de um grupo nominal (locução prepositiva ou preposição composta), como em (2.15).

(2.14) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias**</EM> depois de ter ido ver o futebol.

(2.15) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias**</EM> depois do jogo.

Na construção com a locução prepositiva, distinguimos ainda duas situações: a primeira, como no exemplo (2.15), em que o núcleo do sintagma nominal é um nome qualquer; e uma segunda situação, ilustrada no exemplo (2.16), em que esse sintagma é preenchido por um nome de tempo (voltaremos a este último caso já adiante).

(2.16) O Pedro fez isso <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**alguns dias depois do domingo**</EM>.

Considerámos, por princípio, que não nos cabia propor qualquer análise unificada (e coerente) deste tipo de fenómeno, mas sim determinar com rigor as regras de delimitação das EM.

Assim, uma vez que, nesta fase de desenvolvimento da tarefa do HAREM dedicada à categoria `TEMPO`, tomámos a decisão de excluir as orações subordinadas, apenas utilizamos a informação da conjunção para determinar o atributo `SENTIDO` com que será anotada a EM (ver adiante).

No caso da locução prepositiva, seguimos critério idêntico, excluindo apenas os casos que envolvem um complemento com nomes de tempo, na medida em que estas expressões complexas exigem uma análise mais subtil.

De facto, no caso de expressões complexas como *dois dias depois do Natal*, a questão que se coloca é a de se saber se esta expressão deverá ser considerada como uma só EM ou, então, segmentada em duas subexpressões *dois dias* + *depois do Natal* (obedecendo tanto a

expressão mais longa como ambas as subexpressões aos critérios definitórios mencionados acima). Neste sentido, verifica-se que uma expressão como *dois dias depois do Natal*, ilustrada no exemplo (2.17) é ambígua podendo ter duas leituras distintas, a que correspondem duas análises sintácticas diferentes (e logo diferentes atomizações).

(2.17) Vimo-nos *dois dias depois do Natal*.

(a) *Vimo-nos no dia 27 de Dezembro*
Vimo-nos <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**dois dias depois do Natal**</EM>.

(b) *Vimo-nos durante dois dias, a seguir ao 25 de Dezembro*
Vimo-nos <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**dois dias**</EM> <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL">**depois do Natal**</EM>.

Os critérios adoptados para a segmentação são os definidos em Hagège e Tannier (2007) e que aqui foram reproduzidos:

> Uma expressão temporal complexa **deverá ser dividida** em unidades menores se se verificarem **simultaneamente** os critérios seguintes:
>
> 1. cada expressão componente é sintacticamente válida quando combinada independentemente com o evento que modifica.
> 2. cada expressão componente, combinada com o evento que modifica, está logicamente implicada na expressão complexa. Ou seja, cada combinação "evento mais expressão_temporal_mínima" deve ser logicamente implicada pela combinação "evento + expressão_temporal_complexa".

Ora, no caso da frase ambígua (2.17), o primeiro critério pode aplicar-se tanto na leitura (a) como na leitura (b), acima:

> Vimo-nos *dois dias* (DURACAO).
> Vimo-nos *depois do Natal* (DATA).

mas o segundo critério não se observa, já que o valor de duração está ausente da leitura complexa (a), que acima glosamos. Ainda assim, neste caso, parece-nos que, embora a presença do segundo membro tenha tendência em "forçar" a leitura complexa da expressão temporal (DATA), em última análise, a ambiguidade deverá ficar expressa na anotação a adoptar futuramente.

As expressões de tempo foram organizadas em quatro grandes tipos:

- as expressões de **localização temporal**, de tipo TEMPO_CALEND;
- as expressões de **quantificação temporal**, de tipo DURACAO;
- as expressões de **frequência**, de tipo FREQUENCIA;
- as ET **genéricas**, de tipo GENERICO.

De um modo geral, esta organização clássica das ET conserva, no essencial, as definições do Primeiro HAREM (Cardoso e Santos, 2007)[7], conquanto se tenha procurado, nesta proposta, precisar e definir com maior rigor alguns dos seus aspectos. Em seguida, apresentaremos, de forma sucinta, cada um destes tipos, remetendo o leitor para o texto da proposta (Hagège et al., 2008), que também se encontra reproduzido no anexo B.

2.2.3 TEMPO_CALEND

As entidades de tipo TEMPO_CALEND são expressões que permitem inserir ou localizar o predicado que elas modificam numa linha temporal (como um ponto ou um intervalo). Correspondem aos seguintes subtipos:

- **datas**, sejam elas **absolutas** (fórmulas contendo os três campos ANO-MES-DIA, nas quais até dois campos no máximo podem ser omitidos) ou **referenciais** (ET cuja resolução implica conhecer a data do momento da enunciação, ou conhecer a data de um outro evento que funciona então como referência temporal para a expressão a calcular).
- **horas** (ET com valor de DATA mas com granularidade inferior à unidade *dia*).
- **intervalos** (expressões denotando uma duração no tempo e que têm explicitamente dois limites).

2.2.3.1 Data

As expressões deste subtipo podem representar *datas absolutas* ou *datas referenciais*[8].

Datas absolutas

As ET constituem *datas absolutas* quando contêm a informação necessária para localizar essa data num calendário. Assim, por exemplo, na expressão *em 23 de Outubro de 2007*, a informação está totalmente especificada em relação aos três campos <dia>, <mês> e <ano>; pelo contrário, nas expressões *em 23 de Outubro* e *em 2007*, a informação está apenas parcialmente especificada em relação aos três campos. Apresentam-se de seguida alguns exemplos de ET do tipo TEMPO_CALEND e subtipo DATA:

- Data absoluta completa (campos dia, mês e ano preenchidos):
  Vou viajar <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***no dia 19 de Outubro de 2007***</EM>.
- Data absoluta incompleta (campos dia e mês não preenchidos)[9]:
  Trabalhei em Londres <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">***em 1998***</EM>.

[7] Relativamente ao Primeiro HAREM, são eliminados os tipos PERIODO, CICLICO, que passam, de um modo geral, a estar integrados em TEMPO_CALEND.

[8] A organização da referência das expressões temporais aqui sucintamente apresentada já é, de resto, bem conhecida. Veja-se, entre outros, Gross (1986) com especial referência a advérbios compostos (ou expressões adverbiais multipalavras), sobretudo de natureza idiomática, e Molinier e Levrier (2000), este último a propósito de advérbios de tempo terminados em *–mente* (*futuramente*, *anteriormente*, *posteriormente*, etc.).

[9] As ET com datas em que apenas os campos <dia>, <mês> ou <dia><mês> estão preenchidos (e.g. *no dia 8*, *em Setembro*, *a 8 de Setembro*) são, em rigor, datas referenciais, cujo valor exacto é relativo ao momento da enunciação. Nesse sentido, será necessário modificar o critério que determina se o valor de TEMPO_REF deve ser ABSOLUTO (ver adiante).

Datas referenciais

Também são consideradas como abrangidas pelo subtipo DATA as expressões que exprimem *datas referenciais*, isto é, para as quais é necessário determinar um ponto de referência para poder localizá-las na linha temporal (e.g. *dois dias mais tarde, na quinta-feira passada, ontem, na próxima terça feira*, etc.). Vejamos, agora, os dois tipos de ET referenciais consideradas: as ET que fazem referência ao momento da enunciação e aquelas que se referem ao tempo de um evento presente no discurso. Um exemplo típico desta distinção pode ser dado através dos exemplos (2.18) e (2.19), respectivamente.

(2.18) O Pedro chegou *ontem*.

(2.19) O barco chegou *no dia anterior*.

Nestes dois exemplos, estamos perante ET que permitem localizar no calendário o evento a que estão associadas, respondendo adequadamente à interrogativa *quando?*. Pode-se, pois, associar a estas expressões o valor SUBTIPO="DATA". Contudo, não se trata aqui de datas absolutas mas sim de expressões referenciais cujo valor tem de ser calculado relativamente a outra referência temporal.

No primeiro exemplo, (2.18), esta referência é o momento da enunciação. Com efeito, se a asserção *O Pedro chegou ontem* for produzida no dia 4/12/2007, pode-se inferir que o evento *chegou* ocorreu no dia 3/12/2007. O tempo em que o evento ocorre, neste exemplo, é função do tempo do momento da enunciação (tempo_enunciação – 1 dia). Fala-se, pois, neste caso, de uma *expressão temporal referencial relativa ao momento da enunciação*.

No segundo exemplo, (2.19), embora também se trate de uma data referencial, a sua referência não é o momento da enunciação, já que a localização temporal de *chegou* é independente do momento em que for produzida a asserção. Neste caso, a referência é outra data/evento que aparece no contexto discursivo. A título ilustrativo, considere-se o exemplo (2.20).

(2.20) O barco só devia chegar ao porto *no dia 25 de Novembro*, no entanto chegou *no dia anterior*.

Como se pode ver, a referência da expressão *no dia anterior* é a data do evento da chegada do barco ao porto, que deveria ter ocorrido no dia 25/11. Conhecendo esta referência pode-se então deduzir que o evento *chegou* ocorreu no dia 24/11. Assim, neste caso está-se em presença de uma *expressão temporal com referência textual*, isto é, uma data relativa a uma outra data explícita no texto.

Esta distinção entre data absoluta, data referencial relativa ao momento de enunciação e data referencial relativa a uma referência textual é formalizada através do atributo TEMPO_REF. No caso de datas absolutas, o valor do atributo TEMPO_REF é ABSOLUTO. No caso de datas referenciais, conforme o tipo da referência o valor do atributo TEMPO_REF é, respectivamente, ENUNCIACAO ou TEXTUAL.

Finalmente, no caso de algumas ET referenciais, é ainda possível acrescentar outra informação complementar com vista à normalização das ET. Trata-se dos atributos SENTIDO e VAL_DELTA. O atributo SENTIDO indica se o seu valor temporal se situa cronologicamente *antes*, *em simultâneo* ou *depois* do *tempo de referência*.

O atributo VAL_DELTA tem por valor uma expressão que indica a distância temporal entre o tempo do evento denotado pela expressão temporal e o momento de referência, seja este o tempo da enunciação ou outro, quando esta distância temporal aparece explicitamente no texto (sobre a normalização destas expressões, ver adiante).

Os exemplos (2.21) a (2.24) ilustram o uso dos atributos TEMPO_REF, SENTIDO e VAL_DELTA e alguns dos seus possíveis valores.

(2.21) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO">**a 3 de Janeiro de 1986**</EM>.

(2.22) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S0D2H0M0S0">**dois dias depois**</EM>.

(2.23) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="ANTERIOR" VAL_DELTA="A0M0S0D2H0M0S0">**dois dias antes do Natal**</EM>.

(2.24) O Pedro nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ENUNCIACAO" SENTIDO="ANTERIOR">**na sexta-feira passada**</EM>.

2.2.3.2 Hora

Trata-se de ET com valor de DATA mas com granularidade inferior à unidade *dia* (ver exemplo (2.25)).

(2.25) O Pedro está disponível <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA" VAL_NORM="+--------T1500E--LMA">**às 15:00**</EM>.

A existência deste subtipo de datas pode justificar-se pelo facto de constituírem uma classe natural de expressões, que seguem um conjunto de convenções gráficas particulares, facilmente modelizáveis por uma gramática própria, distinta da dos outros tipos de datas. Neste sentido, a proposta apresentada ao Segundo HAREM conservou esta distinção entre data e hora.

2.2.3.3 Intervalo

Corresponde a uma expressão complexa, isto é, composta por duas ET elementares/simples mas que, semanticamente, formam um única EM, e que tem explicitamente dois limites temporais (um limite inicial e um limite final), como ilustram os exemplos (2.26) e (2.27).

(2.26) Trabalhei em Londres <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**entre 2000 e 2003**</EM>.

(2.27) Trabalhei em Londres <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**de Outubro a Dezembro de 2007**</EM>.

Note-se que, nesta avaliação conjunta, não se levou em consideração a granularidade das expressões de tempo que constituem os limites explícitos do intervalo. Assim, por exemplo, integram este tipo de ET formas com granularidade inferior à unidade *dia*, tal como em (2.28).

(2.28) O escritório fecha para almoço <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**das 12:00 às 14:00 horas**</EM>.

Por outro lado, incluímos ainda no tipo INTERVALO não só expressões complexas com datas, como as dos exemplos acima, mas combinações que exprimem outros valores temporais como, por exemplo, a duração, em (2.29).

(2.29) Vai demorar <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO">**de 3 a 6 meses**</EM>.

As ET do tipo INTERVALO não foram normalizadas nesta avaliação conjunta, dada a complexidade de que se revestem algumas das suas formas, nomeadamente as que combinam ET dos tipos DATA e HORA. Veja-se o exemplo (2.30).

(2.30) O Pedro esteve a fazer isso isso *desde a meia-noite de 5 de Dezembro de 2007 até ao dia de Natal, ao meio-dia*.

Nesse sentido, será de esperar que algumas destas questões venham a ser resolvidas pelas propostas que apresentamos no fim deste capítulo. Tal permitiria igualmente dar também alguns passos no sentido da normalização das ET do tipo INTERVALO.

Além das expressões TEMPO_CALEND, consideraram-se ainda dentro da categoria TEMPO as expressões de **duração** e de **frequência**, de que trataremos já a seguir.

2.2.3.4 Duração

Corresponde a uma expressão TEMPO que se refere a uma duração de tempo contínuo. Ao contrário das datas, trata-se de expressões que não exprimem propriamente a localização (ou calendarização) de um evento, mas sim uma *quantificação temporal*, sendo constituídas por nomes de unidades de medida de tempo e determinantes com função de quantificadores (numerais, por exemplo). Podem, por vezes, ser introduzidas, facultativamente, pela preposição *durante* (encontrando-se também outras preposições) e respondem adequadamente à interrogativa *(prep) quanto tempo?*. Ver exemplos (2.31) a (2.35).

(2.31) Fiquei <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**dois meses**</EM> em Lisboa.

(2.32) O urso fica <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**todo o inverno**</EM> na toca.

(2.33) O Pedro trabalhou <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**várias semanas**</EM> no restaurante.

(2.34) O Pedro trabalhou <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**durante três anos**</EM> na tese.

(2.35) A aplicação da lei será suspensa <EM ID="..." CATEG="TEMPO" TIPO="DURACAO">**por dez anos**</EM>.

2.2.3.5 Frequência

O tipo FREQUENCIA corresponde a expressões TEMPO que exprimem uma repetição de um evento no tempo. Estas expressões respondem adequadamente às interrogativas do tipo *com que frequência?*, como ilustram os exemplos (2.36) a (2.40).

(2.36) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**amiúde** </EM>.

(2.37) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**diariamente** </EM>.

(2.38) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**todos os dias**</EM>.

(2.39) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**duas vezes por semana**</EM>.

(2.40) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**dia sim dia não**</EM>.

Como se pode ver pelos exemplos acima, as ET deste tipo podem ser advérbios simples, derivados de adjectivos (*diariamente*) ou não (*amiúde*), locuções adverbiais mais ou menos cristalizadas (*dia sim dia não*), certas expressões com forma de sintagma nominal (*todos os dias*) e outras construções em torno de nomes como *vez* (*duas vezes por semana*). Incluem-se ainda neste tipo de ET certos advérbios que têm sobretudo um valor aspectual (*frequentemente*, *pontualmente*, *ocasionalmente*, *raramente*). Contudo, a definição deste tipo de ET é ainda insuficiente para dar conta de expressões cujo significado global parece combinar o valor de frequência com o de localização temporal, como acontece em (2.41).

(2.41) A reunião de pais tem lugar *todas as primeiras segundas-feiras de cada mês*.

2.2.3.6 ET genéricas

Trata-se de expressões TEMPO que não se referem a uma data específica embora a expressão linguística integre elementos lexicais que denotam um valor temporal, como nos exemplos (2.42) e (2.43).

(2.42) Adoro <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">**o Verão**</EM>.

(2.43) <EM ID="..." CATEG="TEMPO" TIPO="GENERICO">**Fevereiro**</EM>é o mês mais curto do ano.

Estas expressões genéricas podem, como se sabe, ter um papel relevante no cálculo de referências temporais, pelo que importa identificá-las adequadamente. Por ora, contudo, elas não são normalizadas.

2.3 Normalização

A normalização das datas absolutas e horas, como ilustrado no exemplo (2.44), obedece ao seguinte formato:

<Era><Ano><Mes><Dia>**T**<Hora><Minuto>**E**<ESTACAO>**LM**<limite_aberto>

Onde:

- <Era> corresponde a um caracter que indica se a data é depois ou antes da nossa era;
- <Ano> corresponde a quatro algarismos que representam o valor do ano;
- <Mes> corresponde a dois algarismos que representam o valor do mês;
- <Dia> corresponde a dois algarismos que representam o valor do dia;
- <Hora> corresponde a dois algarismos que representam o valor da hora;
- <Minuto> corresponde a dois algarismos que representam o valor dos minutos;
- <ESTACAO> corresponde a duas letras maiúsculas referentes às estações do ano;
- <limite_aberto> indica se a expressão normalizada de data absoluta introduz um intervalo de tempo com limite anterior ou limite posterior não determinado (em aberto). Os valores respectivos são "A", no caso de limite *anterior* em aberto; ou "P", no caso de limite *posterior* em aberto.

Exemplo:

(2.44) Nasceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+19860103T----E--LM-">**a 3 de Janeiro de 1986**</EM>.

2.3.1 Normalização de datas referenciais

Como já se disse atrás, algumas ET referenciais recebem uma outra informação complementar com vista à sua normalização. Trata-se dos atributos SENTIDO e VAL_DELTA. O atributo SENTIDO indica se o seu valor temporal se situa cronologicamente *antes*, *em simultâneo* ou *depois* do *tempo de referência*. Os possíveis valores do atributo SENTIDO são, pois:

ANTERIOR, POSTERIOR, SIMULT, ANTERIOR_OU_SIMULT, POSTERIOR_OU_SIMULT.

O atributo VAL_DELTA corresponde ao valor temporal que se deve incrementar ou subtrair a partir do tempo de referência para obter o valor temporal do evento associado à expressão temporal a anotar, quando esta distância temporal aparece explicitamente no texto. No caso de esta distância temporal não estar explícita, o valor de VAL_DELTA é omitido. Tal como ilustrado em (2.45), os valores possíveis de VAL_DELTA são representados da maneira seguinte:

A<digitos>**M**<digitos>**S**<digitos>**D**<digitos>**H**<digitos>**M**<digitos>**S**<digitos>

Onde:

- as letras **A**, **M**, **S**, **D**, **H**, **M**, **S** são constantes que devem aparecer nesta ordem e marcam, respectivamente, a posição dos valores de anos, meses, semanas, dias, horas, minutos e segundos.
- os <digitos> à direita das letras constantes correspondem ao número de anos, meses, semanas, dias, horas, minutos e segundos que se devem adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Exemplo:

(2.45) Apareceu <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="POSTERIOR" VAL_DELTA="A0M0S2D0H0M0S0">**duas semanas**</EM> depois da festa.

2.3.2 Normalização da DURACAO

Para expressões de tipo DURACAO, a normalização exprime uma distância temporal representada com o seguinte formato:

A<digitos>**M**<digitos>**S**<digitos>**D**<digitos>**H**<digitos>**M**<digitos>**S**<digitos>

Onde:

- as letras **A**, **M**, **S**, **D**, **H**, **M**, **S** são constantes que devem aparecer nesta ordem e marcam, respectivamente, a posição dos valores de anos, meses, semanas, dias, horas, minutos e segundos;
- os <digitos> à direita das letras constantes correspondem ao número de anos, meses, semanas, dias, horas, minutos e segundos que se devem adicionar ou diminuir à data de referência para obter o valor temporal da expressão anotada.

Exemplo:

(2.46) Fiquei <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="A0M2S0D0H0M0S0">**dois meses**</EM> em Lisboa.

Para terminar esta secção, uma breve nota apenas para indicar que a proposta de normalização das ET ainda *não* contemplou, neste momento, as expressões do tipo FREQUENCIA nem o subtipo INTERVALO do tipo TEMPO_CALEND. Estes dois aspectos deverão ser aprofundados em futuras edições do HAREM. Por um lado, é possível normalizar, pelo menos parcialmente, alguma da informação veiculada pelas expressões de FREQUENCIA, indicando, nomeadamente, entre outros valores, a granularidade do intervalo entre instâncias do evento modificado e o número de repetições desse evento. Por outro lado, no caso dos intervalos, é possível normalizar cada um dos limites temporais.

2.4 A experiência do Segundo HAREM

Com uma primeira versão, nos seus traços gerais, já bastante próxima da versão final, que ficou disponível logo a 18 de Dezembro de 2007, a elaboração, discussão e redacção final da proposta foi um processo longo e complexo que culminou no documento ora disponível no sítio da avaliação conjunta do Segundo HAREM (13 de Abril de 2008). Produziu-se nessa altura (14 de Abril) uma versão dos primeiros 10% da CD do Mini-HAREM anotada segundo as directivas do TEMPO, que foi distribuída aos participantes para treino e discussão[10].

Participaram na pista do TEMPO sete dos dez participantes no HAREM, embora se verifiquem diferenças relativamente à forma como cada um se apresentou[11]:

- seis sistemas com `TIPO`;
- cinco sistemas com `SUBTIPO`;
- dois sistemas com `TEMPO_REF` (tipo de referência para datas referenciais);
- um sistema com a normalização.

Como primeira conclusão a tirar, inevitavelmente, deste perfil de participação, recomenda-se prudência e moderação no desenvolvimento da tarefa para futuras avaliações conjuntas de TEMPO, o que não impede, naturalmente, que se introduzam melhoramentos ou mesmo correcções.

Do ponto de vista dos resultados[12], e de acordo com o modo de avaliação em que todos os sistemas participaram (TEMPO clássico), é possível fazer algumas observações gerais: na tarefa de classificação (cf. figura 2.1), verifica-se que apenas dois sistemas apresentam resultados de precisão consistententemente acima de 0,7 (máximo 0,767); em termos de abrangência, apenas um sistema apresenta valores acima de 0,7 (máx. 0,758), embora duas das respectivas corridas apresentem valores cerca de dez por cento inferiores; já o segundo melhor sistema em abrangência, embora com resultados consistentes, só consegue valores pouco superiores a 0,5 (entre 0,533 e 0,489); finalmente, os mesmos dois sistemas apresentam resultados consistentes em termos de medida F: o primeiro, com valores superiores a 0,7 (máx. 0,748) e o segundo na casa dos 0,6 (máx. 0,618).

Na tarefa de identificação, e como se pode verificar pela figura 2.2, os melhores sistemas apresentam resultados relativos em grande medida semelhantes aos acima relatados (verificam-se os máximos de 0,769 de precisão, 0,758 de abrangência e 0,747 de medida F).

2.5 Próximos passos e perspectivas futuras

Nesta secção, apresentamos os aspectos que, na sequência da experiência do Segundo HAREM, nos parece relevante tratar, em termos de perspectivas de investigação e desen-

[10] Este fragmento anotado faz parte da LÂMPADA - Pacote de Recursos do Segundo HAREM (http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip).

[11] Remetemos o leitor para os capítulos 1 e 3, para uma descrição mais pormenorizada dos cenários de participação dos sistemas e modos de avaliação, bem como para o capítulo 5, que inclui a descrição da avaliação da pista do TEMPO.

[12] Os valores aqui apresentados correspondem aos disponíveis em http://www.linguateca.pt/HAREM, ver Resultados do Segundo HAREM, e são arredondados à terceira casa decimal.

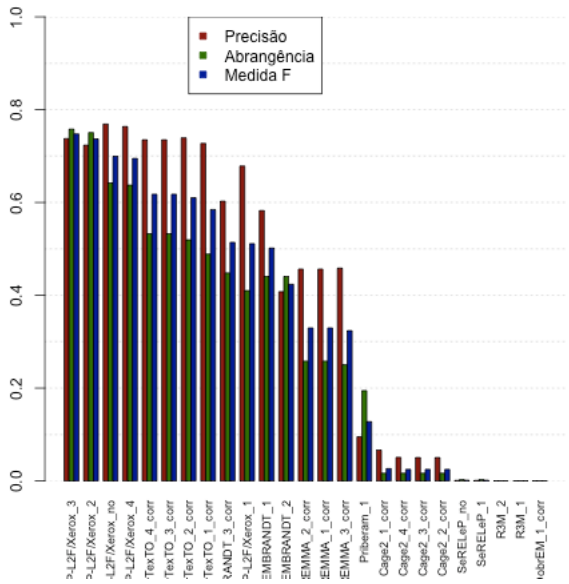

Figura 2.1: Resultados do HAREM clássico no cenário selectivo TEMPO na CD do TEMPO, tarefa de classificação.

volvimento futuros. Duas preocupações norteiam estas sugestões que assim submetemos à apreciação da comunidade de PLN do português:

Em primeiro lugar, corrigir ou melhorar alguns aspectos da proposta actual da avaliação conjunta do Segundo HAREM. Trata-se de observações que fomos recolhendo ao longo do trabalho desenvolvido, bem como várias sugestões recebidas tanto de outros participantes como da parte da organização.

Em segundo lugar, garantir uma continuidade, tanto quanto possível suave, entre as sucessivas edições das avaliações conjuntas de sistemas de REM/TEMPO, por forma a garantir a novos actores uma mais fácil integração neste processo, estabilizando os standards e potenciando os recursos e ferramentas entretanto construídos. Não esquecemos que, nesta edição do Segundo HAREM, parte dos sistemas participantes (ainda?) não integrou todas as dimensões da nossa proposta, nomeadamente aquele que era o seu principal desafio: o de ir além da tarefa de REM e tratar também a normalização das ET. Seria, no mínimo, inadequado fazer evoluir a proposta sem um consenso e participação alargados da comunidade[13]. Neste sentido, as linhas que se seguem podem ser interpretadas como um mapa do caminho para uma futura edição HAREM/TEMPO.

[13] Neste sentido, a equipa L2F/Xerox veria com naturalidade que uma nova avaliação conjunta, Terceiro HAREM, se realizada num prazo relativamente curto, se limitasse para já a repetir a experiência do Segundo HAREM.

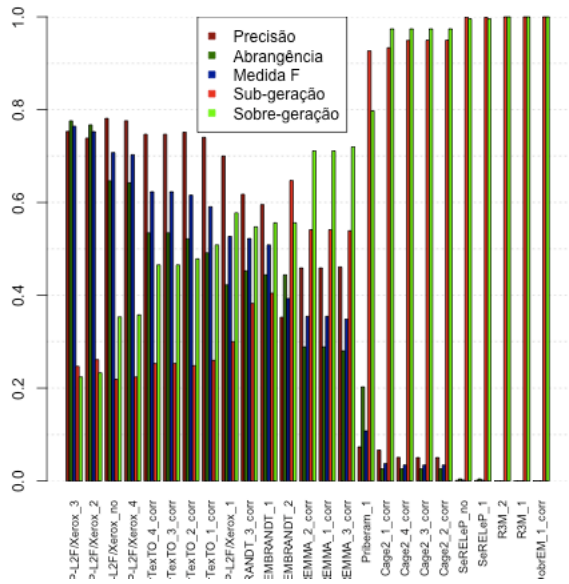

Figura 2.2: Resultados do HAREM clássico no cenário selectivo TEMPO na CD do TEMPO, tarefa de identificação.

2.5.1 TEMPO_CALEND

2.5.2 Novo subtipo=DATA

Propõe-se agregar no tipo `DATA` os actuais subtipos `DATA` e `HORA`. São várias as motivações para esta evolução: em primeiro lugar, semanticamente, ambos os tipos correspondem à localização dos eventos numa linha do tempo, a única diferença entre eles é a granularidade da unidade temporal; a normalização é basicamente a mesma: há campos comuns em cada um dos subtipos e uma representação única irá simplificar a normalização das ET do tipo `INTERVALO` quando os respectivos limites são expressos simultaneamente com datas e horas.

Note-se que uma das motivações principais para conservação da distinção dos subtipos `DATA` e `HORA` prendia-se com as gramáticas (ou regras) usadas para a sua identificação. Na medida em que se pretende orientar a actual proposta no sentido de evoluir para lá da tarefa de REM e passar a incluir também a normalização, não só essa motivação perde alguma da sua razão de ser como se ganha em obter uma normalização uniforme.

Esta alteração implica a (relativamente ligeira) reformulação dos critérios de atomização das ET. Assim, no quadro da actual proposta, considerou-se que nos casos (2.47) e (2.48) se estava em presença de várias ET.

(2.47) Isto aconteceu <EM ... SUBTIPO="DATA">**na sexta-feira**</EM>, <EM ... SUBTIPO="DATA">**23 de Abril de 2008**</EM>, <EM ... SUBTIPO="HORA">**pelas 18:30**</EM>

(2.48) Isto aconteceu <EM ... SUBTIPO="DATA">**na sexta-feira**</EM>, <EM ... SUBTIPO="DATA">**23 de Abril de 2008**</EM>, <EM ... SUBTIPO="DATA">**dia de São Jorge**</EM>, <EM ... SUBTIPO="HORA">**pelas 18:30**</EM>

Nestes exemplos, cada ET é, de acordo com os actuais critérios de atomização, identificada e normalizada separadamente. Contudo, esta forma de representação não é inteiramente adequada, pois trata-se de sequências de ET numa cadeia de aposição, em que cada nova ET precisa ou desenvolve as referências temporais das ET anteriores, pelo que deveriam constituir *uma única referência temporal*. Por outro lado, a estrutura de aposição permite resolver imediatamente alguns dos valores referenciais não absolutos: por exemplo, enquanto a ET *na sexta feira* teria, à partida, um valor refencial relativo ao momento de enunciação, quando integrado nesta sequência apositiva ela é mera informação complementar, dispensando o cálculo da referência temporal, na medida em que se subordina ao valor referencial absoluto da ET de data adjacente, e.g., *23 de Abril de 2008*. Além disso, certas dificuldades de classificação levantadas por ET como *dia de São Jorge*, que poderiam ser incorrectamente classificadas no tipo `GENERICO` podem ser evitadas, já que também esta ET é mera informação adicional à data absoluta adjacente.

A aceitar-se estes argumentos, o critério geral para separar/juntar ET deverá ser alterado de modo a permitir tratar instâncias de `DATA` e `HORA` em aposição como uma única ET, desde que a sua normalização seja complementar:

(2.49) Isto aconteceu <EM ... SUBTIPO="*DATA">**na sexta-feira, 23 de Abril de 2008, pelas 18:30**</EM>

(2.50) Isto aconteceu <EM ... SUBTIPO="*DATA">**na sexta-feira, 23 de Abril de 2008, dia de São Jorge, pelas 18:30**</EM>

em que `*DATA` corresponde ao novo tipo unificado.

2.5.2.1 subtipo=INTERVALO

Propõe-se a normalização das ET do subtipo `INTERVALO`, tais como as apresentadas nos exemplos (2.51) a (2.54), que neste momento não são normalizadas. Para este tipo de situações, como se vê nos exemplos, a normalização poderia ser feita duplicando os pares atributo-valor e dando índices numéricos a cada um dos limites temporais explícitos do `INTERVALO`.

(2.51) O Pedro está de férias <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ABSOLUTO" VAL_NORM1="+----0423T----E--LM" TEMPO_REF2="ABSOLUTO" VAL_NORM2="+----0529T----E--LM">**de 23 de Abril a 29 de Maio**</EM>.

(2.52) O Pedro está de férias <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ABSOLUTO" VAL_NORM1="+ 20090423T - - - - E - - LM" TEMPO_REF2="ABSOLUTO" VAL_NORM2="+ 20090529T - - - - E - - LM">**entre 23 de Abril e 29 de Maio de 2009**</EM>.

(2.53) O Pedro está de férias <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ENUNCIACAO" SENTIDO1="SIMULT" VAL_DELTA1="A0M0S0D0H0M0S0"

TEMPO_REF2="ENUNCIACAO" SENTIDO2="POSTERIOR" VAL_DELTA2="A0M0S1D0H0M0S0">**desde hoje até à próxima semana**</EM>.

(2.54) O Pedro está de baixa <EM ID="..." CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="INTERVALO" TEMPO_REF1="ENUNCIACAO" SENTIDO1="SIMULT" VAL_DELTA1="A0M0S0D0H0M0S0" TEMPO_REF2="ENUNCIACAO" SENTIDO2="POSTERIOR" VAL_DELTA2="A0M0S0D2H0M0S0">**entre hoje e depois de amanhã**</EM>.

2.5.2.2 Novo subtipo=COMPLEXO

Sugere-se a eventual criação de um novo subtipo `COMPLEXO` dentro do tipo `TEMPO_CALEND`, que deverá capturar ET que incluem os conceitos de `DATA` e `VAL_DELTA`:

(2.55) *Faz (hoje, no dia 21 de Dezembro) quinze dias* que isso aconteceu.

(2.56) Isso acontecerá *de (hoje, ontem) a quinze dias*.

2.5.3 DURACAO

Propõe-se que a normalização das ET do tipo `DURACAO` passe a incluir uma unidade menor que o segundo (milissegundos), a fim de permitir o tratamento adequado de, por exemplo, resultados desportivos.

2.5.3.1 tipo=DURACAO subtipo=INTERVALO

O subtipo `INTERVALO` é, na actual proposta, um tipo híbrido pois não integra apenas ET que exprimem uma localização temporal (`TIPO="TEMPO_CALEND"`), desde que apresentem dois limites temporais explícitos, como também abrange expressões de tempo que denotam outras formas de modicação temporal, nomeadamente expressões de `DURACAO`:

(2.57) Isso durou *entre 2 e 3 horas*.

Tal solução não é, pois, inteiramente adequada. Uma solução possível seria que o tipo `DURACAO` passasse a incluir o subtipo `INTERVALO`, por forma a dar conta de situações como as ilustradas no exemplo acima. A normalização deste tipo de intervalos far-se-ia de modo análogo ao dos intervalos com datas (ver acima), através da duplicação de `VAL_NORM` e atribuição de índices aos pares atributo valor:

(2.58) Isso durou <EM ID="..." CATEG="TEMPO" TIPO="DURACAO" SUBTIPO="INTERVALO" VAL_NORM1="A0M0S0D0H2M0S0" VAL_NORM2="A0M0S0D0H3M0S0">**entre 2 e 3 horas**</EM>.

2.5.4 FREQUENCIA

Propõe-se passar a normalizar de forma explícita um determinado conjunto de ET do tipo `FREQUENCIA`. Tomamos como modelo deste tipo de ET expressões como a do exemplo seguinte:

(2.59) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**duas vezes por semana**</EM>.

Para normalização da FREQUENCIA propõe-se usar dois atributos suplementares de EM[14]:

- VAL_QUANT, que indica o número de vezes em que o evento/processo se repete; e
- VAL_MODULO, que representa a *granularidade* dessa frequência.

O primeiro atributo seria preenchido por valores numéricos e o segundo por uma notação semelhante à já usada na normalização da DURACAO:

A<digitos>M<digitos>S<digitos>D<digitos>H<digitos>M<digitos>S<digitos>

Deste modo, a expressão acima ilustrada seria normalizada como em (2.60). Da mesma forma, as ET do tipo FREQUENCIA ilustradas nos exemplos (2.61) a (2.65), passariam a ser normalizadas de acordo com este formato.

(2.60) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="2" VAL_MODULO="A0M0S1D0H0M0S0">**duas vezes por semana**</EM>.

(2.61) O Pedro faz isso <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="1" VAL_MODULO="A0M0S1D0H0M0S0">**semanalmente**</EM>.

(2.62) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="1" VAL_MODULO="A0M0S0D1H0M0S0">**diariamente**</EM>.

(2.63) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="1">**todos os dias**</EM>A0M0S0D1H0M0S0.

(2.64) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="1">**dia sim dia não**</EM>A0M0S0D2H0M0S0.

(2.65) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="1">**todas as semanas**</EM>A0M1S0D0H0M0S0.

Para expressões complexas (agregados temporais) como:

na primeira quinta-feira de cada mês, *quatro domingos seguidos*, *dez dias interpolados*

que incluem tanto o conceito de DATA como de FREQUENCIA, ou para expressões em que nomes como *vez(es)* aparecem determinados por um quantificador indefinido:

(*várias*, *muitas*, *algumas*, *umas poucas*, *poucas*, *bastantes*, *imensas*) *vezes por semana*

ou ainda para expressões em que não é possível determinar com rigor esse quantificador, como sucede na ET *todas as semanas*, sugere-se que só o campo MODULO seja normalizado, como se ilustra nos exemplos (2.66) e (2.67).

(2.66) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_QUANT="not_defined" VAL_MODULO="A0M1S0D0H0M0S0">**na primeira quinta-feira de cada mês**</EM>.

[14] Esta proposta é fortemente inspirada na TimeML (Boguraev et al., 2005).

(2.67) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA" VAL_NORM="not_defined">**algumas vezes por semana** </EM>A0M1S0D0H0M0S0.

Naturalmente, continuariam por normalizar expressões que veiculam valores vagos ou imprecisos, sobretudo os que são expressos por certos adverbiais como *amiúde*, *frequentemente*, *ocasionalmente*, etc.:

(2.68) Vou ver os meus pais <EM ID="..." CATEG="TEMPO" TIPO="FREQUENCIA">**amiúde** </EM>.

2.5.5 Outras sugestões

Além das sugestões acima apresentadas, julgamos que seria oportuno e não demasiado complexo introduzir alguns pequenos melhoramentos na normalização das ET.

2.5.5.1 Not_Norm

Propõe-se a inclusão de uma propriedade que explicite a distinção entre, por um lado, as expressões que, por qualquer razão, não foram normalizadas pelo sistema, das que se definiu como não devendo ser normalizadas de todo. Assim, por exemplo, apenas o advérbio de `FREQUENCIA` *frequentemente* deveria receber este traço, ao contrário de *semanalmente*, que deveria ser normalizado pelos diferentes sistemas.

2.5.5.2 Indefinição (ou vagueza)

Propõe-se a inclusão de uma propriedade que explicite a existência de vagueza em algumas categorias, como acontece com as ET dos exemplos (2.69) a (2.71).

(2.69) O Pedro fez isso *por volta do dia 23 de Abril de 2008*.

(2.70) O Pedro fez isso *perto das 3 da tarde*.

(2.71) O Pedro fez isso *em pouco tempo*.

Em futuras avaliações conjuntas é necessário estender este conceito para tornar mais clara a granularidade da imprecisão temporal da ET.

2.6 Conclusões

Apresentámos neste capítulo a proposta de tarefa de reconhecimento, classificação e normalização de expressões temporais para a segunda avaliação conjunta de sistemas de reconhecimento de entidades mencionadas – o Segundo HAREM. Trata-se de uma proposta de algum modo conservadora na medida em que preserva, embora procure definir com maior precisão, grande parte da estrutura de classificação de ET do Primeiro HAREM.

Ao mesmo tempo, a proposta introduz diversos aspectos inovadores, sobretudo no que diz respeito à delimitação das ET e a normalização das ET, esta última tendo em vista o cálculo de referências temporais. Tratou-se de dar um primeiro passo no sentido de associar as ET aos eventos e estados de coisas que elas modificam, a fim de os ordenar parcialmente, numa sequência cronológica. Contudo, procurámos intencionalmente garantir que

estes aspectos inovadores mantivessem um certo grau de simplicidade, evitando uma excessiva (porque demasiado súbita) descontinuidade com a tarefa do Primeiro HAREM e permitindo uma participação o mais abrangente possível da comunidade do PLN.

Procurámos, além disso, reflectir, ainda que de forma breve, sobre a experiência deste Segundo HAREM. Com base no perfil de participação dos vários sistemas em jogo, parece-nos necessário adoptar prudência e moderação no desenvolvimento da tarefa para futuras avaliações conjuntas de TEMPO, o que não impede, naturalmente, que se introduzam melhoramentos ou mesmo correcções. Do ponto de vista dos resultados, é possível considerar que, de um modo geral, a fasquia do estado da arte, para a classificação de ET, se situa em valores na ordem dos 0,75 para a precisão, abrangência e medida F. Contudo, o conjunto dos sistemas participantes apresenta ainda grandes disparidades nos resultados obtidos, quer entre si, quer entre as diferentes medidas.

Como resultado da experiência deste Segundo HAREM, apresentámos, finalmente, um conjunto de propostas que procuram corrigir ou melhorar aspectos da classificação e normalização das ET, na perspectiva de uma nova avaliação conjunta de entidades mencionadas. Como nota final, referimos que nestas propostas se deixa para um outro ciclo de avaliação o cálculo da referência temporal: não porque não se julgue esta tarefa importante – lembramos ser este o objectivo que pretendemos alcançar com a proposta de normalização das ET –, mas porque consideramos ser necessário e mais proveitoso adoptar uma estratégia de progressão em pequenos (mas firmes) passos, a fim de que se possa manter um grupo de investigadores interessados e activos nesta linha de avaliação.