

Capítulo 3

É tempo de avaliar o TEMPO

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

No capítulo 2 foi apresentada a pista de identificação, classificação e normalização de expressões temporais adoptada no Segundo HAREM, que designaremos daqui em diante como pista do TEMPO, e cujas directivas (Hagège et al., 2008) se encontram republicadas no apêndice B. Uma vez que os autores da proposta referiram desde o início a sua intenção de serem participantes nesta avaliação conjunta, a tarefa de anotação da colecção dourada e de avaliação dos sistemas ficou inteiramente a cargo da organização do Segundo HAREM. O objectivo deste capítulo é então documentar essas duas actividades, dando também conta dos resultados obtidos.

Começamos por discutir as principais questões que tivemos de resolver durante o processo de anotação das expressões temporais na colecção dourada e que se resumem a dois tipos de situações: casos que não nos pareceram suficientemente claros nas directivas e casos que, por verificarem mais do que um critério, tiveram de ser decididos por nós. Embora alguns desses casos tenham sido posteriormente esclarecidos pelos autores da proposta, mantivemos a sua documentação aqui dado que os participantes e outros leitores podem ter as mesmas dúvidas. Além disso, apresentamos critérios adicionais que estabelecemos para situações não previstas nas directivas. Em vários casos, aproveitamos para exprimir o nosso ponto de vista discordante, na perspectiva de enriquecer genuinamente o estudo e processamento temporal do português.

Na próxima secção, fornecemos igualmente dados quantitativos sobre as entidades anotadas com a categoria TEMPO. Em seguida, apresentamos sucintamente os modos de avaliação desta pista (o leitor poderá encontrar informação mais detalhada sobre a avaliação no capítulo 5) e os resultados de desempenho obtidos pelos sistemas. Finalmente, tecemos algumas sugestões sobre novas versões de uma nova pista do TEMPO, quer fazendo uma autocrítica sobre a forma de avaliação como propondo alguns trabalhos futuros.

3.1 Anotação da colecção dourada

O processo de anotação das expressões temporais na colecção dourada (CD) do Segundo HAREM decorreu em duas fases. Numa primeira fase, todas as expressões que verificavam os critérios estabelecidos nas directivas do TEMPO foram anotadas com os atributos do HAREM clássico, ou seja, CATEG, TIPO e SUBTIPO. Essa anotação decorreu em simultâneo com a anotação das entidades pertencentes às restantes categorias previstas nas directivas do HAREM clássico, tal como descrito no capítulo 1.¹

Numa fase posterior, foi seleccionado um subconjunto de documentos, mais precisamente trinta (cf. secção 3.2 para uma descrição detalhada), nos quais se adicionou, às entidades classificadas como TEMPO, os atributos SENTIDO, TEMPO_REF, VAL_DELTA e VAL_NORM, específicos das directivas do TEMPO, e que designaremos como atributos estendidos. Esta colecção de documentos foi baptizada CD do TEMPO.

Tendo em conta que os autores das directivas foram também participantes, o processo de anotação não pôde usufruir da colaboração directa de pelo menos um dos proponentes da pista do TEMPO. Assim, mesmo tendo como material de apoio seis documentos ano-

¹ O HAREM clássico tem, contudo, uma questão que não é clássica no sentido de que o critério de delimitação de entidades temporais, proposto nas directivas do TEMPO, é muito diferente do das outras categorias – como será mais amplamente discutido no capítulo 6.

tados pelos autores das directivas², um dos desafios da anotação feita por nós esteve em esclarecer as diversas dúvidas que foram surgindo, mas que não podiam ser discutidas (explicitadas) de forma directa, visto que não podíamos revelar o conteúdo da CD nem permitir a localização dos documentos correspondentes à colecção dourada na colecção HAREM. Tivemos sempre essa precaução, para não criar uma situação de desigualdade em relação aos restantes participantes.

Na maioria dos casos, as dúvidas com que nos deparámos prenderam-se com o facto de as entidades em análise parecerem não se enquadrar perfeitamente nos critérios previstos nas directivas de anotação do TEMPO, ou então poderem encaixar-se em mais do que um critério quanto à sua classificação. Naturalmente, o facto de ser a primeira vez que as directivas estavam a ser seguidas num processo de anotação contribuiu para que alguns pontos não estivessem ainda suficientemente explicitados.

Contudo, e como já referido, muitas vezes considerámos que as opções tomadas pela proposta do TEMPO não foram as melhores, e vozeamos a nossa crítica no presente capítulo. É muito importante contudo salientar que essa crítica tem como objectivo o futuro, e que durante a anotação tentámos sempre seguir da forma mais próxima possível as directivas acordadas. Ou seja, nunca tentámos anotar as expressões temporais de acordo com a nossa opinião quando esta divergia, mas simplesmente obedecer às directivas.

3.1.1 Opções relativas aos atributos do HAREM clássico de TEMPO

3.1.1.1 Delimitação da entidade quando a expressão temporal verifica os critérios 1 e 2-6

Quando uma expressão temporal verifica o subcritério 2-6, ou seja, quando

um sintagma preposicional cujo núcleo seja uma das palavras *altura*, *tempo*, *momento*, *período*, *era*, etc., quando

- estas palavras forem determinadas por um demonstrativo (por exemplo: *nesse tempo*),
- ou especificados por uma relativa (por exemplo: *na altura em que ela adoeceu*),
- um possessivo (por exemplo: *durante a nossa era*)
- ou modificado por outro sintagma preposicional introduzido por *de* (por exemplo: *durante a era dos dinossauros*)
- ou então por um adjectivo capitalizado (por exemplo: *durante o período Barroco, Cretáceo*, etc.);

(Hagège et al., 2008, reformatação nossa)

estamos perante uma expressão temporal que constitui toda ela (núcleo e respectivos modificadores) uma unidade sintáctica que responde adequadamente às interrogações <prep> quando? ou quando?.

² Estes documentos correspondem a 10% da CD do Mini-HAREM e foram disponibilizados aos participantes como material de treino. A anotação destes documentos foi copiosamente discutida e mutuamente esclarecida entre a organização do HAREM e o grupo do TEMPO em Novembro-Dezembro de 2007, o que permitiu tanto o refinamento das directivas como uma maior clarificação, embora contudo e como descrevemos no presente capítulo, ainda não totalmente suficiente para efectuarmos a anotação sem dúvidas.

Embora da nossa interpretação das directivas, para este caso específico, não fosse claro se se deveria incluir ou não os modificadores, e, de um ponto de vista semântico, a sua não inclusão iria forçar-nos a marcar expressões sem sentido, foi-nos indicado pelos proponentes que os modificadores preposicionais e oracionais não deveriam, de facto, ser tidos em consideração na etiquetagem das expressões temporais, excepto no caso dos modificadores adjectivais em maiúscula, com a seguinte justificação:

se assim não fosse, isso implicaria um processamento sintáctico complexo (que se prende, nomeadamente, com a identificação dos limites temporais das orações subordinadas), desviando-se da proposta de REM,

tal como discutido no capítulo 2.

Assim, nos exemplos (3.1) e (3.2) apenas anotámos o núcleo da expressão temporal, em vez de alargar a anotação também às expressões que representamos em itálico, o que no seu todo nos parecia ser de facto as expressões temporais; no exemplo (3.3), pelo contrário, a expressão temporal inclui então o adjectivo.

(3.1) Muitos milhões acabam, como <EM ID="cver-8" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**no tempo** *de Arafat*, em contas secretas

(3.2) acompanhando «o percurso da decadência, da perdição», do autarca alentejano <EM ID="hub-57257-6" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**até ao momento** *em que aceita dinheiro em troca de um favor*

(3.3) refrões fortes que traduziam o sentimento da juventude <EM ID="hub-77558-120" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**da era Cavaquista**

3.1.1.2 Delimitação da entidade quando a expressão temporal é constituída por DATA e HORA

Quando uma expressão é internamente composta por dois constituintes, um do tipo DATA e outro do tipo HORA, identificámos cada um desses constituintes como EM independentes, mesmo que um deles não se combine isoladamente com o predicado que modifica, como ilustram os exemplos (3.4) e (3.5).

(3.4) O provável primeiro bebé português do ano é do sexo masculino e nasceu <EM ID="hub-71248-191" CATEG="TEMPO" TIPO="TEMPO_CALEND">**aos 30 segundos** <EM ID="hub-71248-192" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**de hoje** na Maternidade Alfredo da Costa

(3.5) <EM ID="aa58069-369" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="HORA">**Às 17h20** <EM ID="aa58069-370" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**de ontem** em Lisboa, a vaga especulativa do mercado do petróleo convergiu com as previsões dos últimos meses

Em casos como estes, embora não se verifique um dos critérios para segmentação de uma expressão complexa em duas entidades³, pareceu-nos que este procedimento se justi-

³ O critério que não se verifica é o das expressões componentes serem ambas sintacticamente válidas quando combinadas com o evento que modificam. No primeiro caso não podemos ter “nasceu *de hoje*” (e parece-nos pouco aceitável ter “nasceu *aos 30 segundos*”), e no segundo não podemos ter “*de hoje* em Lisboa, convergiu”.

ficava porque da nossa interpretação das directivas não era claro se o subtipo `DATA` também englobava expressões complexas que incluíssem hora. Assim, tomando esta opção poderia dar-se mais valorização aos sistemas, nomeadamente se tivessem tido dificuldade em amalgamar estas duas entidades numa.

Naturalmente, poderíamos ter usado a notação dos `ALT` para produzir as duas segmentações na colecção dourada. Não o fizemos, principalmente, porque isso não estava previsto na proposta do `TEMPO`. Estamos agora convencidos que essa solução teria sido muito melhor, em vez de uma decisão arbitrária.

Ao contrário do proposto no capítulo 2 como futura melhoria, de existir apenas um subtipo `DATA*` que permitiria dar conta deste tipo de expressões complexas, defenderíamos uma notação que permitisse o encaixe das entidades `HORA` nas entidades `TEMPO`. Embora concordemos que semanticamente uma `HORA` pode ser uma `DATA`, estas expressões têm de facto uma sintaxe muito diferente, e julgamos que seria mais adequado manter a distinção entre ambas.

3.1.1.3 Classificação como `GENERICO`

Um dos critérios que nos levantou mais dúvidas e que, conseqüentemente, provocou maior discordância entre anotadores foi o da classificação de uma entidade como `GENERICO`⁴: em 92 entidades deste tipo, 22 têm o atributo `COMMENT` preenchido com os valores `2/3` ou `DUVIDA_DIRECTIVASTEMPO`⁵, num total de 1204 expressões temporais das quais 83 foram marcadas da mesma forma.

De acordo com o critério 3, uma expressão temporal deveria ser deste tipo se verificasse um dos subcritérios do critério 2 e não verificasse o critério 1. Ou seja, se a expressão (lexicalmente) contivesse elementos temporais, mas não fosse uma resposta adequada a uma das interrogativas previstas no critério 1: (`<prep>`) *quando?*, (`<prep>`) *quanto tempo?*, (`<haber>`) *quanto tempo?* ou *com que frequência?*.

Considere-se o exemplo (3.6).

(3.6) Lápiz-lazúli, conhecido também como lápis, é uma rocha metamórfica de cor azul utilizada como gema ou como rocha ornamental utilizada desde antes de 7000 a.C. em Mehrgarh na Índia, situado `<EM ID="H2-dhy6432-141" CA-TEG="TEMPO" TIPO="GENERICO">nos dias de hoje` no Paquistão

Embora a expressão *nos dias de hoje* corresponda a um locativo temporal, o critério sintáctico supra-mencionado não parece poder aplicar-se:

**quando é que estava [Mehrgarh] situada no Paquistão? / nos dias de hoje*

Por esse motivo, e de acordo com as directivas, acabámos por atribuir a classificação de `GENERICO` a esta expressão e a outras cujo par pergunta-resposta nos parecia duvidoso ou mal-formado.

⁴ Achamos, a este respeito, que a denominação `GENERICO` não é apropriada e pode induzir em erro, já que genérico tem um sentido concreto bem diferente em linguística e na área do tempo e do aspecto em particular (consulte-se, por exemplo, Krifka et al. (1995), Dahl (1973) e, em português, Lopes e Santos (1993)).

⁵ Como referido no capítulo 1, o valor `2/3` é usado para marcar entidades cuja classificação não resultou do total acordo dos anotadores; usámos `DUVIDA_DIRECTIVASTEMPO` para indicar entidades em que as anotadoras tiveram dúvidas, geralmente associadas a diferentes interpretações possíveis das directivas.

Contudo, houve casos em que apesar de termos achado pouco natural a formulação das interrogações com *<prep> quando?* ou *quando?*, considerámos que esta era mesmo assim mais aceitável do que no caso anterior, como é o caso de expressões temporais que modificam um outro sintagma nominal sem valor temporal.

Veja-se, por exemplo, as expressões *da década de 1920* e *nos anos 1950* na frase (3.7).

(3.7) No Brasil, eles remontam ao século dezenove, com o grupo dos românticos em São Paulo, os grupelhos de poetas simbolistas, os modernistas <EM ID="gtqqq-168" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**da década de 1920**, o grupo antropofágico, os concretistas <EM ID="gtqqq-169" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**nos anos 1950**, o coletivo Rex de artistas na década seguinte

Tanto no primeiro caso, como no segundo, pareceu-nos possível, apesar de questionável, formular o par pergunta-resposta:

*?*os modernistas de quando? / da década de 1920*

*?*os concretistas quando? / nos anos 1950*

Assim, optámos por não classificar estas expressões como GENERICICO.

Como ilustra este exemplo, a questão da aceitabilidade/inaceitabilidade dos pares pergunta-resposta interfere não só com a classificação como GENERICICO, como também com as restantes classificações, pois se o par pergunta-resposta não for aceitável então a expressão temporal tem o tipo GENERICICO, noutros casos a expressão receberá outra classificação (DATA, FREQUENCIA, etc.).⁶

3.1.1.4 Classificação como DURACAO

De acordo com as directivas, as entidades com tipo DURACAO referem “uma duração de tempo contínuo” e correspondem a entidades que exprimem

quantificação temporal, sendo constituídas por nomes de unidades de medida de tempo e determinantes com função de quantificadores (e.g.. numerais). Podem, por vezes, ser introduzidas, facultativamente, pela preposição *durante* e respondem adequadamente à interrogativa (*<prep> quanto tempo?*).

(Hagège et al., 2008)

Como nos pareceu que a leitura estrita deste critério levaria à exclusão de expressões que não contivessem unidades de medida de tempo, e como são dados dois exemplos de expressões deste tipo que não as incluem (*todo o inverno* e *três manhãs*), de facto, acabámos por classificar como DURACAO, expressões que não incluíam unidades de medida temporal.

⁶ Refira-se, ainda, que Jorge Baptista, na sua recensão ao presente capítulo, discordou dos nossos juízos de valor quanto à gramaticalidade/aceitabilidade dos pares pergunta-resposta, até no modo como os pares foram formulados. No caso (3.6), sugeriu a formulação da pergunta usando outro verbo copulativo, o que tornaria, no seu entender, o par mais aceitável:

?**quando é que [Mehrgrah] passou a estar situado no Paquistão? / nos dias de hoje*

Parece-nos pois importante referir que poderá haver mais casos de perguntas que permitam estabelecer que uma dada expressão representa uma data do que aquelas mencionadas nas directivas.

Para tal, teriam de responder adequadamente à interrogação (<prep>) *quanto tempo?* e verificar pelo menos um dos subcritérios 2. Como exemplos, veja-se (3.8) e (3.9).

(3.8) Detroit tem <EM ID="2ght33-10" CATEG="TEMPO" TIPO="DURACAO">**por longo tempo** sido um lugar de referência na imaginação sônica.

(3.9) Passa a viver com a avó Dionísia e as duas tias na Rua da Bela Vista, 17. A mãe e o padrasto também retornam a Lisboa <EM ID="aa87333-155" CATEG="TEMPO" TIPO="DURACAO">**durante um período** de férias <EM ID="aa87333-156" CATEG="TEMPO|TEMPO" TIPO="DURACAO|GENERICO">**de um ano**

3.1.1.5 Classificação de expressões iniciadas por *há*

No caso de expressões temporais iniciadas por *há*, optámos pelo valor durativo sempre que a formulação com *durante* fosse possível, e por tempo de calendário nos casos em que a expressão de tempo respondesse exclusivamente a (<prep>) *quando?*. Vejam-se os exemplos (3.10) a (3.12).

(3.10) o pensamento cartesiano <EM ID="H2-bbb-231" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**de há quatro séculos**

(3.11) o CCB iniciava, <EM ID="Ntyr-78-400" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA">**há quinze anos**, a sua actividade

(3.12) é um projeto que vem sendo realizado <EM ID="bob-14949-607" CATEG="TEMPO" TIPO="DURACAO">**há mais de dois anos**

Em dois casos (um deles ilustrado no exemplo (3.13)), em que ambas as interpretações nos pareceram possíveis, marcámos ambas. Apesar de as directivas do tempo não preverem a marcação de vagueza, pareceu-nos adequado nesta situação tirar partido dessa característica do esquema de anotação do HAREM, em vez de ter de optar arbitrariamente por uma das análises.

(3.13) nesta comemoração de uma data que deve ser pretexto para uma renovação da nossa ligação com o público que conosco habita, todos os dias, <EM ID="Ntyr-78-100" CATEG="TEMPO|TEMPO" TIPO="TEMPO_CALEND|DURACAO" SUBTIPO="DATA|">**de há 15 anos para cá**

Refira-se ainda a propósito do exemplo (3.13) que, de certo modo, vemos toda a sequência *todos os dias, de há 15 anos para cá* como uma expressão temporal complexa que denota um valor de frequência que só está completamente definido se tivermos em conta toda a expressão. Ou seja, esta frequência só é válida *de há 15 anos para cá* e como tal os dois valores de frequência e duração são indissociáveis.⁷

⁷ De acordo com as directivas, contudo, apenas a expressão *todos os dias* deveria ser anotada com o tipo FREQUENCIA.

3.1.1.6 Ausência de anotação relativa a TEMPO

O subcritério 2-2 permite a identificação de expressões temporais no caso de

uma unidade de medida temporal (*dia, mês, trimestre, ano, século, etc.*) ou um advérbio terminado em *mente* derivado destas expressões (*diariamente, semanalmente, mensalmente, etc.*).

(Hagège et al., 2008)

No entanto, não está previsto que adjectivos derivados dos nomes de unidades de medida o sejam. Por esse motivo, não anotámos expressões temporais que envolvessem esses adjectivos, nem os próprios adjectivos, mesmo que a expressão completa tivesse um valor de frequência, como no exemplo (3.14).

(3.14) A partir de maio de 2000 estará sendo lançada a Revista de Direitos Difusos, com mais de 100 páginas e *periodicidade bimestral*

Pareceu-nos que, nesses casos, também não seria aplicável o critério 2-8:

expressões de frequência, como as seguintes: de vez em quando, às vezes, de quando em quando, frequentemente, etc.,

(Hagège et al., 2008)

Um outro caso em que não anotámos como TEMPO foi o de nomes de acontecimentos usados com valor temporal. O problema destas expressões foi que correspondiam a casos em que nos parecia que o sentido era temporal, mas não estavam previstos nas directivas do TEMPO. No entanto, só nos apercebemos disso depois de termos anotado consistentemente esses casos com TEMPO na primeira versão da CD.

Uma vez que não nos parecia muito aceitável adicionar tarde demais esta cláusula, e como iria flagrantemente contra a filosofia do HAREM anotá-los como ACONTECIMENTO se estavam a indicar tempo, anotámos então essas entidades como OUTRO, em vez de termos usado a possibilidade oferecida pelo Segundo HAREM de classificar como CATEG="TEMPO" TIPO="OUTRO".

Assim, não anotámos como pertencendo à categoria TEMPO períodos históricos, como sejam os casos de *Idade Média* (veja-se o exemplo (3.15)) ou *Descobrimientos*.

(3.15) No seguimento do colapso de instituições monásticas e do escolasticismo nos finais da <EM ID="H2-dftre765-102" CATEG="OUTRO">**Idade Média**

No caso de *Idade Média*, em particular, considerámos que não se verificava o critério 2-6, como nos casos de:

“altura, tempo, momento, período, era, etc.”

pois não achamos que a palavra *idade* na expressão *Idade Média* tenha as mesmas propriedades das palavras lá referidas: com efeito, ao contrário delas, *idade* não nos parece poder ser determinada por um possessivo (*nessa idade...*), modificada por uma relativa (*na idade em que...*) ou complementada por um sintagma preposicional (*na idade de...*). Ainda se poderia argumentar que está a ser modificada por um adjectivo em maiúscula, mas trata-se de um composto. Como o núcleo nominal (o próprio nome composto) não está a

ser também ele modificado por um adjetivo, não é modificado por uma relativa, nem é complementado por um outro sintagma preposicional, optámos por não anotar *nos finais da Idade Média* como TEMPO. Em vez disso, conforme já dito, anotámos expressões como esta com a categoria OUTRO.

Finalmente, também não anotámos expressões que, embora verificassem pelo menos um dos subcritérios 2, se encontrassem no contexto de uma pergunta. Nestes casos, não nos pareceu fazer sentido formular um par pergunta-resposta para poder verificar se estávamos perante o critério 1 ou 3. Veja-se, a título de exemplo, a frase interrogativa (3.16).

(3.16) Em que *ano* é que Torquemada foi nomeado Inquisidor Geral?

3.1.2 Opções relativas aos atributos do TEMPO estendido

3.1.2.1 Tensão entre dois tipos de DATA

Por vezes, tomámos opções de delimitação com consequências inesperadas, ou limitadoras, na marcação dos atributos estendidos do TEMPO. Por exemplo, considere-se os exemplos (3.17) e (3.18)

(3.17) o que traduz um crescimento de 2,4%, <EM ID="hub-51467-348" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+—01-T—E-LMP">**a partir de Janeiro do próximo ano**

(3.18) promulgado <EM ID="hub-18050-209" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="+—30T—E-LM">**a 30 desse mês**

O problema destes casos é que são datas que obedecem a ambos os critérios definidos nas proposta do TEMPO como mutuamente exclusivos, e que repetimos aqui:

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).

(Hagège et al., 2008)

Ou seja, os dois critérios definidos são, num primeiro caso, puramente sintácticos (ou mesmo lexicais), e no segundo totalmente semânticos, e nada garante a sua mútua exclusão.

Como se pode verificar, o caso da expressão *a partir de Janeiro do próximo ano* tanto inclui uma referência ao nome de um mês como exige saber em que momento foi enunciada para que se possa proceder à sua resolução. O mesmo se passa com *a 30 desse mês*, que inclui a referência a um dia (30) de um mês cuja localização na linha temporal é fornecida pelo resto do discurso.

Dado que, de acordo com as directivas do TEMPO, a divisão de tais expressões em duas EM temporais não era possível, tivemos de decidir se considerávamos as EM em causa como “referenciais” ou como “absolutas”. A decisão foi arbitrária e recaiu na segunda escolha.

Não nos parece, contudo, que a correcta análise das expressões em questão passasse pela separação em duas partes independentes. Pelo contrário, o que nós defenderíamos era a interpretação da entidade como um todo (que, nos dois casos, implicava que a data não fosse absoluta).

3.1.2.2 Expressões com valor de data sem nenhum dos campos ANO-MES-DIA especificado

De acordo com as directivas, as datas com valor absoluto devem ter pelo menos explicitado um dos campos ANO-MES-DIA. No entanto, pareceu-nos que, tal como no exemplo dado nas directivas, *na era [dos dinossauros]*, que tem valor absoluto sem ter nenhum desses campos explicitados, expressões temporais que não dependessem de outras expressões temporais no texto seriam igualmente anotadas como datas absolutas.

As expressões *No início do século XVI* e *nos anos 90* nas frases (3.19) e (3.20), respectivamente, são exemplos de casos anotados como datas absolutas.

(3.19) <EM ID="hub-66526-557" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="">**No início do século XVI** o rei D. Manuel I ordena uma grande reforma

(3.20) Carlos Gerbase faz parte de uma geração de cineastas que apareceu em Porto Alegre <EM ID="ric-54609-190" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="">**nos anos 90**

O problema com estes casos é que os campos do atributo VAL_NORM não permitem normalizar estas datas. No caso de *nos anos 90* (exemplo (3.20)), temos uma referência à década de 90, mas o ano em causa pode não ser 1990. Por esse motivo, não especificámos o campo ano do atributo VAL_NORM com esse valor. Note-se, além disso, que apesar de ter valor absoluto, por se tratar de uma referência à década de 90, cuja resolução não depende de uma outra data, no contexto em questão não se trata de uma referência a um ponto específico na década de 90, mas sim a um intervalo de tempo⁸. Por exemplo, na frase (3.21), a mesma expressão refere-se a um ponto específico no tempo, por se tratar de um estreia que ocorre num dia particular, ou até a mais pontos se pensarmos que o mesmo filme pode estreiar em vários sítios em datas diferentes.

(3.21) O primeiro filme de Carlos Gerbase estreou <EM ID="ex-903" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="ABSOLUTO" VAL_NORM="">**nos anos 90**

3.1.2.3 Preenchimento de VAL_DELTA e VAL_NORM na ausência total de informação

Quando não existia informação que permitisse preencher pelo menos um dos campos de VAL_DELTA ou VAL_NORM, optámos por preencher esse atributo na colecção dourada com o valor "", em vez de o omitir⁹.

(3.22) <EM ID="aa94781-176" CATEG="TEMPO" TIPO="DURACAO" VAL_NORM="">**Há anos** que se discute se a fotografia da autoria do famoso fotógrafo

⁸ Usamos aqui uma noção mais alargada de intervalo do que a definida nas directivas do tempo, não obrigando a que os limites sejam explícitos.

⁹ As directivas sugeriam que se poderia optar pela omissão do parâmetro em vez do preenchimento com "".

Tabela 3.1: Dados quantitativos sobre a CD do TEMPO

Parâmetro	Valor
Documentos	30
Parágrafos	304
Palavras	12992
Entidades	1508
Entidades vagas	118
Entidades TEMPO	232

(3.23) o miliciano fotografado não é o que foi morto <EM ID="aa94781-192" CATEG="TEMPO" TIPO="TEMPO_CALEND" SUBTIPO="DATA" TEMPO_REF="TEXTUAL" SENTIDO="SIMULT" VAL_DELTA="">naquele dia

Este é um pormenor meramente técnico, mas que documentamos aqui para uma maior clareza na especificação da anotação efectuada.

3.2 O TEMPO em números no Segundo HAREM

No capítulo 1 caracterizámos a CD do Segundo HAREM, a qual inclui 1195 entidades TEMPO (das quais seis são vagas com outra categoria) que foram anotadas com os atributos do HAREM clássico: CATEG, TIPO e SUBTIPO.

A CD do TEMPO, tal como mencionámos anteriormente, foi constituída com o objectivo de adicionar às expressões temporais os atributos estendidos do TEMPO e corresponde a uma sub-colecção de trinta documentos da colecção dourada. Desses documentos, doze correspondem aos documentos que constituem a CD do ReReLEM, e os restantes dezasseis são os primeiros documentos da colecção dourada que não incluem entidades TEMPO marcadas com 2/3 ou DUVIDA_DIRECTIVASTEMPO.

Tínhamos inicialmente previsto anotar apenas 10% dos documentos da colecção dourada com os atributos estendidos do TEMPO e também do ReReLEM, o que corresponderia a doze documentos. No entanto, esses documentos incluíam apenas 85 entidades TEMPO (o que corresponde a menos de 10% das entidades TEMPO) e sobre algumas delas pesavam ainda dúvidas de anotação ou o não total acordo dos anotadores. Optámos, então, por seleccionar mais uma série de documentos que não tivessem essa última característica e que além disso incluíssem cerca de 10% das entidades TEMPO da colecção dourada. Foram igualmente considerados três documentos que permitissem aumentar um pouco a representividade das entidades do tipo HORA, já que o objectivo desta CD era ser usada na avaliação dos atributos estendidos. Acabámos por juntar todos os documentos na CD do TEMPO, cobrindo assim cerca de 19% das entidades TEMPO.

Esta sub-colecção é assim constituída por 304 parágrafos e 12992 palavras; das 1508 entidades anotadas, 232 estão anotadas com a categoria TEMPO (cf. tabela 3.1). Além disso, das 118 entidades vagas, nenhuma envolve a categoria TEMPO e existem 89 sequências delimitadas com ALT, das quais cinco envolvem a categoria TEMPO.

Comparando a distribuição das categorias na CD (já ilustrada no capítulo 1, mas que aqui se reproduz na figura 3.1(a)), e na CD do TEMPO (cf. figura 3.1(b)), podemos ver que as entidades marcadas com a categoria TEMPO correspondem, em ambos os casos, a

cerca de 15% do total de entidades. Notamos, no entanto, que isso não foi um critério que tivéssemos estabelecido à partida.

Podemos igualmente ver-se que a distribuição das categorias na CD do TEMPO é diferente da observada na CD do Segundo HAREM. Por exemplo, *ORGANIZACAO* é a quarta categoria mais frequente na CD do Segundo HAREM, correspondendo a cerca de 14% das entidades, enquanto, na CD do TEMPO, *ORGANIZACAO* é a segunda mais frequente, compreendendo quase 19% das entidades. No que se refere à categoria *TEMPO*, verifica-se que esta é a quarta categoria mais frequente na CD do TEMPO, apesar de ser a terceira mais frequente na CD do Segundo HAREM.

Não temos contudo a certeza de tal constatação ser relevante, a não ser pelo facto de poder ser uma explicação para a avaliação na CD do TEMPO - em relação ao HAREM clássico (ou seja, excluindo a avaliação dos atributos estendidos do TEMPO) - ter resultado numa ordenação diferente dos sistemas em termos de desempenho.

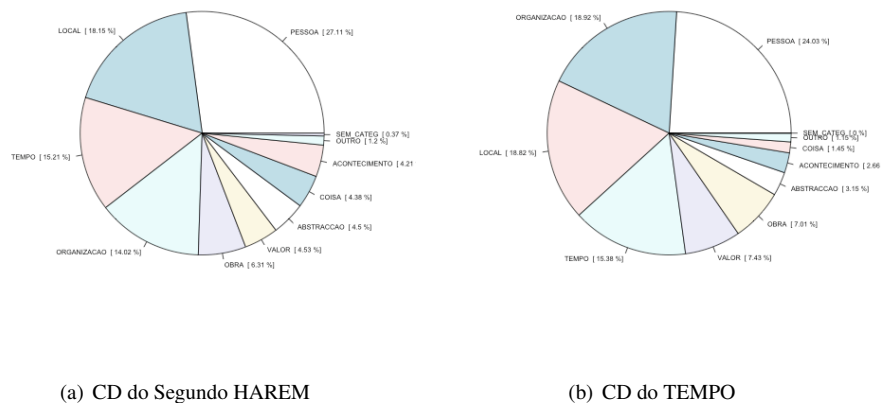


Figura 3.1: Distribuição de categorias na CD do Segundo HAREM e na CD do TEMPO

A distribuição de tipos da categoria *TEMPO* na tabela 3.2 mostra uma prevalência de expressões temporais do tipo *TEMPO_CALEND* (mais de 80%) tanto na CD como na CD do TEMPO, das quais a grande maioria são datas (cerca de 89% na CD e aproximadamente 84% na CD do TEMPO). Talvez por termos forçado a inclusão de mais entidades com o subtipo *HORA* na CD do TEMPO, como referido acima, este subtipo é mais frequente do que o subtipo *INTERVALO* nessa CD do que na CD do Segundo HAREM.

A figura 3.2 ilustra a distribuição dos atributos estendidos na CD do TEMPO. Em particular, a figura 3.2(a) ilustra a distribuição dos valores do atributo *TEMPO_REF*. Como se pode observar, existe um maior uso de expressões temporais com valor absoluto (cerca de 66% das entidades tiveram o atributo *TEMPO_REF* preenchido com valor *ABSOLUTO*) do que com valor referencial: cerca de 25% das vezes o atributo *TEMPO_REF* tem o valor *ENUNCIACAO* e cerca de 9% tem o valor *TEXTUAL*. No que diz respeito ao atributo *SENTIDO* (ver figura 3.2(b)), temos que, na maioria das entidades com valor referencial, esse atributo foi preenchido com

Tabela 3.2: Distribuição de tipos (T) e subtipos (ST) da categoria TEMPO: dentro de parêntesis encontram-se tipos e subtipos de outras categorias que não TEMPO

TIPO	SUBTIPO	CD		CD do TEMPO	
		T	ST	T	ST
TEMPO_CALEND		973		195	
	DATA		873		164
	INTERVALO		63		12
	HORA		37		19
GENERICO		89		12	
FREQUENCIA		71		15	
DURACAO		56		12	
DURACAO (QUANTIDADE)		3		–	
TEMPO_CALEND DURACAO		4		–	
	DATA		4		–
DURACAO GENERICO		1		–	
(EFEMERIDE) GENERICO		1		–	
TEMPO_CALEND (QUANTIDADE)		1		–	
	INTERVALO		1		–
TEMPO_CALEND GENERICO		1		–	
	DATA		1		–

o valor ANTERIOR (41%) e SIMULT (34%), enquanto apenas em 23% dos casos foi preenchido com o valor POSTERIOR.

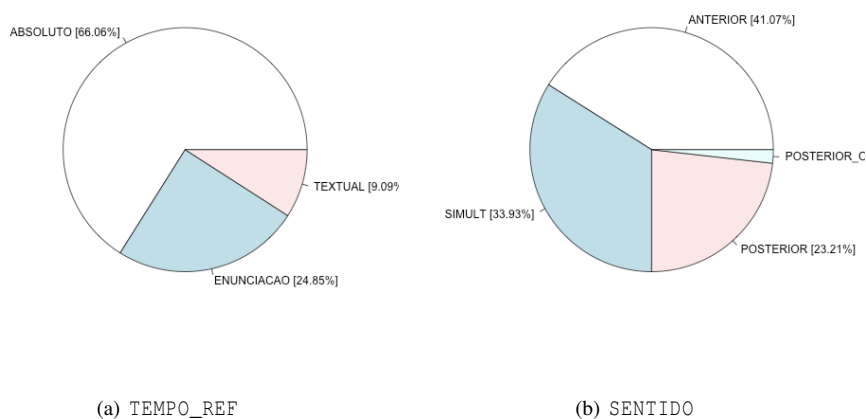


Figura 3.2: Distribuição dos atributos TEMPO_REF e SENTIDO na CD do TEMPO

Quanto aos atributos de normalização, a tabela 3.3 mostra que uma parte significativa das expressões temporais referenciais não explicita a distância temporal ao momento da

Tabela 3.3: Preenchimento dos atributos de normalização VAL_DELTA e VAL_NORM na CD do TEMPO

Atributo	Preenchido	Vazio	Total
VAL_DELTA	36	19	55
VAL_NORM de DURACAO	8	4	12
VAL_NORM de HORA	19	0	19
VAL_NORM de ABSOLUTO	101	8	109

referência, já que cerca de um terço dos valores de VAL_DELTA correspondem à sequência vazia. Numa proporção muito menor, vê-se na mesma tabela que para as data absolutas em cerca de 7% dos casos também não foi possível determinar nenhum dos campos do atributo VAL_NORM.

Finalmente, pode-se observar na figura 3.3, que as datas absolutas cujo campo referente ao ano se encontra especificado se distribuem entre 1131 e 2011. As datas concentram-se sobretudo na passagem do século XV para o século XVI, em meados do século XX e no início do século XXI, o que sugere que os documentos da CD do TEMPO descrevem sobretudo eventos decorridos nesses períodos.

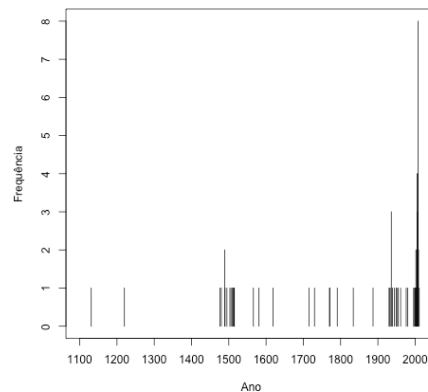


Figura 3.3: Histograma dos anos especificados nas datas absolutas na CD do TEMPO

3.3 Avaliação

Embora a proposta de reconhecimento e normalização de expressões temporais (pista do TEMPO) tenha sido feita de forma independente da proposta de reconhecimento das entidades pertencentes a outras categorias (HAREM clássico), a avaliação dos sistemas no que respeita às entidades TEMPO foi levada a cabo de forma integrada com o HAREM clássico.

Queremos com isto dizer que:

- não desenhamos todo um novo processo de avaliação exclusivo das entidades TEMPO. Pelo contrário, como descrito em pormenor no capítulo 5, integrámos apenas na

sequência de avaliação do HAREM clássico um novo módulo para atribuir uma pontuação adicional às entidades `TEMPO`, no caso dos atributos estendidos do `TEMPO` (`SENTIDO`, `TEMPO_REF`, `VAL_NORM` e `VAL_DELTA`) estarem correctamente preenchidos. Aliás, se não fosse o facto de existirem atributos específicos de `TEMPO`, as entidades `TEMPO` teriam sido avaliadas como se se tratassem de entidades de outra categoria qualquer.

- não separámos a avaliação das entidades pertencentes a outras categorias da avaliação das entidades `TEMPO`. Consequentemente, como vimos no capítulo 1, o HAREM clássico inclui diversos cenários em que as entidades da categoria `TEMPO` foram avaliadas, e, em particular, um dos cenários inclui apenas entidades `TEMPO`. Com ou sem separação, a avaliação da pista do `TEMPO` pode ser vista simplesmente como a avaliação num cenário selectivo constituído apenas por essa categoria.

Esta forma integrada de fazer a avaliação tem a vantagem de permitir ver a tarefa de reconhecimento de entidades mencionadas como um todo, não atribuindo um estatuto especial às entidades `TEMPO` por terem sido o alvo de uma proposta de anotação independente da das entidades pertencentes às restantes categorias.

Por essa razão, mesmo na CD do `TEMPO` fizemos a avaliação sem separar a avaliação do HAREM clássico da avaliação da pista do `TEMPO`.

Uma outra consequência adicional é que usaremos o termo “*avaliação da pista do TEMPO*” para designar apenas a avaliação no cenário `TEMPO`; em todos os outros casos de cenários que contêm a categoria `TEMPO`, referir-nos-emos à “*avaliação das entidades TEMPO*”.

Contudo, como as expressões temporais têm outros atributos, além dos atributos do HAREM clássico, os sistemas foram avaliados, no que respeita às entidades `TEMPO`, de quatro modos distintos:

Clássico, tendo em conta apenas os atributos `CATEG`, `TIPO` e `SUBTIPO` do HAREM clássico;

Estendido completo, tendo em conta todos os atributos;

Estendido sem normalização, ignorando os atributos `VAL_NORM` e `VAL_DELTA`;

Estendido só com normalização, ignorando os atributos `SENTIDO` e `TEMPO_REF`.

A avaliação no modo clássico foi feita tendo como referência tanto a CD do Segundo HAREM completa (cujos resultados no cenário total e nos vários cenários selectivos foram apresentados no capítulo 1), como apenas o subconjunto de documentos pertencentes à CD do `TEMPO` (neste último caso, quer dizer que os atributos estendidos de `TEMPO` não foram tidos em conta na avaliação)¹⁰; os restantes modos eram apenas aplicáveis no caso de se usar a CD do `TEMPO`.

Para cada um dos modos de avaliação usando a CD do `TEMPO`, avaliámos os sistemas no cenário total (ou seja, tendo em conta todas as categorias) e em todos os cenários selectivos de avaliação que incluem a categoria `TEMPO` (ou seja, nos cenários selectivos 2, 4 e 6, cuja descrição se encontra na tabela 1.3). Além disso, avaliámos ainda os sistemas apenas relativamente à categoria `TEMPO`.

Salientamos ainda que a avaliação na CD do `TEMPO` foi feita com base na avaliação estrita de `ALT`, apesar de também poder ser feita a partir da avaliação relaxada de `ALT` (ver

¹⁰ A avaliação no modo clássico foi feita também na CD do `TEMPO` para servir de referência à avaliação tendo em conta os atributos estendidos de `TEMPO`.

Tabela 3.4: Panorâmica da avaliação das entidades TEMPO: CD indica que a avaliação foi feita com a CD do Segundo HAREM e CDT que a avaliação foi feita com a CD do TEMPO

Cenário	Modo			
	Clássico	Estendido	Completo	Sem Norm. Só Norm.
Total	CD e CDT	CDT	CDT	CDT
Selectivos 2, 4 e 6	CD e CDT	CDT	CDT	CDT
TEMPO	CD e CDT	CDT	CDT	CDT

Tabela 3.5: Sistemas participantes na categoria TEMPO

Sistema	TIPO	SUBTIPO	SENTIDO	TEMPO_REF	Normalização
CaGE2					
PorTexTO	x	x			
PRiberam	x	x		x	
REMBRANDT	x	x			
REMMA	x				
SeRELeP	x	x			
XIP-L2F/Xerox	x	x	x	x	x

capítulos 1 e 5 para mais pormenores sobre estas duas formas de avaliação das análise alternativas). Esta escolha foi em parte arbitrária, porque contávamos inicialmente ter feito a avaliação de ambas as formas, o que acabou por não se verificar por falta de tempo. No entanto, como vimos no capítulo 1, não existe uma diferença significativa de desempenho no HAREM clássico entre essas duas formas de avaliação.

A tabela 3.4 sumariza as várias formas de avaliação para a pista do TEMPO.

3.3.1 Sistemas participantes

Os sistemas que participaram na pista do TEMPO e o seu nível de envolvimento na tarefa, ou seja os atributos de TEMPO que foram preenchidos pelos sistemas, encontram-se descritos na tabela 3.5.

Dos dez sistemas participantes no Segundo HAREM, somente três não fizeram reconhecimento de expressões temporais, o que demonstra um claro interesse em reconhecer este tipo de entidades. No entanto, apenas dois sistemas foram além do preenchimento dos atributos do HAREM clássico: o sistema da Priberam atribuiu ainda valores ao atributo TEMPO_REF, e o sistema do grupo proponente (XIP-L2F/Xerox) preencheu todos os atributos.

3.3.2 Resultados

Apesar de no capítulo 1 termos apresentado os resultados do HAREM clássico, que incluem as entidades da categoria TEMPO, não demos destaque à avaliação da categoria TEMPO em particular, pois isso corresponde, como já referimos, à avaliação da pista do TEMPO. Começamos pois por apresentar na figura 3.4 os resultados de avaliação no cenário selectivo constituído pela categoria TEMPO (e atributos TIPO e SUBTIPO) na CD do Segundo HAREM.

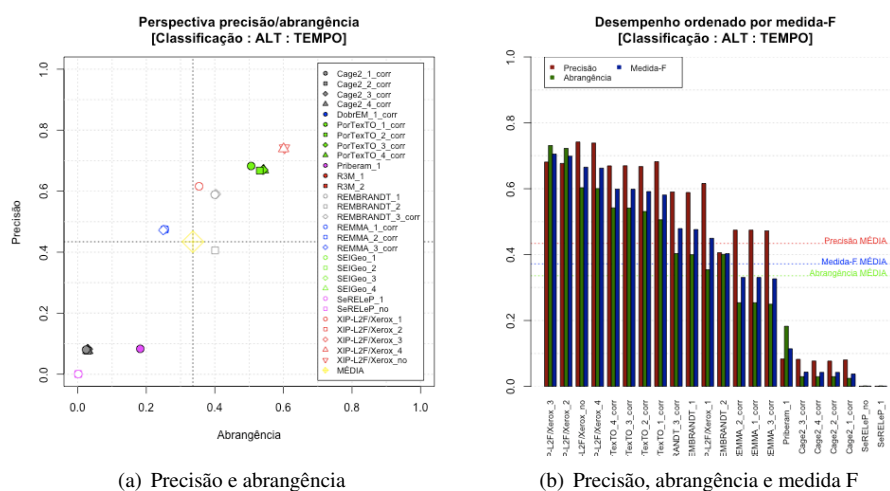


Figura 3.4: Resultados dos sistemas na classificação da categoria TEMPO e respectivos atributos TIPO e SUBTIPO

As quatro melhores corridas pertencem ao sistema XIP-L2F/Xerox, com valores de medida F de 0,7054, 0,6989, 0,6654 e 0,6625, enquanto as quatro corridas seguintes, pertencentes ao sistema PorTextO, têm valores de medida F claramente mais baixos que variam entre 0,58 e 0,60; as restantes corridas obtiveram valores abaixo de 0,479.

Destaca-se ainda que as duas melhores corridas apresentam valores de precisão e abrangência equilibrados: no melhor caso, cerca de 0,68 e 0,73 respectivamente, o que não acontece com as restantes corridas, que apresentam maiores diferenças entre as duas métricas: por exemplo, a terceira melhor corrida tem uma abrangência 0,14 abaixo da precisão que se situa em 0,74.

Centramo-nos agora na análise dos resultados na CD do TEMPO, também no cenário selectivo composto apenas pela categoria TEMPO.

Uma vez que a avaliação das entidades do TEMPO nos modos estendidos tem por base a avaliação no modo clássico, e também para efeitos de comparação com o desempenho na CD do Segundo HAREM, apresentamos em primeiro lugar na figura 3.5(a) o desempenho nesse modo na CD do TEMPO.

Comparando então o desempenho nas duas colecções douradas, podemos concluir que os resultados para a categoria TEMPO no HAREM clássico são melhores para a maioria dos sistemas (cerca de 0,4 mais alto em termos de medida F, no caso da melhor corrida) na CD do TEMPO (figura 3.5(a)) do que na CD do Segundo HAREM (figura 3.4(a)). Em todo o caso, a ordem de desempenho das nove melhores corridas não é alterada e em média estas corridas têm valores de medida F 0,0269 mais altos.

No que diz respeito à avaliação no modo estendido completo (cf. figura 3.5(b)), a comparação com a avaliação clássica na CD do TEMPO (cf. figura 3.5(a)) denota um decréscimo nas classificações de todas as corridas (cerca de 0,09 em média) excepto nas duas melhores corridas do sistema XIP-L2F/Xerox, que subiram ligeiramente, visto que foi o único que tentou atribuir todos os atributos de TEMPO. Isso explica-se porque ao modo estendido

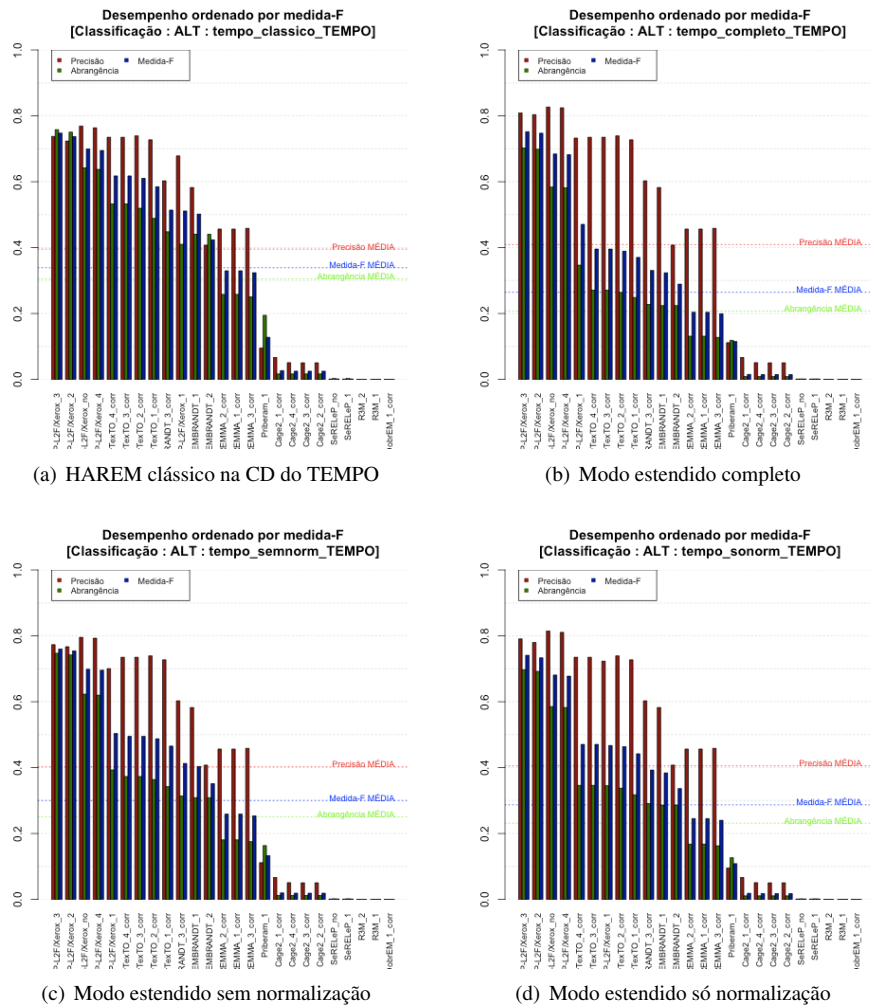


Figura 3.5: Precisão, abrangência e medida F na avaliação da pista do TEMPO (na CD do TEMPO)

Tabela 3.6: Resumo dos resultados da pista do TEMPO na CD e na CD do TEMPO

Modo	XIP-L2F/Xerox					Priberam
	3	2	no	4	1	1
Clássico na CD	0,7054	0,6989	0,6654	0,6626	0,4496	0,1143
Clássico na CD do TEMPO	0,7477	0,7369	0,6997	0,6949	0,5112	0,1277
Estendido completo	0,7518	0,7475	0,6843	0,6821	0,4707	0,1150
Estendido sem normalização	0,7600	0,7543	0,6987	0,6958	0,5035	0,1326
Estendido só com normalização	0,7400	0,7334	0,6811	0,6778	0,4672	0,1085

está associada um maior valor máximo do que no caso de se ter em conta apenas os atributos do HAREM clássico de TEMPO (CATEG, TIPO e SUBTIPO), visto que se atribui uma pontuação adicional às entidades de TEMPO pela existência dos atributos estendidos (TEMPO_REF, SENTIDO, VAL_DELTA, e VAL_NORM), o que faz naturalmente diminuir o valor da abrangência se esses atributos não forem preenchidos.

Mostramos ainda os resultados da avaliação no modo estendido sem normalização (cf. figura 3.5(c)) e só com normalização (cf. figura 3.5(d)). O aspecto que mais se evidencia é que comparando estes resultados com os mostrados para a avaliação no modo estendido completo (cf. figura 3.5(b)), as corridas de sistemas que tentaram resolver a tarefa têm melhores valores de medida F nestes dois modos. Voltamos a frisar que isso se deve, neste caso, a uma diminuição do valor máximo alcançável, pois estão a ser avaliados menos atributos do que no modo estendido completo.

Dado que apenas dois sistemas participaram na atribuição dos atributos estendidos, destacamos na tabela 3.6, o desempenho destes dois sistemas em termos de medida F.

3.4 Sugestões para o futuro da avaliação do TEMPO

Como notas finais abordaremos brevemente algumas questões que esperamos possam contribuir para uma melhor avaliação do TEMPO. Remetemos o leitor para o capítulo 6 para um balanço por pista e também para o capítulo 2 para o balanço dos proponentes desta pista.

3.4.1 Medida de avaliação

Um primeiro aspecto prende-se com a medida de avaliação que propusemos para ter em conta também os atributos estendidos, em torno da qual acabou por não haver qualquer discussão.

Esta medida combina o valor da medida obtido por avaliar os atributos estendidos com o que é obtido por avaliar os atributos do HAREM clássico da entidade. Embora seja possível atribuir pesos diferentes às parcelas dos vários atributos, não chegámos a experimentar variar o valor dos mesmos. No entanto, pensamos que para uma análise mais detalhada dos atributos estendidos poder-se-ia colocar a zero os pesos dos atributos CATEG, TIPO e SUBTIPO. Com esse procedimento, que pode ser facilmente testado no serviço de avaliação SAHARA (consulte-se o apêndice G para mais informações), as expressões temporais passariam então a ser apenas avaliadas em relação aos atributos estendidos.

Além disso, também julgamos que seria útil ter uma forma de avaliação relativa. Nesse caso, seriam avaliados os atributos estendidos apenas de expressões temporais correctamente classificadas quanto aos atributos do HAREM clássico, concentrando assim a análise nas entidades em que existe alguma possibilidade dos atributos estarem correctos (pois com certeza, nas que estão incorrectamente classificadas os atributos estão ausentes ou mal atribuídos). Seria então necessário colocar a zero os pesos dos atributos do HAREM clássico de entidades incorrectamente classificadas (de outro modo, entidades incorrectamente classificadas teriam uma penalização pela classificação espúria).

3.4.2 Estudos empíricos ilumináveis pela LÂMPADA

Tendo em conta que documentámos as várias dúvidas de anotação acrescentando `COMMENT` na colecção dourada disponibilizada na LÂMPADA, o pacote de recursos do Segundo HAREM, pensamos que a análise dessas expressões temporais poderia contribuir para uma melhoria na definição da tarefa. Além disso, serviria igualmente para melhorar a própria anotação.

Mais concretamente, seria interessante:

- refazer a avaliação depois de o grupo do TEMPO corrigir esses casos, ou seja, decidir em cada caso como é que as dúvidas relatadas seriam esclarecidas;
- refazer a avaliação removendo (através da marcação de `OMITIDO`) todos os `COMMENT` relativos ao TEMPO, para ver se em relação a estes casos também houve mais dispersão em relação ao que cada sistema marcou (e se portanto também são os casos mais discriminadores ou difíceis);
- refazer a avaliação depois de escolher e reanotar os casos das preposições, ou seja, apenas incluir nas expressões temporais as preposições que marcam tempo, e excluir das expressões temporais as preposições que vêm, por exemplo, da subcategorização do verbo.

3.4.3 Opiniões diferentes sobre o REM temporal, ou melhor, sobre o RET

Salientamos que, mesmo assim, temos uma opinião diferente em relação a muitas questões linguísticas da proposta do TEMPO. Algumas dessas questões foram sendo mencionadas na secção 3.1, a propósito das dúvidas que tivemos durante o processo de anotação; em [Carvalho e Mota \(2009\)](#), que ainda está em preparação, fazemos uma análise comparativa entre a anotação do TEMPO no Primeiro e no Segundo HAREM, em que mostramos os aspectos positivos e negativos da actual proposta de identificação, classificação e normalização de expressões temporais, e focamos em mais detalhe os aspectos em que estamos em desacordo.

Devemos, no entanto, ao concluir o capítulo, referir uma questão que nos parece central. Do nosso ponto de vista, o reconhecimento de entidades temporais (RET) não é, na sua maior parte, redutível ao REM no sentido em que extravasa em muito a identificação e classificação de entidades únicas e bem denominadas, o que aliás é particularmente visível em português na própria grafia: ao invés de marcar em maiúsculas, as expressões temporais são maioritariamente escritas em minúsculas. Por outro lado, tem também afinidades com a outra tarefa que foi proposta no MUC e que englobámos talvez erradamente

no HAREM, a do REN (reconhecimento de entidades numéricas), e que seria também de explorar e melhorar em futuras avaliações conjuntas para o português.

Estamos contudo convencidos, tal como o grupo do TEMPO, que a área do RET é fascinante e relevante para o processamento da nossa língua, e gostaríamos de a continuar a desbravar em português no futuro.

Agradecimentos

Estamos gratos à Caroline Hagège, Graça Volpes Nunes e Olga Craveiro, bem como ao Jorge Baptista, pelos seus comentários e sugestões que ajudaram certamente a criar um texto mais claro.