Apêndice C

ReRelEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista

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

Cristina Mota e Diana Santos, editoras, *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, Apêndice C, p. 309–317.

Nota das editoras: Este apêndice reproduz a versão 2.2 das directivas para reconhecimento de relações semânticas entre EM, pista do ReRelEM, que foi disponibilizada e actualizada pela última vez no dia 10 de Abril de 2008. Por uma questão de uniformização, colocámos os agradecimentos separadamente numa secção final. O assunto iniciado pelas palavras "A DISCUTIR AINDA", com relação à vagueza, foi apenas resolvido durante a própria anotação da colecção dourada, visto que nenhum participante do ReRelEM se exprimiu sobre essa questão. A decisão tomada está documentada no capítulo 4, secção 4.2.3.

Neste documento preliminar descrevemos uma tarefa piloto que propomos para o Segundo HAREM, e que pretende identificar se existem relações entre as diversas EM de um texto. A inspiração para esta proposta tem várias fontes:

- a existência da tarefa de co-referência no MUC e de "entity-link-tracking" no ACE
- a existência de trabalho sólido sobre a co-referência em português (Collovini et al., 2007)
- a emergência da área "extracção de relações" na extracção de informação (Chu-Carroll e Prager, 2007; Culotta e Sorensen, 2004; Roth e tau Yih, 2004; Zhao e Grishman, 2005)

Por estas razões, e embora tivéssemos de limitar a tarefa para ser realizável, pensamos ser o momento certo para tentar desafiar sistemas de processamento do português para esta tarefa, no âmbito do Segundo HAREM, à imagem da proposta do `TEMPO`.

Como a definição de quais as relações relevantes entre EM é um trabalho altamente subjetivo, e como trata-se de uma tarefa-piloto, prevemos inicialmente a identificação de apenas quatro (ou seis) tipos de relação entre EM, detalhados a seguir:

- identidade (sem `TIPOREL` ou `TIPOREL="ident"`)
- inclusão (`TIPOREL="inclui"` ou `TIPOREL="incluido"`)
- ocorre_em (`TIPOREL="ocorre_em"` ou `TIPOREL="sede_de"`)
- outra (`TIPOREL="outra"`)

Esta proposta foi obtida após marcação exaustiva de alguns textos e discussão alargada sobre a capacidade de consenso generalizado sobre o elenco (mais extenso) das relações originalmente propostas ao grupo mencionado acima.

C.1 Directivas para anotação das relações entre EM

Nos exemplos a seguir, só estarão marcadas as EM que estão ligadas por alguma relação. Para facilitar a leitura, omitimos também a anotação das categorias, tipos e subtipos das EM.

C.1.1 Regras gerais de integração da pista no HAREM

Cada EM recebe uma identificação única (ID), obrigatória no Segundo HAREM.

Para cada EM que apresentar uma relação de co-referência (ou outra relação) com uma outra EM já anotada, deve-se indicar, no campo COREL, a ID da EM relacionada e, em seguida, o tipo de relação entre as EM, no campo TIPOREL.

> Um dos telescópios já está pronto e em funcionamento no <EM ID="L111">**Havaí**</EM>, <EM ID="L165" **COREL="L111" TIPOREL="inclui"**>**EUA**</EM>.

É importante lembrar que só consideramos as relações **entre EM**, isto é relações que envolvem uma EM e pronomes, ou outros tipos de sintagmas nominais, por exemplo, não devem ser anotadas. Ou seja, em

> Batizado de <EM ID="AB60">**Santanaraptor placidus**</EM>, **o fóssil** é o único a ser encontrado no país com restos de tecido mole, como fibras musculares, vasos sanguíneos e pele. Para os paleontólogos, achar esse tipo de evidência equivale a acertar na loteria. "é como se **o dinossauro** tivesse sido enterrado ontem? (...) O nome é uma alusão à região onde **ele** viveu (...)

nem "o fóssil", nem "o dinossauro" nem "ele" devem receber qualquer marca.

Além disso, apenas consideramos relações entre EM **em um mesmo texto**, ou seja, a pista não se refere a relações entre textos diferentes da coleção.

É possível, por outro lado, que uma COREL tenha um ID de uma entidade que ainda não foi mencionada no texto, desde que essa entidade exista. Isso permite que um dado sistema primeiro avalie o texto completo para, em seguida, marcar as relações existentes entre as EM segundo qualquer tipo de algoritmo.

C.1.2 Relações múltiplas de uma dada EM

É naturalmente possível que uma dada EM possua relações diferentes com mais de uma EM. Nesses casos, marcamos as diferentes relações em uma estrutura de lista:

> A actual administração dos <EM ID="471">**Hipermercados Extra**</EM>, presidida por <EM ID="471" **COREL="470" TIPOREL="outra"**>**Abílio Diniz**</EM>, líder do grupo <EM ID="472" **COREL="470 471" TIPOREL="inclui outra"**>**Pão de Açúcar**</EM>,...

C.1.3 Equivalência entre relações

Note-se que não é preciso identificar exaustivamente todas as relações entre todas as EM de um texto. Pelo contrário, se existirem quatro EM com o mesmo referente, basta especificar três relações, e não doze (veja-se Villain et al., 1995).

Da mesma forma, a marcação de uma relação implica a sua inversa. Não é portanto preciso marcar inclui e incluido no mesmo par, ou "outra"duas vezes.

Veja-se um exemplo de duas maneiras equivalentes de anotar a mesma frase:

> Em 9 de Setembro de 1895, foi organizado em <EM ID="15">**New York**</EM> o <EM ID="16" COREL="15" TIPOREL="ocorre_em">**Congresso Americano de Bowling**</EM> ("<EM

ID="17" **COREL="16 15" TIPOREL="ident ocorre_em"**>**ABC**</EM> - <EM ID="18" **COREL="16 15" TIPOREL="ident ocorre_em"**>**American Bowling Congress**</EM>"), sediado em <EM ID="19" **COREL="15 16 17 18" TIPOREL="incluido sede_de sede_de sede_de"**>**Milwaukee**</EM>, com o objectivo de aplicar medidas correctivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

Em 9 de Setembro de 1895, foi organizado em <EM ID="15" **COREL="19" TIPOREL="inclui"**>**New York**</EM> o <EM ID="16" **COREL="15" TIPOREL="ocorre_em"**>**Congresso Americano de Bowling**</EM> ("<EM ID="17" **COREL="16"**>**ABC**</EM> - <EM ID="18" **COREL="16"**>**American Bowling Congress**</EM>"), sediado em <EM ID="19" **COREL="16" TIPOREL="sede_de"**>**Milwaukee**</EM>, com o objectivo de aplicar medidas correctivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

C.1.4 Opcionalidade de marcação de `TIPOREL` no caso de identidade

Para facilitar a tarefa, a relação de identidade é considerada a relação padrão e, por isso, não precisa estar marcada (ou seja, numa EM pode existir `COREL` e não `TIPOREL`):

O presidente em exercício, <EM ID="567">**Mascarenhas Ferreira**</EM>, havia já confirmado(...). (...), mas sim por desinteligências quanto à forma de actuar de <EM ID="867" **COREL="567"**>**Mascarenhas Ferreira**</EM>, a quem alguns dirigentes acusam....

<EM ID="FG51">**João Steiner**</EM>, astrofísico da USP, durante a (...), explicou <EM ID="FG560" **COREL="FG51"**>**Steiner**</EM>

A exceção é para os casos em que há mais de uma relação para uma dada EM, e uma delas é de identidade. Nessas situações a identidade precisa ser marcada através de `TIPOREL="ident"`, para evitar confusão de etiquetas.

Opcionalidade não implica, naturalmente, proibição, o que significa que os sistemas podem marcar sempre `ident` explicitamente se o preferirem.

C.2 Tipos de relações a marcar

A seguir descrevemos os tipos de relação que devem ser anotados. Lembramos novamente que a relação de identidade, por ser considerada padrão, não precisa receber o atributo `TIPOREL`.

C.2.1 Relação de identidade

A relação de identidade ocorre entre EM que pertencem à mesma categoria. Além de marcar como idênticas as EM que têm o mesmo referente, esta relação, porque se refere às EM e não às expressões textuais, vincula também EM relacionadas por abreviaturas, acrônimos, traduções e "nomes alternativos":

<EM ID="1220">**Universidade de Trás-os-Montes e Alto Douro**</EM> (<EM ID="282" **COREL="1220"**>**UTAD**</EM>)

C.2.2 Relação de inclusão

A relação de inclusão é bastante genérica e abrangente, e compreende desde relações entre EM do tipo LOCAL a relações entre EM do tipo ORGANIZACAO e ABSTRACCAO. Quando a entidade descrita por uma EM inclui a entidade descrita por outra, a relação entre as duas EM é marcada como TIPOREL="inclui".

> O <EM ID="119">**Centro de Convenções de Curitiba**</EM>, endereço presente há muitos anos na cidade, escondido na <EM ID="120" **COREL="119" TIPOREL="inclui"**>**Rua Barão do Rio Branco**</EM>.

Quando a relação é inversa, é marcada como TIPOREL="incluido":

> Chama-se "<EM ID="11">**Feira Nova de Outubro**</EM>", é organizada pela Câmara Municipal, e é bem antiga, pois remonta aos finais do século XIV. (...) Complementarmente, para além desta vertente tradicional, a <EM ID="13" COREL="11">**Feira Nova de Outubro**</EM> inclui também um <EM ID="14" **COREL="11" TIPOREL="incluido"**>**Pavilhão de Actividades Económicas**</EM>, onde qualquer empresa pode comercializar e/ou fazer divulgação dos seus produtos.

A relação de inclusão também vincula EM que, embora expressas pela mesma palavra, não apresentam uma relação de identidade, mas antes uma relação entre um elemento de uma classe e a generalidade de uma classe.

> Astrônomos brasileiros esperam fotografar os primeiros planetas fora do Sistema Solar com a ajuda do maior telescópio do mundo, o <EM ID="aa89">**Gemini**</EM> (...). (...) os telescópios <EM ID="AAFG56" **COREL="aa89" TIPOREL="inclui"**>**Gemini**</EM> têm capacidade científica...

A seguir, diversos exemplos em que a relação de inclusão está presente:

- entre ORGANIZACAO e ORGANIZACAO

  > <EM ID="123">**PSD/Vila Real**</EM> O deputado social-democrata Fernando Pereira anunciou ontem a sua candidatura à presidência da <EM ID="435" **COREL="123" TIPOREL="incluido"**>**Comissão Política Distrital de Vila Real do PSD**</EM>. (...) Ao contrário do que seria legítimo pensar, a candidatura de Fernando Pereira não aparece como resposta aos maus resultados obtidos pelo <EM ID="222" **COREL="123" TIPOREL="inclui"**>**PSD**</EM> nas eleições autárquicas.

- entre ABSTRACCAO e ABSTRACCAO

  > (...) as funções que venho ocupando no <EM ID="119">**European Script Fund**</EM> do <EM ID="1690" **COREL="119" TIPOREL="inclui"**>**Programa Media das Comunidades Europeias**</EM> fazem com que a única relação institucional...

- entre LOCAL e LOCAL

(...) havia perdido as grandes batalhas da guerra civil no <EM ID="ff203">**Ribatejo**</EM> (<EM ID="ff204" **COREL="ff203" TIPOREL="incluido"**>**Pernes**</EM>, <EM ID="ff205" **COREL="203" TIPOREL="incluido"**>**Almoster**</EM> e <EM ID="ff206" **COREL="ff203" TIPOREL="incluido"**>**Asseiceira**</EM>)

(...) e refugiara-se com o seu quartel-general no <EM ID="ff210">**Alentejo**</EM>, a única região do <EM ID="ff212" **COREL="ff210" TIPOREL="inclui"**>**Reino**</EM>

Em estudos no <EM ID="DS58">**Terceiro Mundo**</EM>, os cientistas se desobrigariam de fornecer aos doentes o melhor tratamento médico conhecido. (...) O debate surgiu após estudos em <EM ID="QQ87" **COREL="DS58" TIPOREL="incluido"**>**Ruanda**</EM> e na <EM ID="QQ90" **COREL="DS58" TIPOREL="incluido"**>**Tailândia**</EM>

(...) encontrou o fóssil na região da <EM ID="AB78">**Chapada do Araripe**</EM>, <EM ID="AB79" **COREL="AB78" TIPOREL="inclui"**>**Ceará**</EM>

- entre TEMPO e TEMPO (note-se que aqui, devido às especificidades da pista do TEMPO, ao contrário do resto do HAREM, *era dos grandes répteis*, embora completamente em minúsculas, deve ser marcado como EM)

(...), no <EM ID="AB59">**período Cretáceo**</EM> (o último da <EM ID="AB659" **COREL="AB59" TIPOREL="inclui"**>**era dos grandes répteis**</EM>

- entre COISA e COISA

Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o <EM ID="AB80">**Santanaraptor**</EM> ocuparia uma posição no grupo <EM ID="AB90" **COREL="AB80" TIPOREL="inclui"**>**Tyrannoraptora**</EM>, o mesmo do <EM ID="AB850" **COREL="AB90" TIPOREL="incluido"**>**Tyrannossaurus rex**</EM>

C.2.3 Relação de localização, ou de ocorrência em

Esta relação ocorre freqüentemente entre ORGANIZAçõES ou ACONTECIMENTOS, por um lado e LOCAIS, por outro, indica a localização de um evento ou de uma organização em um determinado local. é expressa por TIPOREL="ocorre_em" ou, de maneira inversa, TIPOREL="sede_de".

Embora a designação ocorre_em seja mais apropriada em português para acontecimentos, optámos por ter apenas um nome de relação, visto que a diferença é visível através da categoria a que pertence a entidade relacionada. Leia-se portanto localizada_em quando a relação é entre uma ORGANIZACAO e um LOCAL.

Alguns exemplos desta relação:

- Entre ACONTECIMENTO e LOCAL

Em 9 de Setembro de 1895, foi organizado em <EM ID="15">**New York**</EM> o <EM ID="16" **COREL="15" TIPOREL="ocorre_em"**>**Congresso Americano de Bowling**</EM> (“<EM ID="17" **COREL="16 15" TIPOREL="ident ocorre_em"**>**ABC**</EM> - <EM ID="18" **COREL="16 15" TIPOREL="ident ocorre_em"**>**American Bowling Congress**</EM>”), sediado em <EM ID="19"

COREL="15 16 17 18" TIPOREL="incluido sede_de sede_de sede_de">**Milwaukee**</EM>, com o objectivo de aplicar medidas correctivas contra os excessos de jogatina e aperfeiçoar ainda mais as regras.

A <EM ID="ff001">**Concessão de évora Monte**</EM> ou <EM ID="ff002" COREL="ff001">**Capitulação de évora Monte**</EM> (depois impropriamente chamada de Convenção de évora Monte) foi um acordo assinado entre liberais e miguelistas na pacata vila alentejana de <EM ID="ff005" **COREL="ff001" TIPOREL="sede_de">évora Monte**</EM> (hoje concelho de <EM ID="ff006" **COREL="ff005" TIPOREL="inclui">Estremoz**</EM>), em 26 de Maio de 1834

- Entre `ORGANIZACAO` e `LOCAL`

  A <EM ID="hg65">**IBM Research**</EM>, com o seu quartel general em <EM ID="hgoi76" **COREL="hg65" TIPOREL="sede_de">Yorktown Heights**</EM>, lidera o ranking das publicações americanas na indústria.

C.2.4 Outras relações

Esta relação, marcada por `TIPOREL="outra"`, compreende outros tipos de relação ainda não previstos ou que, até o momento, não nos pareceram relevantes para merecerem uma especificação mais detalhada.

De notar que é importante que, se houver sistemas que marquem explicitamente outras relações que não se incluam nas três anteriores, estamos dispostos a mapeá-las automaticamente nesta categoria, para minimizar o trabalho dos participantes.

É contudo conveniente salientar que a relação "ocorre no mesmo texto que" não se encontra abrangida pela relação `outra`, e que existem portanto casos em que não esperamos que seja natural estabelecer relações entre EM do mesmo texto, ou do mesmo parágrafo.

Exemplos já cobertos pelo nosso trabalho preliminar anteriormente citado (e posto à disposição dos participantes em Dezembro de 2007), mas que classificamos aqui apenas como `outra`, são:

- relação de cargo entre uma `PESSOA` e uma `ORGANIZACAO`

  <EM ID="ex1-39">**Miguel Rodrigues**</EM>, que trabalha nos <EM ID="ex1-40" **COREL="ex1-39" TIPOREL="outra">Serviços Administrativos**</EM>

  <EM ID="115">**Vale Abraão**</EM>, de <EM ID="112" COREL="115" TIPOREL="outra">**Manoel de Oliveira**</EM> teve início com a nomeação de <EM ID="115a">**Pedro Santana Lopes**</EM> para <EM ID="116" **COREL="115a" TIPOREL="outra">secretário de Estado da Cultura**</EM>

- relação de parentesco entre duas `PESSOA`s.

  <EM ID="ex3-12">**D. Miguel I**</EM>, considerado soberano usurpador do trono de sua sobrinha <EM ID="ex3-13" **COREL="ex3-12" TIPOREL="outra">D. Maria da Glória**</EM>

- relação entre um projecto e a `COISA` ou `LOCAL` a que diz respeito

> O <EM ID="119">**Centro de Convenções de Curitiba**</EM>, endereço presente há muitos anos na cidade, escondido na <EM ID="120" COREL="119" TIPOREL="inclui">**Rua Barão do Rio Branco**</EM> agora sendo revitalizada dentro do projeto <EM ID="121" **COREL="120" TIPOREL="outra"**>**Cores da Cidade**</EM> pode ter um impulso que o coloque como ponto de convergência das iniciativas de negócios...
>
> Astrônomos brasileiros esperam fotografar os primeiros planetas fora do Sistema Solar com a ajuda do maior telescópio do mundo, o <EM ID="aa89">**Gemini**</EM> (...). O projeto <EM ID="FG56Oy" **COREL="aa89" TIPOREL="outra"**>**Gemini**</EM>, resultado de um consórcio de sete países, envolve

- relação de baptismo entre um nome (ABSTRACCAO) e o que lhe deu origem

  > O exemplar de <EM ID="AB880">**Santanaraptor**</EM> encontrado (...) O nome é uma alusão à região onde ele viveu (a <EM ID="RR56" **COREL="AB880" TIPOREL="outra"**>**Formação Santana**</EM>).

De notar que também não consideramos como relações de identidade aquelas que ligam a menção do nome (ABSTRACCAO) à referência da entidade, ou seja a relação de identidade pressupõe o mesmo tipo semântico. Por exemplo:

> A <EM ID="ff001">**Concessão de évora Monte**</EM> ou <EM ID="ff002" COREL="ff001">**Capitulação de évora Monte**</EM> (depois impropriamente chamada de <EM ID="ff003" **COREL="ff001" TIPOREL="outra"**>**Convenção de évora Monte**</EM>)
>
> Batizado de <EM ID="AB60">**Santanaraptor placidus**</EM>, o fóssil é o único a ser encontrado no país (...). Outra importante descoberta é que, na cadeia evolutiva dos dinossauros, o <EM ID="AB80" **COREL="AB60" TIPOREL="outra"**>**Santanaraptor**</EM> ocuparia uma posição ...

O fato de o HAREM considerar as EM no contexto implica que entidades relacionadas por metonímia, como *Espanha* no exemplo abaixo, que ora pode ser referida como LOCAL, ora como ORGANIZACAO do tipo ADMINISTRACAO, também deverão estar relacionadas – neste caso, a relação é do tipo outra. Ou, visto de outra maneira, é a relacão do tipo outra que nos garante a presença de um vínculo entre entidades relacionadas por metonímia.

> a retirada para <EM ID="ex3-28" CATEG="LOCAL" TIPO="ADMINISTRATIVO">**Espanha**</EM> para auxiliar a causa carlista de seu primo Don Carlos (pretendente absolutista ao trono da <EM ID="ex3-31" CATEG="ORGANIZACAO" TIPO="ADMINISTRACAO" **COREL="ex3-28" TIPOREL="outra"**>**Espanha**</EM>

C.2.5 Relações entre EM vagas

Outro ponto importante diz respeito a relações entre EM que, embora expressas pelo mesmo mesmo item lexical, seu uso, no contexto, salienta diferentes facetas de seu significado, não sendo possível, ou melhor, necessário, escolher entre elas.

Nos termos do HAREM, trata-se de relações entre EM que são vagas, e que por isso apresentam mais de uma classificacao semântica. Nestes casos, é possível que apenas uma

das facetas participe de uma determinada relação (e, teoricamente, nada impede ainda que outra faceta seja evidenciada em outra relação).

No exemplo abaixo, a EM descrita pelo termo *Concessão de évora Monte* pode ser interpretada como ACONTECIMENTO ou como OBRA (do tipo PLANO). Porém, apenas a faceta ACONTECIMENTO pode apresentar uma relação do tipo SEDE_DE com a EM descrita por *Évora Monte*.

A DISCUTIR AINDA: Neste caso, marcam-se duas relações: aquela identificada recebe o nome da relação, e a outra relação, decorrência da classificação vaga, é anotada como outra:

> A <EM ID="ex3-4" CATEG="OBRA|ACONTECIMENTO" TIPO="PLANO|EVENTO">**Concessão de évora Monte**</EM> ou <EM ID="ex3-5" CATEG="OBRA|ACONTECIMENTO" TIPO="PLANO|EVENTO" **COREL="ex3-4"**>**Capitulação de évora Monte**</EM> (depois impropriamente chamada de <EM ID="ex3-6" CATEG="ABSTRACCAO" TIPO="NOME" **COREL="ex3-4"**>**Convenção de évora Monte**</EM>outra) foi um acordo assinado entre liberais e miguelistas na pacata vila alentejana de <EM ID="ex3-7" CATEG="LOCAL" TIPO="ADMINISTRATIVO" **COREL="ex3-4" TIPOREL="outra sede_de"**>**évora Monte**</EM> (hoje concelho de <EM ID="ex3-9" CATEG="LOCAL" TIPO="ADMINISTRATIVO" **COREL="ex3-7"**>**inclui**</EM>)

Da mesma forma, isso pode acontecer quando a vagueza implica diferenças na delimitação, o que no HAREM é indicado pelas etiquetas ALT.

> Boa parte do poderio militar americano atual foi desenvolvido <ALT><EM ID="oiu65">**durante a era Reagan**</EM>|durante a era <EM ID="xf5">**Reagan**</EM></ALT>. <EM ID="o65pre" **COREL="xf5"**>**Reagan**</EM> gostava particularmente de ...

De notar que ainda não foi tomada nenhuma decisão específica em relação às categorias do TEMPO, visto que a própria definição da tarefa engloba alguns conceitos de co-referência, e porque a sua definição não foi feita por nós.

C.2.6 Quadro-resumo das categorias por tipo de relações a marcar

Tabela C.1: Quadro-resumo das categorias por tipo de relações a marcar

Relação	**Categorias a que se aplica**
Identidade	Todas, exigindo igualdade de CATEG, TIPO e SUBTIPO
Inclusão	Todas menos VALOR, exigindo igualdade de CATEG
Ocorrência ou localização	ACONTECIMENTO e LOCAL; ORGANIZACAO e LOCAL
Outras	Todas, sem qualquer restrição

Agradecimentos

Agradecemos as muitas sugestões e comentários da Renata Vieira, cuja proposta inicial de fazer uma pista de co-referência nos pôs nesta pista, e do David Cruz, Nuno Cardoso e Cristina Mota sobre versões preliminares deste documento. **Aproveitamos para pedir a todos os participantes do HAREM que se pronunciem.**