Avaliação de reconhecimento de entidades mencionadas: princípio de AREM

Cristina Mota, Diana Santos e Elisabete Ranchhod

O reconhecimento de entidades mencionadas (REM), adaptação do inglês "named entity recognition", consiste na identificação e classificação de expressões linguísticas, na sua maioria "nomes próprios", que remetem para um referente específico. Essas expressões contêm um elemento, em geral um nome, que as identifica e por meio do qual são identificadas.

A fim de clarificar esta noção, e de facilitar a compreensão da actividade recoberta por esta área a pessoas menos familiarizadas com ela, começamos por dar, na secção 1, alguns exemplos ilustrativos do que se entende por "entidade mencionada" (EM), exemplificando igualmente em que consiste o seu reconhecimento.

Em seguida, chamamos a atenção para a importância de que se reveste o adequado processamento destas expressões em diversas áreas da engenharia da linguagem (secção 2); apresentamos sucintamente algumas abordagens que têm sido mais frequentemente adoptadas no tratamento de entidades mencionadas (secção 3); e destacamos várias avaliações conjuntas internacionais que tiveram como objectivo o reconhecimento de entidades mencionadas (secção 4).

Por fim, na secção 5, descrevemos a primeira iniciativa da Linguateca na área de avaliação de reconhecimento de entidades mencionadas, de forma a preparar um conjunto de actividades de prospecção e de avaliação cooperativa neste domínio; na secção 6, apresentamos os resultados da sessão do Avalon'2003[1], assim como as perspectivas de futuro que a discussão abriu.

1. O que são entidades mencionadas e o que se entende pelo seu reconhecimento?

Em 1995, na MUC-6 (Message Understanding Conference), a sexta edição de uma conferência dedicada à avaliação conjunta de extracção de informação, foi introduzida a tarefa de reconhecimento de entidades mencionadas (veja-se Hirschman (1998c) para a história da MUC). Desde então, o interesse pelo reconhecimento de entidades mencionadas cresceu significativamente e diversas outras avaliações conjuntas têm sido dedicadas a este assunto, as mais relevantes das quais destacamos na secção 3.

Embora expressões temporais (horas e datas) e expressões numéricas (percentuais e monetárias) tenham feito desde sempre parte desta tarefa, devido ao facto de conterem caracteres não alfabéticos e/ou EMs no seu interior (por exemplo, *€ 30,000, 10 de Março de 2000* ou *23h20, hora de Lisboa*), a parte mais interessante da tarefa (e que tem suscitado maior discussão) tem sido, de longe, a de identificar e classificar, o que, em sentido lato, poderemos chamar nomes próprios e outras entidades afins (ou seja, conjuntos de nomes próprios e de outras palavras que formam na generalidade dos casos uma entidade, como, por exemplo, *ministro da Justiça* ou *administração Bush*).

A categorização mais comum, estabelecida na MUC-6, distingue três tipos de entidades mencionadas[2]: as que se referem a pessoas singulares (antropónimos), colectivas (empresas e organizações) e a lugares (topónimos). No entanto, além destas, existem muitas outras EMs que perfazem um número apreciável de casos: menções a obras diversas (discos, livros ou filmes), a veículos (barcos ou sondas espaciais) e a entidades menos fáceis de incorporar sob uma

[1] Encontro sobre avaliação conjunta que decorreu em Faro em Junho de 2003 como satélite do PROPOR.

[2] Dado que ao longo do capítulo nos preocuparemos essencialmente com o reconhecimento de nomes próprios e entidades afins, deixando de lado as expressões temporais e numerais, usaremos indiferentemente as designações "nome próprio", "entidade mencionada" e "entidade".

designação comum, tais como prémios, medidas políticas, festas e campeonatos, sem contar com as menções a cargos, monumentos ou acontecimentos.

Daremos em seguida exemplos ilustrativos de algumas destas situações, retirados do CETEMPúblico (Rocha & Santos, 2000). Nestes exemplos, as EMs estão destacadas a negrito; uma possível classificação que lhes pode ser atribuída encontra-se dentro de parênteses rectos, no fim de cada excerto.

CP5963-O percurso da **Galileu** pelo sistema joviano ficará completo com 11 órbitas elípticas à volta do planeta e muitas das suas 16 luas. [sonda espacial]

CP6132-Todas as delegações europeias tinham feito a travessia do **Atlântico** no **Queen Mary**, o que lhes permitira concertar posições. [topónimo; barco]

CP32137-«**Madame Bovary**» relida por **Agustina Bessa-Luís**, relida por **Manoel de Oliveira**. [livro; antropónimo; antropónimo]

CP41084-Não se sabe o que é «**The Funeral**», de **Abel Ferrara**, mas a festa de **Manoel de Oliveira**, «**Party**», é um arrepiante e delicioso funeral. [filme; antropónimo; antropónimo; filme]

CP42226-O protocolo anexo ao **Tratado de Maastricht** sobre coesão económica e social é curto, mas diz quase tudo. [tratado]

CP58147-Junto à **Torre de Belém**, uma concentração de povo e "figuras gradas" comemorou o **25 de Abril**, a olhar para o rio, com música e cheiro de festa. [monumento; evento]

CP71346-Esta consolidação reflecte-se no **Pacote Delors II** e no grande tema do alargamento. [programa político]

CP76743-Quem tinha comprado e devia aquela quantia era a **Lopes e Pinheiro, Lda**. [empresa]

De uma forma geral, no tratamento das EMs, a identificação (ou delimitação) das entidades constitui a primeira etapa, à qual se seguirá uma classificação tão operativa quanto possível. Na realidade, identificação, delimitação e classificação estão intimamente relacionadas e podem decorrer em simultâneo.

Ao contrário do que se poderá julgar pelos exemplos anteriores, a primeira etapa levanta desde logo vários problemas (decorrentes, em primeiro lugar, dos diferentes critérios utilizados e, depois, do facto de nela estar quase sempre implícita a posterior classificação). Vejam-se os seguintes exemplos:

O estreito de Magalhães é pouco navegado nos dias de hoje.
A Comissão de Pais da Escola Básica de Montargil convocou uma reunião.
Ao Estado cabe velar pelo bem-estar dos cidadãos.
A Economia é a pasta mais problemática do novo governo

Estes exemplos suscitam imediatamente, ou podem suscitar, entre outras, as seguintes questões: deverá considerar-se como EM a combinação *estreito de Magalhães*, ou apenas *Magalhães*? No segundo exemplo, quantas EMs devem ser identificadas (uma, duas, três ou cinco)? *Estado* é uma entidade mencionada, ou apenas uma palavra, que a convenção ortográfica leva a grafar com maiúscula inicial quando se refere ao Estado português, ou à ideia abstracta de estado, e assim a separa de outras acepções que o substantivo *estado* também tem na língua (por exemplo, o *estado* das ruas em Lisboa é lastimável, o *estado* da Maria inspira cuidado, o *estado da arte* está muito bem descrito)? Deve *Economia*, isoladamente, constituir uma EM, ou a entidade mencionada deverá ser antes *pasta da Economia*, recuperável pelo contexto, se interpretarmos a frase acima como uma redução de *A pasta da Economia é a pasta [...]*?

E se a identificação já levanta diversos problemas, a classificação, em parte dependente da aplicação, é uma tarefa ainda mais complexa e, como se verá ao longo deste capítulo, muito variável de investigador para investigador.

Apresentamos apenas alguns casos que nos parecem consensuais e que mostram que a classificação não pode ser feita independentemente do contexto.

O iate Noruega atracou ontem no porto de Lisboa.
Portugal foi ontem criticado pela falta de contenção monetária.
O Vasco da Gama passou enfim para a primeira divisão.
Alunos do Maria Amália encontram-se em greve.

Festejar o 25 de Abril é um dever de todo o democrata.

Como exemplificam as frases anteriores, um dos problemas do reconhecimento de EMs é que estas expressões são ambíguas, ou seja, à mesma palavra ou sequência de palavras pode ser atribuída mais do que uma classificação. Nestes poucos exemplos podemos desde já observar que:

1. o nome de um país pode designar um barco (Noruega) ou ser usado metonimicamente: o governo de um país (Portugal);
2. um nome de pessoa é dado e passa a identificar também vários tipos de organizações (nos exemplos, um clube desportivo, *Vasco da Gama*, e uma escola, *Maria Amália*);
3. uma data (*25 de Abril*) pode referir um acontecimento histórico, uma entidade com um referente único.

Em muitos destes casos, só é possível tomar decisões correctas de forma automática depois de se ter procedido a uma análise sintáctica do contexto próximo em que as entidades mencionadas ocorrem. Assim, um nome como *Washington*, que a priori pode representar um lugar ou uma pessoa, ganhará um e apenas um valor quando integrado numa estrutura léxico-sintáctica adequada, como acontece nos exemplos seguintes:

Eles deslocaram-se a Washington logo que receberam a notícia. (Lugar)
Washington foi o primeiro presidente dos Estados Unidos. (Pessoa)

Como complemento do verbo *deslocar*, *Washington* faz parte de um complemento locativo (logo, é um lugar); no entanto, na posição de sujeito de uma construção atributiva como "ser presidente" recebe o atributo "humano".

McDonald (1996) propõe os termos evidência interna e externa para separar e medir a influência da forma das EMs vs. o contexto em que se encontram. O autor argumenta que mesmo que seja possível atribuir uma categoria inicial com base em evidência interna (por exemplo, a palavra *Lda.* poderá ser uma evidência interna de que a sequência que a contém é o nome de uma empresa), a decisão final da categoria a atribuir dependerá do contexto em que se encontra.

Naturalmente, existem outros casos em que a análise do contexto imediato (análise frásica) pode não ser suficiente, sendo necessário recorrer a uma análise discursiva. Tome-se como exemplo a frase:

Washington sabia da ameaça, mas preferiu fechar os olhos.

Neste exemplo, *Washington*, na posição de sujeito do verbo *saber*, adquire o traço "humano", mas permite a dupla interpretação: pessoa singular (um indivíduo) e pessoa colectiva (o governo dos Estados Unidos).

Além disso, muitos outros casos há em que *Washington* pode não ser uma EM de nenhum dos três tipos referidos acima, que não esgotam, portanto, o leque de classificações que essa EM pode ter:

A família Washington era detestada em vários estados.

Nesta frase, apesar de *Washington* ser considerado como um antropónimo, este não designa uma pessoa única, mas sim toda uma família. Deverá por essa razão receber uma classificação diferente?

Finalmente, outra questão relacionada com o reconhecimento de entidades mencionadas, e que também esteve ligada a uma das tarefas da MUC, tem a ver com a multiplicidade de EMs (e suas inter-relações). Trata-se da co-referência entre entidades mencionadas. Os exemplos seguintes ilustram esta situação:

A Federação Portuguesa de Andebol emitiu um comunicado condenando a arbitragem do jogo do passado domingo. O líder da Federação confessou-se desgostoso com o sucedido.

O Presidente Jorge Sampaio mostrou-se cauteloso. Sampaio reuniu-se de imediato com vários dirigentes partidários.
A conferência terá lugar no Instituto Superior Técnico. Estando no IST, siga as setas para chegar ao local exacto.

Nos dois primeiros exemplos, *Federação* e *Sampaio* são retomas anafóricas das entidades já referidas na sua forma completa no discurso anterior; no último caso, uma sigla substitui anaforicamente o nome que lhe deu origem (a sua base lexical).

Esta tarefa, apesar de ser extremamente importante e levantar questões problemáticas, sobretudo quando no texto a forma completa (ou base lexical, no caso das siglas) não se encontra suficientemente perto, não foi tida em conta no estudo preliminar que descrevemos mais à frente.

A discussão destes exemplos deverá ter servido, cremos, de breve introdução à problemática do reconhecimento de entidades em português. Não é de forma alguma nossa intenção apresentar aqui uma solução, mas tão somente alertar para o facto de que muitos problemas só se identificam depois de comparar abordagens e, sobretudo, depois de se tentar a sua implementação.

2. Porque é que é preciso identificar EMs?

Talvez a primeira resposta a esta questão seja a de que, para a área de extracção de informação, é importante saber o que se passa com este tipo de entidades, uma vez que o sentido de um texto está frequentemente ancorado nelas. Em geral, os textos (informativos) referem-se mais a entidades e acontecimentos do que a conceitos genéricos. Deste modo, a disciplina de extracção de informação, tal como foi inicialmente concebida, preocupava-se em obter automaticamente e de forma semi-estruturada informação sobre EMs que ocorriam em texto livre.

Contudo, há muitas outras áreas da engenharia da linguagem, que referimos sucintamente em seguida, que beneficiam ou poderão beneficiar de uma extracção correcta destas entidades.

2.1 Processamento sintáctico e semântico de texto

O tratamento adequado das EMs poderá ser de grande utilidade para a análise sintáctica automática de um texto, uma vez que, identificadas previamente como objectos linguísticos particulares, tornam a análise mais simples, ou até mesmo viável. De facto, as sequências de caracteres e de palavras como datas, moradas e nomes próprios, muito frequentes em qualquer tipo de texto mas particularmente abundantes nos textos jornalísticos, constituem unidades sintáctico-semânticas que importa identificar e analisar numa fase precoce do processamento.

Tome-se como exemplo a frase:

O Pereira caiu à água

Se, numa fase anterior à do processamento sintáctico, o nome próprio *Pereira* não for identificado como tal, mas analisado como um nome comum no feminino, isso impedirá a correcta análise da frase. Em particular, o analisador sintáctico ao verificar que o determinante *o* e o nome *Pereira* não concordam em género, como é de regra em português, avaliará a frase como incorrecta.

Por outro lado, uma classificação rigorosa das EMs poderá reduzir ambiguidades (sintácticas e semânticas) e levar a interpretações preferenciais. Considere-se, por exemplo, a frase:

Os jornalistas dirigiram-se a Washington.

Dado que o nome *Washington* pode, pelo menos, ser classificado como nome de pessoa e de lugar, a frase tem igualmente, pelo menos, duas possibilidades de análise: uma em que a interpretação será *os jornalistas interpelaram Washington* e outra em que se considerará *os jornalistas deslocaram-se a Washington*. Mas, se for possível, por análise do contexto discursivo, associar *Washington* a *G. Washington* então a análise preferencial será a primeira.

Inevitavelmente, a própria análise sintáctica e/ou semântica poderá levar a uma alteração das classificações iniciais. Em rigor, a classificação de EMs faz parte integral da análise sintáctico-semântica de um texto. Considere-se, por exemplo, a sequência *Mafalda Reis e Cunha Rego* nas seguintes frases:

Mafalda Reis e Cunha Rego indicou ao Expresso que o dinheiro para as obras camarárias tinha sido obtido por via legal.
Mafalda Reis e Cunha Rego casaram-se com pompa e circunstância na Capela de Sintra.

À partida e dependendo da abordagem, a sequência tanto pode ser considerada como duas EMs coordenadas como ser analisada como uma única EM (Santos, 1999a). A posterior análise sintáctica dos contextos indicará claramente que no primeiro exemplo se trata de uma única EM (sujeito de um verbo no singular), enquanto no segundo caso há duas EMs coordenadas, sujeito de um verbo no plural. Se, contudo, este tipo de sequências ocuparem uma posição de complemento, a sua correcta identificação é bastante mais complexa.

2.2 Recolha de informação

É cada vez mais claro que o tratamento adequado das entidades mencionadas tem toda a relevância para a recolha de informação. De facto, quando um utilizador escreve *Castelo Branco* na caixa de procura, é muito baixa a probabilidade de querer procurar "castelos que sejam brancos", em vez da localidade ou do escritor Camilo de Castelo Branco.

Por outro lado, visualizar uma colecção de documentos (ou de resultados de uma procura) pelo conjunto de entidades mencionadas que inclui é uma forma de explicitar a informação e de guiar o utilizador, o que, em alguns casos, já é uma realidade, como o demonstram Amigó *et al.* (2003) e Kyriakov *et al.* (2003).

Pode assim dizer-se que, de certa forma, o reconhecimento de EMs substitui, ou pelo menos complementa, a tradicional descoberta do tópico de um documento com base em linguagens controladas de indexação.

2.3 Resposta automática a perguntas

Em relação a esta área, a necessidade de identificar entidades sobre as quais se formulam perguntas ou que constituem a resposta procurada é evidente, sobretudo se considerarmos que (por agora) a maior parte das perguntas tratadas pelos sistemas de resposta automática a perguntas (RAP) são factóides, ou seja, referem-se a factos sobre entidades bem determinadas. Veja-se, por exemplo, Greenwood & Gaizauskas (2003) que usam um reconhecedor (e classificador) de EMs para melhorar o desempenho de um tal sistema.

Em relação ao português, na primeira avaliação de RAP para a nossa língua relatada por Rocha & Santos (neste volume), foi possível constatar que 88,5% das perguntas utilizadas se referem explicitamente a entidades mencionadas, enquanto que 60,5% das respostas são ou simplesmente entidades mencionadas ou as incluem.

2.4 Síntese de fala

O reconhecimento de EMs é igualmente pertinente para a síntese de fala. Para se ter uma ideia, comparem-se as diferentes realizações fonéticas de *Luís* e *António* quando formam um único nome próprio (primeiro exemplo) ou quando se trata de dois nomes próprios em sequência (segundo exemplo):

Luís António, que estás tu a fazer?
Luís, António, que estão vocês a fazer?

Além disso, muitas vezes, os nomes próprios conservam características linguísticas que já se perderam na língua, como grafias antigas (mas com pronúncia contemporânea), ou laivos de pronúncia estrangeira, que tornam a sua produção ainda mais complicada do que a das palavras do léxico comum (veja-se o caso de *Paris Saint-Germain*). De facto, não só os nomes estrangeiros têm tendência a ser pronunciados de modo a preservar algumas características da língua de onde provêm – veja-se Trancoso (1995), como casos há em que os nomes próprios são eles mesmos uma mistura de termos do português e de outra língua (caso de *Rei Lear* entre

outros). Segundo a mesma autora, outro aspecto a ter em conta é o da síntese de siglas e acrónimos, que também obedece a regras bem distintas das do léxico comum. Por exemplo, há siglas que, tendo estrutura silábica, tanto podem ser foneticamente realizadas como palavras: *TAP*, como não: *IPO*.

Para uma panorâmica da importância da pronúncia dos nomes em síntese de fala, veja-se Valencia & Yvon (2000).

2.5 Geração de texto

O estudo da forma como este tipo de entidades são referidas é importante para a geração de textos coerentes. Sumarização e criação de textos por medida são duas das áreas em que esta questão é pertinente.

Assim, não espanta que Nobata *et al.* (2002) tenham incorporado um módulo de reconhecimento de EMs no seu sistema de geração de resumos, ao participarem na tarefa "single document summarization" da avaliação DUC 2001 (Document Understanding Conference), obtendo resultados prometedores: não só o desempenho foi melhor do que a média de todos os participantes, como, no que respeita aos critérios de coesão e organização, obtiveram os melhores resultados.

2.6 Tradução automática

A tradução automática é uma das áreas mais complexas em processamento de linguagem natural. Embora a sua qualidade não seja fácil de avaliar (sobre este assunto, veja-se Sarmento *et al.*, neste volume), um dos aspectos que poderá contribuir para a melhorar de forma significativa é o processamento adequado das entidades mencionadas.

Os exemplos seguintes são ilustrativos de alguns problemas que podem surgir em tradução automática se estas unidades lexicais não forem devidamente tratadas.

Os turistas gostaram de visitar a Madeira.
Os turistas gostaram de visitar os Açores.

A Organização dos Países Exportadores de Petróleo emitiu um comunicado.
A Caixa Geral de Depósitos emitiu novos títulos.

No primeiro par de exemplos, a não identificação de *Madeira* como um topónimo, levaria à sua (incorrecta) tradução como nome comum (poder-se-ia pensar que bastava considerar que se um nome comum estiver em maiúscula não se traduz, mas isso levaria a que *Açores* também não fosse traduzido, o que poderia não ser adequado numa tradução para inglês). O segundo par de frases demonstra que a identificação de nomes próprios multipalavra como sequências indivisíveis é fundamental, pois no primeiro caso essa unidade deverá ser traduzida como um bloco, e no segundo caso, a sequência não deverá ser de todo traduzida.

Em Tran *et al.* (2004) é apresentada uma base de dados relacional multilingue de nomes próprios. A grande maioria está em francês, mas uma parte está já associada ao seu equivalente alemão, estando prevista a introdução de nomes próprios em inglês, polaco, russo e serbo-croata. De facto, como Mandl & Womser-Hacker (2005) demonstraram, ao analisar o desempenho dos sistemas que participam no CLEF, a identificação de EMs e a sua correcta tradução também têm uma influência significativa na recolha de informação cruzada, ou interlínguas.

3. Abordagens de reconhecimento de entidades mencionadas

Várias abordagens têm sido adoptadas no reconhecimento de entidades mencionadas. Destacaremos aqui apenas as que parecem ser mais relevantes e que têm sido mais usualmente seguidas.

A abordagem mais simples é a que se baseia na consulta de almanaques ("gazetteers"). Esta abordagem depende fortemente da existência de listas de nomes próprios, sobretudo de antropónimos, topónimos e designações de empresas e organizações. Por exemplo, a principal fonte de informação do sistema Nominator (Ravin & Wacholder, 1996) é a que as autoras designam por "authority lists", que contém, além de nomes próprios, outras palavras que podem

servir para identificar e/ou classificar um nome, como sejam, entre outras, as abreviaturas (*Lda.*, *Inc.*, *Jr.*, etc.).

Convém notar que a extensão e granularidade requeridas por estes almanaques também podem ser objecto de estudo. Com efeito, Stevenson & Gaizauskas (2000) estudaram detalhadamente a contribuição das listas usadas no reconhecimento de EMs, comparando-as com as obtidas através da utilização de corpora já previamente anotados (com EMs).

Uma abordagem diferente é a que se baseia na construção, normalmente "artesanal", de regras intuitivas que dependem apenas do conhecimento de quem as descreve. Tanto Grishman (1995) como Friburger (2002) apresentam sistemas cujas regras são descritas através de autómatos de estados finitos.

O processamento estatístico, tais como os modelos de Markov (Miller *et al.*, 1998) e de máxima entropia (Borthwick, 1999) constituem ainda outra alternativa. Estes métodos fazem normalmente uso de corpora de treino previamente anotados (com EMs).

Por fim, destacamos as abordagens híbridas que tentam tirar partido da combinação de diferentes técnicas. Mikheev *et al.* (1999), autores do sistema com melhores valores de medida F na MUC-7, combinaram gramáticas baseadas em regras e modelos de máxima entropia no reconhecimento de EMs, demonstrando que é possível, com almanaques de dimensões reduzidas, obter bons desempenhos.

É contudo interessante salientar que tipos diferentes de entidades podem levar a (ou necessitar de) abordagens diferentes. Wakao *et al.* (1996) descrevem um estudo dos diversos factores que aumentam a precisão para tipos diferentes de EMs, constatando que, por exemplo, a utilização do contexto melhora significativamente o desempenho do reconhecimento de nomes de organizações, enquanto que para lugares e nomes de pessoas, é mais relevante ter em conta a evidência interna.

4. Avaliação de reconhecimento de entidades mencionadas

Tal como já referimos em secções anteriores, a primeira grande conferência que definiu a tarefa de avaliação de reconhecimento de entidades mencionadas foi a MUC (Message Understanding Conference), na sua sexta edição (MUC-6). Esta série de conferências, realizadas ao longo de dez anos, a partir de 1986, tinha como alvo de avaliação a extracção de informação nas suas diversas componentes (Hirschman, 1998).

Nessa edição, para além de se introduzir a tarefa de reconhecimento de EMs em textos em inglês, caracterizaram-se mais duas tarefas relacionadas: co-referência entre entidades e identificação de constituintes de padrões ("template element"). Na MUC-7, foi acrescentada a tarefa de estabelecimento de relações entre padrões. Esta edição teve ainda a particularidade de decorrer em simultâneo com a segunda edição da conferência Multilingual Entity Task (MET-2), que tinha também como objectivo avaliar o reconhecimento de EMs, mas em línguas diferentes do inglês (neste caso, o chinês e o japonês). A primeira edição da MET, em 1996 (Merchant *et al.*, 1996), tinha tido como línguas de trabalho o inglês e o espanhol.

Outras iniciativas que devem ser destacadas são o programa ACE (Automatic Content Extraction) e as sessões de avaliação da conferência CoNLL em 2002 e 2003. O programa ACE (Doddington *et al.*, 2004), que começou em 1999 com um estudo piloto, tem por objectivo o desenvolvimento de tecnologias de (i) reconhecimento de entidades (não apenas dos seus nomes) - Entity Detection and Tracking (EDT), (ii) reconhecimento de relações entre entidades - Relation Detection and Characterization (RDC), e (iii) extracção de eventos nos quais as entidades participam - Event Detection and Characterization (EDC). Para além de extracção de informação a partir de texto escrito, pretende-se trabalhar com dados obtidos a partir de som e imagem; o inglês começou por ser a língua em estudo, tendo-se começado a anotar o chinês e o árabe, a partir da segunda fase.

A avaliação proposta pelo programa ACE distingue-se das anteriores por três aspectos principais: (i) maior complexidade e nível de detalhe das tarefas propostas (por exemplo, a identificação de entidades envolve também o reconhecimento de grupos nominais, mesmo que não envolvam nomes próprios, como seja *o candidato democrata*); (ii) introdução de novos tipos de entidades: entidades geopolíticas (para dar conta, por exemplo, da utilização dos nomes

de países como organizações), armas, veículos e instalações ("facilities"), onde se incluem monumentos, pontes, fábricas, e outras construções; (iii) sub-tipificação de entidades e de relações entre entidades.

A conferência anual CoNLL (Conference on Computational Natural Language Learning) tem promovido, desde 1999, a avaliação de métodos de aprendizagem na resolução de problemas específicos de processamento de linguagem natural, propondo tarefas comuns ("shared tasks") a serem realizadas pelos sistemas participantes. Em 2002 e 2003 essa avaliação foi dedicada ao reconhecimento "independente da língua" de organizações, pessoas, lugares e outras EMs que não se enquadrassem em nenhuma das anteriores. No primeiro ano, as línguas estudadas foram o espanhol e o flamengo; no segundo, as línguas escolhidas foram o alemão e o inglês (veja-se Sang & Meulder (2002) e Sang (2003) para mais informações).

Todas estas avaliações foram realizadas usando textos escritos, na sua maioria da imprensa. Robinson *et al.* (1999), para a avaliação DARPA/NIST HUB 4, definiram uma tarefa semelhante à das MUC para notícias faladas transcritas ("broadcast text").

Destacamos, por fim, a avaliação TERN (Time Expression Recognition and Normalization), realizada em Setembro de 2004, em conjunto com uma das fases da avaliação ACE, e apoiada por este programa. Esta avaliação conjunta, dedicada apenas ao reconhecimento de expressões temporais (Ferro *et al.*, 2003), tem por objectivo alargar a definição de expressão temporal das MUC, de modo a permitir a identificação de uma maior variedade de expressões e a elaboração de um novo esquema de normalização.

É de lamentar que o português nunca tenha estado incluído em nenhuma das avaliações referidas. De facto, quanto à avaliação do português nesta área, apenas temos conhecimento de dois trabalhos publicados.

Palmer & Day (1997) fazem um estudo quantitativo sobre a ocorrência de entidades mencionadas em textos jornalísticos de diversas línguas, calculando, entre outras medidas a taxa de transferência (percentagem das entidades que ocorrem no texto de treino e que também ocorrem no de teste). É de destacar que, logo a seguir ao chinês, o português, com 61,3%, foi a língua que apresentou maior taxa de transferência. Com base nas várias medidas calculadas, os autores estimaram ainda um limite mínimo de desempenho que para o português foi de 71,3%, o segundo mais alto. Este valor deveria ser usado como ponto de partida para a avaliação de sistemas de reconhecimento de entidades mencionadas.

Para o reconhecimento de entidades mencionadas em português, Bick (2003) descreve uma abordagem linguística que recorre a diferentes fontes de informação (léxico-semântica, padrões e regras morfo-sintácticas) organizadas num sistema multi-camada baseado em gramáticas de restrições. O sistema, avaliado sobre um excerto do CETEMPúblico, identifica seis tipos de entidades, entre as quais: pessoas, organizações e lugares. Dado que a mesma abordagem tinha anteriormente sido adoptada no reconhecimento de EMs em dinamarquês, Bick (2003:214) apresenta resultados comparativos preliminares que apontam para que haja um maior equilíbrio entre as proporções de nomes de pessoas (35,5%), organizações (22,6%) e lugares (34,5%) no CETEMPúblico do que no corpus dinamarquês, que apresenta cerca de 50% de nomes de pessoas e apenas 16% de organizações.

Embora estes resultados sejam difíceis de comparar com os obtidos por Palmer & Day (2003), quanto mais não fosse pelos diferentes dados e classificações utilizados, é de referir que as percentagens relativas das entidades detectadas neste estudo são de 30%, 50% e 19,5%, para nomes de pessoas, organizações e lugares, respectivamente.

5. Actividades preparatórias para avaliação conjunta de REM em português

A Linguateca, através do seu pólo do LabEL, iniciou as actividades de preparação de avaliação conjunta no campo do reconhecimento de entidades mencionadas, organizando um ensaio preliminar em que os participantes foram incentivados a anotar manualmente textos previamente fornecidos ou a rever o resultado dos seus sistemas, de forma a identificar os requisitos mínimos e a divergência que pudesse haver entre várias especificações da "mesma" tarefa.

A mensagem inicial foi enviada para a lista `avalia` em finais de Janeiro de 2003, tendo-se estabelecido como prazo final para envio dos resultados o dia 10 de Março do mesmo ano. No sítio da Linguateca dedicado a AREM, foram colocados os dez primeiros extractos do CETEMPúblico (1291 palavras) e os vinte primeiros extractos do CETENFolha (1344 palavras). De forma a não constranger artificialmente os resultados, não foi estipulada a priori qualquer terminologia, embora os exemplos se inspirassem directamente nas conferências MUC 6 e 7, como é evidente na mensagem inicial, que transcrevemos aqui:

> O primeiro passo a dar será estabelecer e caracterizar as entidades que pretendemos identificar e de que forma serão anotadas. Por exemplo, nas MUC foi pedido aos sistemas participantes que reconhecessem as seguintes entidades (os exemplos são naturalmente de português):
>
> - Nomes próprios de
> - Pessoas (ex: Fernando Pessoa, Maria do Carmo, Sampaio)
> - Organizações (ex: IST, Instituto Superior Técnico, Portugal Telecom)
> - Lugares (ex: Sintra, Serra da Estrela, Minho)
> - Expressões temporais
> - Datas (ex: 24 de Janeiro de 2000, segundo semestre de 1992, anos 60)
> - Horas (ex: meio-dia, 13:40, 4 horas da manhã)
> - Expressões numéricas
> - Monetárias: (ex: 20 milhões de euros, 900 mil contos)
> - Percentuais: (ex: 10.5%, sete por cento)

Logo nesta primeira mensagem, foram apresentados alguns dos problemas sobre os quais teria de haver consenso (e daí avaliação conjunta) para uma definição de "correcto":

- Deve um nome de um estabelecimento comercial (livraria, cinema, discoteca, etc.) ser identificado como uma organização?
- Deverão os sistemas reconhecer entidades estrangeiras, cujos nomes não têm tradução para português, tais como Empire State Building, New York Times, BBC, Manchester United?
- Deve ter-se em conta que, embora linearmente semelhantes, algumas entidades estão encaixadas noutras, como é o caso de Lisboa em 'Crise na faculdade influencia eleições de amanhã para a reitoria da <NOMEPROP TIPO="ORGANIZAÇÃO">Universidade Técnica de Lisboa</NOMEPROP>' por oposição a 'A <NOMEPROP TIPO="ORGANIZAÇÃO"> Polícia Judiciária</NOMEPROP> de <NOMEPROP TIPO="LUGAR"> Lisboa </NOMEPROP> anunciou ontem a conclusão de três inquéritos respeitantes (…)' ?

É-nos grato confirmar que a participação nesta tarefa foi grande, tendo nove investigadores diferentes enviado os textos anotados. Uma primeira análise quantitativa indicou que, no CETEMPúblico, foram identificadas entre 81 a 106 entidades, enquanto que, no CETENFolha, foram identificadas entre 98 a 134 entidades. Visto que alguns participantes não anotaram as entidades numéricas nem temporais (cf. Tabela 1 -1 e Tabela 1 -2), trabalhámos sobretudo os resultados relativos aos nomes próprios.

Tabela 1-1: Anotação das entidades mencionadas no CETEMPúblico por anotador

Anotador	A	B	C	D	E	F	G	H	I
NomeProp	99	81	89	83	87	84	92	90	83
ExpTemporais	0	0	2	0	2	7	2	10	3
ExpNumericas	0	0	0	0	6	9	0	6	4

Tabela 1-2: Anotação das entidades mencionadas no CETEMFolha por anotador

Anotador	A	B	C	D	E	F	G	H	I
NomeProp	102	106	100	98	108	103	111	108	106
ExpTemporais	0	0	0	0	7	10	0	20	7
ExpNumericas	0	0	4	0	6	9	0	5	6

No CETEMPúblico, foram identificados 115 nomes próprios distintos[3], dos quais 63 são consensuais (i.e. foram identificados por todos como tal). Contudo, apenas 30 destes foram classificados de forma idêntica por todos os participantes. No CETENFolha, em 125 nomes próprios identificados, 70 são consensuais. Destes, também apenas 31 obtiveram classificação consensual.

Tabela 1-3: Concordância na identificação e classificação dos nomes próprios: NP1=Número de nomes próprios identificados por pelo menos um anotador; NPT=Número de nomes próprios identificados por todos os anotadores; NPC=Número de nomes próprios igualmente classificados por todos os anotadores; NPT/NP1=Percentagem de nomes próprios que foram identificados por todos os anotadores; NPC/NPT=Percentagem de nomes próprios consensuais que tiveram idêntica classificação

	NP1	NPT	NPC	NPT/NP1	NPC/NPT
CETEMPúblico	115	63	30	54,78%	47,62%
CETENFolha	125	70	31	56,00%	44,29%
Total	240	133	61	55,42%	45,86%

Nos casos em que havia discrepâncias, deveria ter existido interacção com os participantes (até porque, sendo em muitos casos uma anotação manual, deve ter-se sempre em conta o erro humano), mas tal não chegou a ser feito na altura.

Quanto à classificação, a falta de padrão nos nomes dos tipos de entidades levou a uma situação quase caótica, como fica bem ilustrado na Figura 1-1.

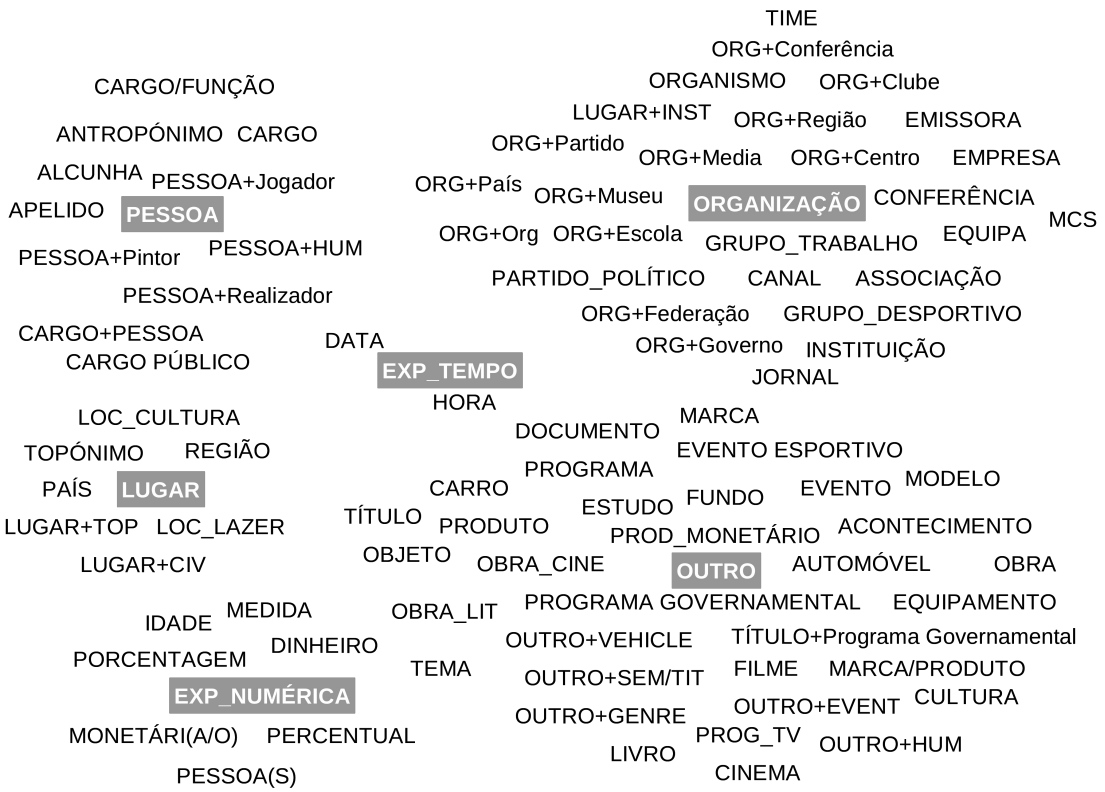

Figura 1-1 . Variedade das etiquetas utilizadas pelos anotadores

As classificações utilizadas foram múltiplas, como se vê na seguinte lista. Para melhorar a legibilidade, agrupámo-las pelo que nos pareceu serem classificações relacionadas (não necessariamente idênticas):

acontecimento evento evento-esportivo outro+evento
alcunha antropónimo apelido

[3] De notar que se tiver havido, por exemplo, duas ocorrências de *Lisboa* nos textos, essa entidade conta como duas, visto que podem ter classificações diferentes.

associação
automóvel carro outro+veículo
canal emissora
cargo cargo-público cargo/função cargo+pessoa
cinema filme obra-cine
conferência
cultura cultural
data
documento
empresa
equipa time grupo-desportivo organização+clube
estudo
fundo
grupo-trabalho
instituição
jornal organização+média mcs
livro obra-lit
loc_cultura loc_lazer
lugar lugar+civ lugar+inst lugar+top país região topónimo
marca marca/produto modelo produto
moeda prod+monetário
objeto equipamento
obra
organismo organização organização+centro organização+conferência organização+escola organização+federação organização+governo organização+org organização+museu organização+país organização+região
outro
partido-político organização+partido
pessoa pessoa+jogador pessoa+pintor pessoa+marca/objecto pessoa+hum pessoa+realizador programa governamental título+programa governamental
prog-tv programa
tema
título outro+sem/tit

Se, por um lado, a situação pode ser quase caricata por parecer não representar diferenças semânticas ou conceptuais, por outro, foi importante mantê-la para não privilegiar nenhuma nomenclatura ou encobrir hipóteses implícitas.

Na verdade, há, pelo menos, duas possibilidades de organizar os diferentes dados (aliás, representadas de forma diferente na lista acima): (i) integrar as categorias usadas em estruturas ontológicas, em que existe uma relação hierárquica entre várias classificações possíveis; (ii) considerar que a tarefa consiste em escolher uma ou mais do que uma categoria de entre os vários tipos, considerados em pé de igualdade (ou seja, uma estrutura plana a nível de classificação). Interpretámos a notação A+B como uma forma embrionária de estrutura hierárquica, enquanto que as outras notações (A-B e A/B) nos parecem corresponder a hierarquias planas.

Veja-se um subconjunto da estrutura hierárquica proposta por Sekine *et al.* (2002), utilizada no processamento do inglês e do japonês, com exemplos em português:

```
TOP
  NAME
    LOCATION # Praça de Espanha, Rossio, Ribeira
      GPE # Palestina, Médio Oriente
        CITY # Lisboa, Vila Real de Santo António
        COUNTY # Trás-os-Montes, Andaluzia
        PROVINCE # Estado, Distrito, Província
        COUNTRY # República Popular da China, Portugal
      REGION # América Latina, Escandinávia, Costa Azul
      …
    ORGANIZATION # Organização das Nações Unidas, OTAN
      COMPANY # Novabase, TAP
      COMPANY_GROUP # Sonae, Grupo Salvador Caetano
```

```
      MILITARY # Força Aérea Portuguesa
      INSTITUTE # FIA, Associação Portuguesa de Linguística
      MARKET # Bolsa de Valores de Lisboa
      POLITICAL_ORGANIZATION #
         GOVERNMENT # Ministério da Defesa, Secretaria de Estado da Cultura
         POLITICAL_PARTY # CDS-PP, Bloco de Esquerda
         PUBLIC_INSTITUTION # CTT
      …
   PERSON # Jorge Sampaio, José Fonseca e Costa, Miguel Esteves Cardoso
      LASTNAME # Sampaio, Fonseca, Costa, Esteves, Cardoso
      MALE_FIRSTNAME # Jorge, José, Miguel, Esteves
      FEMALE_FIRSTNAME # Cristina, Diana, Elisabete
   …
…
```

Os defensores deste tipo de estrutura argumentarão que, embora haja diferenças entre os nomes das folhas (categorias básicas), será possível medir a concordância entre sistemas no que se refere às categorias globais.

Os proponentes de uma estrutura plana, por seu lado, defenderão que, pese embora a semelhança conceptual, haverá propriedades linguísticas que diferenciarão claramente, por exemplo, o nome de uma empresa do nome de um clube de futebol (ambos normalmente considerados como "organizações"), não havendo razão para os colocar sob uma categoria comum. De facto, na generalidade dos casos, o género do determinante que precede um e outro tipo de entidade (masculino para clube e feminino para empresa, cf. *o Estrela da Amadora* e *a Salvador Caetano*) poderá servir para decidir qual a classificação adequada. Saliente-se, no entanto, que existem diversos casos em que não é possível tirar partido desta característica linguística. Para dar um exemplo, em relação às 834 ocorrências de *União de Leiria* no CETEMPúblico, verifica-se que em cerca de metade esta EM é precedida por um determinante feminino.

Retomando a análise dos resultados fornecidos pelos participantes, fizemos algumas experiências para tentar identificar quais as diferenças que poderiam ser facilmente atribuídas a questões de ordem terminológica e quais as que implicavam claramente diferenças de análise. Apresentamos em seguida alguns casos ilustrativos das duas situações:

CP9-2 Tomando a defesa do filme de Chahine, numa sala atulhada, o bastonário dos advogados, Ahmed al-Khawaga, replicou que o realizador egípcio se inspirou na história de José, "mas não se afastou das palavras do **Corão** que evoca, em termos claros, as propostas feitas ao profeta pela esposa do mestre que o comprou à chegada ao Egipto".
CP3-2 O **Audioman** foi desenvolvido na Suíça em apenas sete meses e compõe-se de um microfone e de um altifalante que se podem acoplar facilmente a um computador, devendo ser comercializado ao preço de 290 francos suíços (28 contos).
CP1-5 Não deixa de ser, nos tempos que correm, um certo "very typical" algarvio, cabeça de cartaz para os que querem fugir a algumas movimentações nocturnas já a caminho da ritualização de massas, do género "vamos todos ao **Calypso** e encontramo-nos na Locomia".
CP8-1 A propósito, no Museu da Segunda Guerra Mundial, que aí foi aberto, a história da maior guerra no continente europeu começa com a fotografia de Estaline a cumprimentar o ministro dos Negócios Estrangeiros da Alemanha nazi, ou seja, a guerra começa com a assinatura do **Pacto Molotov-Ribbentrop**.
CF11-1 "O Charade vai concorrer na faixa do Suzuki Swift e do **Twingo**, da Renault", afirma Caparelli.
CF11-2 O Applause, um sedã quatro portas, com motor 1.6, é o carro mais caro da **Daihatsu**.
CF15-3 Começou bem antes do que se previa a batalha pela futura sucessão na Fifa apenas seis meses depois do super-acordo que, nas vésperas da **Copa**, reconduziu o brasileiro João Havelange ao sexto mandato consecutivo.
CP5-1 Além do **Museu do Ar**, o projecto gira em torno do parque temático propriamente dito.

Corão teve as seguintes classificações: "obra literária", "obra", "livro", "título", "produto cultural" e "outro". Podemos considerar que talvez haja apenas uma distinção relevante a fazer entre, por um lado, "título" e, por outro, "obra/livro/obra literária/produto cultural". Uma vez que "título" não representa o tipo da entidade mencionada mas a forma convencional de a mencionar, observa-se uma diferença ontológica importante entre a

classificação das expressões linguísticas (caso do título) e a dos seus referentes no mundo real (caso de obra).

Em CP3-2, *Audioman* foi classificado como "produto", "marca", "objecto", "equipamento" e "outro".[4] Podemos estabelecer o mesmo raciocínio e sugerir que "marca" é a designação que se convencionou para mencionar um "produto/objecto/equipamento".

Em CP1-5, *Calypso* foi classificado como "lugar", "organização", "empresa" e "local de lazer". Neste caso, a classificação reflecte a vagueza inerente ao nome de uma empresa com uma função específica (lazer) localizada num sítio bem definido pelo contexo.

Em CP8-1, *Pacto Molotov-Ribbentrop* foi classificado como "documento", "acontecimento" e "outro". Este é um caso interessante, em que provavelmente a entidade mencionada é simplesmente PACTO. A sua assinatura pressupõe a existência de um documento e consiste ela própria num acontecimento. Claro que "documento" não seria uma classificação adequada se o texto não contivesse a palavra *assinatura*. A classificação PACTO não seria contudo posta em causa pela ausência de documento escrito.

A *Twingo* em CF11-1 foi atribuído o estatuto de "marca", "modelo", "automóvel", "carro", "marca de produto" e "outro". Neste contexto, "automóvel" e "carro" são sinónimos, e o mesmo se poderá dizer de "marca" e "marca de produto", que já discutimos. Existe, contudo, uma subtileza interessante entre "modelo de automóvel" e "automóvel", aliás ilustrada na própria frase em que esta entidade é mencionada.

Daihatsu foi classificado em CF11-2 como "organização", "marca", "empresa" e "lugar". Também neste tipo de situação nos parece, uma vez mais, difícil estabelecer satisfatoriamente as diferenças entre uma empresa e uma marca comercial, com idêntica designação.

Em CF15-3, *Copa* recebeu os atributos "evento", "acontecimento", "organização" e "outro". Ora, embora *Copa* seja um acontecimento desportivo organizado periodicamente, o texto refere-se a uma competição específica.

No último exemplo da amostragem, CP5-1, a entidade *Museu do Ar* foi classificada como "organização", "lugar" e "lugar de cultura", que reflecte diferentes propriedades associadas à entidade.

É ainda de referir que a observação e análise dos resultados obtidos levantou questões e dúvidas sobre as quais convém reflectir.

Vimos claramente que, em português (e provavelmente em todas as línguas), muitos nomes remetem para entidades difíceis de integrar em classificações gerais. Refira-se que, em apenas 30 extractos contendo cerca de 2600 palavras, surgem categorias tais como: pactos, medidas governamentais, programas de televisão, nomes de emissoras, etc. Até que ponto é que essas entidades se encontram em contexto sintáctico adequado (nível da frase) ou mesmo em discurso adequado (nível do texto) de modo a que o seu reconhecimento seja óbvio, é uma questão que se nos coloca. Comparem-se a este propósito as entidades *Hermitage* ou *Louvre* com *Museu do Ar*.

Um dos objectivos de uma avaliação conjunta é precisamente medir o peso relativo destes casos, desde as entidades com classificação trivial e óbvia (bastando constar no almanaque) até aos casos que exigem raciocínio complexo ou conhecimento do mundo por parte de um leitor e que podem mesmo não ser resolúveis sem fontes de conhecimento externas.

Há ainda a salientar que apenas um dos participantes estruturou as anotações. Por exemplo, para a sequência *Escola de Medicina de Harvard*, o anotador identificou toda a sequência como uma entidade mencionada, identificando igualmente *Harvard* como entidade encaixada no seu interior. Todos os outros não fizeram encaixe de entidades mencionadas, indicando sempre a maior sequência possível ou dando várias sem as incluir numa única entidade. Nos extractos do CETEMPúblico, dos dez casos em que não houve concordância do ponto de vista da delimitação da EM, metade teve a ver com a questão do encaixe; no CETENFolha, tal verificou-se em oito dos doze casos.

[4] Foi também atribuída a *Audioman*, julgamos nós por lapso, a categoria "pessoa".

6. Discussão no Avalon2003 e desenvolvimentos futuros

O trabalho descrito na secção anterior foi apresentado no Avalon'2003, e deu azo a uma viva discussão entre os participantes no encontro, envolvendo tanto aqueles que tinham contribuído para a anotação como outros que, não o tendo feito, se mostraram muito interessados no assunto. As duas perspectivas (a hierárquica e a plana) foram apresentadas, sem que houvesse consenso, como aliás seria de esperar.

Outra questão bastante polémica, e que também não reuniu consenso, teve a ver com a escolha dos textos a utilizar. Foram apresentadas três sugestões: (i) utilizar os mesmos textos da avaliação em RI, de modo a ter um recurso reutilizável e mais rico do ponto de vista da anotação (Aires *et al.* 2003), (ii) utilizar a colecção de textos da Web WPT03 que já está a ser marcada com entidades geográficas (Cardoso *et al.*, neste volume), ou (iii) utilizar textos jurídicos, uma vez que um dos grupos participantes já dispunha de textos nessa área (Quaresma & Rodrigues, neste volume).

Muitas outras perguntas colocadas nessa sessão ficaram (nessa altura) sem resposta, até porque o tempo era limitado: se as entidades seriam ou não encaixadas; se cargos, funções, títulos, etc., devem ou não fazer parte da entidade; se palavras do léxico comum que se convencionou grafar com inicial maiúscula e que aparecem isoladas ou em palavras compostas (*Governo*, *visita de Estado*, …), devem ou não ser reconhecidas; que ferramentas usar para auxiliar nos processos de anotação e revisão manual.

Alguns caminhos foram todavia delineados. A decisão mais importante foi a de que se deveria avançar na especificação e ensaio de um subconjunto deste problema, consensualmente definido como a identificação automática de entidades geográficas, ou de lugares. (Destacamos o concomitante início do projecto GREASE (Silva *et al.*, 2004)). Foi também sugerido que, depois desta fase, por alguns considerada excessivamente "tolerante", a organização tomasse uma atitude mais normativa e apresentasse, com base no estudo preliminar aqui descrito, um conjunto de etiquetas que pudessem satisfazer as aspirações dos participantes, de forma a que se pudesse avançar, a não muito longo prazo, com a organização de uma primeira avaliação conjunta em REM.

Esta movimentação em torno das EMs também serviu, além disso, para enriquecer dois dos recursos criados no âmbito da Linguateca: a Floresta Sintá(c)tica (Afonso *et al.*, 2002), de forma a uniformizar e padronizar a identificação de nomes próprios, e o COMPARA (Frankenberg-Garcia & Santos, 2003a), em que as entidades mencionadas passaram a ser também anotadas manualmente, além dos títulos, que já o eram. Ambos os recursos, embora não classificando mas apenas identificando, podem já considerar-se como repositórios deste tipo de entidades, o primeiro em texto jornalístico e o segundo em texto literário (e neste caso, nas duas línguas, português e inglês).

Como considerações finais, não será demais salientar novamente que, mesmo para uma actividade que parece simples, os diferentes investigadores chegaram a resultados distintos e, em alguns casos, não consensuais (ao nível de especificação do que estaria correcto), e que o trabalho de procurar obter especificações consensuais não é insignificante … mas extremamente compensador.

Pós-escrito: Apraz-nos verificar, passados mais de dois anos da escrita do presente artigo, que este documenta um esforço seminal que levou mais tarde à organização do HAREM (**H**AREM é **A**valiação conjunta de **R**econhecimento **E**ntidades **M**encionadas) pela Linguateca – pólos XLDB e Oslo – em 2004/2005 (veja-se Santos *et al.*, 2006). O HAREM constituiu um novo marco na avaliação conjunta da língua portuguesa, tendo as autoras participado de formas diferentes nessa actividade.