Capítulo 2

Estudo preliminar para a avaliação de REM em português

Cristina Mota

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 2, p. 19–34, 2007.

O presente capítulo visa relatar, de forma mais completa do que em Mota et al. (2007), uma actividade de prospecção realizada em 2003 que serviu de inspiração à organização do HAREM. Essa actividade consistiu na anotação manual ou semi-automática de uma pequena série de extractos do CETEMPúblico (Rocha e Santos, 2000), um corpus que integra artigos extraídos de 1500 edições diárias do jornal *Público*, e do CETENFolha, um corpus correspondente de português do Brasil criado com base no jornal Folha de São Paulo, de 1994. O seu principal objectivo foi preparar e motivar a participação numa futura avaliação conjunta dedicada a sistemas de REM, numa tentativa de compreender quais as categorias das entidades que os sistemas deveriam anotar, bem como estabelecer as directivas que deviam ser seguidas. Salienta-se desde já que, embora os participantes pudessem usar um sistema de base que os auxiliasse na anotação, o objectivo não era comparar o desempenho de sistemas mas sim o que os participantes consideravam como correcto. Apresentamos uma descrição da tarefa levada a cabo e uma análise dos resultados.

No âmbito do seu modelo de trabalho, IRA (Informação-Recursos-Avaliação), a Linguateca iniciou em 2002 actividades que visavam promover a avaliação conjunta de sistemas de processamento de linguagem natural. Estas actividades pioneiras para o processamento de textos escritos em português, bem como os seus primeiros resultados, encontram-se documentados em Santos (2002), Santos et al. (2004) e Santos (2007b). Uma das áreas de actuação escolhida foi a do REM, que começou por ficar a cargo do pólo da Linguateca no LabEL. Essa escolha deveu-se ao facto da presente autora, que na altura era colaboradora no pólo, ter já experiência no desenvolvimento de uma ferramenta de reconhecimento de entidades mencionadas para português.

O HAREM veio então no seguimento deste estudo preliminar, no qual em parte se inpirou. No entanto, houve modificações importantes que se encontram discutidas em vários outros capítulos deste livro, e por isso faz sentido documentar este estudo inicial de forma independente. A primeira tentativa de cristalizar esses passos iniciais foi realizada em Mota et al. (2007), mas dadas as restrições de tamanho (uma secção num capítulo de livro), apresentamos aqui uma descrição mais detalhada.

O arranque do processo deu-se no dia 29 de Janeiro de 2003 com o envio para a lista `avalia@linguateca.pt`, uma lista de divulgação para os investigadores interessados em avaliação conjunta, de uma mensagem com uma primeira proposta de avaliação. Essa proposta solicitava aos interessados na avaliação que anotassem manualmente, ou de forma automática combinada com revisão manual, um conjunto de extractos do CETEMPúblico e do CETENFolha. Esses exctractos anotados deveriam ser enviados até ao dia 21 de Fevereiro de 2003, tendo este prazo inicial sido adiado por coincidir com o prazo de submissão de artigos de várias conferências internacionais. Assim, a nova data estabelecida foi dia 10 de Março de 2003. Os exctractos enviados, bem como uma análise preliminar da classificação feita pelos participantes, foram disponibilizados no sítio da Linguateca logo em

29 de Janeiro de 2003	Envio da proposta inicial
10 de Março de 2003	Data limite para envio dos textos anotados
22 de Maio de 2003	Divulgação dos resultados
28 de Junho de 2003	Sessão de trabalho no AvalON 2003
Setembro de 2004	Início do HAREM

Tabela 2.1: Calendário da actividade preparatória.

seguida. A discussão dos resultados e a preparação de uma futura avaliação conjunta teve lugar no AvalON 2003, a 27 de Julho, na Universidade do Algarve. A Tabela 2.1 apresenta um calendário com as etapas desta actividade preparatória.

Neste capítulo, começamos por descrever a tarefa proposta, apresentamos a análise de resultados e, em jeito de conclusão, alguns comentários finais.

2.1 Descrição da Proposta

A proposta enviada sugeria duas linhas de acção a serem seguidas: a criação cooperativa de directivas; e a criação de recursos de avaliação.

Para a primeira linha de acção, numa primeira fase, pretendia-se estabelecer e caracterizar as entidades que os sistemas teriam de identificar, bem como de que forma as entidades deveriam ser anotadas no texto. Foram exemplificadas algumas entidades, adaptando a classificação do MUC (Grishman e Sundheim, 1995; Chinchor e Marsh, 1998) para 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)

Além disso, estabeleceu-se que as entidades deveriam ser marcadas com etiquetas SGML, tendo sido fornecidos exemplos de anotação em contexto, adoptando o esquema de marcação original do MUC, tal como se ilustra na Tabela 2.2.

PESSOA	(...) aquilo que <ENAMEX TYPE="PERSON">Fernando Pessoa</ENAMEX> tão expressivamente denominou (...)
ORGANIZAÇÃO	(...) a <ENAMEX TYPE="ORGANIZATION">Portugal Telecom</ENAMEX> voltou a ultrapassar (...)
LUGAR	(...) vai do <ENAMEX TYPE="LOCATION">Minho</ENAMEX> à região do (...)
DATA	Foi durante o <TIMEX TYPE="DATE">segundo semestre de 1992</ENAMEX> que a inflacção (...)
HORA	(...) se estipula as <TIMEX TYPE="TIME">4 horas da manhã</ENAMEX> como limite de (...)
MONETÁRIA	(...) com <NUMEX TYPE="MONEY">900 mil contos</ENAMEX> a fundo perdido (...)
PERCENTAGEM	(...) aos <NUMEX TYPE="PERCENT">sete por cento</ENAMEX> do capital (...)

Tabela 2.2: Exemplos de utilização de cada uma das etiquetas do MUC em extractos da Parte 20 do CETEMPúblico.

Esta linha de acção resultaria num conjunto de critérios e de recomendações (*directivas*) que deveria igualmente conter exemplos que ilustrassem o que devia e não devia ser marcado. A proposta chamava a atenção para algumas das muitas questões que se poderiam colocar e cuja resposta deveria ser tornada clara nas recomendações:

- Quais os tipos de nomes próprios que os sistemas deveriam ser capazes de identificar (e classificar)? Deveria um nome de um estabelecimento comercial (livraria, cinema, discoteca, etc.) ser identificado como uma organização?
- Os sistemas deveriam reconhecer entidades que incluíssem léxico não português, como por exemplo *Empire State Building*, *New York Times*, *BBC* ou *Manchester United*?
- O que fazer no caso de uma entidade estar encaixada noutra? Por exemplo, deveria *Lisboa* fazer parte do nome da organização, como no caso a), e não ser marcada como nome de lugar, ou deveria ser marcada como tal uma vez que não faz parte do nome da instituição, como no caso b) ?

  a) *(...) Crise na faculdade influencia eleições de amanhã para a reitoria da Universidade Técnica de Lisboa (...)*

  b) *(...) A Polícia Judiciária de Lisboa anunciou ontem a conclusão de três inquéritos respeitantes (...)*

A segunda linha de acção consistia na criação de recursos para a avaliação, que seriam anotados manualmente de acordo com os critérios e a classificação estabelecidos nas recomendações. Esses recursos de avaliação constituiriam uma colecção dourada que se-

ria usada como referência na comparação com os resultados produzidos pelos sistemas a partir do mesmo texto sem anotação.

Dado que estas duas linhas de acção poderiam ser desencadeadas em paralelo, foi então sugerido que se começasse por fazer a anotação de dois pequenos conjuntos de textos. A sua dimensão era pequena, apenas os dez primeiros extractos do CETEMPúblico (versão 1.7) e os primeiros vinte[1] do CETENFolha (versão 1.0), porque o objectivo era sobretudo motivar os investigadores para a tarefa. Apesar de tanto o CETEMPúblico como o CETENFolha serem públicos, os extractos para anotar foram disponibilizados no sítio da Linguateca. Deste modo, todos estariam certamente a usar a mesma versão do conjunto de textos. Alternativamente, também foi sugerido que os participantes, em vez de usarem extractos do CETEMPúblico e do CETENFolha, enviassem os textos que preferissem. Talvez por se ter chamado a atenção para o facto de que esta solução tornaria a comparação de resultados mais difícil, ninguém optou por escolher novos textos.

Findo o prazo de duas a três semanas para anotação, ter-se-ia material suficiente para observar a percepção que cada participante tinha sobre o REM, donde poderiam ser tirados resultados comparativos.

A mensagem enviada sugeria ainda que se adoptasse a classificação do MUC adaptada para português e continha o extracto 26 do CETEMPúblico com todos os nomes próprios anotados, quer estivessem ou não contemplados pela classificação do MUC (ver Figura 2.1).

Depois de ter sido enviada a mensagem inicial, precisou-se um pouco melhor a tarefa, aquando da disponibilização da informação no sítio da Linguateca. O objectivo seria que todas as sequências consideradas pelos participantes como sendo nomes próprios deveriam ser delimitadas com a etiqueta SGML `NOMEPROP`, em que o atributo TIPO deveria ter um dos seguintes valores: `PESSOA`, `ORGANIZACAÇÃO`, `LUGAR` ou `OUTRO`. Em alternativa, em vez de `OUTRO`, poderiam ser usadas etiquetas mais específicas, da escolha do participante.

2.2 Descrição dos textos

Como mencionado acima, foram anotados os primeiros dez extractos da versão 1.7 do CETEMPúblico e os vinte primeiros extractos da versão 1.0 do CETENFolha. As Figuras 2.2 e 2.3 mostram respectivamente a distribuição por semestre e por tópico nos dois conjuntos de extractos.

A variedade de semestres no CETEMPúblico deve-se ao facto de o corpus corresponder a 16 semestres compreendidos entre 1991 e 1998, enquanto o CETENFolha só contém edições do ano de 1994. Naturalmente que o conjunto destes extractos é demasiado pequeno para poder tirar quaisquer conclusões que sejam aplicáveis aos corpora completos.

[1] Foi inicialmente sugerido usar também os primeiros 10 extractos do CETENFolha; no entanto, se assim fosse, o número de nomes próprios dos dois subconjuntos seria muito díspar por isso o número de extractos deste corpus foi duplicado.

```
<ext n=26 sec=soc sem=91b>
<p>
<s>O caso ocorreu numa noite de 1978, na ilha de <NOMEPROP TIPO="LUGAR">
 Carvalo</NOMEPROP>, ao largo da <NOMEPROP TIPO="LUGAR">Córsega
 </NOMEPROP>.</s>
<s>O príncipe jantava com amigos num restaurante deste paraíso para
milionários, quando um grupo barulhento de jovens da alta sociedade
italiana acostou na enseada de
<NOMEPROP TIPO="LUGAR">Palma</NOMEPROP>, ao lado do seu iate, o
<NOMEPROP  TIPO="BARCO">L'Aniram</NOMEPROP>.</s>
<s>Os advogados da defesa sublinharam no processo que este facto perturbou
altamente o "senhor de <NOMEPROP TIPO="LUGAR">Sabóia</NOMEPROP>".</s>
<s>Naquele ano, as <NOMEPROP TIPO="ORGANIZAÇÃO">Brigadas Vermelhas
</NOMEPROP> (<NOMEPROP TIPO="ORGANIZAÇÃO">BR</NOMEPROP>) estavam no
auge da actividade terrorista, o líder cristão-democrata <NOMEPROP
TIPO="PESSOA">Aldo Moro</NOMEPROP> acabara de ser raptado, e o príncipe
-- proibido de entrar em <NOMEPROP TIPO="LUGAR">Itália</NOMEPROP>
desde o exílio do pai em 1946 -- teria mesmo recebido ameaças das
<NOMEPROP TIPO="ORGANIZAÇÃO">BR</NOMEPROP>.</s>
</p>
<t>Uma vida por um barco</t>
<p>
<s>O certo é que, pouco depois, <NOMEPROP TIPO="PESSOA">Vítor-Emanuel
 </NOMEPROP> apercebeu-se que um barco pneumático fora deslocado do seu
iate e atracado ao <NOMEPROP TIPO="BARCO">Cocke</NOMEPROP>, o navio dos
jovens italianos.</s>
<s>"Irritado com este acto de apropriação", foi buscar uma espingarda
<NOMEPROP TIPO="ARMA">US 30</NOMEPROP> semiautomática, utilizada em
safaris, e 31 cartuchos, e dirigiu-se para o <NOMEPROP TIPO="BARCO">Cocke
</NOMEPROP>.</s>
<s>Um dos jovens, <NOMEPROP TIPO="PESSOA">Nicola Pende</NOMEPROP>,
 acorda com um grito:</s>
<s>"Roubaste o meu barco, vais pagar."</s>
<s>Pouco depois, o príncipe aponta-lhe a arma ao ventre.</s>
<s>Na confusão que se segue, parte um primeiro tiro, depois um segundo, e
os dois homens caem ao mar.</s>
</p>
</ext>
```

Figura 2.1: Extracto 26 do CETEMPúblico, anotado pela autora.

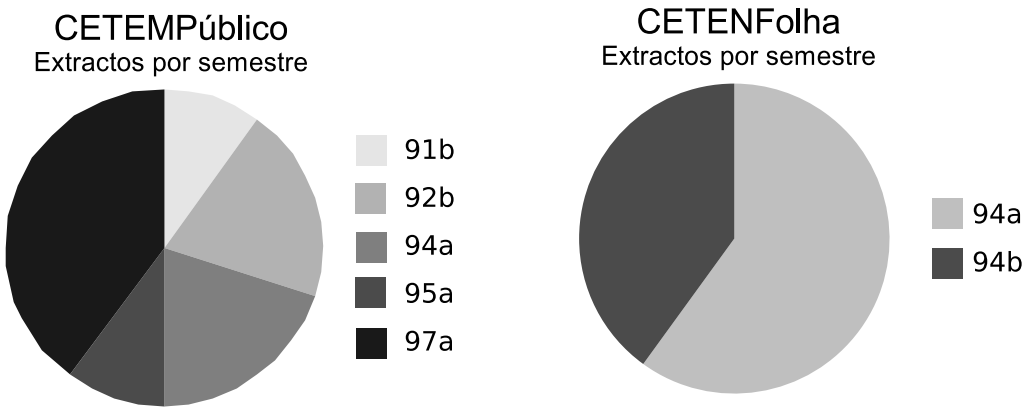

Figura 2.2: Distribuição dos extractos por semestre.

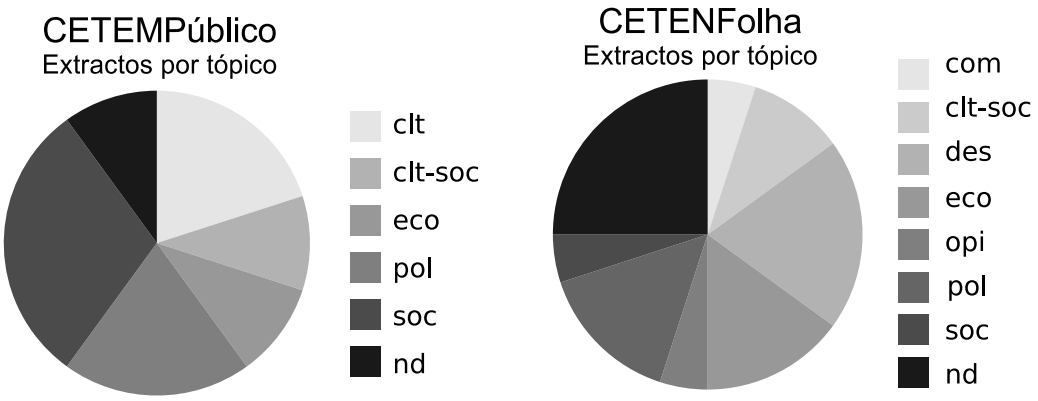

Figura 2.3: Distribuição dos extractos por tópico.

Para além dos extractos não terem sido escolhidos de modo a serem representativos do corpus completo, basta dizer que o semestre com mais extractos no corpus completo é o primeiro semestre de 1992 (92a), que nem sequer se encontra representado no conjunto dos dez extractos seleccionados.

Quanto aos tópicos, o CETENFolha apresenta mais variedade do que o CETEMPúblico. Tal como foi referido anteriormente, inicialmente tinham sido escolhidos também apenas dez extractos do CETENFolha. No entanto, como se pode constatar na Tabela 2.3, em média o subconjunto do CETENFolha apresenta um número significativamente inferior de palavras quer por parágrafo quer por frase, apesar de ter mais do dobro do número de frases e de parágrafos do subconjunto do CETEMPúblico (ver Figura 2.4).

Na Figura 2.4 mostra-se a frequência de várias unidades textuais. Entende-se por *átomo* qualquer sequência de caracteres delimitados pelo espaço; *palavra* são sequências de letras

Número médio	Palavras		Palavras com maiúsculas	
	CETEMPúblico	CETENFolha	CETEMPúblico	CETENFolha
Por parágrafo	82,60	28,37	7,80	3,39
Por frase	25,29	14,36	2,39	1,72

Tabela 2.3: Número médio de palavras e de palavras em maiúsculas por frase e por parágrafo nos dois subconjuntos seleccionados.

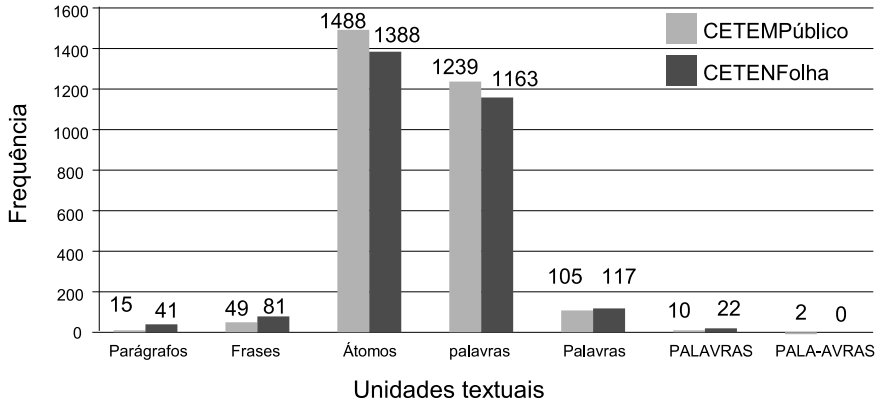

Figura 2.4: Número de ocorrências de várias unidades textuais.

e de caracteres, como o hífen e a barra; *Palavra* é qualquer palavra que comece por uma letra maiúscula; *PALAVRA* é qualquer sequência de letras em maiúsculas e *PALA-AVRAS* qualquer sequência de letras maíusculas e também hífens e barras.

Para se ficar também com uma ideia da variedade das sequências contíguas de palavras em maiúsculas (ou seja, sem considerar que um nome próprio pode conter determinadas palavras que podem não estar em maiúscula, como certas preposições), contabilizou-se o comprimento dessas sequências e o correspondente número de ocorrências (ver Figura 2.5). No CETEMPúblico existem sequências que variam entre comprimento 1 e 6 (não existindo sequências de comprimento 5), enquanto as sequências no CETENFolha variam entre 1 e 3.

2.3 Resultados

Participaram no exercício de anotação manual (ou automática com revisão) 9 participantes/anotadores. Na Tabela 2.4 encontra-se o nome dos participantes e das instituições a que pertenciam na altura.

Os resultados que a seguir se apresentam têm em conta as seguintes noções:

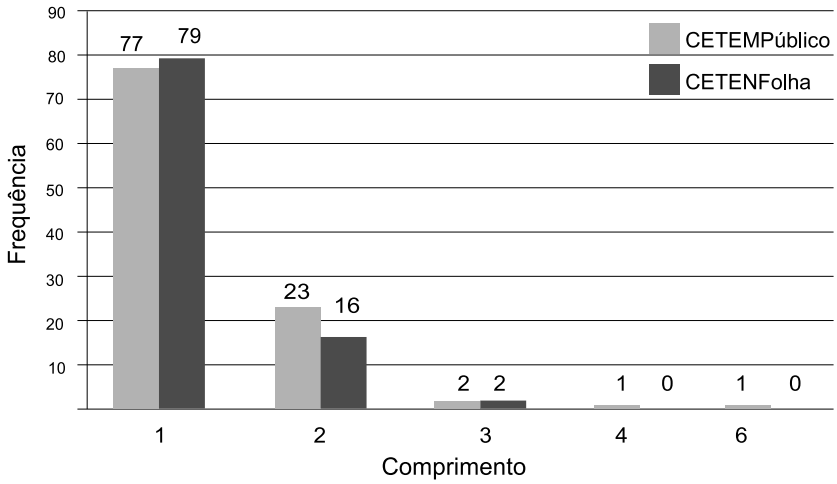

Figura 2.5: Número de ocorrências de sequências de palavras em maiúsculas de comprimentos *n*.

Identificador	Participante	Instituição
AS	Alberto Simões	Linguateca, Pólo do Minho
Prib	Cláudia Pinto	Priberam
CM	Cristina Mota	Linguateca, Pólo do LabEL
DS	Diana Santos	Linguateca, Pólo do SINTEF
EB	Eckhard Bick	Southern Denmark University
LO	Lucelia de Oliveira	NILC
Lab	Paula Carvalho	LabEL
RM	Raquel Marchi	NILC
VM	Vanessa Maquiafavel	NILC

Tabela 2.4: Participantes na tarefa de anotação.

- *entidade* corresponde a qualquer sequência delimitada com etiquetas SGML pelos anotadores;
- *nome próprio* corresponde a uma entidade marcada com a etiqueta `NOMEPROP`;
- *entidade (ou nome próprio) em comum* corresponde a uma sequência identificada por pelo menos um anotador, ou seja, uma sequência identificada consensualmente por um ou mais anotadores. Se para uma mesma sequência um anotador tiver identificado, por exemplo, *secretário de Estado* e outro tiver identificado apenas *Estado*, nenhuma das entidades contribuirá para o total de entidades em comum.

Foram calculadas três medidas de concordância na classificação:

- CE1: concordância relativa ao total de entidades em comum (ou seja, identificadas por pelo menos um anotador);
- CNP1: concordância relativa ao total de nomes próprios em comum (ou seja, identificados por pelo menos um anotador);
- CNPT: concordância relativa ao número total de nomes próprios identificados igualmente por todos os anotadores.

Foram tidos em conta os seguintes aspectos:

1. No caso de CE1 e CNP1, se um anotador não identificou uma entidade que outros reconheceram, essa entidade conta para o total de entidades em comum, mas não para o número de entidades em que há acordo;
2. Não se entrou em linha de conta com a subcategorização feita por alguns anotadores, ou seja, a concordância é relativa apenas à classificação feita usando o atributo `TIPO` ;
3. Dado que um dos anotadores propôs um conjunto bem variado de etiquetas que não contempla algumas das classes inicialmente sugeridas, estabeleceu-se a equivalência entre `ANTROPÓNIMO` e `PESSOA` e entre `TOPÓNIMO` e `LUGAR` (o estabelecimento desta última equivalência obrigou adicionalmente a substitutir a classificação das entidades marcadas originalmente por esse anotador como `LUGAR` por `LUGAR1`);
4. Ignorou-se igualmente que possa haver classes que são equivalentes por classificarem com nomes diferentes o mesmo conjunto de entidades (ou de nomes próprios), ou classes que possam estar completamente contidas noutras;
5. Não foram contabilizadas as entidades identificadas dentro das etiquetas SGML que já se encontravam nos extractos, uma vez que essas etiquetas correspondem a meta-informação estrutural do próprio corpus e como tal não deveriam ter sido analisadas[2].

2.3.1 Identificação de entidades

Como se pode ver na Figura 2.6, no CETEMPúblico foram identificadas de 81 a 106 entidades, enquanto no CETENFolha (Figura 2.7) o número de entidades identificadas variou entre 98 e 134. Destaca-se ainda que três dos nove anotadores identificaram exclusivamente nomes próprios, deixando sem marcação as expressões temporais e numéricas.

Combinando as entidades identificadas por pelo menos um anotador obtêm-se um conjunto de 140 entidades diferentes para o CETEMPúblico e de 163 para o CETENFolha. Desses conjuntos, respectivamente 63 e 70 entidades foram consensualmente identificadas

[2] Esta é uma das situações que mostra a falta de clareza nas instruções dadas aos anotadores.

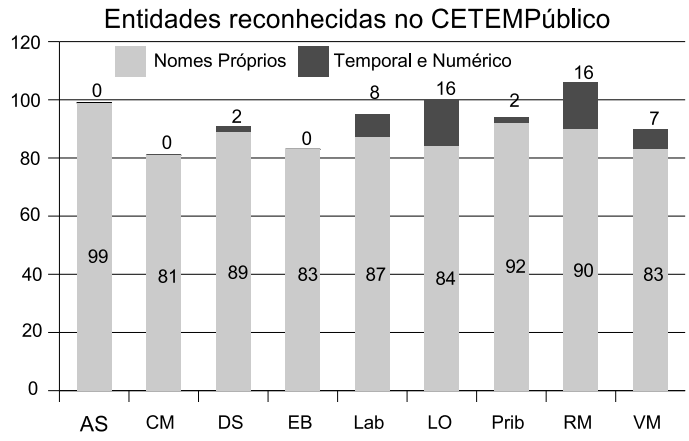

Figura 2.6: Total de entidades identificadas no CETEMPúblico por anotador.

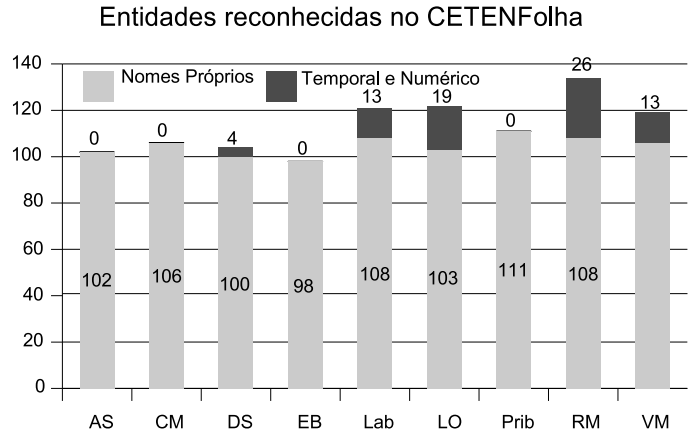

Figura 2.7: Total de entidades identificadas no CETENFolha por anotador.

por todos os anotadores, o que corresponde a 45% de concordância na identificação das entidades no CETEMPúblico e a 42,95% de concordância na identificação das entidades no CETENFolha. Se tivermos em conta apenas os nomes próprios então existe acordo na identificação em respectivamente 54,78% (63 em 115) e 56% (70 em 125) dos nomes distintos.

A lista das entidades comuns – ou seja, que foram identificadas por pelo menos um anotador e que não envolvem encaixe nem sobreposição com outras – e respectiva classificação encontram-se no apêndice B. Estas entidades correspondem a 67,86% (95 em 140) das entidades distintas do CETEMPúblico e a 74,85% (122 em 163) das entidades distintas

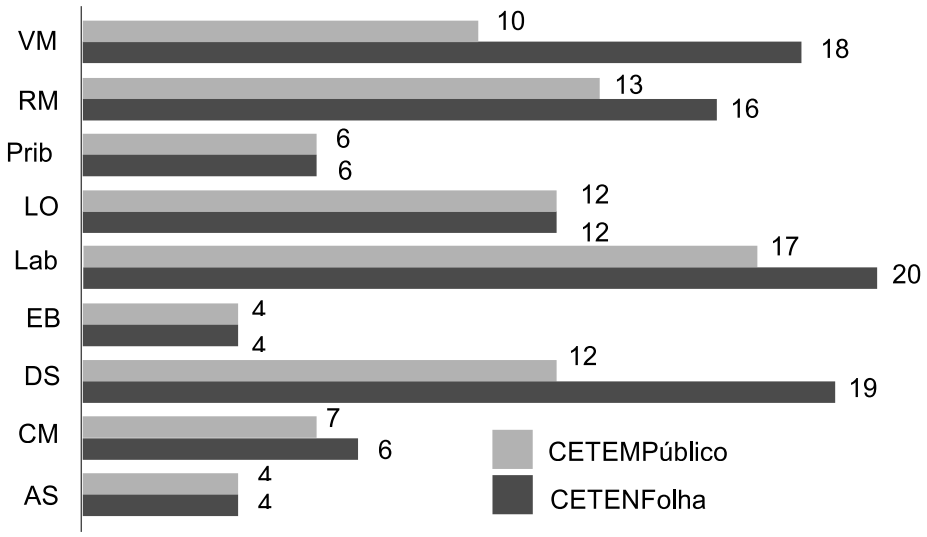

Figura 2.8: Total de categorias diferentes usadas por anotador.

do CETENFolha. O apêndice B também mostra as entidades para as quais não houve consenso na identificação, que também inclui as entidades que foram estruturadas (ou seja, têm outras entidades encaixadas. Apenas um anotador considerou este tipo de entidades.)

2.3.2 Classificação de entidades

Apesar do número de entidades ser bastante pequeno (cerca de uma centena), e de o número de categorias por anotador variar entre 4 e 20 (ver Figura 2.8, de facto, o número de diferentes categorias combinando as categorias de todos os anotadores é substancialmente elevado: 63 categorias no CETEMPúblico e 81 categorias diferentes usadas no CETENFolha. Esta variedade de categorias está bem patente em Mota et al. (2007, Figura 14.1) e que aqui se reproduz na Figura 2.9.

Naturalmente que, dada a variedade de etiquetas, a concordância quanto à classificação foi baixa (ver Tabelas 2.2 a 2.4). Note-se que os valores destas três tabelas não entram em consideração com as entidades que envolvem encaixe ou sobreposição com outras.

Se entrarmos também em consideração com os nomes próprios identificados por todos os anotadores que possam envolver encaixe ou estar sobrepostos com outros então obtemos 47,62% de concordância no CETEMPúblico (30 em 63) e 45,86% de concordância na classificação no CETENFolha (31 em 70).

Figura 2.9: União das etiquetas usadas no CETEMPúblico e no CETENFolha. Salienta-se a negrito as etiquetas originalmente propostas, gravitando à sua volta as etiquetas sugeridas pelos participantes.

	Entidades em comum (E1)	Entidades com mesma classificação	CE1
CETEMPúblico	95	30	31,58%
CETENFolha	122	30	24,59%

Tabela 2.5: Concordância na classificação das entidades comuns (CE1).

	Nomes próprios em comum (NP1)	Nomes próprios com a mesma classificação	CNP1
CETEMPúblico	79	30	37,97%
CETENFolha	98	30	30,61%

Tabela 2.6: Concordância na classificação dos nomes próprios comuns (CNP1).

	Nomes próprios identificados por todos os anotadores (NPT)	Nomes próprios com a mesma classificação	CNPT
CETEMPúblico	59	30	50,85%
CETENFolha	66	30	45,45%

Tabela 2.7: Concordância na classificação dos nomes próprios identificados por todos os anotadores (CNPT).

	AS	CM	DS	EB	Lab	LO	Prib	RM	VM	Média	Desvio Padrão
AS	100	89,47	93,42	90,79	92,11	86,84	93,42	93,42	86,84	90,79	2,63
CM	97,14	100	100	97,14	98,57	92,86	100	100	94,29	**97,50**	2,55
DS	95,95	94,59	100	95,95	97,3	93,24	100	100	93,24	96,28	**2,51**
EB	97,18	95,77	100	100	98,59	94,37	100	100	92,96	97,36	2,58
Lab	89,74	88,46	92,31	89,74	100	93,59	92,31	100	92,31	92,31	3,33
LO	81,48	80,25	85,19	82,72	90,12	100	85,19	95,06	85,19	85,65	*4,53*
Prib	95,95	94,59	100	95,95	97,3	93,24	100	100	93,24	96,28	**2,51**
RM	76,74	76,74	80,23	76,74	83,72	80,23	80,23	100	86,05	*80,09*	3,21
VM	89,19	89,19	93,24	89,19	97,3	93,24	93,24	100	100	93,07	3,73
Média	90,42	*88,63*	93,05	89,78	94,38	90,95	93,05	**98,56**	90,52		
Desvio Padrão	*7,25*	6,47	6,82	6,63	4,97	4,61	6,82	**2,53**	3,54		

Tabela 2.8: Acordo entre pares de anotadores na identificação das entidades no CETEMPúblico.

2.3.3 Quadros comparativos entre pares de anotadores

De modo a perceber até que ponto é que as entidades identificadas por um dado anotador são consensuais, calculámos em relação às entidades que cada um dos anotadores reconheceu a percentagem de entidades identificadas também por cada um dos outros anotadores. As Tabelas 2.8 e 2.9 apresentam esses valores para o CETEMPúblico e para o CETENFolha, respectivamente.

Por exemplo, a célula (CM,DS) na Tabela 2.8 indica que todas as entidades identificadas por CM foram igualmente identificadas por DS; a célula (DS,CM) na mesma tabela indica que das entidades identificadas por DS, 94,5% foram igualmente identificadas por CM. Isto significa que DS identificou todas as que CM identificou e mais algumas. A média e o desvio padrão de uma coluna dão uma ideia de quanto é que o anotador representado na coluna concorda com os anotadores representados nas linhas; a média e o desvio padrão de uma linha indicam quanto é que anotadores representados nas colunas concordaram com a anotação do anotador representado nessa linha. Ou seja, se o desvio padrão for alto para uma linha, isso significa que esse anotador é polémico, pois há uns anotadores que concordam mas outros que discordam muito dele; se o desvio padrão for alto na coluna, isso significa que o anotador discorda mais de uns anotadores do que de outros.

2.4 Comentários finais

Tal como já referido anteriormente, todos os resultados aqui apresentados, incluindo os textos marcados por cada um dos anotadores bem como as entidades integradas em concordâncias, ficaram públicos no sítio da Linguateca antes do encontro presencial ter decorrido.

Aquando dessa sessão, além de como chegar a um consenso quanto à escolha das categorias, foram ainda levantadas mais algumas questões, que ficaram também em aberto

	AS	CM	DS	EB	Lab	LO	Prib	RM	VM	Média	Desvio Padrão
AS	100	97,59	93,98	92,77	97,59	93,98	98,8	96,39	93,98	**95,64**	**2,08**
CM	86,17	100	91,49	91,49	100	95,74	100	97,87	95,74	94,81	4,5
DS	88,64	97,73	100	92,05	97,73	94,32	98,86	96,59	94,32	95,03	3,21
EB	88,5	89,9	93,1	100	98,9	96,6	100	98,9	95,4	95,16	4,09
Lab	79,41	92,16	84,31	84,31	100	95,1	92,16	98,04	95,1	90,07	6,15
LO	74,29	85,71	79,05	80	92,38	100	86,67	94,29	89,52	*85,24*	6,51
Prib	85,42	97,92	90,63	90,63	97,92	94,79	100	96,88	95,83	93,75	4,17
RM	69,64	80,36	74,11	74,11	86,61	83,93	82,14	100	88,39	79,91	6,23
VM	74,29	85,71	79,05	80	92,38	86,67	94,29	89,52	100	*85,24*	*6,51*
Média	69,73	79,65	74,08	85,67	83,08	80,57	81,62	83,7	81,61		
Desvio Padrão	27,09	30,73	28,76	6,61	31,66	30,73	31,4	31,74	30,96		

Tabela 2.9: Concordância entre pares de anotadores na identificação das entidades no CETENFolha.

para a futura realização da avaliação conjunta, nomeadamente:

1. Que sequências considerar como entidades mencionadas? Nomes próprios? Ou também expressões temporais e numéricas?

2. Deveria ser considerada a constituição interna das entidades permitindo a delimitação de entidades encaixadas noutras? Por exemplo, **<EM>** `Escola de Medicina de Harvard` **</EM>** versus **<EM>** `Escola de Medicina de` **<EM>** `Harvard` **</EM> </EM>**.

3. O que fazer com cargos, títulos e funções? Integrá-los na delimitação da entidade como em **<EM>** `Presidente Jorge Sampaio` **<EM>** ou ignorar, pretendendo-se `Presidente` **<EM>** `Jorge Sampaio` **<EM>**? Mas e se o cargo, por exemplo, não começar por maiúscula como em `major Carlos Barbosa`?

4. Atribuir-se-á a etiqueta em função do contexto? Compare-se por exemplo *(...) feira especializada que teve lugar em Basileia(...)* com *(...)chegará o dia em que a Rússia ajudará(...)*.

5. O que fazer quando não é possível decidir? Anotar ou ignorar?

Além disso, delineou-se um primeiro esboço dos passos a tomar na primeira avaliação conjunta de sistemas de REM, no sentido de continuar o trabalho iniciado com a experiência que relatámos:

1. Estabelecer o conjunto de etiquetas e regras de anotação a adoptar;

2. Realizar um nova anotação manual com os mesmos textos usando o novo conjunto de etiquetas, tendo se sugerido a utilização de uma ferramenta auxiliar de anotação, como por exemplo o Alembic Workbench (Day et al., 1997) que facilitaria não só o processo de anotação manual como também o de revisão e comparação das anotações;

3. Seleccionar e preparar os textos. Uma sugestão consistia em utilizar os mesmos textos que fossem utilizados na avaliação de recuperação de informação e sumarização automática, com o objectivo de ter um recurso reutilizável e mais rico;

4. Fazer uma pré-inscrição;

5. Propor um calendário para a avaliação.

Após quatro anos decorridos, penso que as conclusões mais salientes do presente ensaio foram que ele demonstrou indubitavelmente haver interesse da parte da comunidade, mas grande necessidade de consenso, o que talvez tenha motivado os organizadores a tomar uma atitude mais impositiva na condução da própria avaliação conjunta.