# CLEF: Abrindo a porta à participação internacional em RI do português

*Paulo Rocha e Diana Santos*

Neste artigo relatamos o trabalho da Linguateca na adição do português ao CLEF, uma avaliação conjunta internacional de recolha de informação (RI) em várias línguas. Após uma breve descrição desta avaliação conjunta na secção 1, abordamos na secção 2 os motivos e o interesse que tal participação assumiu do ponto de vista da Linguateca. Desta forma, não só damos um contributo para a documentação dos sistemas para os quais já existem recursos de avaliação para o português, como ilustramos detalhadamente mais uma avaliação conjunta, com características muito diferentes das Morfolimpíadas. Na secção 3 explicamos o nosso trabalho e as variadas opções tomadas para produzir dados para a competição, assim como para efectuar a sua avaliação. Finalmente, na secção 4 descrevemos o progresso desde a primeira edição em que o CLEF abrange o português, fazendo também algumas sugestões de melhoria para o futuro.

## *1. O que é o CLEF?*

O CLEF[1] (Peters *et al.*, 2004) começou como uma pista de CLIR ("cross-language information retrieval") no TREC e desenvolveu-se — em estreita cooperação com o TREC, e com as iniciativas japonesas — como um projecto europeu de recolha cruzada ("crosslingual IR"), cujo foco principal é promover a diversidade das abordagens e a multiplicidade das línguas (das línguas dos utentes, e das línguas em que está expressa a informação).

Assim, de 2000 a 2004 houve várias tarefas monolingues para incentivar a investigação em línguas menos usadas, assim como tarefas multilingues e de recolha cruzada. Conforme o interesse dos participantes, tarefas diferentes foram vindo a ser organizadas, tais como a RI interactiva, a RI sobre colecções científicas e a resposta automática a perguntas (RAP). Veja-se, a título de exemplo, as descrições pormenorizadas de cada tarefa nas últimas actas do CLEF (Peters *et al.*, 2005).

Embora a Linguateca já tivesse estabelecido contacto com a organizadora antes, apenas em 2003 foi possível garantir a participação do português na avaliação conjunta de 2004. O presente artigo versa sobretudo sobre essa edição, mas actualizado com os progressos até 2006 quando tal for pertinente.

### 1.1 Participar na organização de uma avaliação conjunta internacional

Dois dos pressupostos mais importantes da actividade da Linguateca são os de que o estudo e processamento do português não são apenas traduções, adaptações ou apropriações dos sistemas desenvolvidos para outras línguas e, por outro lado, de que os investigadores de língua materna portuguesa são e serão sempre os melhores avaliadores do que se faz com a nossa língua (Santos, 1999c).

Por isso, pode parecer contraditório que a Linguateca tenha contribuído para que a segunda avaliação conjunta em que a língua portuguesa participa seja organizada por uma comissão internacional, em que todas as línguas participam em pé de igualdade, sem ter seguido o modelo de juntar primeiro os actores e investigar o problema e a dificuldade, como proposto em Santos (neste volume).

[1] http://www.clef-campaign.org/

Por outro lado, há razões de índole diversa que justificam a nossa opção pela participação em tal avaliação, em vez da criação de raiz de todo um sistema de avaliação:

Do ponto de vista de trabalho envolvido, uma tal solução permite-nos concentrar os nossos poucos recursos precisamente na parte interessante, ou seja, a parte relacionada com o português. Desta vez, não precisamos de nos preocupar com a comunicação e divulgação, com a infra-estrutura e o registo dos participantes, com a manutenção da informação no sítio da rede, ou com outras tarefas essencialmente burocráticas.

Por outro lado, o português é "só" mais uma língua. E vários grupos internacionais (no sentido de que não processam o português) podem e querem treinar e testar os seus algoritmos sobre a nossa língua. Ou seja, fornecendo o português a grupos possivelmente mais avançados (porque já andam nestas lides há mais tempo) podemos também fomentar o trabalho sobre o português a nível internacional (admitindo, como é o caso, que esses participantes nunca teriam tido interesse ou possibilidade de participar numa competição só sobre português)[2].

Além disso, é importante referir que a principal razão de existir do CLEF é precisamente medir e incentivar a procura cruzada: procurar ou perguntar numa língua e obter respostas em colecções em outras línguas. Ou seja, todas as línguas estão em posição de igualdade por definição. Não se trata, pois, de aplicar o "mesmo" modelo a várias línguas, mas pelo contrário ter um modelo que permita lidar com várias línguas.

Em último lugar, uma boa razão para participar — mesmo que continuássemos convencidos de que se deveria começar por fazer avaliações conjuntas só para o português — é que essa seria a única forma de fundamentar a nossa crítica, ou de poder mudar de opinião.

Simplesmente do ponto de vista científico, e deixando agora o terreno da estratégia política da ciência e da forma de melhorar o processamento do português, o facto de toda a competição ser multilingue leva também a novos desafios que envolvem o português, e que não podem ser subestimados. Por exemplo, como mostraremos no resto do artigo, as seguintes questões não são triviais e merecem ser discutidas e investigadas:

- Como avaliar ou ponderar a relevância dos tópicos para um público lusófono?
- Como identificar a relevância dos tópicos lusófonos (ou relacionados com a lusofonia) para um público internacional (por exemplo, outros participantes lidam com o finlandês, o francês, o italiano ou o búlgaro)?
- E, claro, toda a problemática associada à tradução para e do português.

### 1.2 Comparando o CLEF e as Morfolimpíadas

No contexto do presente livro, faz sentido apresentar um breve esboço das diferenças entre a organização do CLEF e das Morfolimpíadas:

Em primeiro lugar, no CLEF há muito menos envolvimento dos sistemas participantes na definição das medidas e do tipo de perguntas, ou mesmo na definição da informação que deva estar na colecção. Isto deve-se, obviamente, ao muito maior número de participantes e ao facto de estes começarem já numa versão "estabelecida" da campanha (e não a primeira).

Em segundo lugar, nem todos os programas usados na avaliação e os recursos para comparação são, por enquanto, acessíveis aos participantes, donde a organização

[2] Sobre este aspecto, note-se que seria necessário que tivessem elementos do grupo que falassem e escrevessem português, o que evidentemente não é impossível, mas será, estatisticamente falando, menos provável.

tem um controlo muito maior sobre os resultados, e a avaliação conjunta torna-se muito menos transparente[3].

Tal pode, contudo, dever-se apenas ao grande número de grupos organizadores e participantes, embora os casos em que várias línguas trabalham em conjunto possam levar a casos intrincados de direitos de autor.[4] Contudo, note-se que, sem as colecções e os julgamentos de relevância em forma de monte ("pool"), os tópicos servem de pouco, e o mesmo se passa com as respostas às perguntas usadas na RAP. Em relação a esta última, Magnini *et al.* (2003) produziram uma colecção multilingue (com perguntas e respostas em três línguas: italiano, holandês e espanhol), com as perguntas e repostas usadas no CLEF 2003, mas sem os ponteiros para os documentos (ou seja, sem a informação de onde encontrar a resposta na colecção).

Para evitar uma tal situação,[5] a Linguateca garantiu, através de contacto com os detentores dos direitos de autor, que tanto a colecção de documentos como os recursos criados seriam disponibilizados logo após decorrerem as diversas avaliações conjuntas, seguindo a filosofia de máxima abertura e de disponibilização pública de todo o seu trabalho. Assim, foi criada a colecção CHAVE[6], actualizada após cada edição, que contém, além dos textos completos do PÚBLICO de 1994 e 1995,

- as listas de tópicos em português, compilados em cooperação com os restantes organizadores do CLEF; assim como as avaliações (binárias) de cada tópico
- as listas de perguntas e respostas em português, compiladas em cooperação com os restantes organizadores do QA@CLEF, com a indicação de um conjunto não-exaustivo de documentos que justificam a(s) resposta(s) para um subconjunto dessas perguntas

## *2. O que significou adicionar o português*

Aliviados da parte burocrática de coordenação, o nosso trabalho consistiu essencialmente na preparação do material para utilização na avaliação, e na avaliação dos resultados.

### 2.1 A colecção

Uma vez que, ao contrário de outras línguas que participaram nas campanhas anteriores do CLEF, não existia ainda uma colecção de textos em português para uso nesta avaliação conjunta, o nosso primeiro passo consistiu na preparação de uma tal colecção. Obedecendo às directivas, essa colecção foi criada com base em texto jornalístico dos anos de 1994 e 1995, neste caso os mesmos textos do diário Público que foram utilizados para a construção do CETEMPúblico (Rocha & Santos, 2000). No ano subsequente, o mesmo processamento foi aplicado às edições de 1994 e 1995 do jornal A Folha de São Paulo. A colecção foi então dividida em documentos, cada um com uma identificação única, marcada da forma ilustrada na Figura 1-1 abaixo. Note-se que não aparecem explicitamente marcados nem títulos nem autores nem o domínio (assunto, secção do jornal) em que foram publicados.

---

[3] Lembramos que, logo que os participantes das Morfolimpíadas entregaram os seus resultados, a lista dourada foi tornada pública, e que pouco tempo medeou entre a sessão final e a disponibilização total dos resultados, programas e recursos para o público em geral.

[4] Por exemplo, o conjunto de tópicos para RI cruzado (e que também é o usado na RI monolingue) foi criado em conjunto, com base no trabalho dos vários grupos, o que poderia levar a que nenhum grupo tivesse possibilidade de os disponibilizar sozinho

[5] À data da escrita do presente artigo, ainda não é possível, a quem não participou no CLEF, obter as colecções referentes à maior parte das outras línguas, embora tal tenha sido prometido em 2004.

[6] http://www.linguateca.pt/CHAVE/

```
<DOC>
<DOCNO>PUBLICO-19950326-091</DOCNO>
<DOCID>PUBLICO-19950326-091</DOCID>
<DATE>19950326</DATE>
<CATEGORY>Diversos</CATEGORY>
<TEXT>
«Os Donos da Bola»
Um erro de «casting»
Ao promover uma informação descomprometida, com o critério da conveniência a
dar lugar ao critério da actualidade, a SIC alterou substancialmente o
panorama da informação televisiva em Portugal. «Os Donos da Bola» surgem,
deste modo, como a primeira tentativa de modificar um cenário -- da informação
desportiva -- próximo da indigência. A ideia foi fazer o primeiro «talk show»
da especialidade em Portugal»; fazer, afinal, diferente e melhor. E bem se
pode dizer que os resultados são claramente positivos. Os diálogos entre Pinto
da Costa e os ex-jogadores do FC Porto Octávio Machado e Fernando Gomes, entre
Valentim Loureiro e Sousa Cintra ou entre o jornalista Marinho Neves e o
funcionário portista Jorge Gomes ficaram na memória dos telespectadores como
excelentes momentos de televisão.
(...)
A presença do famigerado Silva, agressor confesso do jornalista Paulo Martins
da RTP, acabou entretanto por servir de pretexto para um dos momentos mais
interessantes do programa. Ficou claro aos olhos da opinião pública, pela voz
autorizada do ex-conselheiro Sequeira Teles, que os árbitros têm
inclusivamente medo de noticiar nos seus relatórios as pressões e as sevícias
de que são vítimas nos túneis de acesso às cabinas dos estádios portugueses.
Esta foi, aliás, uma das raras alturas em que os convidados não se ficaram
pelas meias palavras. Porque José Silvano, José Guímaro e Marques da Silva, as
estrelas da noite, ficaram-se mais uma vez pela ameaça de um dia revelarem o
que mais uma vez fizeram questão de calar.
Perdeu-se, assim, mais uma oportunidade de fazer luz sobre o reino das trevas.
Jorge Santos
</TEXT>
</DOC>
```

Figura 1-1. Um texto da colecção CHAVE (abreviado pela remoção de um parágrafo)

## 2.2 Tópicos para Recolha de informação (RI)

Uma vez distribuída a colecção aos participantes, o segundo passo consistiu na preparação do material necessário para a avaliação de RI, ou seja, a escolha de aproximadamente quinze tópicos referidos nos documentos da colecção do ano de 1995, não usados nas campanhas dos anos anteriores, e que não fossem demasiado frequentes. Evitámos igualmente tópicos cujo aparecimento nas restantes colecções fosse altamente improvável (digamos, a final da Taça de Portugal em andebol feminino).

Alguns dos tópicos foram escolhidos com o apoio da cronologia do ano de 1995 da Wikipedia[7]. Outros tópicos reflectiam preferências pessoais do primeiro autor deste artigo. Os tópicos podem ainda ser divididos por um lado em tópicos gerais (arte rupestre), e por outro em eventos ocorridos no ano de 1995 (a possível descoberta do túmulo de Alexandre o Grande), incluindo eventos que se repetem todos os anos (tal como a Volta a França). Em Santos & Rocha (2005) sugerimos uma tipologia mais elaborada que levasse em conta a possível utilidade dos tópicos como necessidades de informação. Apresentamos dois exemplos de tópicos na Figura 1-2.

```
<top>
<NUM>PT-06</NUM>
<PT-title>Construção de barragens</PT-title>
<PT-desc>Encontrar documentos sobre construção de barragens</PT-desc>
<PT-narr>Documentos relevantes devem mencionar barragens e outra informação
adicional disponível. Só barragens de construção humana, mesmo que ainda só
planeadas, são relevantes. Documentos que apenas mencionem o nome de uma
barragem e sua localização são irrelevantes.</PT-narr>
```

[7] www.wikipedia.org.

```
</top>

<top>
<NUM>PT-28</NUM>
<PT-title>Vitória de Fernanda Ribeiro</PT-title>
<PT-desc>Documentos sobre a vitória de Fernanda Ribeiro nos 10000m nos
Campeonatos Mundiais de Atletismo em Gotemburgo</PT-desc>
<PT-narr>Os documentos são relevantes se descreverem a final dos 10.000m nos
Mundiais de Atletismo de Gotemburgo, culminando na vitória de Fernanda
Ribeiro. Listas de resultados finais não são relevantes.</PT-narr>
</top>
```

Figura 1-2. Exemplos de tópicos de RI propostos pelo grupo português

Note-se que tentámos ter o cuidado de apresentar, espalhado pela definição, título, e explicação, formas alternativas de referir ou mencionar o assunto, para ajudar tanto quanto possível os sistemas. Apenas não o conseguimos fazer em casos flagrantes em que, dependendo da variante do português, a denominação é completamente diferente, porque tal poderia prejudicar os sistemas ao procurar na colecção portuguesa. Assim, no tópico 217, na Figura 1-3, usámos apenas *sida* e não *aids*.

```
<top>
<num> C217 </num>
<PT-title> Sida em África </PT-title>
<PT-desc> Encontrar documentos discutindo o crescimento da sida em África.
</PT-desc>
<PT-narr> Houve um aumento explosivo da sida em África. Documentos relevantes
discutirão este problema. De particular interesse são documentos mencionando
organizações humanitárias lutando contra a sida em África. </PT-narr>
 </top>
```

Figura 1-3. Um tópico cuja especificação não cobre a variante brasileira

Esses tópicos foram em seguida traduzidos para inglês, e enviados para a coordenadora (Carol Peters), junto com informação adicional para ajudar os organizadores responsáveis pelas outras línguas a procurar essa informação na colecção deles, veja-se a Figura 1-4.

```
<top>
<NUM> PT-18 </NUM>
<EN-title> East Timor Guerrilla </EN-title>
<EN-desc> Find documents on the guerrilla activity in East Timor </EN-desc>
<EN-narr> Relevant documents should mention concrete political or military
activities of the East Timorese guerrilla. </EN-narr>
</top>

Background info: despite the arrest of its leader, Xanana Gusmão, in 1992, the
East Timorese guerrilla did not stop military activities against the
Indonesian occupation.
Number of hits in Público 1995: 10+ documents.
```

Figura 1-4. Um tópico proposto pelo grupo português traduzido para inglês e com informação adicional

Tivemos, em seguida, de verificar o número de ocorrências dos tópicos sugeridos pelos restantes grupos organizadores. Dos 98 tópicos fornecidos pelos seis grupos organizadores, foram colectivamente escolhidos cinquenta, que tivemos que traduzir para português. Para cinco não conseguimos encontrar resposta na colecção portuguesa.

Note-se que a tradução de um dado tópico nem sempre é trivial: Por exemplo, Kluck (2004:15) refere que o tópico holandês “Muisarm” tem de ser traduzido para o castelhano “Síndrome RSI y ratones de ordenador”. Kluck & Womser-Hacker (2002), por seu lado, apresentam diversos exemplos onde o uso de abreviaturas e/ou de termos

culturais específicos ("culture specific terms") apresentam problemas interessantes. Cesária Évora, descrita como a *Diva dos pés descalços*, apareceu na versão holandesa como “barefoot diva” — em língua inglesa, portanto.

### 2.3 Perguntas para resposta automática a perguntas (RAP)

Em paralelo, preparámos o material para o QA@CLEF. A fase inicial consistiu na escolha de 100 perguntas com resposta na nossa colecção.

Segundo as directivas da organização, essas perguntas deviam ter uma resposta nos textos de 1994 e 1995 da colecção, não serem subjectivas (tais como *Quem foi o maior poeta português do século XVI?*), não serem perguntas de confirmação, que exigem uma resposta sim ou não (*Foi Camões que escreveu «Os Lusíadas»?*), não conterem outras perguntas (*Quem era o rei de Portugal quando Camões nasceu?*) nem serem perguntas sobre causas (*Porque é que Camões fugiu para a Índia?*). Além disso, deviam ser excluídas perguntas cuja resposta fosse uma lista (*Mencione obras de Camões*), embora perguntas do tipo *Mencione uma obra de Camões* fossem válidas.

Pelo nosso lado, evitámos perguntas cuja análise sintáctica nos parecesse demasiado complexa, perguntas com respostas demasiado complicadas, e perguntas que considerássemos demasiado artificiais. Tentámos, além disso, ao escolher as perguntas, manter um balanço entre os diferentes tipos de resposta esperados. A distribuição resultante, em que D significa definição, e F factóide, encontra-se na Tabela 1-1.

Tabela 1-1: Distribuição dos tipos de perguntas

| | | |
|---|---|---|
| D | ORGANIZATION | 3 |
| D | PERSON | 8 |
| F | LOCATION | 22 |
| F | MANNER | 2 |
| F | MEASURE | 11 |
| F | OBJECT | 6 |
| F | ORGANIZATION | 9 |
| F | OTHER | 10 |
| F | PERSON | 20 |
| F | TIME | 9 |

Um total de 49 perguntas, ou seja, quase metade, focava assuntos portugueses ou dos outros países de língua portuguesa. A sintaxe das perguntas indica, a seguir ao número, o tipo (F ou D), o subtipo, e depois um conjunto de respostas enumeradas. “LING-etc.” é a identificação de um documento da colecção portuguesa (LING para Linguateca) em que a resposta pode ser encontrada[8].

```
609 F LOCATION   Em que distrito fica Paredes de Coura?
   1 LING-940112-077 Viana do Castelo
626 F PERSON   Quem é a «diva dos pés descalços»?
   1 LING-940121-089 Cesária Évora
647 F ORGANIZATION   Que empresa tem uma refinaria em Leça da Palmeira?
   1 LING-941223-116 Petrogal
```

Nos casos em que as perguntas podiam ser formuladas de formas diferentes e igualmente comuns (tal como *Quantos anos tem X?* e *Que idade tem X?*), tentámos usar diferentes formas em diferentes perguntas, mas sempre favorecendo a que levasse à resposta mais parecida com a que se encontrava na colecção. Veja-se os seguintes exemplos:

[8] Neste artigo apresentamos perguntas em três formatos diferentes: perguntas criadas por nós – com número maior do que 600; perguntas cujas respostas foram traduzidas por nós mas não identificadas com uma justificação, que têm a tradução após a palavra SEARCH; e perguntas na sintaxe final do recurso monolingue, em que, além da tradução da resposta, está a justificação desta na nossa colecção.

```
558 F OTHER Qual a nacionalidade do tenista Sergi Bruguera?
   1            SEARCH[espanhol]
582 F LOCATION De que país é a escritora Taslima Nasreen?
   1            SEARCH[Bangladesh]
```

Arranjar perguntas *Como* foi contudo muito difícil – na lista final, apenas figuram duas. Curiosamente, a maioria das perguntas de MANNER na colecção internacional eram relacionadas com a causa da morte (por assassínio ou "overdose"); a formulação de outras, como discutido em Santos & Rocha (2005), variava conforme a língua: *o que é?* ou *como se faz?*.

```
675 F MANNER Como morreu Pasolini?
   1            SEARCH[assassinado]
676 F MANNER Como se tornou o Brasil tetracampeão de futebol?
   1            SEARCH[através da marcação de pontapés da marca de penalti]
```

Outras questões complicadas surgiram, quer na procura de perguntas válidas, quer na tradução de perguntas propostas pelos outros grupos: o que fazer quando a pergunta é no masculino (genérico) e a resposta é no feminino, ou vice versa? Por exemplo, devemos usar *Quem é a ministra do ambiente sueca?* ou *Quem é o ministro do ambiente sueco?* quando a titular é uma mulher?[9]

Igualmente questionável é se devemos ou não fazer perguntas com pressuposições erradas: *Como se chama a filha do líder chinês Deng Xiaoping?* pressupõe que ele tinha uma única filha (quando na realidade tinha duas), ou pressuposições desconhecidas, como em *Qual o nome da amante de Mussolini?* que indica implicitamente que Mussolini só teve uma amante — ou que se deverá presumir *a única amante conhecida de Mussolini*. Além disso, esta última pergunta afirma explicitamente que a pessoa de que se está à procura é do sexo feminino. O que fazer na tradução para línguas em que tal não é fácil de exprimir? Em inglês, poder-se-ia usar a palavra "mistress", mas perder-se-ia essa informação se fosse usada a palavra "lover".

Em ambos os casos, optámos pela solução que presumimos mais simples para os sistemas actuais, i.e., usámos o feminino na pergunta quando a resposta o exigisse, e tentámos evitar usar formulações com pressupostos incorrectos.

As perguntas obtidas foram, em seguida, traduzidas para inglês e enviadas para os coordenadores. Recebemos, em seguida, as perguntas escolhidas pelos restantes seis grupos (alemão, espanhol, francês, holandês, inglês e italiano[10]), nas línguas originais e na sua tradução para inglês, que foram então por nós traduzidas para português.

Seguimos três regras básicas para a tradução para português dessas perguntas, por esta ordem: optámos por aproximar o mais possível a tradução ao texto existente na nossa colecção, e não a uma tradução literal (*Quantos guarda-costas* em vez *Quantos elementos da escolta*); procurámos formular a pergunta de um modo natural em português (*quando é que o príncipe Carlos e Diana se casaram?* em vez do mais literal *quando se casaram o príncipe Carlos e Diana?*); e, sempre que tal foi possível, traduzimos a partir da língua original, e não a partir da tradução inglesa (por exemplo, *quando rebentou a intifada?*, a partir do texto original italiano, em vez de *quando começou a intifada?*, que seguiria a tradução inglesa que nos foi apresentada).

Também as respostas tiveram de ser traduzidas. Foi necessário adicionar apenas 569 respostas de um total de 600, pois as restantes exigiam respostas numéricas, ou

[9] Note-se que esta é uma questão completamente diferente de perguntas como *Quem foi a primeira presidente da Islândia?* em que o género é relevante, ou *Que ministra acompanhou Mário Soares na visita ao Brasil?* em que não só o género é relevante, como exprime o pressuposto de que houve apenas uma ministra mulher na dita viagem.

[10] Em 2005, a adição do búlgaro e do finlandês elevou este número para oito.

nomes próprios sem versão portuguesa diferente da original. Foram seguidas as mesmas regras na tradução, com especial ênfase na primeira: tentar que a tradução fosse o mais aproximada possível do texto realmente existente na colecção.

No entanto, para algumas perguntas (cf. exemplos abaixo) a variabilidade das respostas era muito elevada, não só porque a pergunta era respondida por várias passagens diferentes da (nossa) colecção (cf. os dois primeiros casos), mas também porque nas colecções originais nos aparecia uma quantidade substancial de respostas (a traduzir) possíveis (dois últimos casos).

```
236 F LOCATION Onde fica o arquipélago de Svalbard?
   1            SEARCH[na Noruega]
   2 941122-055 mais de mil quilómetros acima do Círculo Polar Árctico
   3 941219-032 junto à Noruega
   4 950331-127 Noruega
   5 950607-033 800 quilómetros a norte do continente europeu

330 D ORGANIZATION O que é o FSK?
   1            SEARCH[o serviço de contra-espionagem russo]
   2 940612-093 serviço federal de contra-espionagem
   3 940616-048 agência russa de contra-informação
   4 940706-102 serviço russo de contra-espionagem
   5 941227-020 serviços de contra-espionagem

445 D ORGANIZATION O que é o CERN?
   1            SEARCH[o Laboratório Europeu de Física de Partículas]
   2            SEARCH[o maior laboratório de investigação do mundo]
   3            SEARCH[o Centro Europeu de Pesquisa Nuclear]
   4            SEARCH[laboratório europeu para pesquisa de física nuclear]
   5            SEARCH[o maior centro de pesquisa mundial]
   6            SEARCH[centro europeu para a física de partículas]
   7            SEARCH[organização europeia de pesquisa nuclear]
   8 940220-026 Laboratório Europeu de Física de Partículas

439 F MEASURE Qual o lucro do grupo electrónico e de telecomunicações
finlandês Nokia em 1994?
   1            SEARCH[quatro milhões de markkas]
   2            SEARCH[aproximadamente 1,1 biliões de francos suíços]
   3 951020-040 3,6 mil milhões de markka (mais de 125 milhões de contos)
```

A nossa postura foi a seguinte: sempre que havia respostas na nossa colecção, tentámos adaptar a tradução da resposta; quando havia várias respostas cuja tradução para português era idêntica, descartámo-las. De notar que esta tradução das respostas era importante para sistemas de recolha cruzada, que por exemplo fizessem as perguntas em holandês mas obtivessem os resultados em português.

De todas as tarefas a que nos entregámos, esta foi aquela para a qual houve menos indicações e mais subjectividade;[11] será também a menos útil, dado que em muitos casos as respostas (traduzidas de outras línguas com outras necessidades de informação) são em si mesmas pouco interessantes para um falante de português; senão, vejam-se os exemplos (NIL significa que nenhuma resposta foi encontrada na nossa colecção):

```
174 F OTHER O que significa «Forza Italia!»?
   1            SEARCH[Força, Itália!]
   2            SEARCH[Força Itália]
   3            NIL

202 F OTHER Qual o acrónimo da Amnistia Internacional?[12]
   1            SEARCH[AI]
   2 940113-144 AI
```

[11] Devido à nossa crítica, esta tarefa (tradução das respostas) deixou de ser feita a partir da edição de 2005, o que confirma estas observações.

## 2.4 Avaliação de RI

A avaliação de RI foi feita com recurso ao programa ASSESS (Rogers, 1998), desenvolvido pelo National Institute of Standards and Technology (NIST) americano. O cálculo das medidas utilizadas no CLEF (Braschler & Peters, 2003) baseia-se no resultado deste programa, pelo que é indispensável no processo de avaliação.

Os critérios para avaliação de resultados (em que cada documento podia ser exclusivamente considerado como relevante ou não-relevante, não existindo valores intermédios) foram acordados aquando da escolha geral dos tópicos, tendo cada um dos tópicos a sua própria definição de relevância (o que não impediu alguma discussão, durante a fase de avaliação, sobre a validade de algumas respostas).

A tarefa foi complicada pelo facto de ter sido necessário avaliar um grande número de respostas a cada tópico (entre 128 e 997 documentos, numa média de 446 documentos por tópico; veja-se a Tabela 1-2 abaixo), das quais uma grande maioria irrelevantes: por exemplo, para o tópico 227 (intitulado *A donzela de gelo do Altai*), tivemos que avaliar 476 documentos, nenhum dos quais considerado relevante.

Assim, optámos por fazer a avaliação em duas passagens: uma primeira em que marcámos como relevantes todos os artigos em que detectámos uma centelha de relação com o tópico em questão, e uma segunda passagem em que removemos dessa lista de documentos possivelmente relevantes todos aqueles que não satisfaziam os critérios previamente acordados (e constantes na descrição do tópico ou sua precisão, através de discussão com todos os juízes por correio electrónico).

Tabela 1-2: Número de documentos a avaliar (i.e., número de documentos que pelo menos um sistema considerou relevante) e número de documentos relevantes por tópico

| Tópico | Respostas | Válidas | | Tópico | Respostas | Válidas | |
|---|---|---|---|---|---|---|---|
| 201 | 496 | 14 | 2,8% | 226 | 524 | 4 | 0,8% |
| 202 | 170 | 9 | 5,3% | 227 | 476 | 0 | 0,0% |
| 203 | 369 | 37 | 10,0% | 228 | 514 | 51 | 9,9% |
| 204 | 546 | 7 | 1,3% | 229 | 505 | 189 | 37,4% |
| 205 | 169 | 2 | 1,2% | 230 | 147 | 13 | 8,8% |
| 206 | 200 | 2 | 1,0% | 231 | 834 | 6 | 0,7% |
| 207 | 588 | 12 | 2,0% | 232 | 665 | 29 | 4,4% |
| 208 | 484 | 5 | 1,0% | 233 | 266 | 16 | 6,0% |
| 209 | 635 | 2 | 0,3% | 234 | 558 | 2 | 0,4% |
| 210 | 379 | 5 | 1,3% | 235 | 453 | 1 | 0,2% |
| 211 | 243 | 28 | 11,5% | 236 | 244 | 3 | 1,2% |
| 212 | 609 | 10 | 1,6% | 237 | 387 | 3 | 0,8% |
| 213 | 282 | 56 | 19,9% | 238 | 240 | 1 | 0,4% |
| 214 | 604 | 8 | 1,3% | 239 | 456 | 27 | 5,9% |
| 215 | 350 | 1 | 0,3% | 240 | 557 | 0 | 0,0% |
| 216 | 573 | 0 | 0,0% | 241 | 997 | 58 | 5,8% |
| 217 | 584 | 7 | 1,2% | 242 | 405 | 4 | 1,0% |
| 218 | 128 | 21 | 16,4% | 243 | 383 | 1 | 0,3% |
| 219 | 656 | 2 | 0,3% | 244 | 389 | 1 | 0,3% |
| 220 | 602 | 0 | 0,0% | 245 | 293 | 3 | 1,0% |
| 221 | 296 | 6 | 2,0% | 246 | 307 | 1 | 0,3% |
| 222 | 583 | 3 | 0,5% | 247 | 531 | 8 | 1,5% |
| 223 | 387 | 1 | 0,3% | 248 | 380 | 10 | 2,6% |
| 224 | 549 | 2 | 0,4% | 249 | 349 | 4 | 1,1% |
| 225 | 437 | 1 | 0,2% | 250 | 532 | 2 | 0,4% |

[12] De notar que a palavra *acrónimo* pode ter sido uma tradução/adaptação pouco precisa da pergunta original, visto que em português seria mais natural dizer *sigla*. Seja como for, esta pergunta seria ociosa em português.

| | | | |
|---|---|---|---|
| Total | 22.311 | 678 | 3,0% |

Em suma, tivemos 22.311 documentos para avaliar, dos quais apenas 678 (3,04%) foram considerados relevantes; em apenas cinco casos o número de documentos relevantes excedeu 10%.

### 2.5 Avaliação de RAP

No CLEF 2004, houve três conjuntos de respostas a perguntas em português usando a colecção portuguesa, correspondentes a dois sistemas, cujos resultados globais apresentamos aqui.[13]

A tabela seguinte indica os resultados obtidos. "Correcção" refere-se à percentagem correcta de respostas; das 199 presentes na tarefa monolingue "perguntar em português obtendo respostas em português".

Tabela 1-3: Resultados dos sistemas participantes
R - certa (Right) ; W - errada (Wrong) ; X - aproximada, por excesso ou defeito (ineXact)
U - sem justificação (Unsupported); M - sem resposta (Missing)

| Sistema | R | W | X | U | M | Correcção geral | Correcção | | Respostas NIL | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | factóides | definição | precisão | abrangência |
| PTUE041ptpt | 57 | 125 | 18 | - | 127 | 28,64 | 29,17 | 25,81 | 14,2 | 90 |
| sfnx041ptpt | 22 | 166 | 8 | 4 | 101 | 11,06 | 11,90 | 6,45 | 14,9 | 75 |
| sfnx042ptpt | 30 | 155 | 10 | 5 | 71 | 15,08 | 16,07 | 9,68 | 15,5 | 55 |

Aparentemente os resultados não foram brilhantes, mas é preciso não esquecer que foi a primeira vez que os grupos (e respectivos sistemas) concorreram, e que nenhum dos sistemas tinha sido desenvolvido precisamente para o tipo de texto jornalístico que constituía o alvo do QA@CLEF (para a descrição dos sistemas participantes, veja-se Costa, 2005 e Quaresma *et al.*, 2005).

Na Tabela 1-4 apresentamos os resultados por tipo de pergunta, e um resultado virtual, a que chamamos «combinação», que representa o número de perguntas para as quais pelo menos um dos sistemas conseguiu encontrar uma resposta correcta. Entre parênteses rectos, fornecemos o tipo de pergunta, segundo a classificação da organização. É interessante reparar que não há grande diferença entre definições e factóides, o que leva a crer que as perguntas talvez não fossem tão diferentes como os organizadores à partida pensaram.[14] Por outro lado, perguntas relativas a tempo e a medidas têm um desempenho mais fraco do que as outras.

Tabela 1-4: Apresentação detalhada do desempenho dos sistemas participantes: ORG - organização; PER - pessoa; LOC - local; MAN - modo; MEA - medida; OBJ - objecto; TIM - tempo; OTH - outro

| Conjunto de resultados | respostas correctas | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | definição | | factóide | | | | | | | | total | |
| | org [14] | per [17] | loc [43] | man [4] | mea [23] | obj [6] | org [12] | oth [21] | per [44] | tim [15] | # 199 | % |
| PTUE041ptpt | 3 | 5 | 19 | 1 | 5 | 1 | 4 | 3 | 14 | 2 | 57 | 28.64 |
| sfnx041ptpt | | 2 | 4 | | 3 | 1 | 2 | 3 | 7 | | 22 | 11.06 |
| sfnx042ptpt | 1 | 2 | 8 | | 4 | 2 | 2 | 4 | 7 | | 30 | 15.08 |
| combinação | 3 | 6 | 25 | 1 | 5 | 3 | 4 | 6 | 19 | 2 | 74 | 37.18 |
| % comb | 21,4 | 35,3 | 58,1 | 25,0 | 21,7 | 50,0 | 33,3 | 28,6 | 43,2 | 13,3 | | |

[13] Para as edições seguintes, vejam-se as actas respectivas.
[14] Para uma problematização das definições, veja-se também Santos & Cardoso (2005).

Não é contudo trivial comparar desempenhos de sistemas de RAP. De facto, um dos problemas apontados por Voorhes & Tice (2000b) na organização do TREC Q&A é o facto de não ser possível usar os resultados de uma competição (de um sistema) para aumentar ou criar uma colecção reutilizável sem intervenção humana, ao contrário do TREC (procura de documentos) cujo resultado é o número ou identificação do documento.

De facto, muitas vezes os sistemas participantes no TREC Q&A (e no QA@CLEF) produzem um excerto que contém a resposta, além da resposta, excerto esse, além disso, que varia de sistema para sistema, conforme as técnicas de extracção de informação utilizadas. Por exemplo, vejam-se os seguintes casos concretos no QA@CLEF (que obtiveram a classificação benevolente de "X"):

```
P: Quem é Yves Saint-Laurent?
R: costureiro como Sonya Rykel Christian Dior Nina Ricci Hermés Azzaro
Cacharel Guy Laroche Balenciaga Paco Rabanne Ted Lapidus Pierre Balmain

P: Quem é o presidente da UEFA?
R: Lennart Johansson presidente UEFA um

P: O que é o CERN?
R: Laboratório Europeu de Física das Partículas enquanto observação
```

Isto faz com que a avaliação, mesmo decidindo por uma benevolência máxima, não seja fácil, como tentaremos ilustrar aqui. Por vezes, questões muito finas são relevantes, tais como preposições ou artigos, geralmente deitados fora pelos sistemas:

```
P: O que é a UNICEF?
R: organização Nações Unidas
```

Ora a resposta está certa (mas incompleta) se a lermos como *uma organização das Nações Unidas* (a organização das Nações Unidas para a infância) mas está evidentemente errada se a lermos como *a Organização das Nações Unidas* = ONU.

```
P: Onde era o campo de concentração de Auschwitz?
R: sul Polónia
```

Da mesma forma, esta resposta pode significar *sul da Polónia* e nesse caso está certa, ou *ao sul da Polónia* (portanto fora da Polónia) e nesse caso estaria errada.

```
P: O que é a EUA?
R: European Universities Association
```

Mas a pergunta *O que são os EUA?* teria uma resposta diferente (Estados Unidos da América)

```
P: O que é o PC do B?
```

Aqui seria certamente impossível retirar a contracção *do* e obter o mesmo referente, visto que o *Partido Comunista do Brasil* e o *Partido Comunista Brasileiro* são duas formações políticas diferentes.

Outras questões, associadas à semântica de algumas expressões, se põem:

```
P: Quem é Leonor Beleza?
R1: ex-ministra da Saúde
R2: deputada
```

Na pergunta anterior, com R1 e R2 as respostas consideradas correctas, como classificar *ministra da Saúde*? Ou seja, deve considerar-se como incompleta (ou

simplesmente errada) uma resposta que exclua o prefixo *ex*? Dadas todas estas incertezas, apesar das diversas tentativas de criar um sistema julgador automático (Breck *et al.*, 2000; Voorhes & Tice, 2000b), não foi possível ainda obter uma metodologia que resolvesse este problema (e que concordasse, portanto, com os juízes humanos).

Outra das questões interessantes e não triviais que se nos pôs é a de como julgar a comprovação de uma resposta. Por um lado, era óbvio que alguns dos sistemas tentam apenas verificar a sua resposta (obtida provavelmente da rede) na colecção em jogo, seguindo a sugestão de Brill *et al.* (2001), o que pode dar origem a questões complicadas, tais como "resposta certa mas não comprovada" (ou seja, o documento pode mencionar todos os termos envolvidos, mas não afirmar explícita ou implicitamente uma resposta ao que se pergunta); ou "resposta certa mas muito improvável" (veja-se o caso de *Moscou* como capital da Rússia). De facto, a palavra *Moscou* apenas aparece uma única vez em toda a colecção do Público, numa entrevista a Oscar Niemeyer (mas provavelmente aparece mais frequentemente do que *Moscovo* na rede, dado ser a forma correcta na norma brasileira).

Além disso, é possível encontrar vários casos de "resposta errada e nem sequer plausível". Aqui dá-se o caso, aliás muito comum, de o sistema simplesmente produzir uma resposta que não se enquadra no tipo de pergunta:

```
P: Qual a área da Baixa Saxónia?
R: 67 mil votos

P: Qual foi o primeiro filme sonoro português?
R: Leitão de Barros
```

Finalmente, ainda é possível o caso patológico de um sistema apresentar como justificação um documento contendo uma resposta certa (embora diferente da esperada), para justificar **outra** resposta. No caso seguinte, o texto, incluído na Figura 1-1 acima, refere-se a um ex-jogador do FC Porto:

```
P: Quem é Fernando Gomes?
R: presidente da Câmara
Justificação: Os diálogos entre Pinto da Costa e os ex-jogadores do FC Porto
Octávio Machado e Fernando Gomes, entre Valentim Loureiro e Sousa Cintra ou
entre o jornalista Marinho Neves e o funcionário portista Jorge Gomes ficaram
na memória dos telespectadores como excelentes momentos de televisão.
```

Noutros casos, é delicado decidir até onde ir na cadeia de raciocínio (sabendo que tal raciocínio não só não foi feito pelo sistema como é duvidoso que algum dia o venha a ser). Assim, no primeiro ano, para compensar ao máximo os sistemas participantes, foram consideradas certas (e justificadas) as seguintes respostas, que exibem uma benevolência inegável por parte dos avaliadores:

```
P: Que vende Fausto ao Diabo?
R: alma
Justificação: um dos actores principais desta nova versão anos 90, ganhou
também um Tony pelo seu papel de jovem jogador de basebol que vende a alma ao
Diabo. «Damn Yankees» é uma comédia musical sobre a paixão dos americanos -- o
basebol -- inspirada no mito de Fausto. Ou seja , um reformado fã dos
Washington Senators que não faz outra coisa senão ver jogos na televisão e
gritar «Damn Yankees» (porque os New York Yankees ganham jogo após jogo) vende
a alma ao Diabo para se transformar no melhor jogador de basebol e ajudar a
sua equipa

P: Quem é Christo?
R: artista
```

```
Justificação: «Christo vestiu o Reichstag , Fialho veste Cristo!», é também um
mote que lhe ocorre, comparando-se ao artista que recentemente cobriu o
Parlamento alemão,

P: Quem escreveu "Mein Kampf"?
R: Hitler
Justificação: Hitler, no sétimo capítulo do «Mein Kampf», transforma-se no
porta-voz extremista de um ressentimento mais globalmente partilhado

P: Qual é a capital da Rússia?
R: Moscovo
Justificação: O Presidente da Câmara de Moscovo exprimiu ontem receios de que
sejam cometidos «actos terroristas» no 1º de Maio, mas afirmou que a capital
da Rússia estava «preparada para enfrentar os extremistas».
```

## *3. Alguns comentários finais e balanço*

Após participar na organização do CLEF, chegámos a um conjunto de sugestões a serem introduzidas em futuras avaliações de RAP em português, publicadas em Vallin *et al.* (2005) e Santos & Rocha (2005), que resumimos na próxima secção. A própria dinâmica do CLEF levou a indiscutível progresso em relação ao processamento computacional da língua portuguesa, como tentaremos demonstrar na secção seguinte.

### 3.1 Sugestões de melhoria para avaliação de RAP

Visto que há várias formas diferentes de fazer a mesma pergunta, poderia ter vantagem fornecer um conjunto de formas alternativas de formular a mesma pergunta em vez de uma pergunta única.

Em segundo lugar, parece-nos que responder a perguntas de confirmação pode ser muito útil, sobretudo porque, psicologicamente, pode ser mais fácil confiar num sistema que confirma algo que já sabemos do que num que produz uma resposta sobre um assunto do qual não temos conhecimento nenhum. Portanto, o desempenho de sistemas que respondam a perguntas de confirmação como as seguintes também deveria ser avaliado, ainda que a justificação de uma resposta negativa possa ser mais complexa do que uma positiva: *Oslo é a capital da Noruega? Jorge Sampaio é o presidente da República Portuguesa? O PSD ganhou as últimas eleições?*

Quanto ao método de avaliação, parece-nos relevante distinguir, no caso das respostas vazias, os casos em que o sistema não encontrou resposta (que lá estava) e os casos em que não encontrou resposta porque lá não estava.

Finalmente, parece-nos essencial separar os casos de resposta incompleta dos casos de resposta completa dentro de texto irrelevante (e, por vezes, contendo respostas incorrectas), ambos de momento classificados como “X” (inexacta). Não só porque, para melhorar o desempenho nestes dois casos, são precisas acções quase complementares, mas porque, de um ponto de vista de um utilizador humano, têm graus de utilidade muito distintos. De facto, numa avaliação de RAP parece-nos que uma outra categoria — inspirada em Richardson & Braden-Harder (1993) — deveria ser levada em conta: a resposta ajuda ou não ajuda a pessoa que formulou a pergunta? Veja-se Vallin *et al.* (2005) para uma primeira tentativa de classificar respostas segundo estes parâmetros.

Todas estas observações, contudo, foram-nos suscitadas pela prática, e pela necessidade concreta de responder a perguntas e de avaliar sistemas que o fazem. Note-se que, embora só tenhamos conhecimento de três grupos portugueses que participaram no CLEF’2004 em português, vários outros participantes houve que usaram a parte portuguesa para testar as capacidades multilingues dos seus sistemas.

### 3.2 Evolução e balanço da presença do português no CLEF

A colaboração da Linguateca no CLEF não parou em 2004, encontrando-se neste momento o português no terceiro ano consecutivo nesta avaliação conjunta. Em 2005, a presença do português alargou-se a três novas pistas: recolha de informação geográfica com tópicos em português sobre duas colecções jornalísticas, em inglês e alemão, GeoCLEF (Gey *et al.*, 2005); ImageCLEF (procura de imagens com base nas legendas, Clough *et al.*, 2005) e WebCLEF (Sigurbjörnsson *et al.*, 2005), com tópicos sobre as páginas governamentais europeias num conjunto diversificado de línguas.

Além disso, consolidou-se a investigação e a participação nas duas tarefas repetidas, nomeadamente RI e RAP: Com efeito, as pistas de RI *ad hoc*, tradicionalmente as mais concorridas do CLEF, reuniram um total de 32 experiências sobre a colecção em português, com nove grupos a participar na tarefa monolingue e oito na tarefa bilingue com o português como língua-alvo (seis inglês-português, três espanhol-português, e um francês-português).

Quanto à RAP (Vallin *et al.*, 2005), assistiu-se à entrada em cena de mais um grupo concorrendo na tarefa monolingue (e a continuação dos anteriores grupos participantes), com resultados francamente positivos, tendo-se revelado um dos melhores sistemas participantes no QA@CLEF daquele ano. Dada a adição da colecção brasileira, e o facto de não haver uma forma objectiva de avaliar a dificuldade das perguntas (comparadas com as dos anos anteriores), a variação dos resultados dos grupos "veteranos" (num caso, claramente positiva, no outro, ligeiramente negativa) é, no entanto, mais difícil de explicar. Pela primeira vez, registou-se a presença de um sistema de RAP bilingue, usando o português como língua-alvo e o inglês como língua-base.

A presença, desde 2004, do português em pistas tão diversas, e o crescimento do número de grupos que levam em consideração o português na sua participação permitem-nos concluir que o português ocupa actualmente um lugar de relevo no CLEF, equiparável a línguas mais habituadas a este tipo de fóruns, lugar esse que, esperamos, incentiva quer o desenvolvimento de novos sistemas para o processamento da nossa língua, quer a participação dos sistemas já existentes em processos de avaliação externa indispensáveis ao seu desenvolvimento. Além disso, o facto de a Linguateca ter um papel na organização do CLEF permite não só chamar a atenção dos outros organizadores para questões que interessam à nossa língua, como acautela os interesses do português em várias decisões, muitas delas de ponta a nível científico. Refira-se por exemplo o facto de termos contribuído quer para o desenho de novas tarefas piloto, quer para a (re)formulação de algumas das opções (GeoCLEF, RAP, etc.).

Parece-nos, pois, ser possível fazer um balanço positivo desta experiência: ajudou a desenvolver os sistemas participantes, aumentou a presença do português no panorama internacional, e contribuiu para pôr a nu várias áreas interessantes que é preciso investigar para o português, além de criar primeiros recursos públicos, passíveis de extensão futura.