AvalON: uma iniciativa de avaliação conjunta para o português

Diana Santos
Paulo Rocha
Linguateca

Neste artigo descrevemos a primeira iniciativa de avaliação conjunta no âmbito do processamento computacional da língua portuguesa, a que chamámos AvalON (**Aval**iação: **O**rganização **N**ão-dirigida).

Avaliação conjunta é um processo de avaliação em que os critérios se estabelecem de comum acordo entre os participantes, que definem, para uma dada área de aplicação, um conjunto de tarefas e os critérios de sucesso para avaliar a execução dessas tarefas.

Embora tenha, em conjunto com outras formas de avaliação, o objectivo de medir a qualidade de um sistema, tem, ao contrário de uma avaliação por juízes, examinadores, ou painel de peritos, a garantia de uma total abertura em relação aos critérios e à explicitação dos objectivos – e não deixa margem para diferenças de opinião na atribuição do resultado. Isto porque apenas se avalia conjuntamente aquilo que os participantes conseguem concordar numa primeira fase (a definição do problema), processo esse que é, aliás, extraordinariamente valioso para definir uma área científica e estimar a sua maturidade.

A diferença em relação à avaliação "normal" que se faz em todos os projectos de engenharia informática (e, portanto, em todo o processamento computacional da língua que se preze) é que a medição do progresso é muito mais objectiva, assim como se consegue também uma estimativa da dificuldade do problema em si, e não apenas da sua resolução.

O modelo de avaliação conjunta apareceu nos EUA ainda nos anos oitenta do século passado, e em breve se tornou um motor e um exemplo na área do processamento da língua, falada ou escrita. Veja-se Hirschman (1998), Gaizauskas (1998), Voorhes (2000), assim como a lista compilada em Soares (2002b).

O modelo expandiu-se pouco depois para a Europa (veja-se Mariani (1998), Paroubek e Blasband (1999) e Chibout et al. (2000)) e para a Ásia. Para o português, foi considerado uma das formas de avançar significativamente, por ocasião do debate alargado associado à criação do Livro Branco (1999), em Santos (1999a; 1999b).[1] Também, devido ao facto de os cidadãos da nossa língua se encontrarem espalhados pelo mundo, há uma necessidade maior de juntar esforços através de um conhecimento mútuo só possível através de competições amigáveis.

[1] Além de a organização deste tipo de iniciativas ter sido preconizada no documento preparatório, foi mencionada em várias intervenções no debate.

Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, APL, 2002, pp. 693-704.

A tentativa de organização nesse sentido foi, desde o início, um dos objectivos da Linguateca[2], e o primeiro passo nesse sentido para o português foi dado no início de 2002 com a criação de um sítio na rede dedicado à avaliação conjunta e o pedido de inscrição de todos os interessados, juntamente com a indicação das áreas em que pretendiam participar.

Após uma discussão electrónica frutuosa, com a respectiva apresentação dos vários intervenientes, teve lugar um encontro preparatório na Universidade de Faro, que reuniu um número considerável de investigadores na área, tanto de Portugal como do Brasil. Esse encontro serviu para lançar as bases de trabalho em três áreas, assim como para pôr em contacto pessoalmente muitos dos investigadores, e lançar vários tópicos importantes de discussão. Além disso, serviu para esclarecer melhor o modelo e para confrontar, de viva voz, algumas divergências sobre maneiras de proceder, como só é possível fazer presencialmente. O processo de organização de uma iniciativa deste género (e sendo apenas ainda um encontro preparatório) exige uma enorme dose de engenharia social, como todos os membros de comunidades virtuais (listas electrónicas, comissões de programa, co-autoria remota, etc. etc.) bem o sabem.

O modelo de avaliação conjunta que nós, organizadores do encontro preparatório (ou seja, Jorge Baptista, Alexsandro Soares e os autores do presente artigo), tentámos transmitir, se, por um lado, era directamente inspirado nas avaliações conjuntas "tradicionais" (MUC, TREC, Parseval, SUMMAC, etc. – veja-se Soares (2002b) para uma lista de referências cabais à maioria destes "evaluation contests"), tentou também adequar-se à realidade da língua portuguesa.

Em particular, conhecíamos bem[3] a falta total ou quase total de padrões de avaliação no trabalho existente, a falta de comunicação, citação ou reconhecimento dos pares, e mesmo a inexistência, sequer, de documentação em relação a eventuais avaliações efectuadas. Além, claro da escassez (ou ausência) de recursos que pudessem ser usados publicamente para efeitos de avaliação.

Sabíamos também da fragmentação da comunidade em grupos localizados dispersamente (em países diferentes), além disso formados em tradições académicas diferentes, com consequente falta de massa crítica – veja-se Gago (1999) a esse respeito. Era também patente a quase ausência de dados sobre o português em conferências internacionais de renome, cf. Branco (1999), perversamente acoplada à insistência de alguns investigadores no processamento da língua em geral, mesmo trabalhando sobre o português.

---

[2] A Linguateca, www.linguateca.pt, é a sucessão natural do projecto Processamento Computacional do Português, que deu origem a um centro de recursos distribuído para a língua portuguesa. As actividades da Linguateca não se desenrolam apenas no SINTEF em Oslo, mas também no Departamento de Informática da Universidade do Minho em Braga, no Laboratório de Engenharia da Linguagem (LabEL) em Lisboa, e no Centro de Linguística da Universidade do Porto (CLUP) no Porto.

[3] Veja-se a radiografia da área, Santos (1999a) já mencionada.

A situação da comunidade científica da área era, na altura, desoladora. Alguns dos factores melhoraram (e a Linguateca tem-se vindo a esforçar para colmatar alguns dos problemas, quer informando e tornando acessível o trabalho dos vários grupos, quer criando recursos ou facilitando a sua disponibilização), mas nunca duvidámos que só em conversa, em conjunto, cada área poderia vir a avançar significativamente, ou seja, era preciso pôr as pessoas a trabalhar em conjunto, "forçá-las" a olhar para o trabalho dos outros, para se obter, de facto, comunicação e fertilização cruzada.

O facto de não haver praticamente nenhuma área do processamento do português (escrito) que se pudesse considerar suficientemente madura para não precisar de avaliações conjuntas levou-nos, ao contrário do que seria de esperar, a sermos mais ambiciosos e a tentar iniciar as actividades de avaliação globalmente.

Além disso, estendemos o âmbito da avaliação conjunta também a recursos (tais como léxicos e corpora) por duas razões: uma percentagem significativa dos grupos que se consideram do processamento computacional da língua desenvolvem ou dedicam-se a recursos; por outro lado, a avaliação destes, embora certamente problemática, não deve ser descuidada (veja-se Santos e Rocha (2001) e Santos e Gasperin (2002) para algumas propostas de avaliação de corpora).

As categorias mais escolhidas, por ordem de número de inscritos em cada e ponderadas pelo peso de cada inscrito, em Janeiro de 2003, encontram-se na tabela 1.

Tabela 1: Área, peso e número de inscritos

Área	Peso	Número
Análise morfológica	7,02	45
Léxicos monolingues	6,30	48
Corpora anotados	6,04	46
Recuperação de informação	5,64	29
Análise sintáctica	5,01	38
Extracção de informação	3,71	31
Léxicos bilingues ou multilingues	3,55	34
Análise semântica	3,47	28
Tradução automática	3,00	21
Extracção de relações semânticas	2,92	25
Separação de frases	2,82	18
Tesauros	2,76	25
Extracção de termos	2,68	22
Sistemas de diálogo	2,51	16
Sumarização	2,43	12
Corpora de fala	2,16	19
Indexação	2,11	14
Dicionários de sinónimos	1,99	21
Análise textual	1,86	16

Correcção ortográfica	1,81	20
Desambiguação de sentidos	1,75	16
Geração morfológica	1,73	16
Classificação de texto	1,70	16
Identificação de nomes próprios	1,67	15
Corpora alinhados	1,55	19
Diálogo falado	1,54	12
Identificação de língua escrita	1,53	10
Terminologias monolingues	1,52	17
Leitura automática	1,48	12
Identificação de língua falada	1,48	10
Correcção sintáctica	1,47	15
Classificação de termos	1,42	14
Resolução de anáforas	1,42	10

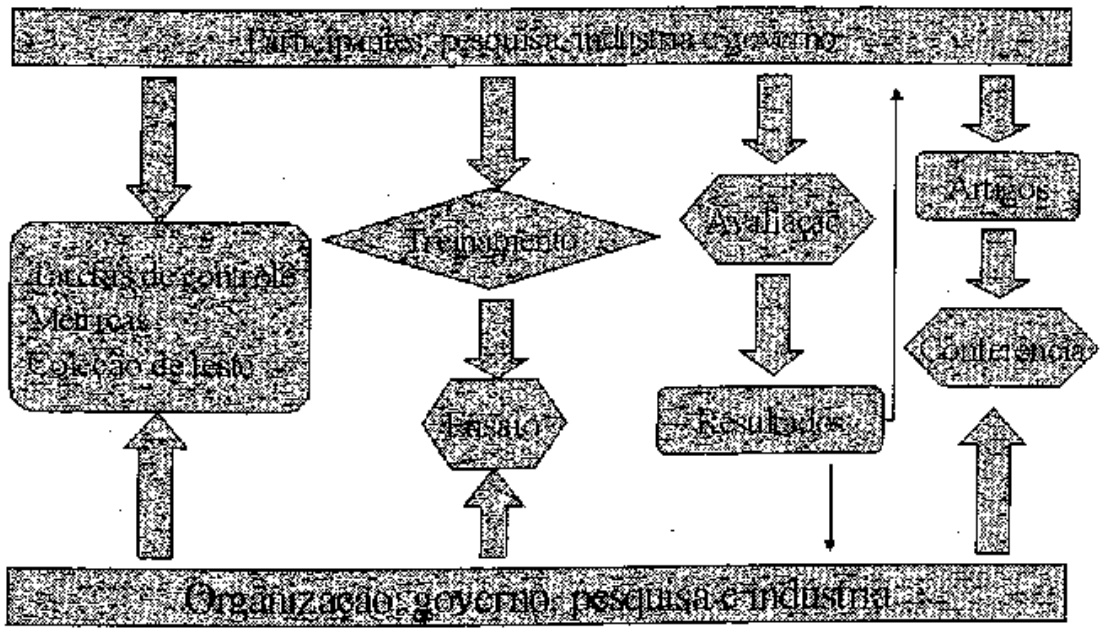
Mas explicitemos um pouco melhor o modelo de avaliação conjunta, começando pelos seus vários pressupostos:

- A avaliação é uma parte importante do processo científico.
- O processamento computacional de uma língua é uma actividade científica.
- Toda a ciência tem aplicação, como afirma Kuhn (1962).
- A experimentação, e a avaliação, são difíceis (para todas as áreas).
- A avaliação é uma actividade de investigação no seu pleno direito. (De facto, até se poderia dizer, de ordem superior.)
- A ciência deve contrariar processos pouco lícitos entre os seus praticantes; veja-se Decoo (2002).
- O processamento da linguagem natural devia ser estruturado, não através das disciplinas subsidiárias às quais se vai inspirar, como a morfologia, a sintaxe, a semântica, mas sim pelos seus domínios de aplicação (ou problemas que tenta resolver), como argumentado em Santos (2000).

Este último item é a razão por que tentámos (com maior ou menor sucesso, variando com as sub-áreas e as tradições já arreigadas) classificar as avaliações conjuntas por tipo de aplicação. Em alguns casos, contudo, tal não foi possível, por existirem suficientes sistemas reclamando-se de uma tarefa linguística específica, como é o caso dos analisadores sintácticos ou morfológicos. De facto, quanto mais gerais (multi-usos, de aplicação generalizada) são os sistemas, mais difícil é desenhar experiências e tarefas em que o seu desempenho possa ser facilmente medido, visto que os requisitos de tais sistemas genéricos podem ser diametralmente opostos dependendo da aplicação.

O modelo da avaliação conjunta por nós sugerido é apresentado graficamente na Figura 1, extraída de Soares (2002c), e tem os seguintes ingredientes principais. 1) Uma ou várias tarefas cuja execução seja compreensível (executável) por um ser

Fases da Avaliação Conjunta



humano; 2) uma solução ou forma de avaliar com que todos concordem (mesmo que, se for por exemplo baseada em juízes, se aceite ou espere uma margem de discordância tão pequena quanto possível); 3) uma forma de comparar automaticamente, ou pelo menos rigorosamente, os vários resultados. Além disso, é preciso que se diga, envolve sempre um grande esforço organizativo. Daí ser costume executar um ensaio antes da avaliação conjunta propriamente dita.

Exemplo detalhado: Morfolimpíadas

No presente artigo apresentamos a questão da avaliação de analisadores morfológicos, a que chamámos Morfolimpíadas – inspirados pelas Morpholympics de Hausser (1994). Pareceu-nos ser a área mais simples e, ao mesmo tempo, aquela que congregava mais sistemas existentes – não apenas planeados ou em fase embrionária – e, daí, uma proposta concreta de como prosseguir foi desde logo apresentada em Faro em Junho de 2002.

Desde aí, tivemos ampla ocasião de reconhecer que fomos excessivamente optimistas e ambiciosos, como se verá de seguida. Por outro lado, consideramos que o desenrolar do ensaio, com as suas muitas vicissitudes, foi contudo positivo por dar ocasião aos participantes de se conhecerem mutuamente e discutirem várias opções. Estamos convencidos de que aumentou consideravelmente a consciencialização das outras abordagens e sistemas, assim como a familiaridade com um processo de avaliação conjunta, que nunca é trivial.

Eis o processo seguido: Começámos por sugerir um modelo concreto de avaliação através de uma página da rede publicada em Março de 2002 para discussão, prosseguindo através da prospecção dos investigadores e sistemas interessados em participar, através da lista electrónica **avalia**, e culminando com a apresentação de uma sugestão mais concreta apresentada e discutida em grupo em Faro, e continuando o debate electronicamente numa lista criada para esse efeito, **analex**. Um anúncio público foi feito em Julho de 2002 apelando à participação num ensaio, com uma calendarização definida, com início em Setembro. Esse ensaio teve como um dos seus pontos altos a reunião presencial de (alguns) participantes no Porto a 1 de Outubro de 2002. O envio dos resultados, assim como a disseminação das conclusões do ensaio seguir-se-ia em breve, culminando com a definição da execução da verdadeira competição. Alguns primeiros resultados são apresentados aqui.

O ensaio tinha como principal objectivo a rodagem de toda a máquina organizativa, assim como a verificação de algumas hipóteses. Em primeiro lugar, verificar se o desenho da competição permitia de facto comparar e avaliar sistemas; em segundo lugar, se era realizável o que era exigido aos participantes e à organização, e quais os problemas possíveis; em terceiro lugar, avaliar algumas opções mais específicas.

O ensaio

Todos os participantes deveriam ir buscar uma colecção de textos suficientemente grande para não ser possível a sua revisão manual, executar os seus sistemas sobre essa colecção de textos e colocar o resultado num local pré-determinado (através de FTP).

Antes disso, todos os grupos (e quem quer que estivesse interessado no problema de avaliar uma análise morfológica), ajudariam a compilar o que nós chamámos (por analogia com os *golden standard* usados na literatura de avaliação em inglês) a *lista dourada*, um conjunto de formas com solução associada, que os participantes considerassem interessantes serem analisadas e comparadas.

Os textos criados pela organização deviam reflectir esta compilação, de forma a que todas as formas da lista dourada se encontrassem por eles dispersas, pelo menos uma vez. Para testar qual a melhor maneira de distribuir os textos, usámos três formatos alternativos: texto seguido, texto atomizado, e lista de formas. Além disso, tentámos garantir uma dispersão máxima de tipos de textos utilizados, usando os corpora AC/DC (veja-se Santos e Sarmento (este volume)) e outros recursos a que tínhamos acesso, desenvolvendo um conjunto de programas (os "castores") que seleccionavam extractos constituídos por algumas frases, e em que a forma a testar se encontrava aleatoriamente situada dentro de cada um desses extractos.

A tabela 2 ilustra o conteúdo textual dos textos de teste. De notar que um dos objectivos de usarmos uma grande variedade de textos era observar se era possível medir diferenças entre os sistemas de acordo com o tipo de texto, donde a compilação manteve (como informação metalinguística, não distribuída aos participantes antes da execução dos sistemas) a classificação de cada extracto segundo variante, género literário, forma de publicação, tipo de autor, se é tradução, etc.

Tabela 2: informação sobre o tipo de texto

Tamanho	Palavras	Textos
total	39.850	199
origem brasileira	16.132	82
origem portuguesa	21.728	113
origem africana	1.390	4
jornais	23.823	118
texto literário original	836	3
texto literário tradução	3.117	18
rede / email	3.333	19

A compilação cooperativa da lista foi extremamente complexa, porque logo ali se verificaram pontos de vista claramente distintos naquilo que se considerava "correcto" e mesmo do âmbito da morfologia (definida operacionalmente como classificação de uma forma no que se refere a campos como género, número, tempo, classificação gramatical (PoS), etc., sem entrar em conta com o contexto).

Após um processo complicado de harmonização e definição dos campos e dos valores possíveis (a partir do resultado dos sistemas), obtivemos uma lista dourada final (refira-se que houve mais de vinte versões, e que todos os participantes corrigiram e suplementaram as várias entradas, as suas e as dos outros).

Tabela 3: descrição quantitativa da lista dourada

formas	205	
análises	394	
formas só com uma análise	112	
formas com duas análises	44	
formas com três análises	23	
formas com quatro análises	17	
média de análises por forma	1,92	
análises como verbo e seu peso[4]	68	0,647
« «nome « « « «	93	0,523
« «nome próprio « « « «	11	0,572
expressões com várias palavras	6	
clíticos & contracções	9	
formas não padrão[5]	6	
palavras com hífen[6]	10	

[4] O peso verbal (nominal, ...) é a proporção das análises verbais (nominais, ...) das formas que têm pelo menos uma análise verbal (nominal, ...).

[5] Esta categoria engloba, entre outras, estrangeirismos, erros ortográficos conhecidos e neologismos.

[6] Inclui verbos com clíticos e palavras compostas, tais como *luso-franco-suíço* e *Trás-os-Montes*.

Foi decidido que a própria lista dourada utilizada no ensaio ficaria apenas do conhecimento dos participantes até às Morfolimpíadas; apresentamos contudo alguma informação sobre a sua constituição na tabela 3.

Em relação ao comportamento dos sistemas no que se refere aos itens constantes da lista dourada, e notando que o ensaio não tinha como objectivo classificar ou avaliar os sistemas, apenas verificar se o modo de proceder permitiria, eventualmente, avaliá-los (e a muitos outros, mais tarde), a tabela 4 apresenta, para os cinco/seis[7] sistemas participantes, o número de formas conhecidas e desconhecidas.[8]

Tabela 4: cobertura simples da lista dourada

sistema	1	2	3	4	5	6
reconhecidos	159	164	176	157	155	
não reconhecidos	11	9	0	24	24	44

De forma a poder comparar automaticamente os vários sistemas, foi preciso convertê-los todos para um mesmo formato neutro, através de programas que chamámos "zebras" e que tinham de processar (e traduzir para um formato comum) entradas que diferiam consideravelmente, como se pode ver pelos seguintes exemplos de formas simples devolvidas por cinco sistemas diferentes, respectivamente para as formas *matemática*, *patas*, *percas*, *reunir* e *folhas*:

'matemática'.
['matemática', 'CAT','n', 'NUM','s', 'GEN','f', 'TAN','t2'].
patas 0:lex(pata, [CAT=nc,G=f,N=s], [], [N=p], []), lex(pato, [CAT=nc,G=m,N=s], [], [G=f,N=p], [])
percas
[perca] N F P <ich> <qu>
[perder] V PR 2S SUBJ VFIN <vt> <vr> <de^vrp> <por^vp> <B-Rare>
reunir,reunir.V:R:U1s:U4s:U3s:W:V1s:V4s:V3s
folhas=<S.F.PL.N.[]?.?.[folha]0.#V.[INT.PRONOM.TD.][PRES.TU.]N.[][folhar]0.>

A tabela 5, a partir dos resultados dos sistemas, indica quantas formas da lista dourada foram analisadas por cada sistema, e dessas, quantas o sistema reconheceu como palavras portuguesas e conseguiu dar uma análise. Note-se, já aqui, que há diferenças consideráveis no que se pretende que um analisador morfológico desempenhe: Alguns sistemas usam heurísticas para dar sempre uma análise, outros

[7] Um dos sistemas (a que chamamos sistema 6) era, não um analisador morfológico, mas simplesmente um reconhecedor / verificador ortográfico; daí haver menos informação a ele relativa.

[8] Como se poderá verificar, os valores das várias tabelas foram calculados sobre versões diferentes da lista dourada. O seu interesse é mostrar o tipo de diferenças presentes, mesmo ao nível do simples reconhecimento, que seria trivial.

esforçam-se por apenas reconhçcer formas de acordo com uma dada gramática (tal obviamente prende-se com os objectivos dos sistemas maiores para os quais os analisadores morfológicos foram desenhados); o que se pode dizer aqui é que uma avaliação conjunta passa apenas por formas sobre as quais há um consenso, mas a dimensão desse consenso quando nos debruçamos sobre texto real é também muito interessante de medir, e um subproduto desta iniciativa.

Tabela 5: comparação sobre as 200 formas da lista dourada

sistema	1	2	3	4	5	6
total de análises	170	173	176	181	179	
reconhecidas	159	164	176	157	155	
não reconhecidas	11	9	0	24	24	44

Resultados do mesmo tipo podem ser calculados para o total das formas presentes nos textos. Apresentamos na tabela 6 o que se passou em relação aos textos fornecidos em formato de texto seguido, ou corrido. Esta panorâmica ilustra eloquentemente, na nossa opinião, as diferenças em relação à atomização que os diferentes sistemas fazem de um mesmo texto. Note-se que tal observação em relação ao português já tinha sido publicada sobre dois sistemas em Santos e Bick (2000) e que aqui se torna a confirmar com seis sistemas.

Tabela 6: comparação sobre os textos todos

sistema	1	2	3	4	5	6
formas diferentes	10654	10342	10248	10186	10447	-
» » conhecidas (1)	9563	9687	10248	9723	9408	
» » desconhecidas	10,2%	6,3%	0	4,5%	9,8%	1750
análises diferentes de (1)	17822	15989	16594	13578	11865	
análises por forma	1,86	1,65	1,62	1,40	1,26	
total de análises	33605	73367	70263	57651	69627	
total de conhecidos	32408	72017	70263	57062	66852	
desconhecidos	1179	1350		589	2775	2349

A função de comparação é, contudo, muito mais elaborada do que estes primeiros números deixam transparecer. Por um lado, tem interesse medir a concordância por campo morfológico, ou seja: os sistemas diferem em categoria gramatical, ou simplesmente no lema ou no número que atribuem a uma dada forma? Por outro lado, uma das causas de diferença é exactamente um número diferente de interpretações morfológicas de uma mesma forma reconhecida. Por último, não devem ser consideradas penalizadoras (ou favorecedoras) diferenças consistentes derivadas de opções fundamentais de cada sistema. Entre estas mencionem-se duas grandes divisões: tratamento ou não da derivação (e como especificar o lema); e diferente identificação do que é um particípio passado.

De facto, é preciso para cada comparação pormo-nos na "pele" dos sistemas e fazer uma comparação justa. A descrição pormenorizada do tipo de diferenças e como lidámos com esta situação será objecto de um outro artigo, por limitações de espaço.

Outros casos: RI, TA e corpora anotados

A iniciativa de avaliação conjunta para o português não se resume, nem se esgota, na avaliação de analisadores morfológicos. Com efeito, ao mesmo tempo que a primeira proposta foi descrita, mais duas, sobre recolha de informação (cf. Aires (2002)) e sobre corpora anotados foram lançadas; e no encontro em Faro iniciou-se um processo de avaliação de tradução automática e aplicações afins (nomeadamente que relacionassem mais do que uma língua).

De momento, encontra-se também em progresso uma iniciativa para avaliar aspectos parcelares de análise sintáctica (ou em contexto), tais como reconhecimento de entidades citadas, identificação de localizações espaciais e temporais, desambiguação da categoria gramatical, etc. Por outro lado, poder-se-ia também pensar em utilizar a metodologia proposta por Gaizauskas et al. (1998) para criar mais facilmente recursos para avaliação, eventualmente usando a Floresta Sintá(c)tica (Afonso et al. 2002a; 2002b) como matéria prima.

Agradecimentos

Estamos gratos à resposta da comunidade, que foi avassaladoramente positiva, e que demonstrou que tínhamos provado a nossa intenção de a servirmos. Estamos ainda mais gratos aos participantes (e oradores) no Encontro Preparatório em Faro, e aos participantes no ensaio das morfolimpíadas que acima descrevemos.

O trabalho descrito neste artigo foi o resultado de um esforço conjunto de uma equipa maior e, muito particularmente, muita da motivação e descrição do modelo de uma avaliação conjunta foi inspirada nos textos de Alexsandro Soares (2002a; 2002c).

Toda a equipa da Linguateca, aliás, continua em força a trabalhar no AvalON e merece pois o nosso agradecimento.

Referências

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002a ""Floresta sintá(c)tica": a treebank for Portuguese", in Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, pp. 1698-1703.

Afonso, Susana, Eckhard Bick, Renato Haber & Diana Santos. 2002b "Floresta sintá(c)tica: um treebank para o português", in Anabela Gonçalves & Clara Nunes Correia (orgs.), *Actas do XVII Encontro da Associação Portuguesa de Linguística* (Lisboa, 2-4 de Outubro de 2001), APL, pp. 533-45.

Aires, Rachel. 2002 "Avaliação em recuperação de informação", http://acdc.linguateca.pt/aval_conjunta/aval_RI.html.

Branco, António Horta. 1999 "Uma Perspectiva de Progresso para a Engenharia da Linguagem: contribuição para o debate preparatório do Livro Branco", 25 de Março de 1999, http://www.linguateca.pt/branco/antonio.html.

Chibout, Karim, Joseph Mariani, Nicolas Masson & Françoise Néel (eds.) 2000 *Ressources et évaluation en ingénerie des langues*. Paris/Bruxelles: De Boeck & Larcier, 2000.

Decoo, Wilfried, with a contribution by Jozef Colpaert. 2002 *Crisis on Campus: Confronting Academic Misconduct*. Cambridge, Mass. & London: The MIT Press.

Gago, José Mariano. Intervenção no debate público de 17 de Abril de 1999, no Forum Picoas, http://www.linguateca.pt/branco/transcricao/Gago.html.

Gaizauskas, Robert. 1998 "Evaluation in language and speech technology". *Computer Speech and Language*, **12** (4) (1998), pp.249-62.

Gaizauskas, R., M. Hepple & C. Huyck. 1998 "Modifying Existing Annotated Corpora for General Comparative Evaluation of Parsing", *Workshop on Evaluation of Parsing Systems, at the 1st International Conference on Language Resources and Evaluation (LREC'98)* (Granada, 1998).

Hausser, Roland. 1994 "The coordinator's final report on the first Morpholympics". *LDV-Forum* **11**(1), 1994, pp. 54-64; reprinted in Roland Hausser (ed.), *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics 1994*, Tübingen: Max Niemeyer Verlag, 1996, pp.167-81.

Hirschman, Lynette. 1998 "The evolution of Evaluation: Lessons from the Message Understanding Conferences", *Computer Speech and Language* **12** (4) (1998), pp.281-305.

Thomas S. Kuhn. 1962 *The Structure of Scientific Revolutions*. Chicago, Illinois: University of Chicago Press, 1962, 2nd edition, 1970, 3rd edition 1996.

Livro Branco do Desenvolvimento Científico e Tecnológico Português (1999-2006) 1999 Observatório das Ciências e das Tecnologias, Ministério da Ciência e da Tecnologia, Portugal.

Mariani, J. 1998 "Some evaluation-based language engineering actions for French", *Computer Speech and Language*, **12** (4) (1998), pp.307-16.

Oksefjell, Signe & Diana Santos. 1998 "Breve panorâmica dos recursos de português mencionados na Web", in Vera Lúcia Strube de Lima (ed.), *Anais do Terceiro Encontro de Processamento da Língua Portuguesa (Escrita e falada), PROPOR'98* (Porto Alegre, 3-4 novembro 1998), pp.38-47.

Paroubek, Patrick & Marc Blasband (eds.) 1999 ELSE LE4-8340, Evaluation in Language and Speech Engineering: Executive Summary of a Blueprint for a General Infrastructure for Natural Language Processing Systems Evaluation Using Semi-Automatic Quantitative Black Box Approach in a Multilingual Environment. June 10 1999, Version 3.2, http://www.limsi.fr/TLP/ELSE/FullXreportXver302.htm.

Santos, Diana. 1999a "Processamento computacional da língua portuguesa: documento de trabalho". SINTEF, Oslo, versão base de 9 de Fevereiro; revista a 13 de Abril, http://www.linguateca.pt/branco/.

Santos, Diana. 1999b "Disponibilização de corpora através da WWW", in Palmira Marrafa & Maria Antónia Mota (orgs.), *Linguística Computacional: Investigação Fundamental e Aplicações: Actas do I Workshop sobre Linguística Computacional da Associação Portuguesa de Linguística* (Lisboa, 25-27 de Maio de 1998), Colibri, pp. 323-346.

Santos, Diana & Elisabete Ranchhod. 1999 "Ambientes de processamento de corpora em português: Comparação entre dois sistemas", in *Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR* (Évora, 20-21 de Setembro 1999), pp. 257-268.

Santos, Diana & Eckhard Bick. 2000 "Providing Internet access to Portuguese corpora: the AC/DC project", in Maria Gavriladou, George Carayannis, Stella Markantonatou, Stelios Piperidis & Gregory Stainhaouer (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC2000* (Athens, 31 May-2 June 2000), ELRA, pp.205-210.

Santos, Diana. 2000 "Introdução ao processamento de linguagem natural através das aplicações", in Elisabete Ranchhod (ed.), *Tratamento das Línguas por Computador. Uma introdução à linguística computacional e suas aplicações*, Lisboa: Caminho, pp. 229-59.

Santos, Diana & Paulo Rocha. 2001 "Evaluating CETEMPúblico, a free resource for Portuguese", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* (Toulouse, 9-11 July 2001), pp.442-449.

Santos, Diana & Caroline Gasperin. 2002 "Evaluation of parsed corpora: experiments in user-transparent and user-visible evaluation", in Manuel González Rodríguez & Carmen Paz Suárez Araujo (eds.), *Proceedings of LREC 2002, the Third International Conference on Language Resources and Evaluation* (Las Palmas de Gran Canaria, Spain, 29-31 May 2002), ELRA, pp.597-604.

Santos, Diana & Luís Sarmento. este volume. "Projecto AC/DC: acesso a corpora / disponibilização de corpora".

Soares, Alexsandro S. 2002a "Razões para se avaliar o processamento computacional do português", http://acdc.linguateca.pt/aval_conjunta/aval_porque.html.

Soares, Alexsandro Santos. 2002b "Conferências de avaliação conjunta realizadas", http://acdc.linguateca.pt/aval_conjunta/outras_aval_conj.html.

Soares, Alexsandro Santos. 2002c. "Avaliações conjuntas: Visão geral", apresentação, http://acdc.linguateca.pt/aval_conjunta/Faro2002/intro_avalconj.html.

Voorhes, Ellen M. 2001 "Philosophy of IR Evaluation", in Carol Peters (ed.), *Results of the CLEF 2001 Cross-Language System Evaluation Campaign: Working Notes for the CLEF 2001 Workshop* (3 September, Darmstadt, Germany).