CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Repositório Comum

Resumo da actividade da Linguateca de 16 de Dezembro de 2006 a 31 de Dezembro de 2008
Coordenado por Luís Costa

Este relatório divide-se em três secções principais que correspondem às três vertentes do modelo IRA (informação, recursos e avaliação) seguido pela Linguateca, apresentando no final algumas estatísticas tentando dar uma medida do impacto da Linguateca no período a que se refere o relatório. Todo o trabalho aqui descrito é a continuação lógica do trabalho iniciado em 2000, e relatado pormenorizadamente em [265, 266]. O presente relatório foi redigido por vários membros do projecto, cujos nomes aparecem sob as secções respectivas.

1. Avaliação

1.1 CLEF

1.1.1. GeoCLEF
Paula Carvalho e Diana Santos

Tal como aconteceu em 2006, a Linguateca participou activamente na organização do GeoCLEF 2007 e 2008. Em concreto, esteve envolvida na criação dos tópicos e na avaliação da relevância dos documentos portugueses devolvidos pelos sistemas participantes em ambas as avaliações. No ano de 2007, a Linguateca foi ainda responsável pelo trabalho de edição dos artigos do GeoCLEF, que foram integrados no volume da Springer, publicado em 2008, organizando 11 revisores e as respectivas recensões. Participou igualmente nos Encontros do CLEF: em 2007, em Budapeste, na Hungria [75, 78], e em 2008, em Aarhus, na Dinamarca [77, 79].

A Linguateca esteve ainda envolvida na tradução/localização do sistema DIRECT para o português, bem como na melhoria da usabilidade e das funcionalidades deste sistema, através das sugestões que foi fazendo.

No que respeita à participação dos sistemas nesta avaliação, é de destacar a participação do pólo do XLDB, no ano de 2007 [16, 26].

A organização do CLEF 2008 decidiu não dar continuidade a esta pista. Porém, o GikiP, uma tarefa-piloto da responsabilidade da Linguateca, que decorreu sob o escopo do GeoCLEF 2008, foi alargada e rebatizada como GikiCLEF, sendo uma das pistas contempladas no CLEF 2009, como referimos na secção seguinte.

1.1.2. GikiP: piloto de avaliação conjunta de resposta automática a perguntas sobre a Wikipédia
Nuno Cardoso

Prosseguindo com a participação activa na co-organização de eventos internacionais de avaliação conjunta na área do processamento da língua, a Linguateca organizou uma pista piloto de avaliação conjunta intitulada GikiP, que esteve sob a chancela da pista GeoCLEF do CLEF, que versava a recolha de informação cruzada geográfica. O GikiP teve por objectivo avaliar sistemas que produziam respostas com ênfase especial sobre o domínio geográfico, sobre a Wikipédia (alemã, inglesa e portuguesa).

A organização desta pista envolveu a divulgação, a definição da tarefa, a criação dos tópicos, a especificação das directivas de avaliação, a proposta e implementação de medidas de avaliação, o julgamento das respostas, a criação dos programas de avaliação e a apresentação dos resultados à comunidade do CLEF e internacionalmente, culminando nos artigos [107,112] .

Além disso, foi aceite a continuação desta avaliação, rebatizada como GikiCLEF, para o próximo ano no âmbito do CLEF, que se encontra neste momento em curso, com uma lista de co-organizadores já apreciável, veja-se www.linguateca.pt/GikiCLEF. Associada a esta nova iniciativa esteve já a apresentação no workshop do GIR 2008, em Napa Valley, CA, EUA [108].

1.1.3. Resposta automática a perguntas

Paulo Rocha

Realizaram-se as seguintes actividades no âmbito da organização da parte portuguesa da tarefa de resposta automática a perguntas:

- Discussão com os restantes membros da organização sobre os moldes de realização do QA@CLEF em 2007 e 2008 e respectivas directrizes.
- Selecção, em cada edição, das 200 perguntas a realizar aos sistemas na pista monolingue portuguesa.
- Tradução para português das perguntas a realizar aos sistemas nas pistas bilingues com o português como fonte.
- Avaliação das respostas da pista monolingue portuguesa e das pistas bilingues com o português como alvo.
- Escrita da secção portuguesa dos artigos gerais sobre o QA@CLEF para as working notes de 2007 e 2008 e colaboração no resto dos artigos [42, 59]; reescrita das secções referidas desses artigos para publicação internacional [43, 60].
- Participação nos workshops do CLEF 2007 e CLEF 2008.

1.2 Segundo HAREM

Cláudia Freitas

O HAREM é uma iniciativa da Linguateca cujo objectivo é a avaliação de sistemas que realizam o reconhecimento de entidades mencionadas (REM) em textos de língua portuguesa.

No Encontro do Primeiro HAREM, em Julho de 2006, ficou claro o interesse da comunidade de PLN/REM na continuação das atividades do HAREM, e em 2007-2008 a Linguateca organizou o Segundo HAREM (http://www.linguateca.pt/HAREM/).

Além da continuação e do aperfeiçoamento das características inovadoras do HAREM, como o tratamento da vagueza na análise de EM e a possibilidade de diferentes segmentações (e análises) para a mesma entidade mencionada (EM), o Segundo HAREM contou com duas novas tarefas: uma avaliação especial para a categoria TEMPO e a classificação de relações semânticas entre EM.

A avaliação especial da categoria TEMPO foi proposta por um grupo de participantes e foi incorporada pela organização do Segundo HAREM, seguindo o princípio de avaliação conjunta. Como nenhum dos proponentes da tarefa quis participar da organização, preferindo a função exclusiva de participante, coube à organização a tarefa de anotar os textos de acordo com as directivas propostas pelo grupo e de criar as medidas e programas de avaliação específicos para o TEMPO.

A tarefa de classificação de relações semânticas entre EM, o ReRelEM, foi uma tarefa-piloto proposta pela própria organização, e teve como objectivo avaliar sistemas que, além da identificação e classificação de EM, reconhecem e classificam relações semânticas entre EM em um texto. É uma tarefa inovadora não apenas no âmbito do PLN/REM para a língua portuguesa, mas do PLN/REM para qualquer língua. Para a execução da tarefa, coube à organização a elaboração de directivas, material de treino, colecção dourada e medidas de avaliação.

O Segundo HAREM teve 10 sistemas participantes, 3 dos quais também participantes na tarefa ReRelEM e 7 que também participaram na pista do TEMPO.

O Segundo HAREM contou com uma equipa de 5 membros da Linguateca, que interagiu muito ativamente na execução das tarefas necessárias a um contexto de avaliação como o HAREM.

Especificamente, foram produzidos para cada uma das pistas (HAREM clássico, TEMPO e ReRelEM):

- Material de exemplo e de treino
- Colecções douradas
- Directivas detalhadas (excepto para a pista do TEMPO, em que as directivas foram propostas por Hagege et al. [223])
- Programas de avaliação, e respectiva documentação
- Programa de auxílio à anotação, o Etiquet(h)arem, e respectiva documentação
- Programa de visualização de relações semânticas para o ReRelEM
- Lâmpada, um pacote com todos os recursos criados no Segundo HAREM e publicamente disponível em http://www.linguateca.pt/HAREM/PacoteRecursosSegundoHAREM.zip
- Sa(h)ara, uma interface web que permite a avaliação remota de novas saídas de sistemas que desejem ser avaliados nas tarefas do Segundo HAREM.

Assim como nas edições anteriores do evento, o Segundo HAREM culminou com o Encontro do Segundo HAREM, em Setembro de 2008 (ver secção 3.7), que reuniu organizadores, participantes e interessados no reconhecimento de entidades mencionadas.

Em Novembro de 2008 foi disponibilizada a LÂMPADA, um pacote com os recursos finais do Segundo HAREM.

Está em curso a edição de um livro sobre o Segundo HAREM, com as comunicações apresentadas durante o evento [134].

2. Recursos

2.1 AC/DC

Diana Santos

Neste projecto incorporaram-se três novos corpos: CHAVE (dando acesso à versão anotada pelo PALAVRAS [264] desta colecção), Vercial (contendo textos literários portugueses disponibilizados pelo projecto Vercial) e CDHAREM (dando acesso à colecção dourada do primeiro HAREM, também anotada pelo PALAVRAS). Também os seguintes corpos foram aumentados ou melhorados: ConDiv (integração da parte dos textos de moda) e Museu da Pessoa (integração de entrevistas do Museu da Pessoa brasileiro).

Além disso, foi reformulada a interface (de HTML para PHP) pelo Paulo Rocha, introduzindo alguma sofisticação adicional, documentação e exemplos específicos para cada corpus, e foram criadas versões novas da anotação (pelo PALAVRAS) de praticamente todos os corpora. Também foi dado acesso ao lado português do COMPARA através do AC/DC, de forma experimental.

Publicou-se um artigo e foi feita uma apresentação pública sobre o AC/DC, assim como este foi objecto de um tutorial no TaLC [89, 182 , 188].

2.2 Catálogo de publicações da Linguateca e SUPeRB

Luís Cabral

Entre 2007 e 2008, o catálogo de publicações sofreu alterações com a integração do SUPeRB (http://www.linguateca.pt/superb/), uma ferramenta desenvolvida no âmbito do mestrado do Luís Miguel Cabral [10], que veio auxiliar a inserção e validação de novas entradas de publicações no catálogo através de meios automáticos.

O SUPeRB apresenta um leque de funcionalidades de pesquisa e extracção de referências bibliográficas e de importação ou exportação de diversos formatos, tais como o BibteX, o EndNote e o RIS.

Permite ainda a gestão das entradas através de uma interface Web, multilingue, permitindo a edição por qualquer pessoa num formulário ou em formatos como o BibTeX, por exemplo, bem como a inserção de meta-informação.

O SUPeRB permite a validação prévia de todas as publicações inseridas ou editadas, sendo este passo também efectuado numa interface Web.

Foi também implementada uma funcionalidade de gestão de listas, facilitando a geração de páginas com listas de publicações, altamente configuráveis, que facilita a gestão das várias listas referentes a projectos distintos.

Para além da gestão das publicações da Linguateca, e das publicações relacionadas com o processamento computacional do português, o SUPeRB, facilitou a gestão das referências bibliográficas de três livros publicados pela Linguateca. O catálogo contém actualmente um total de cerca de 2400 publicações (466 publicações da Linguateca), 200 conferências, livros, revistas e outros eventos, de mais de 2700 autores distintos.

Ao longo deste período, foi também disponibilizado o código fonte dos módulos que permitem o processamento de referências bibliográficas (http://www.linguateca.pt/superb/docs/).

Publicaram-se dois artigos sobre o SUPeRB [12, 13], um dos quais foi apresentado no evento "Applications of Portuguese Speech and Language Technologies" associado ao PROPOR 2008.

2.3 COMPARA

Ana Frankenberg-Garcia, Diana Santos e Luís Costa

Durante este período o COMPARA sofreu uma série de melhorias: foi concluída a revisão textual de inserção de marcas de parágrafo; iniciou-se uma revisão pormenorizada dos textos digitalizados por terceiros, com o objectivo de garantir a qualidade da digitalização de todo o corpus; e deu-se início à sistematização do registo dos erros detectados nas edições impressas que formam a base do COMPARA. Procedeu-se também à aquisição (ou empréstimo em bibliotecas) de versões impressas dos textos em falta.

Relativamente ao sítio do COMPARA, efectuou-se uma reformulação completa do mesmo que incluiu a actualização da documentação e ajudas, bem como o desenvolvimento de novas funcionalidades. Criou-se também um arquivo dos textos impressos, ficheiros de erros e autorizações do COMPARA no servidor e um programa de criação do corpus texto a texto, de modo a dar conta das autorizações que expiram em datas diferentes.

No âmbito da revisão da anotação gramatical da parte portuguesa do corpo, concluiu-se o processo de revisão dos adjectivos e dos substantivos, completou-se grande parte da revisão dos verbos, assim como se efectuou a anotação semântica do COMPARA, em ambas as línguas, no campo da cor [69, 120].

Introduziu-se anotação morfossintáctica dos textos ingleses com o CLAWS 7, reviram-se os lemas e iniciou-se a revisão da anotação morfossintáctica (categorias FW, NN2, NP2 e VV2). Elaborou-se uma lista de categorias prioritárias na revisão e documentação sobre a revisão da anotação gramatical inglesa [121, 122].

Deu-se continuidade ao processo de documentação de todas as decisões inerentes à anotação gramatical em português. Actualizaram-se as instruções de digitalização dos textos. Foram totalmente revistos e actualizados os tutoriais de utilização do corpus em versão portuguesa e inglesa , assim como criado mais um tutorial avançado, em português.

Contactaram-se todos os detentores de direitos de autor dos textos contidos no COMPARA, dando conta de uma descrição actualizada do projecto e do fim do financiamento, e solicitando a liberação dos textos para utilização por terceiros após 31 de Dezembro de 2008.

Finalmente, realizaram-se várias iniciativas de divulgação do COMPARA tais como a elaboração de um desdobrável e de várias apresentações e cursos sobre o COMPARA, assim como escritos, publicados ou enviados para publicação vários artigos sobre o mesmo [45, 46, 47, 48, 49, 50, 68, 92, 99, 102, 110, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 192, 194, 195].

2.4 Corpógrafo

Sérgio Matos, Anabela Barreiro e Luís Cabral

O Corpógrafo (http://www.linguateca.pt/Corpografo), actualmente na versão 4.0, é composto por um conjunto de ferramentas para análise de corpora e construção de bases de dados de terminologia, fraseologia e lexicografia. Já se inscreveram no Corpógrafo cerca de 1600 utilizadores, sendo que cerca de 440 o fizeram nos últimos dois anos.

Na versão 4 do Corpógrafo, foram acrescentadas novas funcionalidades como por exemplo a cópia de termos entre bases de dados terminológicas, incluindo as relações já definidas entre os termos copiados e a pesquisa por textos na internet, a partir de termos gravados nas bases de dados terminológicas, oferecendo um método muito eficiente para acrescentar novos textos especializados aos corpora já existentes. Foram também feitas algumas correcções ao nível da interface com o utilizador, e actualizadas as taxonomias utilizadas na organização e classificação dos ficheiros, corpora e bases de dados (terminológicas e fraseológicas) criados pelos utilizadores. Para tornar estas classificações mais úteis, criou-se também a possibilidade de os utilizadores acrescentarem os seus próprios itens às taxonomias (dentro da sua área de trabalhos pessoal).

Na pesquisa por concordâncias, foi adicionada a possibilidade de pesquisar por etiquetas gramaticais (em português, inglês e francês), através da integração do sistema de processamento de corpora NooJ (http://www.nooj4nlp.net/). É possível estender esta funcionalidade a outras línguas, descarregando e instalando os respectivos dicionários.

Foram também adicionadas novas funcionalidades para a criação de bases de dados de expressões lexicais e de fraseologia, incluindo a definição de relações léxico-semânticas e discursivas. O sistema NooJ é também aqui utilizado, permitindo pesquisar os corpora usando etiquetas gramaticais para encontrar expressões lexicais / fraseológicas.

Outro recurso adicionado foi a possibilidade de criar corpora paralelos, através do alinhamento de textos ao nível da frase, utilizando o alinhador do Corpus Workbench (http://cwb.sourceforge.net/). Os corpora paralelos criados podem ser pesquisados, sendo os resultados apresentados em duas colunas de texto.

Além dos corpora paralelos, é também possível obter concordâncias em corpora comparáveis, a partir dos termos ou itens lexicais/fraseológicos equivalentes de tradução. O resultado são concordâncias para o termo (ou item lexical) e o equivalente de tradução definido pelo utilizador, apresentadas em duas colunas, permitindo verificar se os dois se aplicam nos mesmos contextos e se

são, de facto, equivalentes.

Foi também criada uma versão instalável do Corpógrafo, permitindo ultrapassar algumas dificuldades verificadas durante a instalação manual do sistema e as consequentes limitações na sua distribuição (http://www.linguateca.pt/Corpografo/descarregar/). Esta versão foi divulgada em listas de discussão sobre processamento computacional do português.

Foi também actualizada a documentação. O tutorial e as janelas de ajuda foram actualizadas com a descrição das funcionalidades de processamento de textos paralelos, alinhamento de ficheiros e modificação de alinhamentos, criação de corpos paralelos e pesquisas em corpos paralelos revistos.

Finalmente publicaram-se vários artigos e fizeram-se várias apresentações sobre o Corpógrafo [73, 74, 82, 176, 177, 179].

2.5 CorTrad

Diana Santos

O CorTrad (http://nilc.icmc.usp.br/dispara/CorTrad/) é um projecto conjunto com o COMET, http://www.fflch.usp.br/dlm/comet, (Universidade de São Paulo, Stella Tagnin) em que a Linguateca providencia e melhora o DISPARA enquanto a USP cria e revê os corpos linguísticos. Além disso temos a parceria do NILC, http://nilc.icmc.usp.br/, que aloja o serviço, e que assegurará a sua continuação após o termo da Linguateca.

Enquanto do ponto de vista da disponibilização de recursos públicos na rede o CorTrad não pareceria merecer uma nova rubrica -- afinal de contas, o ConDiv, parceria com o Augusto Soares da Silva, é apenas descrito sob o AC/DC --, este projecto vem de facto inovar do ponto de vista técnico e das funcionalidades oferecidas em relação a corpos paralelos.

Com efeito, o CorTrad permite o acesso a várias fases da tradução de um mesmo texto, permitindo assim a comparação não só da tradução final com o original, mas também o estudo das várias alterações que a própria tradução sofreu.

Estão previstos três tipos de texto no CorTrad (português-inglês): a) uma parte literária, contendo de momento contos escritos em inglês e traduzidos para português, b) uma parte técnica, que por agora contém apenas textos de culinária traduzidos para inglês, e c) uma parte de divulgação científica, que ainda não foi integrada.

Embora este seja o projecto mais novo da Linguateca (após contactos em Maio de 2006 no PROPOR, apenas começou efectivamente em Junho de 2008) conseguimos pelo menos uma adaptação do DISPARA a um novo domínio e necessidades de procura diferentes, assim como a definição de um modo de trabalho que permite a continuação do projecto após a Linguateca acabar.

2.6 Esfinge

Luís Costa

O Esfinge participou tanto em 2007 como em 2008 na tarefa de RAP (resposta automática a perguntas) monolingue português do CLEF. Uma parte dos esforços foram dirigidos à resposta aos novos desafios propostos pela organização do evento: tópicos e anáforas nas perguntas (perguntas mais complicadas) e o uso da Wikipedia para procurar e justificar as respostas (mais dificil encontrar as respostas). Adicionalmente investigou-se uma abordagem "Várias Perguntas/Várias Respostas" para RAP.

Nesse sentido desenvolveram-se um módulo para resolver a co-referência nas perguntas (que cria igualmente expressões de pesquisa alternativas para seleccionar os textos relevantes para responder às mesmas) usando a análise sintáctica do PALAVRAS [264], um módulo para procurar textos relevantes na Wikipedia e um módulo para escolha da melhor resposta de entre um conjunto limitado

de respostas candidatas. Investigaram-se também o uso de vários tipos de padrões de pesquisa de passagens de texto relevantes para as perguntas (bem a como a combinação destes) e a extracção semi-automática de padrões para obtenção de respostas.

No que diz respeito a 2007, escreveu-se um artigo [11], apresentou-se um poster e reformulou-se o primeiro artigo para publicação internacional [14]. Já em 2008, escreveram-se dois artigos, um dos quais foi apresentado oralmente no workshop do CLEF 2008 [40] e o outro apresentado como poster no PROPOR 2008 [39]. O primeiro dos artigos referidos foi reformulado para publicação internacional que acontecerá já em 2009 [41].

2.7 Floresta Sintá(c)tica

Cláudia Freitas e Paulo Rocha

Realizaram-se as seguintes actividades no âmbito da Floresta Sintá(c)tica:

- Criação de uma nova página do projecto Floresta Sintá(c)tica.
- Actualização e expansão da documentação técnica e linguística, e passagem ao formato HTML.
- Criação de um mini-guia da Floresta e actualização e tradução para inglês de um breve tutorial (visita guiada) sobre o sistema de busca da Floresta.
- Mudança do sistema de busca Águia para Coimbra e sua adaptação às mudanças de formato previamente decididas; manutenção do sistema.
- Criação de um novo motor de busca em árvores sintácticas, o Milhafre.
- Criação de tutorial, exemplos e documentação relativa ao novo motor de busca.
- Correcções/modificações automáticas sobre o material da Floresta Sintá(c)tica.
- Correcções no Bosque, e disponibilização das suas novas versões (actualmente 8.0) em diversos formatos.
- Busca de maior homogeneidade no tratamento de determinados fenómenos lingüísticos (flutuação entre adjectivos e substantivos; numerais; locuções adverbiais; expressões multi-vocabulares; datas; prefixos e elipses).
- Criação de novas etiquetas.
- Criação de um programa de alinhamento de novas etiquetas no Bosque (i.e., nos ficheiros revistos manualmente), o que possibilita a inserção dessas novas etiquetas, relativas principalmente a estruturas sintácticas mais complexas.
- Criação de programas para validação dessas novas etiquetas.
- Inserção / revisão manual de cerca de 15.000 novas etiquetas.
- Criação dos novos corpora Selva e Amazônia, e consequente disponibilização ao público.
- Criação de documentação relativa aos novos corpora.
- Revisão de parte da Selva.
- Divulgação do novo material da Floresta em três workshops, uma conferência e um artigo de periódico [53, 55, 168, 169, 170].

2.8 PAPEL

Hugo Oliveira e Diana Santos

O PAPEL (Palavras Associadas - Porto Editora/Linguateca) é um projecto que pretende construir, de forma automática ou quase automática, uma ontologia lexical para o português, através de informação extraída do Dicionário da Língua Portuguesa da Porto Editora [263].

Neste projecto, iniciado apenas em Setembro de 2007 com a contratação de Hugo Oliveira, mas

que pretendia dar continuação ao trabalho iniciado pela bolsa de doutoramento do Nuno Seco interrompida em Setembro de 2006, foram desenvolvidas as seguintes actividades, em colaboração com o Núcleo de Investigação e Desenvolvimento da Porto Editora:

- Actividades exploratórias da base lexical da Porto Editora (PE), incluindo uma primeira ideia dos n-gramas mais frequentes nas definições.
- Foi implementado um chart parser em Java para analisar as entradas no dicionário da PE disponibilizado em (http://linguateca.dei.uc.pt/index.php?sep=recursos).
- Foi desenvolvida uma primeira gramática que pretendia extrair relações de causa/efeito entre palavras.
- Foi definido e testado o workflow que liga as três localizações do projecto PAPEL, ou seja, o pólo de Coimbra da Linguateca, o pólo de Oslo da Linguateca, e a Porto Editora, no Porto.
- Foi desenvolvido um programa para facilitar a comparação de grupos de relações (por exemplo dois grupos obtidos com diferentes versões das gramáticas).
- Melhorias à gramática para a extracção da relação de causa/efeito e criadas novas gramáticas/regras, respectivamente para as relações de: parte-de, finalidade-de, hiperonímia, local-origem-de e produtor-de.
- Foram avaliadas manualmente um conjunto das relações obtidas com ajuda de um programa para esse fim.
- Artigo aceite e apresentado no PROPOR 2008 (Aveiro) [64].
- Outras apresentações no CMS WinterMeeting (Coimbra) e Workshop on Language Resources for Teaching and Research (Porto) [174, 175].
- Foram escritos quatro relatórios sobre o PAPEL, após revisão da literatura [61, 62, 65, 66].

2.9 Port4NooJ

Anabela Barreiro

O Port4NooJ é um módulo linguístico desenvolvido no sistema NooJ, tendo em vista o processamento automático do português. Este módulo é constituído por recursos linguísticos (dicionários e gramáticas), os quais incluem também uma extensão bilingue, permitindo a sua utilização em aplicações como, por exemplo, a tradução automática do português para o inglês. O dicionário de cobertura geral foi construído com base nos recursos lexicais do sistema de tradução automática OpenLogos (http://logos-os.dfki.de/). Embora o Port4NooJ tenha sido desenvolvido no âmbito do trabalho de doutoramento de Anabela Barreiro, financiado pela Fundação para a Ciência e a Tecnologia (Refª SFRH/BD/14076/2003), entre o dia 1 de Setembro e o dia 31 de Novembro de 2008, o desenvolvimento e melhoria deste recurso foi financiado pela Linguateca. O Port4NooJ está a ser usado no Corpógrafo, nomeadamente para pesquisas e obtenção de concordâncias mais sofisticadas do ponto de vista linguístico. O sistema está disponível no sítio da Linguateca (http://www.linguateca.pt/Repositorio/Port4Nooj/) e descrito em [8].

2.10 ReEscreve

Anabela Barreiro

O ReEscreve é um sistema de ajuda e sugestão à escrita que utiliza mecanismos de parafraseamento e edição de texto [6]. Permite alterar, simplificar ou clarificar textos e pode ser utilizado em modo interactivo ou em modo automático. Pode auxiliar na escrita de linguagem técnica ou na preparação de textos para a tradução automática (a pré-edição contribui para a obtenção de melhores traduções). O

trabalho foi desenvolvido no âmbito do doutoramento da Anabela Barreiro, financiado pela Fundação para a Ciência e a Tecnologia (Refª SFRH/BD/14076/2003). No âmbito do contrato de prestação de serviços com a Linguateca, o sistema foi convertido num serviço público e está alojado em http://www.linguateca.pt/ReEscreve/

2.11 WPT-05

David Cruz

As colecções WPT consistem nas recolhas da web portuguesa criadas pelo grupo de investigação XLDB/Lasige da Faculdade de Ciências da Universidade de Lisboa.

A WPT-05 consiste na última versão da recolha e engloba as páginas obtidas entre Fevereiro de 2005 e Setembro de 2006 e apresenta-se como uma actualização da colecção WPT03 disponibilizada em 2004 pela Linguateca [15].

A nova recolha estará disponível em duas versões: uma versão em RDF/XML e a outra em formato ARC.

A versão RDF/XML contém cerca de 9,5 milhões de documentos e tira partido da tecnologia RDF para a representação de duplicados e hierarquias entre páginas, apresentando o conteúdo dos documentos filtrado de forma a conter exclusivamente o texto presente no documento. Esta versão consiste num recurso interessante para diversas áreas de investigação, nomeadamente a linguística, a sociologia e a computação.

A versão ARC contém cerca de 10,5 milhões de documentos no estado em que foram recolhidos, i.e. sem pós-processamento dos documentos. A colecção encontra-se no formato ARC do Internet Archive e tem interesse, entre outros, para as áreas da computação e do arquivamento da web.

3. Instrução e disseminação

3.1 Formação de recursos humanos

Diana Santos

De notar que esta rubrica corresponde ao trabalho dos vários orientadores no âmbito (total ou parcial) do trabalho na Linguateca, e não apenas ao trabalho da redactora da rubrica.

Em 2007, a seguinte tese de mestrado foi concluída e defendida na Faculdade de Engenharia da Universidade do Porto:

- Luís Miguel Cabral. SUPERB - Sistema Uniformizado de PEsquisa de Referências Bibliográficas. Tese de Mestrado. Faculdade de Engenharia da Universidade do Porto. 2007. [10]

Em 2008 a seguinte tese de doutoramento foi concluída e defendida na Faculdade de Engenharia da Universidade do Minho, em Braga:

- Alberto Simões. Extracção e Generalização de Recursos de Tradução com base em Dicionários Probabilísticos. Tese de Doutoramento. Faculdade de Engenharia da Universidade do Minho , Universidade do Minho. Março de 2008. [127]

Marcirio Chaves efectuou a prova de qualificação para o doutoramento, exigida pela Universidade, a 22 de Fevereiro de 2007 [33], e entregou a sua tese em Dezembro de 2008, com consequente defesa em 2009:

- Marcirio Silveira Chaves. Uma metodologia para construção de ontologias e integração de conhecimento geográfico. Universidade de Lisboa. [36]

A tese de doutoramento de Anabela Barreiro na Faculdade de Letras da Universidade do Porto encontra-se na sua fase final:

- Anabela Marques Barreiro. Make it simple with paraphrases: Automated paraphrasing for authoring aids and machine translation.

Em Abril de 2007 foi iniciado um doutoramento no âmbito da Linguateca, o de Nuno Cardoso (Universidade de Lisboa), sobre reformulação de pesquisas na área de recolha de informação geográfica:

- Nuno Cardoso. A Semantic-flavored Query Reformulation for Geographic Web Search Engines (título provisório)

Também se iniciou no fim de 2008 o doutoramento do Hugo Gonçalo Oliveira na Universidade de Coimbra, sobre a construção de ontologias lexicais para o português, na continuação do trabalho iniciado no âmbito do PAPEL:

- Hugo Gonçalo Oliveira. Extracção automática de uma ontologia lexical para o português. (título provisório)

Também alguma orientação foi dada a uma tese de mestrado em integração de sistemas de resposta automática a perguntas e serviços Web semânticos, de Yuzhu Yang (Universidade de Oslo), no domínio geográfico, que contudo não foi ainda entregue.

3.2 Outras actividades de orientação e formação

Diana Santos

Em 2007 tiveram lugar o quarto e o quinto simpósios doutorais da Linguateca, ambos no pólo de Lisboa do XLDB, respectivamente em Março e em Outubro, cujo desenrolar pode ser visto em http://www.linguateca.pt/documentos/SimposioDoutoralMarco2007.html e http://www.linguateca.pt/documentos/SimposioDoutoralOutubro2007.html.

Além disso, foi dado um curso sobre Avaliação na *European Summer School on Language, Logic and Information* (ESSLLI 2007) (Dublin, 6-17 de Agosto de 2007), suportado em conjunto por esta escola, pelo SINTEF e pela Linguateca [184].

Em Junho de 2008 a Linguateca organizou o TaLC at TALC, um tutorial sobre o uso dos corpos de língua portuguesa da Linguateca no ensino do português [162, 170, 177, 188], e em Novembro de 2008 foi dado um curso avançado sobre o COMPARA na Segunda Escola Brasileira de Linguística Computacional, em São José do Rio Preto [192], suportado por esta escola e pela Linguateca.

3.3 Livro do Primeiro HAREM

Nuno Cardoso e Diana Santos

A edição do livro sobre o Primeiro HAREM foi concluída em Novembro de 2007, com a sua edição digital [133]. O processo envolveu uma revisão cruzada entre autores e editores de todos os capítulos do livro, a formatação do documento, a requisição de um identificador ISBN e a sua divulgação pelas listas de interesse em processamento computacional do português. Em 2008, o livro foi editado em versão impressa pela FCCN, tendo sido distribuído gratuitamente aos participantes do Encontro do Segundo HAREM , do PROPOR e do Encontro Linguateca: 10 anos em Setembro de 2008.

3.4 Livro do Segundo HAREM

Cristina Mota

Aquando da realização do encontro do Segundo HAREM, em Aveiro no dia 7 de Setembro de 2008, os participantes foram convidados a enviar artigos sobre a sua participação tendo em vista a criação do livro sobre o Segundo HAREM. O envio da primeira versão dos artigos decorreu até ao dia 12 de Outubro de 2008, momento a partir do qual se deu início ao processo de revisão cruzada entre os vários autores. Foram igualmente convidados revisores externos à avaliação conjunta e constitui-se uma comissão de revisão com sete revisores. Já foram recebidas todas as versões finais dos artigos e encontramo-nos já na última fase de edição, antes de estabelecer os contactos necessários de modo a que o livro venha a ser editado digitalmente e de forma impressa.

3.5 Livro alusivo ao encontro Linguateca: 10 anos

Luís Costa

Este volume resultou da iniciativa da Linguateca de chamar as comunidades das áreas do processamento computacional do português, do ensino da língua ou da tradução usando meios computacionais, entre outras, a pronunciar-se sobre a nossa actividade e sobre os nossos falhanços e sucessos, no encontro Linguateca: 10 anos.

Este volume foi assim organizado alfabeticamente a partir dos resumos alargados enviados à organização do encontro. A edição concluiu-se em Novembro de 2008, com a sua edição digital [136]. O processo envolveu uma ligeira revisão dos editores (que se limitaram a recomendações de terminologia e a garantir a coerência das citações), a formatação do documento, a requisição de um identificador ISBN e a sua divulgação pelas listas de interesse em processamento computacional do português.

3.6 Encontro Linguateca: 10 anos

Luís Costa

O encontro Linguateca: 10 anos realizou-se no dia 11 de Setembro de 2008, na Curia, Aveiro, como encontro satélite do PROPOR 2008.

Pretendíamos com este encontro chamar a comunidade da área a quem servimos, ou seja, as pessoas envolvidas no processamento computacional da nossa língua, em linguística computacional do português, no ensino da língua ou na tradução usando meios computacionais, entre outros, a pronunciar-se sobre a nossa actividade e sobre os nossos falhanços e sucessos, por ocasião de um marco significativo na nossa existência: o terminar da Linguateca como fora concebida até aí.

Queríamos também discutir se haveria alguma forma de continuar e/ou melhorar o que tínhamos feito, assim como pedir um balanço, pelo exterior, dos nossos contributos.

Nesse sentido, em Fevereiro de 2008 fez-se uma chamada de comunicações para o encontro em que se pediram resumos de propostas de comunicação descrevendo ou comentando trabalho em processamento da língua portuguesa relacionado de alguma forma com a Linguateca.

Aproveitámos também o encontro para fazer um balanço da actividade da Linguateca, quer sob uma perspectiva global da visão de cada pólo, representado pelo sénior responsável, quer sob uma perspectiva de apresentar pela primeira vez, de uma forma estruturada, todos os trabalhos de investigação (a nível de doutoramento) em curso ou terminados sob a chancela (mesmo que parcial ou não oficial) da Linguateca.

Consideramos que o encontro foi muito rico por ter congregado tanto intervenções de carácter predominantemente científico como apresentações de cariz de política científica ou de balanço (todas elas acessíveis do sítio do encontro, http://www.linguateca.pt/Linguateca10anos/), e por ter permitido a discussão e o debate presencial.

3.7 Encontro do Segundo HAREM

Cláudia Freitas

Em 7 de Setembro de 2008, no Campus da Universidade de Aveiro, aconteceu o Encontro do Segundo HAREM, planejado como um evento satélite do PROPOR 2008.

O Encontro teve quatro objectivos principais, que acreditamos terem sido integralmente cumpridos:
a) apresentar o Segundo HAREM
b) apresentar os sistemas participantes do Segundo HAREM
c) discutir a continuação do HAREM
c) divulgar o HAREM para além dos sistemas participantes

Com relação ao primeiro ponto, foram feitas quatro apresentações, três delas relativas ao trabalho da organização.

Quanto ao segundo ponto, todos os nove participantes do Segundo HAREM apresentaram seus sistemas, e ainda houve espaço para a apresentacão de um décimo sistema, que não participou do HAREM, mas continuou como observador do evento.

Ao fim do Encontro houve uma sessão intitulada “Discussão sobre o futuro do HAREM”, em que foi feito um balanço do Segundo HAREM, bem como foram discutidos o interesse em uma continuação da inciativa e formas de viabilizar essa continuação. Todos os presentes foram unânimes em declarar o interesse na continuação do HAREM, mas poucos se manifestaram positivamente no sentido de assumir a organização do Evento, reconhecendo todo o trabalho necessário à sua realização.

Ainda nessa última sessão foram discutidas questões relativas à publicação das Actas do Segundo HAREM, como a definição de prazos e processo de revisão.

Por fim, o grande número de inscritos no encontro – 29, exceptuando-se os membros da organização – confirma o interesse da comunidade científica em iniciativas como a do HAREM, corroborando o que já havia ficado explícito no debate ao fim do encontro.

Todas as apresentações estão disponíveis na página do HAREM:
http://www.linguateca.pt/HAREM/

3.8 “TaLC at TaLC”
Diana Santos

Os pólos de Oslo, Coimbra e do COMPARA organizaram um encontro satélite do TaLC 2008, denominado "TaLC at TaLC: Teaching and Linguateca's (Portuguese language) Corpora", sobre os corpos portugueses da Linguateca e a sua aplicação no ensino do português no dia 3 de Julho de 2008. Veja-se o material de ensino em [162, 170, 177, 188].

3.9 “Workshop on Language Resources for Teaching and Research”
Sérgio Matos

O pólo do Porto da Linguateca organizou o “Workshop on Language Resources for Teaching and Research”, que decorreu no dia 23 de Abril de 2008, na Faculdade de Letras da Universidade do Porto. Este workshop serviu para apresentar vários trabalhos desenvolvidos no âmbito da Linguateca, dirigidos tanto à área de investigação como ao ensino, e permitiu também uma partilha de ideias entre os presentes. Foram feitas nove apresentações [140, 169, 175, 176, 179, 195, 201, 259, 261], sendo o resultado final do wokshop bastante positivo, tendo permitido aos vários apresentadores a divulgação do seu trabalho, assim como a discussão e troca de ideias. Houve também participação de alguns elementos da comunidade académica, o que permitiu não só a apresentação destes trabalhos e recursos como possíveis ferramentas para o ensino e investigação, que era um dos objectivos principais do workshop, como também ouvir as suas opiniões sobre a investigação desenvolvida e como esta pode ser melhor orientada para os interesses e necessidades desta comunidade.

3.10 “Um passeio pela Floresta Sintá(c)tica”
Cláudia Freitas e Paulo Rocha

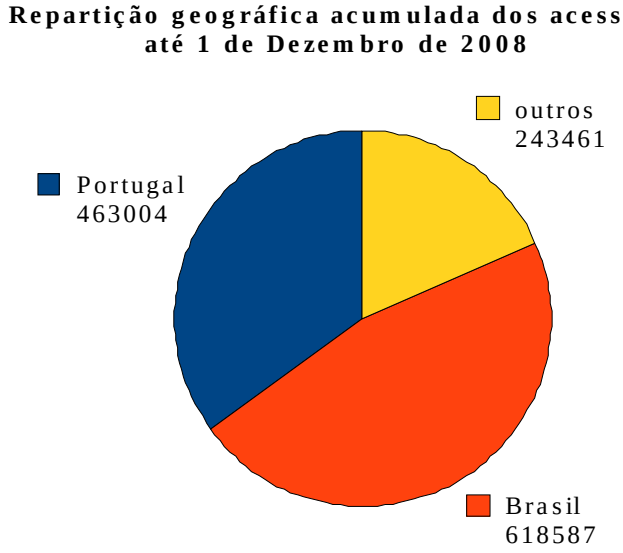
Foi realizada uma acção de divulgação da Floresta Sintá(c)tica em Coimbra no dia 28 de Setembro de 2007. Destinada principalmente a linguistas e informáticos, o evento, chamado "Um Passeio pela Floresta Sintá(c)tica", contou com 7 palestras que, além da Floresta, apresentaram também a Linguateca, o projecto AC/DC e o Pólo de Coimbra [168, 172, 182, 185, 186, 187, 240].

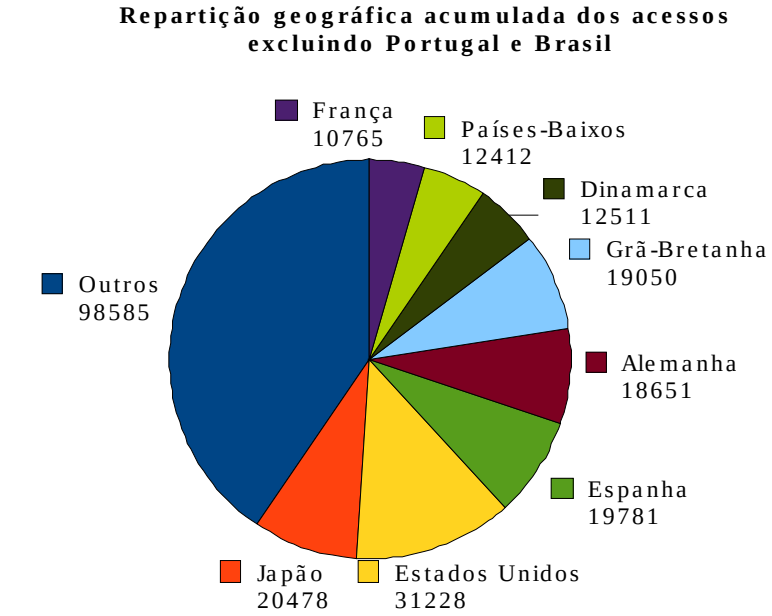
4. Medidas de impacto (Catálogo e estatísticas)

Luís Cabral

O número de visitas às páginas da Linguateca, de 1 de Janeiro de 2007 a 1 de Dezembro de 2008, foi de 3.574.523 (no total, desde o início, atinge os 6.641.243).

Para dar uma ideia do impacto internacional do nosso portal apresentamos nas figuras seguintes a distribuição do perfil geográfico dos visitantes das nossas páginas desde o início do projecto (os valores referem-se apenas aos acessos para os quais foi possível determinar a sua origem, 27,9% do total de acessos):

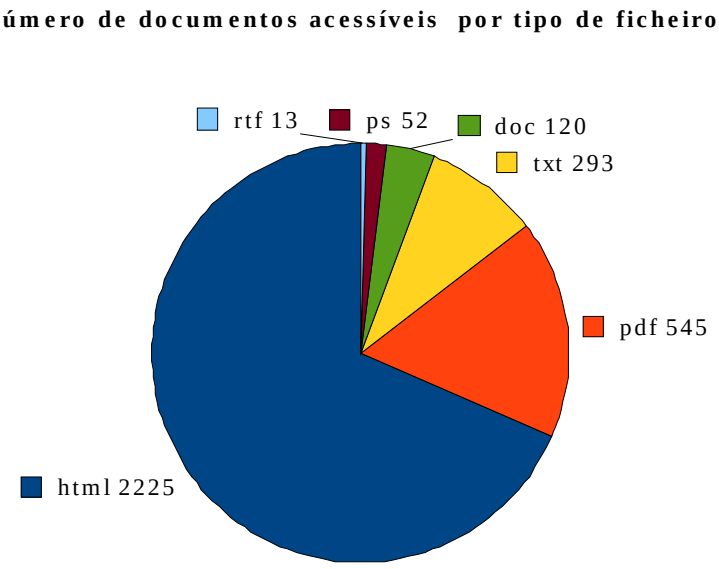




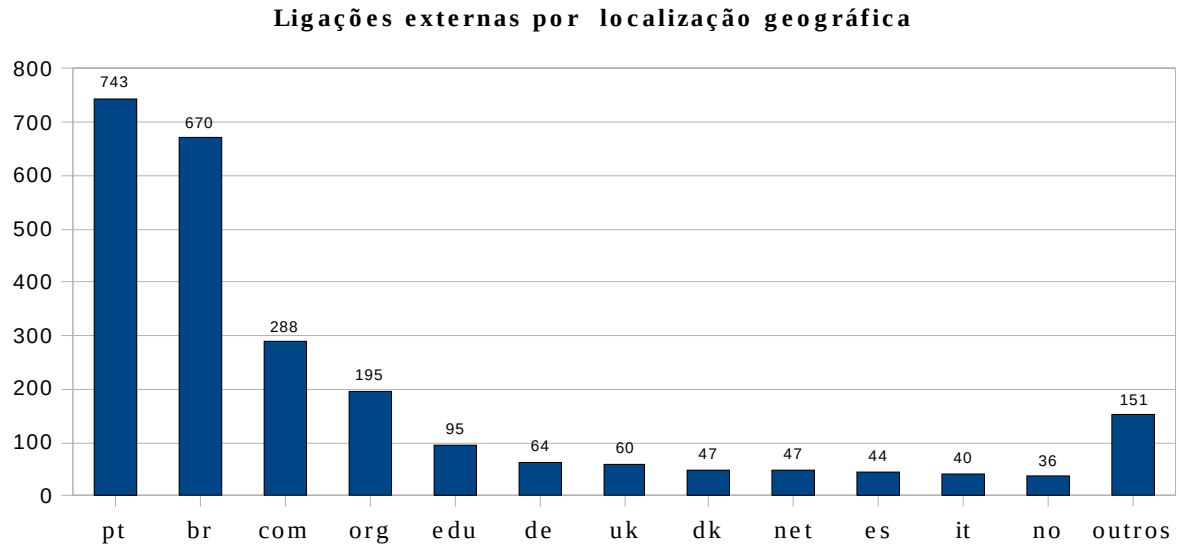
Apresentamos também o número de pedidos dos recursos que exigem registo durante o período de 2007-2008 (o valor entre parênteses refere-se ao número de pedidos desde o início da disponibilização de cada um dos recursos)

Corpógrafo: 559 (1595)
CETEMPúblico: 111 (529)
CETENFolha: 102 (349)
CHAVE: 34 (70)
WBR-99: 4 (28)
GeoNet-PT01: 13 (28)
WPT-03: 0 (11)

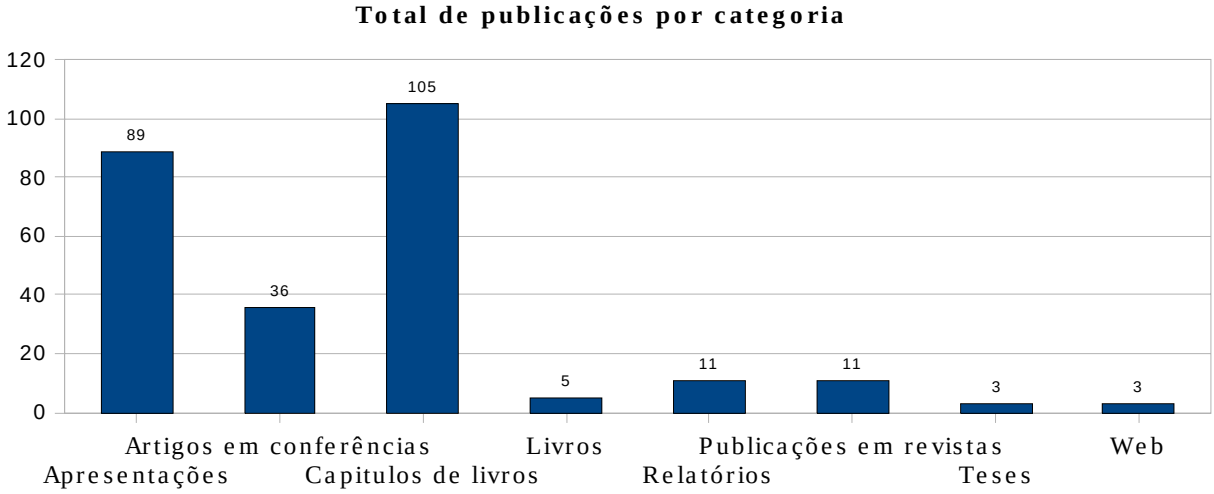
O número actual de documentos acessíveis no sítio da Linguateca é de 3248, constituindo 2242 documentação própria da Linguateca. A figura seguinte pormenoriza o tipo de documentos disponibilizados.



O número de ligações externas incluídas nos nossos catálogos é de 2554, e a distribuição geográfica das localizações para que apontam encontra-se na figura seguinte:



Finalmente, a distribuição das publicações (listadas na secção seguinte) por categoria, seguindo as normas do catálogo de publicações da Linguateca, e por categoria, encontra-se na figura seguinte:



De notar que algumas da publicações, embora apenas tenham sido publicadas em 2007 ou 2008, referem-se a trabalho produzido em anos anteriores, como por exemplo as publicações no livro "Avaliação conjunta: um novo paradigma no processamento computacional da lingua portuguesa" [132].

5. Lista de publicações

As publicações da Linguateca no período a que se refere o presente relatório foram listadas com o auxílio do SUPeRB [10], e são apresentadas por ordem alfabética.

Listamos na secção 5.2 as publicações criadas no âmbito da Linguateca (por exemplo incluídas em livros ou material de ensino editado ou criado pela Linguateca) mas não por pessoas com vínculo formal à Linguateca, para as distinguir claramente.

5.1 Publicações e apresentações da Linguateca

5.1.1. Artigos em conferências, livros, revistas, livros e teses

[1] Rachel Aires. "Uso de marcadores estilísticos para a busca na Internet em português". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 87-92.

[2] José João Almeida & Alberto Simões. "XML::TMX - Processamento de Memórias de Tradução de Grandes Dimensões". In José Carlos Ramalho, João Correia Lopes & Luís Carriço (eds.), *XML: Aplicações e Tecnologias Associadas (XATA2007)* (FCUL, Lisboa, 15-16 de Fevereiro de 2007), Universidade do Minho, pp. 83-93.

[3] José João Almeida & Alberto Simões. "Jspellando nas morfolimpíadas: Sobre a participação do Jspell nas Morfolimpíadas". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 83-89.

[4] José João Dias de Almeida. "RENA - Reconhecedor de Entidades". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 157-172.

[5] Anabela Barreiro & Susana Afonso. "Construção da lista dourada para as primeiras Morfolimpíadas do português". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 107-118.

[6] Anabela Barreiro. "ParaMT: a Paraphraser for Machine Translation". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), *Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008)* Vol. 5190, (Aveiro, Portugal, 8-10 de Setembro de 2008), Springer Verlag, pp. 202-211.

[7] Anabela Barreiro. "Novas Ferramentas e Recursos Linguísticos para a Tradução Automática: Por ocasião d'O Fim do Início de uma Nova Era no Processamento da Língua Portuguesa". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 13-23.

[8] Anabela Barreiro. "Tutorial introdutório do Port4NooJ". 21 de Novembro de 2008. http://poloclup.linguateca.pt/Port4NooJ/Tutorial.pdf

[9] Eckhard Bick, Diana Santos, Susana Afonso & Rachel Marchi. "Floresta Sintá (c)tica: Ficção ou realidade?". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no*

processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 291-300.

[10] Luís Miguel Cabral. SUPeRB - **S**istema **U**niformizado de **Pe**squisa de **R**eferências **B**ibliográficas. Tese de Mestrado. Faculdade de Engenharia da Universidade do Porto. Março de 2007.

[11] Luís Miguel Cabral, Luís Fernando Costa & Diana Santos. "Esfinge at CLEF 2007: First steps in a multiple question and multiple answer approach". In Alessandro Nardi & Carol Peters (eds.), *Working Notes for the CLEF 2007 Workshop (CLEF 2007)* (Budapeste, Hungria, 19-21 de Setembro de 2007), s/pp.

[12] Luís Miguel Cabral, Diana Santos & Luís Fernando Costa. "SUPeRB: Building bibliographic resources on the computational processing of Portuguese". In PROPOR 2008 Special Session: Applications of Portuguese Speech and Language Technologies (Curia, Portugal, 10 de Setembro de 2008).

[13] Luís Miguel Cabral, Diana Santos & Luís Fernando Costa. "SUPeRB - Gerindo referências de autores de língua portuguesa". In VI Workshop Information and Human Language Technology (TIL'08) (Vila Velha, ES, Brasil, 28-29 de Outubro de 2008).

[14] Luís Miguel Cabral, Luís Fernando Costa & Diana Santos. "What happened to Esfinge in 2007?". In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras & Diana Santos (eds.), Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Berlin :Springer, Lecture Notes in Computer Science, 2008, pp. 261-268.

[15] Nuno Cardoso, Bruno Martins, Daniel Gomes & Mário J. Silva. "WPT 03: a primeira colecção pública proveniente de uma recolha da web portuguesa". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 279-288.

[16] Nuno Cardoso, David Cruz, Marcirio Chaves & Mário J. Silva. "The University of Lisbon at GeoCLEF 2007". In Alessandro Nardi & Carol Peters (eds.), Working Notes for the CLEF 2007 Workshop (CLEF 2007) (Budapeste, Hungria, 19-21 de Setembro de 2007), s/pp.

[17] Nuno Cardoso & Mário J. Silva. "Query Expansion through Geographical Feature Types". In 4th Workshop on Geographic Information Retrieval (GIR 2007) (Lisboa, Portugal, 9 de Novembro de 2007), ACM press.

[18] Nuno Cardoso, Mário J. Silva & Marília Antunes. "Validação Estatística dos resultados do Primeiro HAREM". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 59-77.

[19] Nuno Cardoso & Diana Santos. "Directivas para a identificação e classificação semântica na colecção dourada do HAREM". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 211-238. Documento original publicado no sítio do HAREM a 29 de Março de 2006. Republicado como Relatório técnico DI-FCUL TR-06-18 : Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa, Novembro de 2006

[20] Nuno Cardoso, Diana Santos & Rui Vilela. "Directivas para a identificação e classificação morfológica na colecção dourada do HAREM". In Diana Santos & Nuno Cardoso (eds.),

Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 239-244. Republicação do Relatório técnico DI-FCUL TR-06-19 : Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa. Novembro de 2006

[21] Nuno Cardoso, Mário J. Silva & Bruno Martins. "The University of Lisbon at CLEF 2006 Ad-Hoc Task". In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke & Maximilian Stempfhuber (eds.), Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers. Berlin / Heidelberg :Springer, Lecture Notes in Computer Science, 2007, pp. 51-56.

[22] Nuno Cardoso & Diana Santos. "To separate or not to separate: reflections about GIR practice". In Proceedings of NMEIR 2008 : ECIR 2008 Workshop on novel methodologies for evaluation in information retrieval (Glasgow, Reino Unido, 30 de Março de 2008), s/pp.

[23] Nuno Cardoso. "Apêndice H: SAHARA - Serviço de Avaliação HAREM Automático". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[24] Nuno Cardoso. "Novos rumos para a recuperação de informação geográfica em português". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos. Linguateca, 2008, pp. 71-85.

[25] Nuno Cardoso, Mário J. Silva & Diana Santos. "Handling Implicit Locality Evidence for Geographic IR". In ACM 17th Conference on Information and Knowledge Management (CIKM 2008) (Napa Valley, CA, USA, 26-30 de Outubro de 2008).

[26] Nuno Cardoso, David Cruz, Marcirio Chaves & Mário J. Silva. "Using Geographic Signatures as Query and Document Scopes in Geographic IR". In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras & Diana Santos (eds.), Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers 5152, 2008, Berlin : Springer. Lecture Notes in Computer Science, pp. 802-810.

[27] Paula Carvalho & Cláudia Freitas. "Apêndice F: Exemplário do Segundo HAREM". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[28] Paula Carvalho & Hugo Gonçalo Oliveira. "Apêndice G: Manual de Utilização do Etiquet (H)AREM". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[29] Paula Carvalho & Cláudia Freitas. "Apêndice E: Critérios de ALT". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[30] Paula Carvalho, Hugo Gonçalo Oliveira, Diana Santos, Cláudia Freitas & Cristina Mota. "Segundo HAREM: Modelo geral, novidades e avaliação". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[31] Paula Carvalho & Cristina Mota. "Sobre o TEMPO no Segundo HAREM". (2009). [A ser enviado]

[32] Marcirio Chaves, Catarina Rodrigues & Mário J. Silva. "Data Model for Geographic

Ontologies Generation". In José Carlos Ramalho, João Correia Lopes & Luís Carriço (eds.), XML: Aplicações e Tecnologias Associadas (XATA2007) (FCUL, Lisboa, 15-16 de Fevereiro de 2007), Universidade do Minho, pp. 47-58.

[33] Marcirio Silveira Chaves. Uma metodologia para construção de ontologias e integração de conhecimento geográfico. Prova de qualificação. Faculdade de Ciências, Universidade de Lisboa. 22 de Fevereiro de 2007.

[34] Marcírio Chaves. "Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o SEI-Geo no Segundo HAREM". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[35] Marcirio Chaves. "Criação e expansão de geo-ontologias, dimensionamento de informação geográfica e reconhecimento de locais e seus relacionamentos em textos". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos. Linguateca, 2008, pp. 49-59.

[36] Marcirio Silveira Chaves. Uma Metodologia para Construção de Geo-Ontologias. Doutoramento. Faculdade de Ciências, Universidade de Lisboa. Dezembro de 2008.

[37] Luís Costa, Paulo Rocha & Diana Santos. "Organização e resultados morfolímpicos". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 15-33.

[38] Luís Costa. "Resumo da actividade da Linguateca de 16 de Dezembro de 2006 a 31 de Dezembro de 2007". Janeiro de 2008. Com a colaboração (por ordem alfabética) de Ana Frankenberg-Garcia, Cláudia Freitas, Diana Santos, Hugo Oliveira, Luís Costa, Luís Miguel Cabral, Paulo Rocha e Sérgio Matos

[39] Luís Fernando Costa & Luís Miguel Cabral. "Answering Portuguese Questions". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008) Vol. 5190, (Aveiro, Portugal, 8-10 de Setembro de 2008), Springer Verlag, pp. 228-231.

[40] Luís Fernando Costa. "Esfinge at CLEF 2008: Experimenting with answer retrieval patterns. Can they help?". In Francesca Borri, Alessandro Nardi & Carol Peters (eds.), Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop (Aarhus, Dinamarca, 17-19 de Setembro de 2008), s/pp.

[41] Luís Fernando Costa. "Using Answer Retrieval Patterns to Answer Portuguese Questions". In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas & Viviane Petras (eds.), Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 2009, Springer.
[Prelo]

[42] Pamela Forner, Anselmo Peñas, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe & Erik Tjong Kim Sang. "Overview of the CLEF 2008 Multilingual Question Answering Track". In Francesca Borri, Alessandro Nardi & Carol Peters (eds.), Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop (Aarhus, Dinamarca, 17-19 de Setembro de 2008), s/pp.

[43] Pamela Forner, Anselmo Peñas, Iñaki Alegria, Corina Forascu, Nicolas Moreau, Petya

Osenova, Prokopis Prokopidis, Paulo Rocha, Bogdan Sacaleanu, Richard Sutcliffe & Erik Tjong Kim Sang. "Overview of the CLEF 2008 Multilingual Question Answering Track". In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas & Viviane Petras (eds.), Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 2009, Springer.

[44] Ana Frankenberg-Garcia. "COMPARA - Instruções para a revisão do Reconhecimento Óptico de Caracteres (ROC) e etiquetagem preliminar". 4 de Novembro de 2007.

[45] Ana Frankenberg-Garcia. "Lost in Parallel Concordances". In W. Teubert (ed.), Corpus Linguistics: critical concepts in linguistics. Londres :Routledge, 2007, pp. 176-190. Anteriormente publicado em G. Aston, S. Bernardini and D. Stewart (eds.) Corpora and language learners. Amsterdam: John Benjamins, 2004.

[46] Ana Frankenberg-Garcia. "Suggesting rather special facts: a corpus-based study of distinctive lexical distributions in translated texts". Corpora 3.2 (2008).
[Prelo]

[47] Ana Frankenberg-Garcia. "Raising teachers' awareness to corpora". In Natalie Kubler (ed.), Selected papers from the 7th TaLC conference" (preliminary title). Amesterdão :Rodopi, 2008.
[Prelo]

[48] Ana Frankenberg-Garcia. "Compilação e uso de corpora paralelos". In Stella E. O. Tagnin & Oto Araújo Vale (eds.), Avanços da Lingüística de Corpus no Brasil. São Paulo, Brasil :Editora Humanitas/FFLCH/USP, 2008, pp. 117-136.

[49] Ana Frankenberg-Garcia. "Are translations longer than source texts? A corpus-based study of explicitation". In A. Beeby, P. Rodríguez & P. Sánchez-Gijón (eds.), Corpus use and learning to translate (CULT) : An Introduction 2008, Amesterdão and Filadélfia : John Benjamins. Republicação de. "Are translations longer than source texts? A corpus-based study of explicitation". In Third International CULT (Corpus Use and Learning to Translate) Conference 2004
[Prelo]

[50] Ana Frankenberg-Garcia. "Compiling and using a parallel corpus for research in translation". Cambridge Scholars Publishing, 2009.
[Enviado]

[51] Cláudia Freitas, Diana Santos, Paula Carvalho & Hugo Gonçalo Oliveira. "Apêndice C: ReRelEM - Reconhecimento de Relações entre Entidades Mencionadas. Segundo HAREM: proposta de nova pista". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[52] Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho & Cristina Mota. "Relações semânticas do ReRelEM: além das entidades no Segundo HAREM". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[53] Cláudia Freitas, Paulo Rocha & Eckhard Bick. "Floresta Sintá (c)tica: Bigger, Thicker and Easier". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008) Vol. 5190, (Aveiro, Portugal, 8-10 de Setembro

de 2008), Springer Verlag, pp. 216-219.

[54] Cláudia Freitas & Susana Afonso. "Bíblia Florestal: Um manual lingüístico da Floresta Sintá (c)tica". 29 de Setembro de 2008. http://linguateca.dei.uc.pt/Floresta/BibliaFlorestal/

[55] Cláudia Freitas, Paulo Rocha & Eckhard Bick. "Um mundo novo na Floresta Sintá (c)tica - o treebank para Português". Calidoscópio - Revista de Pós Graduação em Lingüística Aplicada da Unisinos, Rio Grande do Sul 6.3 (2008).
[Prelo]

[56] Claudia Freitas, Diana Santos, Hugo Gonçalo Oliveira, Paula Carvalho & Cristina Mota. "Detection of relations among named entities: report of a shared task". In Semantic Evaluations: Recent Achievements and Future Directions, NAACL-HLT 2009 Workshop (SEW2009) (Boulder, Colorado, USA, 4 de Junho de 2009). Artigo para ser submetido
[A ser enviado]

[57] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Andres Montoyo, Giorgio M. Di Nunzio & Nicola Ferro. "Challenges to Evaluation of Multilingual Geographic Information Retrieval in GeoCLEF". In Workshop on Evaluation of Information Access (EVIA) May 15 (Tóquio, Japão, 15 de Maio de 2007), s/pp.

[58] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio M. Di Nunzio & Nicola Ferro. "GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview". In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke & Maximilian Stempfhuber (eds.), Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers. Berlin / Heidelberg :Springer, Lecture Notes in Computer Science, 2007, pp. 852-876.

[59] Danilo Giampiccolo, Pamela Forner, Anselmo Peñas, Christelle Ayache, Dan Cristea, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu & Richard Sutcliffe. "Overview of the CLEF 2007 Multilingual Question Answering Track". In Alessandro Nardi & Carol Peters (eds.), Working Notes for the CLEF 2007 Workshop (CLEF 2007) (Budapeste, Hungria, 19-21 de Setembro de 2007), s/pp.

[60] Danilo Giampiccolo, Pamela Forner, Anselmo Peñas, Christelle Ayache, Dan Cristea, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu & Richard Sutcliffe. "Overview of the CLEF 2007 Multilingual Question Answering Track". In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras & Diana Santos (eds.), Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Berlin :Springer, Lecture Notes in Computer Science, 2008, pp. 200-236.

[61] Hugo Gonçalo Oliveira, Paulo Gomes & Diana Santos. "PAPEL - Trabalho relacionado e relações semânticas em recursos semelhantes". Departamento de Engenharia Informática, FCTUC, CISUC. Dezembro de 2007.

[62] Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes & Nuno Seco. "Extracção de relações a partir de dicionários: Breve história". CISUC, SINTEF. Agosto de 2008.

[63] Hugo Gonçalo Oliveira, Cristina Mota, Cláudia Freitas, Diana Santos & Paula Carvalho. "Avaliação à medida no Segundo HAREM". In Cristina Mota & Diana Santos (eds.),

Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[64] Hugo Gonçalo Oliveira, Paulo Gomes & Diana Santos. "PAPEL: a dictionary-based lexical ontology for Portuguese". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008) Vol. 5190, (Aveiro, Portugal, 8-10 de Setembro de 2008), Springer Verlag, pp. 31-40.

[65] Hugo Gonçalo Oliveira & Paulo Gomes. "Utilização do (analisador sintáctico) PEN para extracção de informação das definições de um dicionário". CISUC. Dezembro de 2008.

[66] Hugo Gonçalo Oliveira & Paulo Gomes. "Apresentação das relações extraídas do Dicionário da Porto Editora". CISUC. 2008.
[Prelo]

[67] Hugo Gonçalo Oliveira & Nuno Cardoso. "SAHARA: a Web-service for HAREM Named Entity Recognition Evaluation". In 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009) (Atenas, Grécia, 30 de Março a 3 de Abril de 2009).
[Enviado]

[68] Susana Inácio, Diana Santos & Rosário Silva. "COMPARAndo cores em português e inglês". In Sónia Frota & Ana Lúcia Santos (eds.), Artigos seleccionados do XXIII Encontro da Associação Portuguesa de Linguística (APL) (Évora, 1-3 de Outubro de 2007), 2008, pp. 271-286.

[69] Susana Inácio & Diana Santos. "Documentação da anotação morfossintáctica da parte portuguesa do COMPARA". Dezembro de 2008.

[70] Bernardo Magnini, Danilo Giampiccolo, Pamela Forner, Christelle Ayache, Valentin Jijkoun, Petya Osenova, Anselmo Peñas, Paulo Rocha, Bogdan Sacaleanu & Richard Sutcliffe. "Overview of the CLEF 2006 Multilingual Question Answering Track". In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke & Maximilian Stempfhuber (eds.), Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers. Berlin / Heidelberg :Springer, Lecture Notes in Computer Science, 2007, pp. 223-256.

[71] Belinda Maia & Anabela Barreiro. "Uma experiência de recolha de exemplos classificados de tradução automática de inglês para português". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 205-216.

[72] Belinda Maia, Rui Sousa Silva, Anabela Barreiro & Cecília Fróis. "N-grams in search of theories". In Proceedings of the 6th International Conference of Practical Applications in Language and Computers (PALC 2007) (Lodz University, Lodz, Polónia, 19-22 de Abril de 2007).

[73] Belinda Maia & Sérgio Matos. "Corpógrafo V4 - Tools for Researchers and Teachers using Comparable Corpora". In Pierre Zweigenbaum, Éric Gaussier & Pascale Fung (eds.), LREC 2008 Workshop on Comparable Corpora (LREC 2008) (Marraquexe, Marrocos, 31 de Maio de 2008), European Language Resources Association (ELRA), pp. 79-82.

[74] Belinda Maia. "Corpógrafo V4 - Tools for Educating Translators". In Elia Yuste Rodrigo (ed.), Topics in Language Resources for Translation and Localisation. Amesterdão/Filadélfia :John Benjamins Pub. Co. 2008, pp. 57-70.

[75] Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker & Xing Xie. "GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview". In Alessandro Nardi & Carol Peters (eds.), Working Notes for the CLEF 2007 Workshop (CLEF 2007) (Budapeste, Hungria, 19-21 de Setembro de 2007), s/pp.

[76] Thomas Mandl, Fredric Gey, Giorgio di Nunzio, Nicola Ferro, Mark Sanderson, Diana Santos & Christa Womser-Hacker. "An evaluation resource for Geographical Information Retrieval". In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) (Marraquexe, Marrocos, 28-30 de Maio de 2008), European Language Resources Association (ELRA), s/pp..

[77] Thomas Mandl, Paula Carvalho, Giorgio Maria di Nunzio, Nicola Ferro, Fred Gey, Ray Larson, Diana Santos & Christa Womser-Hacker. "GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview". In Francesca Borri, Alessandro Nardi & Carol Peters (eds.), Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop (Aarhus, Dinamarca, 17-19 de Setembro de 2008), s/pp.

[78] Thomas Mandl, Fredric Gey, Giorgio Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker & Xie Xing. "GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview". In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras & Diana Santos (eds.), Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Berlin :Springer, Lecture Notes in Computer Science, 2008, pp. 745-772.

[79] Thomas Mandl, Paula Carvalho, Fredric Gey, Ray Larson, Diana Santos & Christa Womser-Hacker. "GeoCLEF 2008: the CLEF 2008 Cross-Language Geographic Information Retrieval Track Overview". In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas & Viviane Petras (eds.), Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 2009, Springer.
[Prelo]

[80] Bruno Martins, Mário Silva & Marcirio Chaves. "O sistema CaGE no HAREM - reconhecimento de entidades geográficas em textos em língua portuguesa". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 97-112.

[81] Bruno Martins, Nuno Cardoso, Marcirio Silveira Chaves, Leonardo Andrade & Mário J. Silva. "The University of Lisbon at GeoCLEF 2006". In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke & Maximilian Stempfhuber (eds.), Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers. Berlin / Heidelberg :Springer, Lecture Notes in Computer Science, 2007, pp. 986-994.

[82] Sérgio Matos & Belinda Maia. "NooJ and Corpógrafo A New Partnership". In Proceedings of the 2008 International NooJ Conference (NooJ'08) (Budapeste, Hungria, 8-10 de Junho de 2008).

[83] Cristina Mota, Diana Santos & Elisabete Ranchhod. "Avaliação de reconhecimento de

entidades mencionadas: princípio de AREM". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 161-176.

[84] Cristina Mota. "R3M, uma participação minimalista no Segundo HAREM". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[85] Cristina Mota, Paula Carvalho, Cláudia Freitas, Hugo Gonçalo Oliveira & Diana Santos. "É tempo de avaliar o tempo". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[86] Viviane Moreira Orengo & Diana Santos. "Radicalizadores versus analisadores morfológicos: Sobre a participação do Removedor de Sufixos da Língua Portuguesa nas Morfolimpíadas". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 91-104.

[87] Elisabete Ranchhod & Cristina Mota. "Unidades lexicais multipalavra, um osso duro de roer: Sobre a participação do LabEL nas Morfolimpíadas". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 75-80.

[88] Paulo Rocha & Diana Santos. "CLEF: Abrindo a porta à participação internacional em avaliação de RI do português". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 143-158.

[89] Paulo Rocha & Diana Santos. "Disponibilizando a <OBRA>Colecção Dourada</OBRA> do <ACONTECIMENTO>HAREM</ACONTECIMENTO> através do projecto <LOCAL|ORGANIZACAO|ABSTRACCAO>AC/DC</LOCAL|ORGANIZACAO|ABSTRACCAO>". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 307-326.

[90] Diana Santos. "Breves explorações num mar de língua". Ilha do Desterro 52 (2007), pp. 127-150. Florianópolis. ISSN: 0101-4846.

[91] Diana Santos. "Avaliação conjunta". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 1-12.

[92] Diana Santos & Ana Frankenberg-Garcia. "The corpus, its users and their needs: a user-oriented evaluation of COMPARA". International Journal of Corpus Linguistics 12 (2007), pp. 335-374.

[93] Diana Santos, Nuno Cardoso & Nuno Seco. "Avaliação no HAREM: Métodos e medidas". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 245-282. Republicação de Relatório técnico DI-FCUL TR-06-17 : Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa. Novembro de 2006

[94] Diana Santos. "O modelo semântico usado no Primeiro HAREM". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 43-57.

[95] Diana Santos & Nuno Cardoso. "Balanço do primeiro HAREM e perspectivas de trabalho futuro". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 87-94.

[96] Diana Santos & Nuno Cardoso. "Breve Introdução ao HAREM". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 1-16.

[97] Diana Santos, Eckhard Bick, Raquel Marchi & Susana Afonso. "A Floresta Sintá (c)tica como recurso para estudar a língua portuguesa". In Regina Márcia Gerber & Vera Vasilévski (eds.), Um percurso para pesquisas com base em corpus. Florianópolis - SC :Editora da UFSC, 2007, pp. 149-160.

[98] Diana Santos. "A tradução na sociedade do conhecimento OU Tradução: uma tecnologia humana de ponta OU Ciência E Tradução". In Actas do IX Seminário de Tradução Cientïfica e Técnica em Língua Portuguesa 2007, Fevereiro de 2007, Lisboa : União Latina.

[99] Diana Santos. "Perfect mismatches: Result in English and Portuguese". In Margaret Rogers & Gunilla Anderman (eds.), Incorporating Corpora: The Linguist and the Translator. Clevedon :Multilingual matters, 2008, pp. 217-242.

[100] Diana Santos & Luís Costa. "QolA: fostering collaboration within QA". In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke & Maximilian Stempfhuber (eds.), Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers 4730, 2007, Berlin / Heidelberg : Springer. Lecture Notes in Computer Science, pp. 569-578.

[101] Diana Santos & Alberto Simões. "Portuguese-English word alignment: some experiments". In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) (Marraquexe, Marrocos, 28-30 de Maio de 2008), European Language Resources Association (ELRA), s/pp..

[102] Diana Santos, Rosário Silva & Susana Inácio. "What's in a colour? Studying and contrasting colours with COMPARA". In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) (Marraquexe, Marrocos, 28-30 de Maio de 2008), European Language Resources Association (ELRA), s/pp..

[103] Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira, Paula Carvalho & Cristina Mota. "Segundo HAREM: Balanço e perspectivas de futuro". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[104] Diana Santos, Paula Carvalho, Cláudia Freitas & Hugo Gonçalo Oliveira. "Apêndice A: Segundo HAREM: directivas de anotação". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[105] Diana Santos. "Enquadramento e historial do Segundo HAREM". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[106] Diana Santos, Cláudia Freitas, Hugo Gonçalo Oliveira & Paula Carvalho. "Second HAREM: new challenges and old wisdom". In António Teixeira, Vera Lúcia Strube de Lima, Luís Caldas de Oliveira & Paulo Quaresma (eds.), Computational Processing of the Portuguese Language, 8th International Conference, Proceedings (PROPOR 2008) Vol.

5190, (Aveiro, Portugal, 8-10 de Setembro de 2008), Springer Verlag, pp. 212-215.

[107] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling & Yvonne Skalban. "Getting geographical answers from Wikipedia: the GikiP pilot at CLEF". In Francesca Borri, Alessandro Nardi & Carol Peters (eds.), Cross Language Evaluation Forum: Working Notes for the CLEF 2008 Workshop (Aarhus, Dinamarca, 17-19 de Setembro de 2008), s/pp.

[108] Diana Santos & Nuno Cardoso. "GikiP: Evaluating geographical answers from Wikipedia". In 5th Workshop on Geographic Information Retrieval (GIR'08) (Napa Valley, CA, USA, 30 de Outubro de 2008), pp. 59-60.

[109] Diana Santos. "Corporizando algumas questões". In Stella E. O. Tagnin & Oto Araújo Vale (eds.), Avanços da Lingüística de Corpus no Brasil. São Paulo, Brasil :Editora Humanitas/FFLCH/USP, 2008, pp. 41-66.

[110] Diana Santos. "Exercícios e exemplos do uso do COMPARA para estudos contrastivos". Dezembro de 2008.

[111] Diana Santos & Nuno Cardoso. "Fostering named entity research: the HAREM evaluation contest". Computational Linguistics (2009).
[Enviado]

[112] Diana Santos, Nuno Cardoso, Paula Carvalho, Iustin Dornescu, Sven Hartrumpf, Johannes Leveling & Yvonne Skalban. "GikiP at GeoCLEF 2008: Joining GIR and QA forces for querying Wikipedia". In Carol Peters, Tomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J.F.Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas & Viviane Petras (eds.), Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers 2009, Springer.
[Prelo]

[113] Diana Santos. "Caminhos percorridos no mapa da portuguesificação: A Linguateca em perspectiva". Linguamática (2009).
[A ser enviado]

[114] Luís Sarmento, Anabela Barreiro, Belinda Maia & Diana Santos. "Avaliação de Tradução Automática: alguns conceitos e reflexões". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 181-190.

[115] Luís Sarmento. "Ferramentas para experimentação, recolha e avaliação de exemplos de tradução automática". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 193-203.

[116] Luís Sarmento. "O SIEMÊS e a sua participação no HAREM e no Mini-HAREM". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 173-189.

[117] Luís Sarmento. "A first step to address biography generation as an iterative QA task". In Carol Peters, Paul Clough, Fredric C. Gey, Jussi Karlgren, Bernardo Magnini, Douglas W. Oard, Maarten de Rijke & Maximilian Stempfhuber (eds.), Evaluation of Multilingual and Multi-modal Information Retrieval - 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006. Alicante, Spain, September, 2006. Revised Selected papers. Berlin / Heidelberg :Springer, Lecture Notes in Computer Science, 2007, pp. 473-482.

[118] Nuno Seco, Nuno Cardoso, Rui Vilela & Diana Santos. "A arquitectura dos programas de avaliação do HAREM". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 283-306.

[119] Nuno Seco. "MUC vs HAREM: a contrastive perspective". In Diana Santos & Nuno Cardoso (eds.), Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca, 2007, pp. 35-41.

[120] Rosário Silva, Susana Inácio & Diana Santos. "Documentação da anotação relativa à cor no COMPARA". 14 de Novembro de 2008. http://www.linguateca.pt/COMPARA/DocAnotacaoCorCOMPARA.pdf

[121] Rosário Silva & Ana Frankenberg-Garcia. "COMPARA's English annotation with CLAWS C7: revision criteria". Novembro de 2008.

[122] Rosário Silva & Ana Frankenberg-Garcia. "Anotação inglesa do COMPARA com o CLAWS C7: critérios de revisão". Novembro de 2008.

[123] Alberto Simões & José João Almeida. "Avaliação de alinhadores". In Diana Santos (ed.), Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal :IST Press, 2007, pp. 219-228.

[124] Alberto Simões, Rúben Fonseca & José João Almeida. "Makefile::Parallel Dependency Specification Language". In Anne-Marie Kermarrec, Luc Bougé & Thierry Priol (eds.), Parallel Processing, 13th International Euro-Par Conference (Euro-Par 2007) (Rennes, France, 28-31 de Agosto de 2007), Berlin/Heidelberg : Springer. Lecture Notes in Computer Science, pp. 33-41.

[125] Alberto Simões & José João Almeida. "Parallel Corpora based Translation Resources Extraction". Procesamiento del Lenguaje Natural 39 (2007), pp. 265-272. SEPLN.

[126] Alberto Simões. "Segmentação bilingue com base na marker hypothesis". In César Analide, Paulo Novais & Pedro Henriques (eds.), Simpósio Doutoral em Inteligência Artificial (SDIA 2007) (Guimarães, Portugal, Dezembro de 2007), pp. 135-144.

[127] Alberto Simões. Extracção de Recursos de Tradução com base em Dicionários Probabilísticos de Tradução. Tese de Doutoramento. Faculdade de Engenharia da Universidade do Minho, Braga. Março de 2008.

[128] Alberto Simões. "Extracção de recursos de tradução". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos. Linguateca, 2008, pp. 7-11.

[129] Alberto Simões & José João Almeida. "Bilingual Terminology Extraction based on Translation Patterns". Procesamiento del Lenguaje Natural 41 (2008), pp. 281-288. SEPLN.

[130] Rui Sousa Silva. "Listas de frequência de palavras como marcadores de estilo no reconhecimento de autoria". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos. Linguateca, 2008, pp. 93-97.

[131] Rui Vilela. "WebJspell, an Online Morphological Analyser and Spell Checker". Procesamiento del Lenguaje Natural 39 (2007), pp. 291-292. SEPLN.

[132] Diana Santos (ed.). Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. Lisboa, Portugal : IST Press. 2007. ISBN: 972-8469-60-8.

[133] Diana Santos & Nuno Cardoso (eds.). Reconhecimento de entidades mencionadas em

português: Documentação e actas do HAREM, a primeira avaliação conjunta na área. Linguateca. 2007.

[134] Cristina Mota & Diana Santos (eds.). Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca. 2008.

[135] "Apêndice D: Segundo HAREM: texto em minúsculas dentro de EM". In Cristina Mota & Diana Santos (eds.), Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM. Linguateca, 2008.

[136] Luís Costa, Diana Santos & Nuno Cardoso (eds.). Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos. Linguateca. 2008.

[137] Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Doug W. Oard, Anselmo Peñas, Vivien Petras & Diana Santos (eds.). Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Berlin : Springer. Lecture Notes in Computer Science. 2008. ISBN: 978-3-540-85759-4.

5.1.2. Apresentações da Linguateca

[138] Rachel Aires. "Uso de marcadores estilísticos para a busca na Internet em português". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[139] José João Dias de Almeida. "Para quê gastar 3 dias em trabalho manual quando podemos fazer (em 3 meses) uma ferramenta que o automatize? ou Lutando por um NLP operating system". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[140] Anabela Barreiro. "Port4NooJ: Portuguese Linguistic Module and Bilingual Resources for Machine Translation". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[141] Anabela Barreiro. "Novas Ferramentas e Recursos Linguísticos para a Tradução Automática: Por ocasião d'O Fim do Início de uma Nova Era no Processamento da Língua Portuguesa". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[142] Eckhard Bick. "The PALAVRAS parser and its Linguateca applications - a mutually productive relationship". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[143] Luís Miguel Cabral, Diana Santos & Luís Costa. "SUPeRB: using an automated publication helper in 9012 at SINTEF ICT". (SINTEF, Oslo, 21 de Novembro de 2008).

[144] Nuno Cardoso. "Diz que é uma espécie de survey sobre Query Expansion". *IV Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 20 de Março de 2007).

[145] Nuno Cardoso. "Novos rumos na expansão de pesquisas para RIG". *V Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 4 de Outubro de 2007).

[146] Nuno Cardoso. "REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[147] Nuno Cardoso. "Novos rumos para a recuperação de informação geográfica em português". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[148] Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota, Diana Santos & Cláudia Freitas.

"Segundo HAREM: Modelo geral, novidades e avaliação". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[149] Marcirio Chaves. "Sistema de Extração e Integração de Conhecimento Geográfico - SEI-Geo". *IV Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 20 de Março de 2007).

[150] Marcirio Chaves. "SEI-Geo: Primeiros Resultados". *V Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 4 de Outubro de 2007).

[151] Marcírio Chaves. "Geo-ontologias para reconhecimento de relações entre locais: a participação do SEI-Geo no Segundo HAREM". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[152] Marcirio Chaves. "Criação e expansão de geo-ontologias e dimensionamento de informação geográfica e reconhecimento de locais e seus relacionamentos em textos". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[153] Luís Fernando Costa & Diana Santos. "Question answering systems: a partial answer". (SINTEF, Oslo, 22 de Janeiro de 2007).

[154] Luís Costa & Luís Miguel Cabral. "Medindo a Linguateca". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[155] Ana Frankenberg-Garcia. "Using a parallel corpus in translation practice and research". *IV Ibero-American Conference on Translation and Interpretation ((IV CIATI))* (São Paulo, Brasil, 14 de Maio de 2007).

[156] Ana Frankenberg-Garcia. "Que referências consultam os estudantes de tradução?". *IV Ibero-American Conference on Translation and Interpretation ((IV CIATI))* (São Paulo, Brasil, 14 de Maio de 2007).

[157] Ana Frankenberg-Garcia. "O que é o COMPARA?". (Lisboa, Portugal, Julho de 2007). Apresentação do COMPARA na FCCN

[158] Ana Frankenberg-Garcia. "Compilação e uso de corpora paralelos". (Universidade de São Paulo, SP, Brasil, 3-5 de Setembro de 2007). Visiting lecturer

[159] Ana Frankenberg-Garcia. "COMPARA". (Cascais, Portugal, Outubro de 2007). Apresentação do COMPARA para professores da Escola Secundária Ibn Mucana

[160] Ana Frankenberg-Garcia. "Building a Parallel Corpus for Translation Research and much more". (Universitat Jaume I, Castellón, Espanha, Novembro de 2007).

[161] Ana Frankenberg-Garcia, Pedro Sousa, Rosário Silva & Susana Inácio. "Things a parallel corpus can do". *Actas do 8th Teaching and Language Corpora Conference (TaLC 2008)* (Lisboa, Portugal, 3-6 de Julho de 2008), pp. 482.

[162] Ana Frankenberg-Garcia. "COMPARA". *TaLC at TaLC: Teaching and Linguateca's (Portuguese language) Corpora* (ISLA, Lisboa, 3 de Julho de 2008).

[163] Ana Frankenberg-Garcia. "A utilização de um corpus paralelo de português e inglês nos estudos linguísticos e literários". (Palestra convidada a Universidade Estadual do Rio de Janeiro, Rio de Janeiro, Brasil, 20 de Agosto de 2008).

[164] Ana Frankenberg-Garcia. "A utilização do corpus COMPARA nos estudos linguísticos e literários". *Postgraduate Seminar in English Language and Literary Studies* (Universidade de São Paulo, Brasil, 26 de Agosto de 2008).

[165] Ana Frankenberg-Garcia. "COMPARA: 9 anos de desenvolvimento". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[166] Ana Frankenberg-Garcia. "Compiling and using a parallel corpus for Translation Studies".

International Seminar. "New Trends in Corpus Linguistics for Language Teaching and Translation Studies. In Honour of John Sinclair" (Granada, Espanha, 22-24 de Setembro de 2008). Palestra convidada

[167] Ana Frankenberg-Garcia. "Utilização do corpus COMPARA na prática tradutória e nos estudos de tradução". (Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, 18 de Novembro de 2008).

[168] Cláudia Freitas. "Alguns caminhos da Floresta". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).

[169] Claudia Freitas. "Exploring Portuguese Syntax with Floresta Sintáctica". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[170] Cláudia Freitas. "A Floresta Sintáctica no Ensino de Português". *TaLC at TaLC: Teaching and Linguateca's (Portuguese language) Corpora* (ISLA, Lisboa, 3 de Julho de 2008).

[171] Cláudia Freitas, Diana Santos, Hugo Gonçalo Oliveira & Paula Carvalho. "ReRelEM: Relações Semânticas no Segundo HAREM". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[172] Paulo Gomes. "Apresentação do Pólo de Coimbra". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).

[173] Paulo Gomes. "Linguateca: Polo de Coimbra - Plantando Florestas e Criando Papel". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[174] Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes & Nuno Seco. "PAPEL: A lexical ontology for Portuguese". *Cognitive & Media Systems Winter Meeting 2008 (CMS 2008)* (Coimbra, Portugal, Fevereiro de 2008).

[175] Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes & Nuno Seco. "PAPEL: A lexical ontology for Portuguese". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[176] Belinda Maia. "Applications of Corpógrafo and NooJ to Teaching and Research". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[177] Belinda Maia. "Corpógrafo". *TaLC at TaLC: Teaching and Linguateca's (Portuguese language) Corpora* (ISLA, Lisboa, 3 de Julho de 2008).

[178] Belinda Maia. "Alice no País das Maravilhas ou as aventuras e desventuras de uma linguista no mundo do PLN". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[179] Sérgio Matos, Anabela Barreiro & Belinda Maia. "Corpógrafo and NooJ: using linguistic resources to obtain aligned concordances from corpora". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[180] Cristina Mota. "É tempo de avaliar o TEMPO". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[181] Cristina Mota. "R3M, uma participação minimalista no Segundo HAREM". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[182] Paulo Rocha. "Arte dos corpora em português: o projecto AC/DC". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).

[183] Diana Santos. "A Linguateca, e em especial a Floresta, o PAPEL e o HAREM".

(Universidade de Coimbra, Coimbra, Portugal, 4 de Abril de 2007).

[184] Diana Santos. "Evaluation in natural language processing". *European Summer School on Language, Logic and Information (ESSLLI 2007)* (Trinity College, Dublin, Irlanda, 6-17 de Agosto de 2007).

[185] Diana Santos. "Aplicações de uma floresta sintáctica". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).

[186] Diana Santos, Eckhard Bick & Susana Afonso. "Floresta Sintá (c)tica: apresentação e história do projecto". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).

[187] Diana Santos. "Linguateca: sete anos devotados ao processamento computacional do português". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).

[188] Diana Santos. "Corpos linguísticos da Linguateca: apresentação". *TaLC at TaLC: Teaching and Linguateca's (Portuguese language) Corpora* (ISLA, Lisboa, 3 de Julho de 2008).

[189] Diana Santos. "The Linguateca experience: What can be reused? Or, what have we learned?". *Presentation at SINTEF* (Refsnes Gods, Noruega, 4 de Setembro de 2008).

[190] Diana Santos, Hugo Gonçalo Oliveira, Cláudia Freitas, Cristina Mota & Paula Carvalho. "Segundo HAREM: Balanço e perspectivas de futuro". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[191] Diana Santos. "Linguateca 10 anos: festejo ou luto?". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[192] Diana Santos. "Curso avançado de estudos contrastivos usando o COMPARA como ferramenta". *Módulo na EBraLC, Segunda Escola Brasileira de Linguística Computacional* (Universidade Estadual Paulista - UNESP - Campus de São José do Rio Preto, 3-5 de Novembro de 2008).

[193] Diana Santos. "Linguística com corpos na era da internetização". *VII Encontro de Lingüística de Corpus* (Universidade Estadual Paulista - UNESP - Campus de São José do Rio Preto, 6-7 de Novembro de 2008). Palestra convidada

[194] Rosário Silva, Susana Inácio & Diana Santos. "Colouring COMPARA: contrastive and monolingual colour studies in English and Portuguese". *Presentation at the conference of the American Association for Corpus Linguistics* (Provo, UT, EUA, Março de 2008).

[195] Rosário Silva, Susana Inácio & Diana Santos. "Colouring COMPARA: contrastive and monolingual colour studies in English and Portuguese". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[196] Mário J. Silva. "Pólo XLDB da Linguateca: 4 anos". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[197] Alberto Manuel Brandão Simões. "Extracção de Recursos de Tradução com base em Dicionarios Probabilísticos". *IV Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 20 de Março de 2007).

[198] Alberto Simões. "Extracção de Recursos de Tradução com base na Hipótese das Palavras-Marca". *V Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 4 de Outubro de 2007).

[199] Alberto Simões. "Makefile::Parallel - Uma ferramenta para paralelização de processos". *V Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 4

de Outubro de 2007).

[200] Alberto Manuel Simões. "Automatic Extraction of Translation Resources from Parallel Corpora". *Engenharia'07. e as cidades* (Universidade do Minho, Braga, Portugal, 22 de Outubro de 2007).

[201] Alberto Simões. "Examples Extraction for Machine Translation". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[202] Alberto Simões. "Extracção de recursos de tradução". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[203] Rui Sousa Silva. "Listas de frequência de palavras como marcadores de estilo no reconhecimento de autoria". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

5.2 Publicações e apresentações no âmbito da Linguateca

5.2.1. Artigos em livros

[204] Rachel Virgínia Xavier Aires & Sandra Maria Aluísio. "As avaliações atuais de sistemas de busca na Web e a importância do usuário". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 247-254.

[205] Carlos Amaral, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto & Tiago Veiga. "Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[206] Christian Nunes Aranha. "O Cortex e a sua participação no HAREM". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 113-122.

[207] Jorge Baptista. "Alguns comentários sobre a lista dourada". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 121-131.

[208] Eckhard Bick. "Morfologia com sintaxe debaixo de olho: Sobre a participação do PALMORF nas Morfolimpíadas". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 53-67.

[209] Eckhard Bick. "Functional Aspects on Portuguese NER". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 145-155. Este artigo foi publicado pela Springer, na série LNAI, vol. 3960, ISBN-10: 3-540-34045-9

[210] Daniela Braga & Miguel Sales Dias. "Os recursos da Linguateca ao serviço do desenvolvimento da tecnologia de fala na Microsoft". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 29-33.

[211] Mírian Bruckschen, José Guilherme Camargo de Souza, Renata Vieira & Sandro Rigo. "Sistema SeRELeP para o reconhecimento de relações entre entidades mencionadas". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[212] Sara Candeias & Fernando Perdigão. "Conversor de grafemas para fones baseado em regras para português". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 99-104.

[213] Nuno Cardoso. "REMBRANDT - Reconhecimento de Entidades Mencionadas Baseado em Relações e ANálise Detalhada do Texto". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[214] Olga Craveiro, Joaquim Macedo & Henrique Madeira. "PorTexTO: sistema de anotação/extracção de expressões temporais". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[215] Liliana Ferreira, António Teixeira & João Paulo da Silva Cunha. "REMMA - Reconhecimento de entidades mencionadas do MedAlert". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[216] Liliana Ferreira & António Teixeira. "Linguateca e Processamento de Linguagem Natural na Área da Saúde: Alguns Comentários e Sugestões". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 43-48.

[217] Óscar Ferrández, Zornitsa Kozareva, Antonio Toral, Rafael Muñoz & Andrés Montoyo. "Tackling HAREM's Portuguese Named Entity Recognition task with Spanish resources". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 137-144.

[218] Milena Uzeda Garrão & Maria Carmelita Padua Dias. "Uma abordagem estatística para a identificação de colocações verbais usando o projeto AC/DC em www.linguateca.pt". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 65-69.

[219] Marco Gonzalez, Leonardo Cavalheiro Langie & Vera Lúcia Strube de Lima. "Avaliação de Sistemas de recuperação e categorização de textos: métodos e aplicações". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 231-245.

[220] Caroline Hagège. "Segmentação e análise morfológica do português com recursos léxicos limitados: Sobre a participação do Smorph nas Morfolimpíadas". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 35-40.

[221] Caroline Hagège, Jorge Baptista & Nuno Mamede. "Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre o INESC-L2f e a Xerox". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[222] Caroline Hagège, Jorge Baptista & Nuno Mamede. "Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e

o futuro". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[223] Caroline Hagège, Jorge Baptista & Nuno Mamede. "Apêndice B: Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[224] Ricardo Ueda Karpischeck. "As Morfolimpíadas e a avaliação da verificação ortográfica: Sobre a participação do br.ispell nas Morfolimpíadas". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 69-73.

[225] Vera Lúcia Strube de Lima. "Dez anos de convivência: um apanhado geral quanto ao uso dos recursos da Linguateca no Programa de Pós-Graduação em Ciência da Computação da PUCRS - Brasil". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 105-107.

[226] Ronaldo Teixeira Martins & Maria das Graças Volpe Nunes. "Dos processos de individuação e de categorização lexical: sobre a participaçao do ReGra nas Morfolimpíadas". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 43-50.

[227] Bruno Martins & Mário J. Silva. "O HAREM e a avaliação de sistemas para o reconhecimento de entidades geográficas em textos em língua portuguesa". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 79-86.

[228] Bruno Martins. "O sistema CaGE e a participação no Segundo HAREM". In Cristina Mota & Diana Santos (eds.), *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. Linguateca, 2008.

[229] Agostinho Monteiro, Julio Barbas & Nuno C. Marques. "Utilização da Programação Declarativa para processamento do CETEMPúblico". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 1-5.

[230] Cristina Mota. "Estudo preliminar para a avaliação de REM em português". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 19-34.

[231] Cristina Mota & Max Silberztein. "Em busca da máxima precisão sem almanaques. O Stencil/NooJ no HAREM". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 191-208.

[232] Maria das Graças Volpe Nunes. "Relato sobre a parceria Linguateca-NILC". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 61-63.

[233] Paulo Quaresma & Irene Rodrigues. "Avaliação de sistemas interactivos de recuperação de informação em bases de texto jurídicas". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 257-265.

[234] Violeta Quental. "10 anos de Linguateca - depoimento". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 109-113.

[235] Lucia Helena Machado Rino & Thiago Alexandre Salgueiro Pardo. "A Coleção TeMário e a avaliação de sumarização automática". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 267-176.

[236] Mário Silva, Bruno Martins & Miguel Costa. "Avaliação conjunta de recuperação de informação da Web portuguesa". In Diana Santos (ed.), *Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa*. Lisboa, Portugal :IST Press, 2007, pp. 133-140.

[237] Augusto Soares da Silva. "O corpus CONDIV e o estudo da convergência e divergência entre variedades do português". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 25-28.

[238] Thamar Solorio. "MALINCHE: A NER system for Portuguese that reuses knowledge from Spanish". In Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*. Linguateca, 2007, pp. 123-136.

[239] Lílian Figueiró Teixeira & Rove Luiza de Oliveira Chishman. "Um estudo do córpus COMPARA: a semântica dos compostos nominais". In Luís Costa, Diana Santos & Nuno Cardoso (eds.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos*. Linguateca, 2008, pp. 35-41.

5.2.2. Apresentações

[240] Eckhard Bick. "Computational Syntax". *Encontro Um passeio pela Floresta Sintá (c)tica* (Coimbra, Portugal, 28 de Setembro de 2007).

[241] Daniela Braga & Miguel Sales Dias. "Os recursos da Linguateca ao serviço do desenvolvimento da tecnologia de fala na Microsoft". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[242] Mírian Bruckschen, Renata Vieira & José Guilherme Camargo de Souza. "Sistema SeRELeP para o reconhecimento de relações". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[243] Sara Candeias & Fernando Perdigão. "Conversor de grafemas para fones baseado em regras para português". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[244] Nuno Cardoso. "Geographic IR Challenges". (Oslo, Noruega, 4 de Dezembro de 2007).

[245] Olga Craveiro. "PorTexTO: sistema de anotação/extracção de expressões temporais". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[246] Liliana Ferreira & António Teixeira. "REMMA - Reconhecimento de entidades mencionadas do MedAlert". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[247] Liliana Ferreira & António Teixeira. "Linguateca e Processamento de Texto Livre na Área da Saúde: Alguns Comentários". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[248] Milena Uzeda Garrão & Maria Carmelita Padua Dias. "Uma abordagem estatística para a identificação de colocações verbais usando o projeto AC/DC em wwwlinguateca.pt". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[249] Caroline Hagège, Jorge Baptista & Nuno Mamede. "Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre o INESC-L2f e a Xerox". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[250] Caroline Hagège, Jorge Baptista & Nuno Mamede. "Identificação, classificação e normalização de expressões temporais do português: A experiência do Segundo HAREM e o futuro»". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[251] Vera Lúcia Strube de Lima. "Dez anos de convivência: um apanhado geral quanto ao uso dos recursos da Linguateca no Programa de Pós-Graduação em Ciência da Computação da PUCRS - Brasil". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[252] Bruno Martins. "O sistema CaGE e a participação no Segundo HAREM". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[253] Pedro Mendes. "Adaptação do sistema de reconhecimento de entidades mencionadas da Priberam ao HAREM". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 1 de de 2008).

[254] Agostinho Monteiro, Julio Barbas & Nuno C. Marques. "Utilização da Programação Declarativa para processamento do CETEMPúblico". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[255] Maria das Graças Volpe Nunes. "Relato sobre a parceria Linguateca-NILC". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[256] Violeta Quental. "10 anos de Linguateca - depoimento". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[257] Diana Santos. "Avaliação: ensaio de mais algum material para o ESSLLI'07". *IV Simpósio Doutoral da Linguateca 2007* (Faculdade de Ciências - Universidade de Lisboa, 20 de Março de 2007).

[258] Cícero Nogueira dos Santos & Ruy Luiz Milidiú. "Reconhecimento de Entidades Mencionadas usando ETL". *Encontro do Segundo HAREM* (Universidade de Aveiro, Portugal, 7 de Setembro de 2008).

[259] Max Silberztein. "NooJ's Text Annotation Structure". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[260] Augusto Soares da Silva. "O corpus CONDIV e o estudo da convergência e divergência entre variedades do português". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

[261] José Paulo Tavares. "NooJ Latin module". *Workshop on Language Resources for Teaching and Research* (Faculdade de Letras da Universidade do Porto, 23 de Abril de 2008).

[262] Lílian Figueiró Teixeira & Rove Luiza de Oliveira Chishman. "A representação semântica de nomes compostos: Um estudo no COMPARA". *Linguateca: 10 anos* (Curia, Portugal, 11 de Setembro de 2008).

5.3 Outras fontes citadas no presente relatório

[263] Dicionário PRO da Língua Portuguesa. Porto Editora, Porto, 2005.

[264] Eckhard Bick. The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. PhD thesis, Aarhus University, 2000.

[265] Diana Santos. "Relatório Linguateca 2000-2003". Setembro de 2003.

[266] Diana Santos. "Resumo da actividade da Linguateca de 15 de Maio de 2003 a 15 de Dezembro de 2006". Linguateca. Dezembro de 2006. Com a colaboração (por ordem alfabética) de Alberto Simões, Ana Frankenberg-Garcia, Belinda Maia, Luís Costa, Luís Miguel Cabral, Luís Sarmento, Marcirio Chaves, Mário J. Silva, Nuno Cardoso, Paulo Gomes e Rui Vilela

6. Indicadores de realização física

	Unidade: em número
A- Publicações	
Livros	5
Capítulos de livros	105
Artigos em revistas internacionais	11
Artigos em revistas nacionais	0
B- Comunicações	
Em congressos científicos internacionais	50
Em congressos científicos nacionais	40
C- Relatórios	11
D- Organização de seminários e conferências	7
E- Formação Avançada	
Teses de Doutoramento	1 terminada, 4 em curso
Teses de Mestrado	1 terminada, 1 em curso
Outra	
F- Modelos	
G- Aplicações computacionais	12

H- Instalações Piloto

I- Protótipos laboratoriais

J- Patentes

L- Outros (discriminar)

Recursos	9
Serviços na rede	7

Recursos: CETEMPúblico, CETENFolha, CHAVE, WBR 99, GeoNet-PT-01, Floresta, COMPARA, Port4NooJ, LÂMPADA

Aplicações: Corpógrafo, Esfinge, GKB, PLN::Base, Águia, Ambiente de avaliação do HAREM, Alinhador/validador de novas etiquetas do Bosque, DISPARA, SUPeRB, SEI-Geo, Chart parser para analisar entradas no dicionário da PE, DIRECT para português (tradução/localização)

Serviços: AC/DC, COMPARA, Floresta, Busca no Catálogo, SUPeRB, CorTrad, ReEscreve